The Central Word Register of the Danish Language

Thomas Widmann
tw@dsn.dk
The Danish Language Council
(Dansk Sprognævn)

eLex 2023

28th June 2023

Introduction

Problems

Lexicographic and computational linguistic resources often lack compatibility or have challenging licences, making it difficult to reuse them.

Introduction

Problems

Lexicographic and computational linguistic resources often lack compatibility or have challenging licences, making it difficult to reuse them.

The problem is more pronounced for smaller languages.

Introduction

Problems

Lexicographic and computational linguistic resources often lack compatibility or have challenging licences, making it difficult to reuse them.

The problem is more pronounced for smaller languages.

In the context of Danish, many electronic resources exist but they lack a unified approach, common identifiers and reasonable licensing terms, leading to difficulties in language technology development.

Introduction

Our Solution

The solution? A shared database key system, similar to Denmark's *Det Centrale Personregister* (CPR).

The solution? A shared database key system, similar to Denmark's *Det Centrale Personregister* (CPR).

We introduce a new resource framework: The Central Word Register (Danish: *Det Centrale Ordregister*: COR).

The solution? A shared database key system, similar to Denmark's *Det Centrale Personregister* (CPR).

We introduce a new resource framework: The Central Word Register (Danish: *Det Centrale Ordregister*: COR).

We assign unique identification numbers to all lemmas and word forms in Danish.

The solution? A shared database key system, similar to Denmark's *Det Centrale Personregister* (CPR).

We introduce a new resource framework: The Central Word Register (Danish: *Det Centrale Ordregister*: COR).

We assign unique identification numbers to all lemmas and word forms in Danish.

The basic register (COR_1, also known as COR-K), launched in September 2022, is accessible at `ordregister.dk`.

Structure and Components of the COR

The Orthographical Foundation: *Retskrivningsordbogen*

Retskrivningsordbogen is the official Danish language orthography reference published by the Danish Language Council.

Structure and Components of the COR

The Orthographical Foundation: *Retskrivningsordbogen*

Retskrivningsordbogen is the official Danish language orthography reference published by the Danish Language Council.

Regular updates keep the dictionary current with the latest changes in Danish orthography.

Structure and Components of the COR

The Orthographical Foundation: *Retskrivningsordbogen*

Retskrivningsordbogen is the official Danish language orthography reference published by the Danish Language Council.

Regular updates keep the dictionary current with the latest changes in Danish orthography.

Semantics and etymology are not considered in determining what constitutes a lemma.

Structure and Components of the COR

The Orthographical Foundation: *Retskrivningsordbogen*

Retskrivningsordbogen is the official Danish language orthography reference published by the Danish Language Council.

Regular updates keep the dictionary current with the latest changes in Danish orthography.

Semantics and etymology are not considered in determining what constitutes a lemma.

COR_1 is based on *Retskrivningsordbogen*.

Structure and Components of the COR

From *Retskrivningsordbogen* to COR

The basic register, COR_1, can be seen as an enhanced and optimised version of *Retskrivningsordbogen* for natural language processing.

Structure and Components of the COR

From *Retskrivningsordbogen* to COR

The basic register, COR_1, can be seen as an enhanced and optimised version of *Retskrivningsordbogen* for natural language processing.

Key differences between the two:

1. COR_1 is designed for computer programs, while *Retskrivningsordbogen* is designed for humans.

Structure and Components of the COR

From *Retskrivningsordbogen* to COR

The basic register, COR_1, can be seen as an enhanced and optimised version of *Retskrivningsordbogen* for natural language processing.

Key differences between the two:

1. COR_1 is designed for computer programs, while *Retskrivningsordbogen* is designed for humans.
2. COR_1 offers more comprehensive coverage of inflected forms.

Structure and Components of the COR

From *Retskrivningsordbogen* to COR

The basic register, COR_1, can be seen as an enhanced and optimised version of *Retskrivningsordbogen* for natural language processing.

Key differences between the two:

1. COR_1 is designed for computer programs, while *Retskrivningsordbogen* is designed for humans.
2. COR_1 offers more comprehensive coverage of inflected forms.
3. Unlike *Retskrivningsordbogen*, COR_1 can be used without restrictions.

Structure and Components of the COR

From *Retskrivningsordbogen* to COR

The basic register, COR_1, can be seen as an enhanced and optimised version of *Retskrivningsordbogen* for natural language processing.

