# The Central Word Register of the Danish language

**Thomas Widmann**

The Danish Language Council

**Abstract**

*Det Centrale Ordregister* ("The Central Word Register", *COR*) is a unique and innovative lexical database for the Danish language. Developed by the Danish Language Council, the Society for Danish Language and Literature and the Centre for Language Technology at the University of Copenhagen, with funding from the Agency for Digital Government, the COR assigns unique identification numbers to every lemma and form of the Danish language.

At the heart of the COR lies *Retskrivningsordbogen*, the official orthographical dictionary of Danish, which provides the foundation for the unique identification numbers. The Danish Language Council will update this basis whenever the orthography changes, publishing the changes compared to the previous version, ensuring that the COR will always reflect the orthography of the day while ensuring that existing resources will continue to function even when the orthography changes.

The COR is divided into three levels, with Level 1 corresponding to the orthographical dictionary, Level 2 encompassing additional resources from professional language bodies and Level 3 comprising all other resources, with no restrictions on who can contribute. Version 1.0 of Level 1 was released by the Danish Language Council in September 2022. The Society for Danish Language and Literature and the Centre for Language Technology are currently working on adding a semantic component on Level 2.

The primary goal of the COR is to create a common key that enables more efficient reuse of language resources, similar to the way Denmark's Central Person Register (CPR) allows different databases containing information about the inhabitants of Denmark to communicate with one another.

The COR database can be easily accessed through a downloadable CSV file or an API, allowing developers to retrieve ID numbers, lemmas, and forms in either CSV or JSON format, providing a great example of invisible lexicography.

The project also opens up new possibilities for historical lexicography, as the Danish Language Council intends to make its previous orthographical dictionaries available in COR format, enabling users to track the evolution of the language over time, to study historical texts in a more accurate way and to modify NLP software to work on such texts.

Another topic is the development of COR linkers (programs that will assign the correct COR number to every word in a text) and how these are effectively solving the problems of part-of-speech tagging and homograph resolution at once. An example of a COR linker is the Danish Language Council's CLINK project.

Another aspect of the COR is the ability to use crowdsourcing in lexicography. Users can contribute their own data and insights, simply by publishing their data with added COR ID numbers. This fosters greater collaboration and enables the creation of a plethora of rich, dynamic resources for the Danish language.

Finally, the article will explore the benefits and potential applications of the COR and discuss the exciting possibilities this creates for the future of the Danish NLP and language research.

**Keywords:** lexical database; orthography; Danish language; historical lexicography

## 1. Introduction

A common challenge when working with lexicographic and computational linguistic resources is the lack of compatibility. Each resource has its own approach to issues such as homonym resolution, part-of-speech tags and lexical coverage. Furthermore, licensing issues can make it exceedingly difficult to determine which resources can be legally reused in a project.

This problem is particularly pronounced for smaller languages, as the initial cost of undertaking any computational linguistic project becomes increasingly prohibitive for smaller actors to initiate.

Although numerous electronic resources for Danish exist—including machine-readable dictionaries, corpora, and taggers—reusing them can be challenging because they are not based on the same fundamental resources, nor do they share database keys or similar attributes. Consequently, the development of language technology for Danish has become more difficult than necessary.

The solution has been known in other areas for years: using a shared database key that facilitates the merging of diverse databases. For example, Denmark has a system called the *Centrale Personregister* (CPR), which assigns a unique identification number to each resident of Denmark. This system offers significant practical benefits; for instance, when an individual changes their address, they only need to inform the local council, and all relevant parties (e.g., the tax authorities, the health system, and the bank) are notified automatically.

Inspired by the CPR, we decided to address this issue by creating a new resource framework: The Central Word Register (Danish: *Det Centrale Ordregister*: COR).

The COR was supported by the Danish Agency for Digitisation, and the project involves the Danish Language Council, the Society for Danish Language and Literature, and the Centre for Language Technology at the University of Copenhagen. It assigns unique identification numbers to all lemmas and word forms in Danish. The Danish Language Council is responsible for the basic register, comprising orthography and morphology for the vocabulary covered by *Retskrivningsordbogen*, the Danish Orthographic Dictionary (Dansk Sprognævn, 2012). This basic register, which we will call $\text{COR}_1$ in the following, was launched in September 2022 and is accessible at ordregister.dk.

