Linking Danish Parser Output to a Central Word Repository - From Morphosemantic Disambiguation to Unique Identifiers

Eckhard Bick
Department of Language, Culture, History and Communication
University of Southern Denmark
eckhard.bick@gmail.com

Abstract

This paper describes and evaluates a grammatically informed linking system that assigns unique identifiers (UIDs) from a central word repository (COR) to running Danish text. To do so, the system's algorithm matches the annotation of a morphosyntactic and semantic parser (DanGram) to corresponding information in the word registry, using a scoring method and disambiguated grammatical tags such as lemma, POS, inflection and semantic class. In addition to ordinary words, the linker also assigns UIDs to the parts of compounds and multi-word expressions. For mixed Danish text, the linker assigned correct UIDs to 97.8% of all non-name, non-number words. Linking failures were caused either by parser errors (0.3%) or COR gaps (1.9%) rather than by the matching tool itself (< 0.1%).

1 Introduction

Despite ongoing advances in natural language processing (NLP), integrating different resources remains a recalcitrant problem, not least due to differences in tokenization, lemmatization and tag definitions and granularity. While the latter has been addressed – at least at the morphosyntactic level – by the Universal Dependencies initiative (e.g. Nivre, 2015), resource differences in terms of lexical granularity are often overlooked, even in well-resourced languages. Thus, it is not trivial to which degree differences in etymology, pronunciation and meaning, inflection paradigms or spelling variation should warrant separate lexicon entries or – at the tagger/parser level – different lemmas or sub-categorization. The problem is compounded by the fact that state-of-the-art systems, while getting more and more accurate, still inherit a predefined and unquestioned lemma granularity from their training data, making it difficult to mount a language technology (LT) pipeline with modules created with different training data, or different morphological analyzers. A possible solution is agreeing – for a given language – on a shared lexical inventory of both lemmas and inflected forms, with unique identifiers (UIDs) for each entry. For Danish, the COR word repository (Dideriksen et al., 2022) is such a resource. However, while conceptually sound, the COR registry itself is still only half of (LT) heaven, as long as it isn't aligned with other resources and shared between tools. Notably, taggers, parsers and semantic analyzers need to be able to link their analyses to such a central repository. In this paper, working with output from the DanGram parser[1], we will show how different morphological and semantic tags from a parser pipeline can be used to link a wordform to a unique identifier, handling matching and disambiguation in an integrated fashion.

2 COR

COR (Det Centrale Ordregister) is a new lexical resource that assigns unique IDs to Danish words[2]. The resource is being developed by the Danish Language Council (Dansk Sprognævn[3], DSN) in cooperation with the Danish Society for Language and Literature (DSL[4]) and Copenhagen University's Center for Language Technology (CST[5]). In its first, level-1 edition, COR covers the content of the official Danish spelling dictionary[6]. Each word ID (1a-c) consists of dot-separated parts - a first part for the lemma and a second part for inflection. A third part is reserved for spelling variation[7].

[1] https://visl.sdu.dk/visl/da/parsing/automatic/parse.php

[2] The targeted word classes are the closed and inflecting POS classes, with predictable limitations for proper nouns, numerical expressions, abbreviations and punctuation-based "words" (e.g %, smileys), as well as dialectal and spoken forms.

[3] https://dsn.dk

[4] https://dsl.dk

[5] https://cst.ku.dk/english/

[6] https://dsn.dk/ordboeger/retskrivningsordbogen/

[7] In principle, this includes historical variants and current spelling made obsolete by a future spelling reform

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 93–101
September 18–22, 2023.

```
1a) COR.37309.200.01 hoste (to cough)
    vb, inf, act
1b) COR.37309.203.01 hoster (coughs)
    vb, pr, act
1c) COR.38283.200.01 hoste (to host)
    vb, inf, act
```

