Michael Nguyen and Peter Juel Henrichsen

THE LOCUM HYPHEN

A Formal Approach to the Lexicalization of Multiword Expressions with Rich Internal Semantics

Abstract We present a study of Danish multiword constructions containing one or more hyphens, such as *gas- og vandmester* ('gas- and water.repairman'; 'plumber'), *ilt- og brintatomer* ('oxygen- and hydrogen atoms') and *haveborde og -stole* ('garden tables and -chairs'). Although materially analogous, such constructions exhibit different semantics, falling — as we shall argue — into two distinct groups ("locum" vs. "pseudo locum" hyphen constructions). This study employs the COR_1 database (Jervelund et al., 2012; Widmann, 2024) and draws on the CLINK formal framework for computational lexicography (Henrichsen, 2023; 2024a). The aim of this paper is to demonstrate how linguistic analysis of such multiword expressions with rich internal semantics can benefit from methods of computational lexicography – and vice versa.

Keywords hyphen; multiword expressions; language technology; structural ambiguity; punctuation; COR_1

1. Introduction

This paper presents a lexicographical study on Danish (written) multiword expressions (MWEs) containing hyphens preceded or followed by a blank space. Examples are *gas- og vandmester* ('gas- and water.repairman', 'repairman of gas and water'; 'plumber'), *ilt- og brintatomer* ('oxygen- and hydrogen atoms') and *haveborde og -stole* ('garden tables and -chairs'). We introduce the term *locum phrase* for a subclass of such phrases, viz. those that can be transformed into a synonymous phrase by replacing its hyphen(s) with lexical material occurring in the phrase. The hyphens of a locum phrase are termed *locum hyphens*. As we shall see, not all MWEs with hyphens are locum hyphens.

As lexical base, we employ the newly published Centrale Ordregister (Central Word Register alias "COR_1", cf., Henrichsen, 2024a), a machine-readable projection of Retskrivningsordbogen, the official Danish Orthographic Dictionary (Jervelund et al., 2012). Retskrivningsordbogen (RO) defines the Danish orthographic norm and is published by The Danish Language Council (Act 1997). The Danish public sector (including the school system) is obliged to follow the norm. RO consists of a traditional dictionary (approximately 64,000 lemmas with morphological information) and a manual of 62 articles with the rules of Danish punctuation (comma marking, above all), abbreviation, capitalization, compounding and more.

COR_1 is specially aimed at the Danish NLP (natural language processing) sectors, serving as the lexical foundation of applications for spell and grammar checking, speech technology, machine translation, chatbots, AI and so forth.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

As in most European languages, the hyphen grapheme (RO § 57) has several functions in Danish orthography. In this paper, we focus on the use of the hyphen in phrases where it replaces a part of a word (see § 57.2), for instance:

(1) vin- og ølflasker
wine- and beer.bottles
'wine bottles and beer bottles'

As mentioned, we refer to such phrases as *locum phrases*. Lexicalizing MWEs such as locum phrases is challenging, not only in COR_1 but in computational lexicography in general.

In this paper, we discuss the semantic interpretation of the locum phrase as opposed to the *pseudo locum phrase* – a phrase that is materially identical to the locum phrase. Our analysis points to some possible improvements of the rule formulations in RO regarding both types of phrases.

2. RO and COR_1

Each word form in COR_1 carries a unique ID in this format:

COR. <lemma index> . <inflexion index> . <variant index>

COR_1 covers all RO lemmas (including a number of MWEs) and also contains sub-verbal lexemes (e.g., affixes, symbols and digits) and non-verbal lexemes (e.g., punctuation marks). Thus, all text elements with an individual semantic contribution (words and punctuation marks alike) carry their own COR-IDs:

Table 1: Samples from the machine-readable COR_1 dictionary

Lexeme	COR-ID	Gloss
bord	COR.44636.120.01	'table'
borde	COR.44636.122.01	'tables'
og	COR.00099.970.01	'and'
-	COR.XP.017	('hyphen', as used in this paper)

As for the rules of RO, they are worded rather informally in order to be user-friendly; thus, intuitive definitions are preferred over formal specifications. Often, large numbers of examples are provided from which the reader can infer how the rules are to be applied. In Section 3, we discuss in greater detail the locum hyphen and also introduce the pseudo locum hyphen.