In this article, we will first describe the structure of the COR, outline the basic resource's structure, and demonstrate how new COR resources can be added. We will then explore various lexicographic applications, with a particular focus on the Danish Language Council's website $\text{RO}^{\text{hist}}$, which enables comparisons of different historical orthographic dictionaries. Subsequently, we will discuss COR linkers (programs that automatically assign COR identification numbers to all words in a running text) and, finally, examine invisible

lexicography and crowdsourcing. Our aim is to provide readers with both the motivation to begin utilising the COR and the practical skills to do so.

## 2. Structure and Components of the COR

### 2.1 The Orthographical Foundation: *Retskrivningsordbogen*

*Retskrivningsordbogen* (Dansk Sprognævn, 2012) is the official reference for Danish language orthography. Published by the Danish Language Council (Dansk Sprognævn), it serves as the primary authority on Danish orthography in accordance with the Danish Orthography Act (LBK 332).

The dictionary is regularly updated to reflect the latest changes in Danish orthography, ensuring it remains current and accurate. The most recent edition was published in 2012; however, new words are added annually to its online version, keeping it up-to-date with contemporary language usage. The latest update was in November 2022.

The categorisation of the basic vocabulary in $\text{COR}_1$ into lemmas is based on *Retskrivningsordbogen*. As a result, it follows the same principle for what constitutes a lemma (Dansk Sprognævn, 2012: 13f):

> Opdelingen i opslagsord er principielt uafhængig af ordenes betydning. Det bevirker at ord med forskellig betydning er slået sammen i ét opslagsord hvis de i øvrigt har samme stavemåde, udtale, ordklasse og bøjning, og hvis de indgår i sammensætninger på samme måde.[1]

From this quotation, it is evident that neither the semantics nor the etymology is considered when determining what a lemma is.

The COR can be regarded as an enhanced and optimised version of *Retskrivningsordbogen*, specifically tailored for natural language processing purposes. Building upon the foundation provided by the dictionary, the COR aims to facilitate and improve the development and utilisation of Danish language technologies. However, there are some key differences between the two:

1. *Retskrivningsordbogen* is designed for humans; the COR is designed for easy use by computer programs.
2. *Retskrivningsordbogen* does not include all inflected forms (and of the ones that are present in the data, only a few are displayed in the book); the COR offers more comprehensive coverage.
3. *Retskrivningsordbogen* comes with the restriction that it cannot be used to create dictionaries; the COR can be used without any restrictions.
4. *Retskrivningsordbogen* contains a good number of usage examples; the COR has none.
5. *Retskrivningsordbogen* has references to its rule appendix; the COR has none.
6. *Retskrivningsordbogen* has more and longer glosses than the COR.[2]

[1] "In principle, the division into headwords is independent of the words' meaning. This leads to the merger of words with distinct meanings into a single headword, provided they share the same spelling, pronunciation, word class, and inflection, and if they are employed in compounds in the same way."

[2] The COR provides glosses solely for the purpose of disambiguating homographs. For instance, `COR.70558` *kalk* carries the gloss "et mineral" ("a mineral"), whereas `COR.77824` *kalk* is glossed "krus el. bæger"

## 2.2 The structure of the COR

### 2.2.1 Unique Identification Numbers

In the following, we will first describe the structure of $COR_1$ and then discuss the general structure for other COR resources.

In $COR_1$, all lemmas found in *Retskrivningsordbogen* and their forms are assigned unique identification numbers. These consist of the prefix 'COR' followed by a 5-digit index number indicating the specific lemma. For example, `COR.56746` corresponds to the lemma *avocado*.

To specify a particular form of a lemma, a three-digit grammatical code is appended to indicate the part of speech and inflection of the word. For example, `COR.56746.111` corresponds to the singular definite form of this common-gender noun, i.e., *avocadoen/avokadoen*. A list of these grammatical codes can be found on ordregister.dk.

In addition to the lemma and grammatical code, a two-digit code is added to indicate orthographical variation. This ensures that each ID number is unique. For example, `COR.56746.111.02` corresponds to the form *avokadoen*. (Both forms, *avocadoen* and *avokadoen*, are co-official, and neither is preferred.)