Homographs are regarded as distinct based on surface markers rather than etymology or semantics proper. Thus, distinguishing criteria are part of speech (POS), grammatical gender (2a-b), pronunciation (1c with English [o]) and differences in inflection paradigms, e.g. different plural forms or not allowing a plural at all. Here, traditional etymological or sense distinctions are often captured implicitly rather than explicitly. For instance, the missing plural is typical for +mass (-countable) semantic classes such as substances, liquids and materials. Thus, because the word *træ* ('tree') does not inflect in the plural when meaning 'wood', most Danish tree names have a separate COR entry as a type of wood (3a-b).

```
2a) COR.47455.110.01 brud (bride)
    n, utr , sg, idf
2b) COR.48668.120.01 brud (rupture)
    n, neu, sg, idf
3a) COR.56312.120.01 bøgetræ (beech tree)
    n, neu, sg, idf
3b) COR.59335.120.01 bøgetræ (beech wood)
    n, neu, sg, idf
```

In addition to these implicit semantic distinctions, COR does have a semantic dimension, as it offers short definitions for ambiguous words, illustrating the semantic reach of a given entry. Also, at level 2, external semantic resources can be linked to COR (Nimb et al., 2022), for instance the existing Danish wordnet, DanNet (e.g. Pedersen et al., 2009) or the Danish Framenet[8] (Bick, 2011). However, as will be discussed in more detail in section 4, sense mapping between such resources and a primarily morphological resource like COR is not always a one-to-one mapping, but may involve a many-to-one sense lumping.

3 DanGram

DanGram is a rule-based, modular parsing system, using the Constraint Grammar (CG) formalism (Bick and Didriksen, 2015). For progressive linguistic annotation levels, contextual rules are used to map and disambiguate different types of token-based tags. Input to the morphosyntactic CG is provided by a pattern-based tokenizer and a lexicon-based morphological analyzer. The former establishes multi-word expressions (MWEs) covering e.g. names and complex equivalents to function words. The latter handles inflection, affixation and compound analysis[9]. After morphosyntactic annotation, another CG module assigns dependency relations based on syntactic function tags. At higher levels, extensive semantic lexica are used to support rules for named entity recognition (NER) and word sense disambiguation (WSD), as well as framenet structures and semantic role annotation.

In native format, each token will receive a readings line containing tags for the different annotation levels in space-separated, type-marked fields, or, in export format, as xml attributes. For instance, the lemma field is marked by a [...] bracket, syntactic function by a @-prefix and semantic roles by '§'. Apart from lemma, POS and inflection, the relevant fields for identifying the correct UID in COR are the semantic fields, in angular brackets, e.g. <H...> (human classes), <tool> or <food>, as well as framenet tags of the type <fn:know> or <fn:increase>.

4 ID Linking

4.1 Tag conversion

The linking program described here has a two-way purpose - on the one hand making it possible to enrich DanGram output with lexical information from future resources built around COR (e.g. dictionaries or encyclopedias), and on the other supporting users who want to build text processing applications around COR or to apply their COR-linked ontologies to e.g. news text for information retrieval by using the DanGram parser. The new tool has been implemented as an independent module, to be run after DanGram and working with the output of the parser as is, adding additional COR tags for matchable words. These tags colntain the COR identifier number (UID) as well as the lemma, POS and inflection tags provided by COR for this ID, with the same uppercase, English abbreviations used by DanGram itself, for better comparability. As a default, the inserted tags have the format <UID:lemma:tags>, with dots between tags, e.g. <COR.49032.115.01:lærer:N.UTR.S.DEF.GEN>

[8] https://framenet.dk

[9] For maximal (productive) coverage, the analyzer works with lemmas and morphemes, not a closed fullform list.

for the word *lærerens* ('the teacher's'). If the UID is the only desired information, DanGram tagging can be ignored, and the UID appended to tokens in running text, e.g. ***katten**_40150.111* ***åd**_38929.206* ***musen**_74798.111* (the cat ate the mouse).

In principle, a simple tag filter would allow the Linker to work with other parsers than DanGram, as long as they provide the same type and granularity of tagging. However, while the linker itself is robust enough to work with (filtered) input from other parsers, the quality of the latter would, obviously, have a bearing on the final result. Thus, a lack of tag types, in particular an absence of semantic, compound and MWE analysis, would not break the linker, but negatively affect performance, as would using a parser with a lower tagging accuracy than DanGram for the standard tags (POS/inflection).