3. The Locum Hyphen

In Danish, compounds and derivatives are typically written as one text word without any blank spaces or hyphens, e.g., *havebord* ('garden table') and *brugervenlig* ('user-

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

friendly'). According to the rules of RO, a hyphen may substitute a part of a word (§57.2) (authors' translation):

> "A hyphen is used to show that two or more compounds or derivatives have a common part which is only mentioned once:"[1]

This common part of the word is recoverable from the immediate syntactic context, more specifically from a conjunctional phrase (the locum phrase). Example (2) below is from RO (§57.2).

(2) haveborde og -stole
garden.tables and -chairs
'garden tables and garden chairs'

The hyphen substitutes *have* ('garden'), thereby qualifying as a locum hyphen. The phrase can be expanded without a change of semantics: *haveborde og havestole* ('garden tables and garden chairs'). Note furthermore that both the substituted part (*have*, 'garden') and the non-substituted part (*borde*, 'tables') are lexical morphs[2].

3.1 Complex Cases of the Locum Hyphen

However, a number of complications arise in other cases. See (3)-(5), all likewise from RO, §57.2:

(3) over- eller underskud
over- or under.shot
'surplus or deficit'

(4) im- og eksport
im- and ex.port
'import and export'

(5) A- og B-skat
A- and B-tax
'A-tax and B-tax'

In (3), the hyphen substitutes *skud*, which has no semantics in and of itself, and so it is no morph, and thus has no lexical entry.

In (4), both the substituted *port* and the non-substituted *im* are likewise not morphs, and thus not represented in any lexical entry.

[1] "Bindestreg bruges for at vise at to eller flere sammensatte eller afledte ord har en fælles del som kun bliver nævnt én gang:"

[2] We follow Henrichsen (2024a) in using the term *morph* for any text element that has an individual semantic contribution. The term includes but is not limited to morphemes, punctuation marks and linking elements.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In (5), what is substituted is not a string of letters but a morph + another hyphen (i.e., *-skat*). This is revealed by the expansion of the phrase: *A--skat og B-skat*. Double hyphens are, however, not used in Danish orthography.

Another complicated case is given in (6) (not mentioned in §57.2):

(6) køb.s.pris og -dato
buy.LINK.price and -date
'(the) price of buying and (the) date of buying'

The hyphen substitutes *købs*, but *købs* has no lexical entry. This is because the *s* is not part of the morph *køb* 'buy' but a so-called linking element, which appears in many compounds in Danish (see e.g., Bauer, 1975, pp. 222–245).

3.2 Introducing the Pseudo Locum Hyphen

Although the examples (3)–(5) from §57.2 are complex cases for a parser, they (as well as example (6)) are in principle consistent with the translated rule passage above.

See, however, the MWEs in (7) and (8), taken from the lexical entries of RO:

(7) gas- og vandmester
gas- and water.repairman
'repairman of gas and water'; 'plumber'

(8) social- og sundhedsassistent
social- and health.LINK.assistant
'social and health care assistant'

Expanding (7) shows that the hyphen does not substitute any part of a word: ??*Gasmester og vandmester* ('repairman of gas and repairman of water'). Although the expanded phrase is structurally well-formed, there is no such thing as a *gasmester* ('repairman of gas') or a *vandmester* ('repairman of water').

Furthermore, it can be observed that the two conjuncts can neither be switched around (??*vand- og gasmester*, 'water- and gas.repairman'), nor can the conjunction be replaced with another conjunction (??*gas- eller vandmester*, 'gas- or water.repairman'). The same restrictions apply to (8).

Examples like (7) and (8) should be distinguished from superficially similar examples like (9) (a constructed example, not a lexical entry of RO):

(9) ilt- og brintatomer
oxygen- and hydrogen.atoms
'oxygen atoms and hydrogen atoms'

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In contrast to (7)–(8), the hyphen in (9) does in fact substitute a part of a word, as the expanded phrase shows: *iltatomer og brintatomer* ('oxygen atoms and hydrogen atoms'). In other words, it is a locum hyphen. Furthermore, the two conjuncts can be switched around, and the conjunction can be replaced with another conjunction.

There is thus a sharp contrast between on the one hand (7)–(8) and on the other hand (9). The hyphens in (7)–(8) are **not** real locum hyphens: they do not substitute a part of a word but only appear to do so. They are **pseudo** locum hyphens.