The ID numbers are arbitrary and are not assigned alphabetically. The lemma indices in the Retskrivningsordbogen range from 0 to 99,999, and they are not assigned based on alphabetical order. For practical reasons, the interval is divided by word class. For example, adjectives have indices between 15,000 and 29,999, and nouns have indices between 40,000 and 99,999. However, this division is not a formal requirement, and other COR resources are not expected to follow this pattern.

Here are the actual contents of $COR_1$ for *avocado*:

The grammatical code in column 4 exhibits a one-to-one correspondence with the second part of the numerical code. For instance, 110 consistently translates to *sb.fk.sg.ubest*. The final column displays a 1 if the form is derived from the dataset underpinning *Retskrivningsordbogen* and is consequently part of the official norm. Conversely, a 0 signifies that the form has been auto-generated, and users should exercise caution when utilising these forms.

("mug or cup"); the former is the same in *Retskrivningsordbogen*, but the latter bears the glos "krus el. bæger til altervin" ("mug or cup for sacramental wine") in the dictionary – the latter half has been omitted from the COR because it is not needed for disambiguation. Additionally, `COR.82322` *kalkbrud* "limestone quarry" has a gloss in the dictionary to aid the user identify the word, but it does not have one in the COR because the lemma has no homographs.

| ID | lemma | gloss | gram. code | form | norm |
|---|---|---|---|---|---|
| COR.56746.110.01 | avocado | – | sb.fk.sg.ubest | avocado | 1 |
| COR.56746.110.02 | avocado | – | sb.fk.sg.ubest | avokado | 1 |
| COR.56746.111.01 | avocado | – | sb.fk.sg.best | avocadoen | 1 |
| COR.56746.111.02 | avocado | – | sb.fk.sg.best | avokadoen | 1 |
| COR.56746.112.01 | avocado | – | sb.fk.pl.ubest | avocadoer | 1 |
| COR.56746.112.02 | avocado | – | sb.fk.pl.ubest | avokadoer | 1 |
| COR.56746.113.01 | avocado | – | sb.fk.pl.best | avocadoerne | 1 |
| COR.56746.113.02 | avocado | – | sb.fk.pl.best | avokadoerne | 1 |
| COR.56746.114.01 | avocado | – | sb.fk.sg.ubest.gen | avocados | 1 |
| COR.56746.114.02 | avocado | – | sb.fk.sg.ubest.gen | avokados | 1 |
| COR.56746.115.01 | avocado | – | sb.fk.sg.best.gen | avocadoens | 1 |
| COR.56746.115.02 | avocado | – | sb.fk.sg.best.gen | avokadoens | 1 |
| COR.56746.116.01 | avocado | – | sb.fk.pl.ubest.gen | avocadoers | 1 |
| COR.56746.116.02 | avocado | – | sb.fk.pl.ubest.gen | avokadoers | 1 |
| COR.56746.117.01 | avocado | – | sb.fk.pl.best.gen | avocadoernes | 1 |
| COR.56746.117.02 | avocado | – | sb.fk.pl.best.gen | avokadoernes | 1 |

Other COR resources ought to adhere to the same general syntax, which includes:

1. The resource name.
2. The lemma id.
3. Any required subdivisions. It is not necessary for these to match $COR_1$; the need for subdivisions, along with their quantity and digit count, is specific to each resource. However, this information must be explicitly detailed on the website.

### 2.2.2 The COR resource landscape

There are three levels of COR resources:

- Level 1 corresponds to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
- Level 2 will contain a plethora of resources from professional language environments in Denmark, specifically members of the Danish Language Council's board of representatives. Additional resources will be included over time. At the present time, it comprises a resource containing supplementary lemmas from the Danish Dictionary (published by the Society for Danish Language and Literature [DSL]); this resource is called `COR.EXT`. Level 2 will also feature a semantic extension to the basic register produced by DSL and the Centre for Language Technology at the University of Copenhagen (CST). For more information on their work developing this semantic component, see Nimb et al. (2022). Prefix: `COR.NAME` (where `NAME` is an alphanumeric identifier).
- Level 3 encompasses all other resources without restrictions. Any relevant project can be assigned a prefix and an ID range if one contacts the Danish Language Council. Prefix: `COR.OPEN.NAME` (where `NAME` follows the same rules as above).