It should be noted that even with a correct UID link, parser and COR tags will not necessarily match one-on-one. For instance, POS-mapping may be many-to-one (e.g. 3 DanGram pronoun classes, but only 1 in COR), and DanGram lemmas may have a number extension, superfluous in COR, given the latter's UIDs. Also, DanGram marks "not genitive" as nominative, while COR only specifies the genitive. The automatic linker program has to be robust enough to work in spite of such mismatches.

4.2 The matching algorithm

The basic linking algorithm first looks up each non-number, non-punctuation token in the COR database, acquiring a list of possible UIDs with their respective lemmas and inflection tags. Next, for each UID item on the list, the linker tries to match lemma and tags to equivalent tags found in the DanGram annotation for the word in question, computing a matching score. The reading with the highest score will get its UID selected and linked. In the straight-forward cases, POS and/or inflection will decide the issue. The word form *vise*, for instance, has four readings, and three meanings, in both COR and DanGram (DG), with matching lemma and POS, and a few morphological extra-tags in DanGram: NOM (nominative) and the portmanteau tags nG and nD for under-specified gender and definiteness, respectively[10].

```
4a) COR.30363.200.01, vise, V, INF, AKT
    DG: [vise] V INF AKT ('show')
4b) COR.46620.110.01, vise, N, UTR, S, IDF
    DG: [vise] N UTR S IDF NOM ('tune')
4c) COR.16117.302.01, vis, ADJ, S, DEF
    DG: [vis] ADJ nG S DEF NOM ('wise')
4d) COR.16117.303.01, vis, ADJ, P
    DG: [vis] ADJ nG P nD ('wise')
```

In the case of an adjective singular reading, for instance (e.g *den vise mand* – 'a wise man'), the correct (third) UID will receive 3 points - for lemma, pos and number -, while the adjective plural reading (fourth) will get only 2 points, for lemma and pos. The noun reading (second) will get 1 point, for number, and the verb reading (first) will fail on all tags, scoring 0. The inserted linking tag will then contain the highest-scoring UID and its COR tags.

4.3 Homograph levels and COR adaptation

The case of *vise* ('show', 'tune', 'wise') could be called a level-1 homograph in the sense that its meaning can be resolved by making use of lemma, POS, grammatical gender and inflection only. However, COR also contains about 400 cases of word-forms that are level-2 homographs, with two (or more) meanings that can be differentiated only by resorting to their pronunciation or inflectional paradigm as a whole (cp. section 2). As neither of the latter is marked in writing, but rather a manifestation of what is really a semantic feature (such as plural-less inflection paradigms for +mass nouns), the linker program has to make use of semantic clues provided and contextually disambiguated by the parser[11]. For about half of the level-2 homographs, DanGram itself distinguishes between two (or more) numbered sub-lemmas based on etymology or major meaning differences matching the COR distinction. In these cases, DanGram's semantic tags are simply bound to the individual sub-lemmas, as in the three noun options in the readings cohort for *ret* in (5).

```
(5)
"ret" <aquant> ADV ('rather')
"ret" <jshape> <jappro> ADJ
    ('right', 'straight')
"ret-1" <f-right> <conv>
    ('right [to]', '[the] law')
"ret-2" <food-c-h> N ('dish')
"ret-3" <inst> N ('court')
```

[10] In context, DanGram will specify these through agreement rules, but they still won't match a COR tag.