We thus make the distinction between locum hyphens and pseudo locum hyphens, and correspondingly between locum hyphen phrases and pseudo locum hyphen phrases. There might, however, be several types of pseudo locum hyphen phrases. See (10) (from §57.2):

(10) de seks- til syvårige
the six- to seven.year.olds
'the six to seven year olds'

The expanded phrase is not synonymous with (10) and is hardly meaningful at all (??*De seksårige til syvårige*, 'the six year olds to seven year olds'). Thus, the hyphen is not a locum hyphen but a pseudo locum hyphen.

However, in contrast to (7) and (8), the conjunction in (10) may be replaced with another conjunction (*og* 'and' or *eller* 'or'), giving rise to a well-formed semantic expression (although a different one from that of (10)). Given the differences in replaceability, we may correspondingly distinguish between different types of pseudo locum hyphen phrases – or at least note that they have different properties.

In any case, the pseudo locum hyphens in (7), (8) and (10) all seem to have a linking (or gluing) function: They link complex compounds consisting of more than one text word. Thus, the relevant article of RO would instead seem to be §57.7a and §57.7.b (here reformulated and simplified):

When the first part of a compound or a derivative consists of more than one word, a hyphen is inserted between the first part and the second part.

Following this rule, then, the hyphen in (10) should instead be placed as in (11):

(11) de seks til syv-årige
the six to seven.year.olds
'the six to seven year olds'

The hyphen links the complex first part *seks til syv* with the second part *årige*. The same would seem to apply to the hyphens (7)–(8).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Finally, in some cases it is not clear whether a hyphen is a locum hyphen or a pseudo locum hyphen. The hyphen in (12) is ambiguous between the two, as shown in the two translations:

(12) salg.s- og leveringsbetingelser
sell.LINK- and delivery.conditions
'conditions for sale and delivery'
'conditions for sale and conditions for delivery'

The complicated cases in (3)–(5), the pseudo locum hyphens in (7), (8) and (10) and ambiguous cases such as (12) suffice to show what kind of challenges a parser faces when analyzing MWEs involving hyphens. In Section 4, we present a formal analysis to tackle some of these challenges.

4. The Formal Analysis of MWEs Involving Hyphens

COR_1 uses the CLINK lexical template (Henrichsen, 2023; 2024a), based on categorial grammar (the Lambek calculus). Readers unfamiliar with the "slash-style grammar" will find Wood (1993) and Henrichsen (2024b) useful guides.

The basic data structure in categorial grammar is the so-called sequent, having the form ANTECEDENT ==> CONSEQUENT, where ANTECEDENT is a list of categories and CONSEQUENT is the hypothetical PoS of the phrase to be proven. Each category containing a slash ($y\backslash x$ or x/y) requires an argument y either to its left side ($y\backslash x$) or to its right (x/y) in order to produce a resulting category x. Table 2 provides an example.

Table 2: The CLINK lexical template

gas- og vandmestrene ('the-plumbers$_{pl.def}$')		
[gas][-]	110 110\119	==> 119
[og]	$X\backslash X/X$	==> $X\backslash X/X$
[vand][][mestrene]	119 $Y\backslash Z/Z$ 113	==> 113
	119 $X\backslash X/X$ 119 $Y\backslash Z/Z$ 113	==> *Consequent*

The main sequent (the formula under the line) is satisfied for X=119, Y=119, Z=113, and *Consequent*=113, proving that *gas- og vandmestrene* ('the gas- and water.repairmen'; 'the plumbers') has the syntactic function of a noun in uter.plur.def (corresponding to the inflexion index 113). The proof also has a semantic projection. The COR_1 homepage at ordregister.dk provides an interactive PoS table relating the category indices of COR_1 (such as 110, 113, 119) and the inflexional forms of Danish morphology.

As a special feature of the categorial grammar analysis, each proof step (i.e., each application of a slash category with an adjacent category) is coordinated with a semantic composition, see details in Henrichsen, 2024a). As a result, compositional

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

ambiguities appear in the form of alternative lambda-formulae, as exemplified in Table 3. Categories and semantic proxies are imported from COR_1 (thus *mestre* has the COR-id COR.51525.112).