Besides a name, each resource is allocated a series of unique ID numbers. These numbers should be utilised in combination with existing ones in other resources on the same or lower levels. For example, any COR resource that indexes *avocado/avokado* should ideally use the number 56746 for it. A hypothetical dictionary of common spelling errors called `COR.OPEN.ORTHERROR` should thus define the common misspelling *advokado* as `COR.OPEN.ORTHERROR.0056746` (preferably padding with extra zeros like this to match other ID numbers in its number series). Existing ID numbers should be used for perfect matches and mere orthographic variation; resources should use their own numbers primarily for non-existing lemmas and ones that do not correspond one-to-one with an existing entry. For example, if a resource needs to index *avocado* and *avokado* separately, it should allocate new numbers to both; the same applies if a resource needs to index *avocado/avokado* together with the antiquated term for this, *advokatpære*.

### 2.2.3 Relations

In the Central Word Register (COR), *relations* act as a mechanism to establish connections between lemmas and word forms, clarifying their associations with one another.

These relations facilitate the organisation and search for data within COR, enabling users and developers of language technology tools to trace connections between lemmas and word forms in order to identify related linguistic components.

Various types of relations can be defined, including:

| Abbreviation | Definition |
|---|---|
| fus | fusion of two or more COR indexes |
| rep | replaced by one or more COR indexes |
| spl | split into two or more COR indexes |
| sms | compound of two COR indexes (for compound words) |
| hyr | hypernym (superordinate concept) for two or more COR indexes |
| hyp | hyponym (subordinate concept) for another COR index |
| rim | rhyme (for rhyming dictionaries) |

The examples above demonstrate the versatility of relations. Each resource can define its own relations; the basic register does not currently use any.

To exemplify this, consider the modern lemmas *fjeder* "(metal) spring" and *fjer* "feather"; they share the same etymology, and as recently as in the orthographical dictionary of 1923 (Glahder, 1923), there was free variation between the forms *Fjeder* and *Fjer* regardless of the meaning. To add this 1923 dictionary to COR, we would resolve this issue by creating a new ID number (in the following 4008020) and adding information about its relation to the two modern entries.

| COR-id | lemma | form | relation |
|---|---|---|---|
| COR.70131 | fjeder | fjeder | |
| COR.70759 | fjer | fjer | |
| COR.DRO1923.4008020.x.01 | Fjeder | Fjeder | fus:70759+70131 |
| COR.DRO1923.4008020.x.02 | Fjeder | Fjer | fus:70759+70131 |
| COR.DRO1923.0070131 | | | rep:4008020 |
| COR.DRO1923.0070759 | | | rep:4008020 |

Here, `rep:4008020` means "this ID number has been replaced by 4008020", and `fus:70759+70131` means "this ID number is a fusion of 70759 and 70131".

The same applies if two historical lemmas correspond to one modern one – for example, the current dictionary only has one lemma *skade* for both the bird (the magpie) and the fish (the common skate) because, as mentioned above, neither semantics nor etymology is taken into account when determining what a lemma is; however, in the 1955 dictionary (Dansk Sprognævn, 1955) there were two corresponding lemmas:

| COR-id | lemma | glosse | relation |
|---|---|---|---|
| COR.45662 | skade | en fugl; en fisk | |
| COR.RO1955.4011080 | skade | en fugl | rep:45662 |
| COR.RO1955.4011081 | skade | en fisk | rep:45662 |
| COR.RO1955.0045662 | | | spl:4011080+4011081 |

By establishing such relations, the COR can effectively manage the connections between historical and modern lemmas, enhancing the overall organisation and retrieval of linguistic data.

## 3. Accessing and utilising the COR

COR's master register and certain other resources can be accessed in two ways:

The entire register can be downloaded as a CSV file from ordregister.dk. This allows for working with, among other things, the master register offline and integrating it into one's own systems.

There is also an online interface that can be used to search the master register's data and access information on lemmas and word forms. This information can be displayed in HTML or accessed from a program in either CSV or JSON format. This interface can be found at the same address: ordregister.dk.

For instance, the following three lines of Python will look up the lemma given an ID number:

```
url = "https://ordregister.dk/id/COR." + str(id) + ".json"
data = json.loads(urlopen(url).read())
word = data['lemma']
```

Most other resources will only have lists of defined lemmas and forms available on ordregister.dk. There will then be a link to the URL from which the resource can be accessed (if it is publicly available).