[11] Pronunciation variation without a difference in meaning (e.g. regional variation) does not lead to different word IDs in COR

```
"rette" <vt> V IMP ('correct!')
```

However, even without a sub-lemma, the remaining COR homographs can be matched, too - because DanGram in these cases assigns (and disambiguates!) the different semantic class tag on the same lemma. This is the case for the adjective *large*, which means 'big' with an English pronunciation (semantic class <jsize>), and 'generous' with a French pronunciation (semantic class <jpsych>), or the verb *hænge* ('hang'), which changes past tense inflection depending on transitivity and meaning. Here, the linker exploits DanGram's framenet tags, distinguishing between the intransitive <fn:spatial_configuration> (past tense *hang*) and the transitive <fn:put_spatial> (past tense *hængte*)[12]. For the linker to be able to use level-2 distinctions, however, they had to be entered into the COR database manually[13]. Thus, the COR version used by our linker program has been "lexicographically" enriched with additional information/tagging[14], adding DanGram sub-lemmas and their semantic classes (6), or just the latter (7), to all level-2 homographs in COR. These will then be matched to DanGram output by the linking algorithm in the same fashion as ordinary tags.

```
6a) COR.56686.110.01,brok-1,<sick>,N...
    ('hernia')
6b) COR.55539.110.01,brok-2,<sem-s>,N...
    ('complaining')
7a) COR.71663.120.01,marsvin,<Aich>,N...
    ('porpoise')
7b) COR.77141.120.01,marsvin,<Azo>,N...
    ('guinea pig')
```

The word ret (5) is an example where a three-way lemma distinction in DanGram has to be matched onto a two-way distinction in COR[15]. In this case, the fused sub-lemmas (*ret-2* and *ret-3*) are not used, because COR's lemma slot is a 1-item slot. Still, the distinction (and the link) will work based on semantic tags alone (8b).

```
8a) COR.43157.110.01,ret-1,
    <f-right><conv><f-cog>, N ...
    ('right [to]','law','[being] right')
8b) COR.43153.110.01,ret,
    <inst><food-c-h>, N ... ('dish')
```

4.4 Multi-part tokens

A special challenge for the linker were multi-part tokens with no equivalent entry in COR. Rather than ignoring such tokens as unlinkable, we opted to perform part-by-part, multiple linking, in order to facilitate NLP tasks such as machine translation, multi-lingual alignment or lemma-driven corpus searches.

For Danish, this issue is of particular importance, as productive compounding is an important aspect of Danish morphology. The process may involve morphological changes for the first part of a compound, such as stemming or the insertion of fuge letters, and a hyphen is only used in special cases. Over 10% of Danish tokens in running text involve compounding or affixation. In our evaluation text (section 5), 1.8% of tokens were words with a live compound analysis and no direct match in COR. An additional 1.4% were words without a COR match, but with a compound lexicon entry in DanGram.

In addition to compounds, tokenization can introduce multi-word expressions (MWEs) by fusing words that syntactically or semantically function as close-knit units. Lexically, an MWE makes sense where its meaning is not transparent from its parts. On the other hand, MWEs create compatibility issues, as there are no authorized closed lists available, and many NLP systems perform tokenization simply by space separation. Therefore, part-by-part linking is useful, as it allows the end user to easily (re)create fully COR-linked "space tokens" by splitting DanGram's MWEs in the Linker's output.

Both DanGram and COR contain closed-class MWEs, but DanGram contains more (table 2), because they help the parser to simplify syntactic

[12]Depending on the semantic type of linked object, prepositions and particles, DanGram distinguishes between nine further framenet meanings for this verb, all of which are grouped into the two COR meanings by the linker in a many-to-one mapping.

[13]It is a matter of interpretation, if this is seen as an enrichment of COR, or as a lookup-filter that is really a part of the linker program. As new words and loan words tend to enter a language with one, well-defined meaning, future level-2 homograph additions to core are unlikely, but they would need to be treated manually, with a linguist selecting those DanGram features necessary to make the homograph distinctions in COR.