Table 3: Lexical information for three COR_1 lexemes

Form	Category	Semantic proxy
gas	110	62857
vand	120	41032
mestre	112	51525

Alternative semantic projections of *gas- og vandmestre*:

51525(62857 & 41032)

51525(62857) & 51525(41032)

These two semantic analyses (simplified here for readability) correspond to the two paraphrases (i) 'repairmen of gas and water' or (ii) 'repairmen of gas and repairmen of water'. These are obviously two different interpretations, as can be appreciated by the singular form *gas- og vandmester*: As mentioned, there is no such thing as a repairman of gas nor a repairman of water; a repairman of gas and water (i.e., a plumber), however, exists. It is not surprising that the singular phrase frequently occurs in ordinary Danish texts (actually even more frequently than the plural). In contrast, *ilt- og brintatomer* ('oxygen and hydrogen atoms') practically never occurs in the singular, as opposed to the plural, as indeed predicted by the semantic analysis. How could a single atom be of the class oxygen **as well as** of the class hydrogen?

5. Discussion

Our study has revealed a hitherto undescribed (or at least unpublished) classification of the MWEs with hyphens as they appear in ordinary Danish text. As shown, these constructions show considerably more complex semantic patterns than implied by their orthographic forms.

For instance, MWEs like *gas- og vandmester* may have non-compositional semantics, making them vulnerable to e.g., misguided translations (into e.g., 'repairman of gas and repairman of water', in place of 'repairman of gas and water').

Inspired by the German typology of hyphens (see for instance Fleischer & Barz, 1995, pp. 92, 122–125, 131–135, 142–143), we suggest that a Danish typology be developed. For instance, the function of the locum hyphen is very different from that of the hyphen in *B-skat* ('B-tax'), which links *B* and *skat*. We consider the present paper as a step in that direction.

We are currently working on the proper inclusion in COR_1 of the locum hyphen and the pseudo locum hyphen, which in turn will facilitate the development of "intelligent"

text processing, including grammar checking and machine translation. As a valuable side effect of our formal approach, we may be able to suggest revisions to the RO rule articles with regard to hyphenation.

6. Summary

In this paper, we have discussed MWEs with hyphens, both in relation to the relevant rules and lemmas of Retskrivningsordbogen and in relation to some of the related challenges of NLP (natural language processing). Importantly, we have pointed to the hitherto unnoticed problem of the pseudo locum hyphen.

We have briefly presented a formal analysis based on categorial grammar and the lambda-calculus and shown how it can represent the semantics of MWEs with hyphens and their potential ambiguities.

Generally, formalization projects like the one reported here have often revealed indeterminacies in Retskrivningsordbogen, such as those related to the (pseudo) locum hyphen. Using formal methods as those employed in this paper makes it possible to pinpoint schisms between normative orthographic rules and the morpho-semantic reality.

References

Act on Danish Orthography (1997). Retsinformation, LOV nr. 332 14/05/1997.

Bauer, L. J. (1975). *Nominal compounds in Danish, English and French.* [Doctoral dissertation, University of Edinburgh].

Fleischer W., & Barz, I. (1995). *Wortbildung der deutschen Gegenwartssprache*, 2nd edition. Max Niemeyer.

Henrichsen, P. J. (2023). Det Centrale Ordregister. Et indeks over det danske sprog – en gave til dansk sprogteknologi. In L. Holmer, G. Horn, H. Landqvist, P. Nilsson, E. Nordgren, & E. Sköldberg (Eds.), *Nordiska Studier i Lexikografi 16* (pp. 113–126) Nordiska Forening i Lexikografi, Lund Universitet.

Henrichsen, P. J. (2024a). Make each morph count. A new approach to computational lexicography for text processing. In K. Š. Despot, A. Ostroški Anić, & I. Brač, *Proceedings of the 21st EURALEX International Congress, Lexicography and Semantics* (this issue).

Henrichsen, P. J. (2024b). Tekstordet som grammatisk domæne. *Ny forskning i grammatik*, *31*, 53–70.

Jervelund, A. Å., Schack, J., Jensen, J. N., & Andersen, M. H. (2012). *Retskrivningsordbogen*, 4th edition. Alinea.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Widmann, T. (2024). The Central Word Register of the Danish Language. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023). Proceedings of the eLex 2023 conference* (pp. 91–103). Lexical Computing CZ.

Wood, M. M. (1993). *Categorial Grammars*. Routledge.

Contact information

Michael Nguyen
The Danish Language Council
mn@dsn.dk

Peter Juel Henrichsen
The Danish Language Council
pjh@dsn.dk