# 4. Applications of the COR

## 4.1 Historical Lexicography

The Danish Language Council's RO$^{hist}$ project (rohist.dk) is a search engine that allows users to compare Danish orthographical dictionaries from 1872 to 2012.

Work is ongoing to expand RO$^{hist}$ with all Danish historical orthographical dictionaries and other orthographic resources, such as normative textbooks. For example, we are currently converting Ove Malling's textbook *Store og gode Handlinger af Danske, Norske og Holstenere* (Malling, 1777) into a dictionary that can be added to RO$^{hist}$ (cf. Hartling & Widmann (2020)).

The ID numbers in COR$_1$ correspond to the latest edition of *Retskrivningsordbogen*, but we plan to also assign COR numbers to the historical dictionaries in RO$^{hist}$. These dictionaries will be level 2 resources and will thus have their own prefix and ID number range.

The same ID number will be reused if the lemma is the same, even if the spelling has changed. A word like *fråse* (Dansk Sprognævn, 2012) will therefore have the same COR number as *frådse* in Dansk Sprognævn (1996) and as *fraadse* in Grundtvig (1872):

| COR id | lemma |
|---|---|
| COR.37337 | fråse |
| COR.RO2001.37337 | fråse |
| COR.RO1996.37337 | frådse |
| COR.DHO1872.37337 | fraadse |

This will simplify the implementation of RO$^{hist}$ considerably. When searching for *fråse* in the future RO$^{hist}$, one would simply need to find the COR number (here 37337) and then determine whether this is defined in the historical dictionaries through a simple lookup.

The existing links between the dictionaries in RO$^{hist}$ will form the basis for this work. We will therefore take the relational database underlying RO$^{hist}$, analyse the data, and assign historical COR ID numbers based on this analysis. This also means that if an error is found in RO$^{hist}$ – for example, if a historical spelling has been linked to the wrong lemma – one would simply need to correct the COR ID number in the historical dictionary where the error occurred.

## 4.2 COR Linkers

In corpus linguistics and other computational linguistic applications, developing programs that assign the correct COR id (including the grammatical code) to each word in a text is essential. These programs are called COR linkers. With a text COR-linked, generating a part-of-speech-tagged text becomes straightforward since all necessary information is contained in the grammatical code. Moreover, COR linking allows for disambiguation of

homographs and makes the annotated text suitable for various NLP tasks, such as spelling and grammar checking, speech synthesis, machine translation, and dialogue systems.

Consider the Danish noun phrase "to kendte russiske historikere" ("two renowned Russian historians") as an example of how to COR-link a text. Three of the four words have more than one potential match in $COR_1$ (in the table below, the correct link (match) is marked in bold):

| **Token** | **Meaning in English** | **Ambiguous COR ids** |
|---|---|---|
| to | two | **COR.01528.600.01** (numeral) |
| | | COR.30835.200.01 (vb. inf.) |
| | | COR.30835.209.01 (vb. imp.) |
| kendte | renowned (plur.) | COR.18159.302.01 (adj. sing. det.) |
| | | **COR.18159.303.01** (adj. plur.) |
| | | COR.30330.206.01 (vb. past act.) |
| | | COR.30330.214.01 (vb. past part. sing. det.) |
| | | COR.30330.215.01 (vb. past part. plur.) |
| russiske | Russian (plur.) | COR.22261.302.01 (adj. sing. det.) |
| | | **COR.22261.303.01** (adj. plur.) |
| historikere | historians | **COR.58774.112.01** (noun plur.) |

At the Danish Language Council, a COR linker project called CLINK is currently being developed. It is an input-output automaton that accepts a tokenised text (or a full corpus) as input, expands the input to its maximal COR-linking, filters away irrelevant links (for homographic tokens only), and delivers a minimally linked version as output. CLINK uses several strategies to achieve an output as close to the optimal linking as possible.

The fundamental idea is to use three different analytical strategies: LSYN (local syntax), CTXT (context), and FREQ (frequency), each implemented as a stand-alone module. Input and output formats are the same: Each module reads a well-linked text as input and writes a well-linked text as output. CLINK modules can only *remove* links but cannot alter the input otherwise. Intuitively, each time a well-linked text passes through a CLINK-module, some of its lexical ambiguity is eliminated, altering the decision basis for the following iteration.