[14]This way, for all non-trivial cases (i.e. where POS feature matching is not sufficient), the decision of what constitutes a linking match - and which features to target - has been taken by a linguist. In other words: what is automatic, is not the meaning/definition, but the matching

[15]In principle, DanGram could be used to enrich COR in such cases. However, the two resources are maintained independently and COR has a policy of following the official Danish spelling dictionary and not implementing purely semantic distinctions without pronunciation or paradigmatic support. Therefore, feedback to COR resulting from the work on our DanGram linker has so far only targeted simple errors and inconsistencies in the resource

structure; *i hvert fald* ('in any case'), for instance, is a shared MWE, while *i eftermiddags* ('yesterday afternoon') is DanGram-only. Open-class MWEs are very rare in COR and are limited to a few foreign expressions (e.g. *quiche lorraine*), place names (*Sankt Petersborg*) and first parts of hyphen-compounds (*dag til dag-levering* 'day-to-day delivery'). DanGram, on the other hand, annotates all complex named entities as MWE (e.g. person/company names, institutions and addresses), as well as anatomical expressions, species names and foreign MWE nouns based on pattern matches (e.g. when including English colour words). Because of this discrepancy between DanGram and COR, the linker is set to ignore MWE names without a complete COR match, as well as other "live" (i.e. heuristic, pattern-based) MWEs[16].

For the linker program, we used the same core strategy for matching compounds and MWEs: Failing a full match, the multi-part token is split into its components[17], which are then looked up in COR individually, using a prioritized matching order. COR contains about 8,000 separate UIDs for compound first parts and 75 prefix tags, which will get the highest priority in compound look-ups (COMP for the former, in 9a and 9c, or PREF for the latter). After that, first parts are looked up with the lemma (or sublemma) and POS provided by DanGram (9b). Failing that, or if DanGram only provides "prefix" as POS, they will be looked up as nouns, adjective or without POS, in that order. Second parts are looked up using the inflected fullform stripped of the first part, plus the provided part-lemma (or, in 9c, sublemma). For Danish compounds, the second part inherits POS, inflection tags and semantics from the overall analysis of the word, so ordinary tag scoring (cp. section 4.1) can be used (e.g. P in 9a, DEF in 9b and the <act-d>[18] semantic tag in 9c). For first parts, no separate semantic tag is provided by DanGram, so in a few cases (where there is polysemy but no sublemma), there is a theoretical risk of unresolved ambiguity.

[16]Using DanGram's <heur> tag to block part-by-part matching attempts for these MWEs.

[17]In the absence of a hyphen or space separator, we used DanGram's compound analysis, which provides first and second lemmas (or sublemmas), normalizing first parts as lemmas, independently of their morphological manifestation (cp. fuge-s in 9b and 9c).

[18]The action tag <act-d> represents one of several meanings of brud-2, each linked to another semantic tag and disambiguated by DanGram. In the modified COR entry all options are listed, and a match for any one of them will select *brud-2* ('rupture' etc.) rather than *brud-1* ('bride', 'weasel')

COR link tags for compounds are added to DanGram output in the same fashion as for single words, but with one, consecutively numbered, link tag for each part:

```
9a) havvindmøller  [havvindmølle]
('offshore turbine' - 'sea+windmill')
<1:COR.59371.129.01:hav:N.NEU.#COMP>
<2:COR.88335.112.01:vindmølle:N.UTR.#P.IDF>
<N:hav+vindmølle> <good-compound> <build>
    N UTR #P IDF NOM

9b) nervøsitetsindikatoren
[nervøsitetsindikator]
('fear gauge' - 'nervousness+indicator')
<1:COR.85108.110.01:nervøsitet:N.UTR.S.IDF>
<2:COR.98639.111.01:indikator:N.UTR.S.#DEF>
<N:nervøsitet~s+indikator> <good-compound>
    <ac> N UTR S #DEF NOM

9c) ægteskabsbrud [ægteskabsbrud]
('adultery' - 'marriage+infringement')
<1:COR.43176.129.01:ægteskab:N.NEU.#COMP>
<2:COR.48668.120.01:brud: <f-phys>.
    <event>.#<act-d>.<Lh>.N.NEU.S.IDF>
<N:ægteskab~s+brud-2> #<act-d>
    N NEU S IDF NOM
```

If one or more compound parts do not have a corresponding entry in COR at all (i.e. not even with a different POS), a dummy ID '0' and a dummy tag string 'X' is used instead. For noun or root parts, such gaps are relatively rare, but may occur, if the part in question is itself a compound (10a, *børne|litteratur* – 'child literature') or an MWE (***en=til=en**-programmet* – 'the one-on-one program'). A more serious problem is COR's limited coverage of prefixes (75 entries) and suffixes (6 entries). As long as the missing affix exists as a full-word entry, this will be used as a fall-back, but that is not possible for some otherwise quite productive prefixes like *special-* ('special', 10b) or suffixes like *-mæssig* ('-related', 10c).

```
10a) børnelitteraturfestival [=]
('child literature festival')
<1:COR.0:børnelitteratur:X>
<2:COR.97204.110.01:festival:N.UTR.S.IDF>
<N:børnelitteratur+festival> <occ>
N UTR S IDF NOM

10b) specialgeotekniske [special..nisk]
('specialized geotechnical')
<1:COR.0:special:X>
```

```
<2:COR.22830.302.01:geoteknisk:ADJ.S.DEF>
<F:special+geoteknisk> <jdomain>
ADJ nG P DEF NOM

10c) momsmæssig [momsmæssig]
('VAT-related')
<1:COR.41058.119.01:moms:N.UTR.COMP>
<2:COR.0:mæssig:X>
<N:moms+mæssig><jtype> ADJ UTR S IDF NOM
```

Unless they match as a whole (11a), MWEs are also looked up part by part (11c). But unlike compounds, there is no lemma or POS available for MWE parts, only the individual tokens from the MWE chain. Also, unlike English noun chains, Danish MWEs have a more varied (and un-tagged) internal syntactic structure, so it is unsafe to let the last part inherit POS or other tags from the MWE as a whole. Our matching algorithm has to reflect this lack of (safe) information. Safest are the separate "in-MWE" UID entries listed by COR for some words (270). Although the MWEs themselves are not provided in these "in-MWE" entries, there is no COR ambiguity across MWEs, so if an MWE matches such an entry, it is assumed to be a correct link, even if the string also exists in COR as a full word. The "in-MWE" entry *rette*[19], for instance, can be used for the 2nd part of the MWE *med rette* ('justifiably', literally 'with right'), discarding the verb infinitive reading 'to correct' (11b).

```
11a) frem=for [=] ('rather than')
<COR.04976.930.01:frem=for:MWE>
<complex> PRP

11b) med=rette [=] ('justifiably')
<1:COR.04087.960.01:med:MWE-PART>
<2:COR.04080.960.01:rette:MWE-PART>
<complex> ADV

11c) i=stedet=for [=] ('instead of')
<1:COR.00852.880.01:i:PRP>
<2:COR.44318.121.01:sted:N.NEU.S.DEF>
<3:COR.00093.880.01:for:PRP>
<complex> PRP
```

If no "in-MWE" entry is found, the linker then looks for ordinary entries (11c), beginning with prepositions and articles, followed by other function word classes, and finally the content word classes, nouns first. As an exception, adjective matches are prioritized higher than nouns for first parts, because Danish NP word order places adjectives to the left of nouns. This matching hierarchy correctly handled the typical adverbial MWEs of the type PRP+N+PRP (e.g. *i stedet* for -'instead of'), but failed for about sixty[20] more idiosyncratic closed-class MWEs, where one (or sometimes two) parts were POS-ambiguous and resolved incorrectly, e.g. *om* in the MWE conjunction *om ikke* ('if not'), where the ordinary POS hierarchy would have chosen a preposition reading for om rather than the correct conjunction reading. This was solved by adding a small POS lookup table for problematic closed-class MWEs. The table is used after "in-MWE" matches, but before ordinary POS matches.

5 Evaluation

To evaluate both overall performance and linking accuracy, we generated random excerpts from DSL's general period corpus *Korpus 2010*[21], covering five different text types for lexical diversity: blog, parliament, special interest home page, general news and financial news, with 11,099 raw tokens in all. The texts were annotated with DanGram both morphosyntactically and semantically, i.e. including framenet annotation and word sense disambiguation for nouns and named entities. After DanGram's name and MWE tokenization there were 8,399 parse tokens (incl. 1,112 punctuation tokens).