The LSYN module makes congruence-based decisions (sentence-internally), while CTXT is based on semantic triggers and long-distance associations. FREQ uses lookups in a frequency table (including bigrams and trigrams); it always outputs a minimally linked text (with a single link per token), guaranteeing a recall of 1.0 (but usually a less than satisfactory precision). In contrast, the recall of the other modules depends critically on the amount of 'triggering' contexts in the input text, and they typically show a very high precision at the expense of a low recall. Hence, FREQ is located as the last module in the CLINK pipeline (as the fallback strategy), ensuring that the output is indeed minimally linked.

As the input and output formats are the same in all modules, the modules can be swapped freely. One can also insert new modules. For instance, the possibility of creating an AI-based module (using TensorFlow) is currently being investigated.

### 4.3 Crowdsourcing and Invisible Lexicography

Companies and individuals can effortlessly contribute their own resources to the COR by applying for a unique prefix and number series, and then incorporating COR ID numbers into their data as previously described. This process will gradually transform COR into a vast, largely crowdsourced resource, particularly if new contributions are distributed under an open-source licence.

We encourage those who have requested a prefix and number series to publish their lemma lists on ordregister.dk, simplifying the process of finding relevant data. For instance, by visiting the website, one can identify who has defined data for ID 56746 (*avocado/avokado*).

The development of COR exemplifies the concept of invisible lexicography (using lexical data without users realising they are employing a "dictionary") by making lexical data machine-readable and integrating it seamlessly into various contexts. By assigning unique identification numbers to every lemma and form in the Danish language, COR provides a common key that facilitates more efficient reuse of language resources.

Fundamentally, COR is a developer-oriented feature with the potential to impact a broad range of user-facing applications, such as spellcheckers, translation services, and search engines. However, many users interacting with these tools may be unaware of the underlying database or the efforts involved in creating it.

COR represents an exciting advancement in the field of lexicography and language technology. By rendering lexical data machine-readable and accessible to developers, COR has the potential to revolutionise the way we process and analyse language. It also offers new opportunities for collaboration and crowdsourcing, as users can contribute their own data and insights to the database.

We hope that many will release COR-linked corpora and additional resources that further enhance the overall utility of COR.

## 5. Benefits and Applications of the COR

The COR offers a variety of advantages and potential applications within the realm of Danish language research and natural language processing. In this section, we shall outline the primary benefits of the COR and its multifarious applications.

**Enhanced Resource Reusability and Collaboration:** The COR encourages different resources to use the same lemma ID numbers, potentially adding *relations* to further describe the relationships between them. This approach will hopefully lead to enhanced resource reusability and collaboration.

**Support for Historical Lexicography:** The Danish Language Council's intention to make previous orthographical dictionaries available in COR format will enable users to trace the evolution of the language over time. This capability allows for more precise study of historical texts and the adaptation of NLP software to work on such texts.

**Efficient POS Tagging and Homograph Resolution:** The development of COR linkers, exemplified by the Danish Language Council's CLINK project, assigns the

correct COR number to every word in a text, thereby addressing part-of-speech tagging and homograph resolution concomitantly. This development simplifies language analysis and significantly bolsters the accuracy and efficiency of NLP applications.

**Crowdsourcing in Lexicography:** The COR permits users to contribute their own data and insights by publishing resources with added COR ID numbers. This approach encourages broader community participation in the development of the Danish language, culminating in a more comprehensive and diverse database that benefits both researchers and NLP practitioners.

**Uncomplicated Access and Integration:** The COR database can be accessed via a downloadable CSV file or an API, allowing developers to effortlessly retrieve ID numbers, lemmas, and forms in either CSV or JSON format. This streamlined access promotes the concept of "invisible lexicography", enabling seamless integration with a variety of applications and tools.

In conclusion, the COR provides a groundbreaking foundation for the Danish language, augmenting collaboration, streamlining processes, and promoting further research and development. The benefits and applications of the COR extend beyond academia, opening up new possibilities for natural language processing, historical analysis, and the future of Danish language studies.

## 6. Future Prospects and Conclusion

As we have demonstrated throughout this article, the Central Word Register (COR) offers significant benefits and potential for applications Danish language research and natural language processing. In this concluding section, we will briefly discuss future prospects for the COR and summarise the key points of the article.

### 6.1 Future Prospects

The future of the COR project promises several exciting developments and prospects, which are outlined below.

**Expansion of Semantic Components:** As part of the COR project, a semantic component is being developed (Nimb et al., 2022), which will further enrich the database and allow for more sophisticated linguistic analyses and applications.

**Development of Additional Tools and Applications:** As the COR continues to evolve and expand, new tools and applications are expected to be developed. These may include advanced COR linkers, state-of-the-art natural language processing utilities and other innovative language technologies that will further enhance its utility and encourage its widespread adoption in language research and technology.

**More COR Resources:** With the ongoing development and promotion of the COR project, we anticipate a significant increase in the number of COR-tagged resources, stemming from both our own efforts and the collaborative contributions of the wider community through crowdsourcing initiatives.

**Integration with Other Language Projects:** The COR's potential for integration with parallel projects in other languages offers the possibility of creating shared,

unified linguistic resources with other languages, particularly the other North Germanic languages. Such a resource could significantly advance language research and technology in the region, fostering greater collaboration and understanding among researchers and practitioners working in these languages.

In summary, the future prospects of the COR project are bright, with the potential for significant advancements in linguistic research and natural language processing technologies. The ongoing development of semantic components, tools, applications, and resources will further solidify the COR's position as a vital and innovative resource in the world of language research and technology.

### 6.2 Conclusion

In conclusion, Det Centrale Ordregister (COR) represents a groundbreaking initiative in the field of Danish language studies, addressing the challenges of resource compatibility and promoting greater collaboration, efficiency, and innovation. Through the establishment of a shared database key and a multi-level structure, the COR has the potential to significantly impact not only academic research but also the broader landscape of natural language processing and language technology.

With promising future prospects, including the addition of semantic components, integration with parallel projects, and the development of new tools and applications, the COR is poised to become an indispensable resource for researchers, practitioners, and enthusiasts alike. By providing both the motivation and practical skills to engage with the COR, we hope to contribute to a vibrant and thriving community of Danish language research and development.

## 7. Acknowledgements

I am grateful to my colleagues at the Danish Language Council (Dansk Sprognævn), as well as our partners, the Society for Danish Language and Literature (DSL, Det Danske Sprog- og Litteraturselskab) and the Centre for Language Technology (CST, Center for Sprogteknologi, KU) for their collaboration and dedication in tackling the challenge of creating the COR. I would also like to express my sincere appreciation to the Digitalisation Agency (Digitaliseringsstyrelsen) for their generous financial support, which has been instrumental in facilitating the realisation of this ambitious project.

## 8. References

Dansk Sprognævn (1955). *Retskrivningsordbog*. Copenhagen: Gyldendal.

Dansk Sprognævn (1996). *Retskrivningsordbogen*. Copenhagen: Aschehoug, 2 edition.

Dansk Sprognævn (2012). *Retskrivningsordbogen*. Copenhagen: Alinea, 4 edition.

Glahder, J. (1923). *Dansk Retskrivningsordbog. Udgivet af Undervisningsministeriets Retskrivningsudvalg*. Copenhagen: Gyldendal.

Grundtvig, S. (1872). *Dansk Haandordbog med den af Kultusministeriet anbefalede Retskrivning*. Copenhagen: C. A. Reitzel.

Hartling, A.S. & Widmann, T. (2020). Den første ortografiske rettesnor for dansk – fra læsebog til ordbog: Malling (1777) på http://rohist.dsn.dk. In Y. Goldshtein,

I.S. Hansen & T.T. Hougaard (eds.) *18. Møde om Udforskningen af Dansk Sprog*. Institut for Kommunikation og Kultur, Aarhus University.

LBK 332 (1997). Lov om dansk retskrivning. LBK nr. 332. URL https://www.retsinformation.dk/eli/lta/1997/332.

Malling, O. (1777). *Store og gode Handlinger af Danske, Norske og Holstenere*. Copenhagen.

Nimb, S., Pedersen, B.S., Sørensen, N.C.H., Flörke, I., Olsen, S. & Troelsgaard, T. (2022). COR-S – den semantiske del af Det Centrale OrdRegister (COR). *LexicoNordica*, 29, pp. 75–97.