Tables 1 and 2 show, for each relevant part of speech, the percentage of tokens that could be automatically linked to COR, both for ordinary tokens (table 1) and for multi-word-expressions (table 2). For the closed word classes (PRP, ART, PRON and K) and for adverbs (ADV), coverage was 100% in both cases.

Among the open word classes, verbs had a better coverage (99.6% for full matches) than nouns (97.1%) and adjectives (97.4%). Also, the latter had a greater share of out-of-vocabulary (OOV) compounds, that had to be matched part-by-part, which led to a certain amount of partial matches (first or second part only) . Unmatchable parts were, for instance, prefixes or (***u***|*kontrolleret* 'uncontrolled'), suffixes (*moms*|***mæssig*** – 'VAT-related), names (***Pisa***|***testen*** – 'the Pisa test') or numerical

[19] The form is an archaic dative of the noun *ret* ('right'), that does not exist in modern Danish outside of fixed expressions, and therefore does not have an ordinary inflection entry in COR.

[20] When checking all of DanGram's closed-class MWEs

[21] https://korpus.dsl.dk/resources/details/korpusdk.html

Table 1: Coverage for non-MWE tokens, direct or through compound parts (%)

POS[22]	direct (full)	comp full	comp partial	all full	all partial
N	86.9	10.2	1.8	97.1	98.9
V	99.3	0.4	0.1	99.6	99.7
ADJ	96.9	4.8	1.9	97.4	99.3
ADV	100	-	-	100	-
PROP	25.5	0	2.8	25.5	28.4
PRP	99.9	-	-	99.9	-

parts (*63-årig* '63-year-old), abbreviations (*C20-indekset* – 'the C20 index') or English parts. In a few cases, DanGram provided a 2-way compound split where one of the parts was itself a compound that couldn't be matched (*børne|litteratur||festival* – 'childrens's literature festival'). Finally, proper nouns and numerals had a low coverage simply because COR contains only 700 proper nouns – all place names – and only numerals that are written with letters. Overall coverage for non-punctuation was 97.4%, or 99% when not counting proper nouns.

For closed-class MWEs there was full coverage (table 2), but as DanGram contains more MWEs than COR, only about 1/3 were direct MWE matches (2nd column), the rest were part-by-part matches (3rd column). In absolute terms, the difference is most marked for MWE prepositions, and least marked for MWE adverbs.

Table 2: Coverage for closed-class MWE tokens, as a whole or part-by-part (%)

POS	MWE as a whole (full)	MWE all parts	MWE partial	all full
ADV	45.9	54.1	0	100
PRP	19.6	80.4	0	100
PRON	8.3	91.7	0	100
K	25.0	64.3	0	100
All	35.7	64.3	0	100

Obviously, in addition to coverage, accuracy is important, and because of sense and paradigm ambiguities, and especially for the major word classes, nouns and verbs, a link to a COR entry with the right part-of-speech is not necessarily correct. We therefore checked all links manually for possible errors[23]. Here, a distinction should be made between text-to-COR performance, including DanGram annotation errors propagating as COR-link errors, on the one hand, and linking-only errors on the other, i.e. correct DanGram analyses still leading to a wrong COR entry. The latter type of errors proved to be extremely rare (< 0.1%, first parts in 1 MWE and 1 compound, plus 1 misspelling), but even text-to-COR accuracy was satisfactory, given the fact that linking failures were mostly due to gaps in COR rather than analysis or linking failures (table 3).

Table 3: Text-to-COR - DanGram errors (column 2, rows 2-6), linking errors and COR gaps

Error type	row sums	link match	non-COR class	link error	COR gap
POS error	18:	12	4	1	1
morph error	2:	2			
sem-class error	2:	2			
tokeniz. error	4:	1	3		
comp. error	2:			1	1
link error only				1	
COR error					76
no COR			325		80
Column sums	28	17	332	3	158
% of words	0.3	0.2	4.0	0.0	1.9

4% of all words were outside of COR's scope (numbers, numerical expressions and most proper nouns[24], while 1.9% were COR gaps that could be addressed by improving COR. Of these, about half had no COR entry at all, half were missing an entry for the correct POS, but offered another ID for the word form in question, that could be used as a fall-back[25]. Linking-relevant DanGram errors amounted to only 0.3%, mostly POS errors, but also a few tokenization, inflection and compound analysis errors, as well as two higher-level, semantic subclass errors. A quarter of the DanGram errors concerned non-COR word classes, in

[23]This was carried out as a double-pass inspection, in-house, by one specialist, facilitated by the fact that COR has definition fields for ambiguous entries

[24]DanGram has a high precision for these word classes, and there were only two cross-class false positives, both wrongly tagged PROP - one adjective (that could have been linked) and one noun (not in COR).

[25]This POS gap problem concerned only a few, but frequent word forms. For instance, der was not listed as a relative pronoun, but only as an adverb, and a couple of common adverbs (*sådan* 'this way' and *meget* 'very') were only listed as adjectives.

most of the others (0.2%, i.e. 2/3 of the DanGram errors) the Linker simply ("correctly") assigned a corresponding, wrong UID link. In combination, COR gaps (1.9%), DanGram errors (0.3%) and pure linking errors (< 0.1%) amounted to a text-to-COR failure rate of 2.2%.

6 Conclusion

We have shown how the output of a morphosyntactic and semantic parser with compound analysis (DanGram) can be linked to unique word identifiers by matching annotation tags such as lemma, POS and semantic class with corresponding information in the target resource (COR). In a random text evaluation, 97.8% of all non-number, non-name words could be matched to a correct COR entry. As most of the link failures were not caused by the linking mechanism as such, but by coverage issues, performance should automatically increase with future editions of COR. Parser errors were a smaller issue, and here, too, future improvements should automatically translate into better linking.

Limitations

The good performance of the parser is unlikely to be evenly distributed and likely to be lower if evaluated separately for level-2 homographs only. Given the fact that DanGram uses the same rule-based strategy for both morphosyntax and WSD, alternative methods for this sub-task, in particular word embeddings (Iacobacci et al., 2016), should be compared, possibly by boot-strapping training data with DanGram output. Depending on the applicational uses of COR, it would make sense to add a kind of "encyclopedic" section for proper nouns, for instance by assigning UIDs to (Danish) Wikipedia entries, allowing a more integrated use of the resource in tasks like information extraction. For many applications it would also be extremely useful to link spelling variations and frequent misspellings to the underlying, correct COR entry[26]. Ultimately, of course, it is a design or resource allocation decision whether normalization should be addressed "live" at the parser level, as is the case for DanGram, or whether it (also) should be supported lexically in COR.

[26]For frequent variants, COR's third UID field, reserved for historical spelling changes, could be used for this purpose. For a wider, unsystematic, inventory of spelling errors, linking an external resource would make more sense

Ethics Statement

As it does not use any training or personal data, questionnaires or user logs, this work does not raise any ethical concerns regarding GDPR. As a rule-based system it also does not need much computing power, neither during development nor as a service, making for a very small environmental footprint. In the same vein, no underpaid student or Mechanical Turk labour has been exploited to produce training data.

Acknowledgements

We appreciate the work that has gone into building COR, and are grateful to DSN for making the resource publicly available.

References

Eckhard Bick. 2011. A FrameNet for Danish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 34–41, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Eckhard Bick and Tino Didriksen. 2015. CG-3 — Beyond Classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Christina Dideriksen, Peter J. Hansen, and Thomas Widmann. 2022. Det Centrale Ordregister. *Nyt fra Sprognævnet*, Oktober 2022.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Sanni Nimb, Bolette S. Pedersen, Nathalie C. Hau Sørensen, Ida Flörke, Sussi Olsen, and Thomas Troelsgaard. 2022. COR-S – den semantiske del af Det Centrale OrdRegister (COR). *Lexico Nordica*, 29:75–97.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*, volume 9041, pages 3–16. Springer, Cham.

Bolette S. Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual

dictionary. In *Language Resources and Evaluation (2009)*, pages 269–299. ELRA.