Key differences between the two:

1. COR_1 is designed for computer programs, while *Retskrivningsordbogen* is designed for humans.
2. COR_1 offers more comprehensive coverage of inflected forms.
3. Unlike *Retskrivningsordbogen*, COR_1 can be used without restrictions.
4. *Retskrivningsordbogen* contains usage examples and references to its rule appendix, which COR_1 does not.

Structure and Components of the COR

From *Retskrivningsordbogen* to COR

The basic register, COR_1, can be seen as an enhanced and optimised version of *Retskrivningsordbogen* for natural language processing.

Key differences between the two:

1. COR_1 is designed for computer programs, while *Retskrivningsordbogen* is designed for humans.
2. COR_1 offers more comprehensive coverage of inflected forms.
3. Unlike *Retskrivningsordbogen*, COR_1 can be used without restrictions.
4. *Retskrivningsordbogen* contains usage examples and references to its rule appendix, which COR_1 does not.
5. *Retskrivningsordbogen* has more and longer glosses than the COR_1.

Structure and Components of the COR

Unique Identification Numbers

In COR_1, all lemmas from *Retskrivningsordbogen* and their forms are assigned unique IDs.

Structure and Components of the COR

Unique Identification Numbers

In COR_1, all lemmas from *Retskrivningsordbogen* and their forms are assigned unique IDs.

Each ID is composed of a prefix 'COR' and a 5-digit index number for the lemma.

Structure and Components of the COR

Unique Identification Numbers

In COR_1, all lemmas from *Retskrivningsordbogen* and their forms are assigned unique IDs.

Each ID is composed of a prefix 'COR' and a 5-digit index number for the lemma.

A three-digit grammatical code specifies a particular form of a lemma.

Structure and Components of the COR

Unique Identification Numbers

In COR_1, all lemmas from *Retskrivningsordbogen* and their forms are assigned unique IDs.

Each ID is composed of a prefix 'COR' and a 5-digit index number for the lemma.

A three-digit grammatical code specifies a particular form of a lemma.

An additional two-digit code indicates orthographical variation.

Structure and Components of the COR

Unique Identification Numbers

In COR_1, all lemmas from *Retskrivningsordbogen* and their forms are assigned unique IDs.

Each ID is composed of a prefix 'COR' and a 5-digit index number for the lemma.

A three-digit grammatical code specifies a particular form of a lemma.

An additional two-digit code indicates orthographical variation.

The ID numbers are arbitrary and not assigned alphabetically.

Structure and Components of the COR

Unique Identification Numbers

In COR_1, all lemmas from *Retskrivningsordbogen* and their forms are assigned unique IDs.

Each ID is composed of a prefix 'COR' and a 5-digit index number for the lemma.

A three-digit grammatical code specifies a particular form of a lemma.

An additional two-digit code indicates orthographical variation.

The ID numbers are arbitrary and not assigned alphabetically.

The lemma indices range from 0 to 99,999, divided by word class for practicality.

Structure and Components of the COR

Unique Identification Numbers: an Example

```
COR.97230.110.01   donut   sb.fk.sg.ubest   donut         1
COR.97230.110.02   donut   sb.fk.sg.ubest   doughnut      1
COR.97230.111.01   donut   sb.fk.sg.best    donutten      1
COR.97230.111.02   donut   sb.fk.sg.best    doughnutten   1
...
```

The grammatical abbrivation exhibits a one-to-one correspondence with the grammatical (three-digit) code.

Structure and Components of the COR

Unique Identification Numbers: an Example

```
COR.97230.110.01   donut   sb.fk.sg.ubest   donut          1
COR.97230.110.02   donut   sb.fk.sg.ubest   doughnut       1
COR.97230.111.01   donut   sb.fk.sg.best    donutten       1
COR.97230.111.02   donut   sb.fk.sg.best    doughnutten    1
...
```

The grammatical abbrivation exhibits a one-to-one correspondence with the grammatical (three-digit) code.

The final column displays 1 if the form is part of the official norm, 0 if auto-generated.

Structure and Components of the COR

Unique Identification Numbers: Other Resources

Other COR resources should adhere to a similar syntax:

Structure and Components of the COR

Unique Identification Numbers: Other Resources

Other COR resources should adhere to a similar syntax:

1. The resource abbreviation, starting with `COR`.

Structure and Components of the COR

Unique Identification Numbers: Other Resources

Other COR resources should adhere to a similar syntax:

1. The resource abbreviation, starting with `COR`.
2. The lemma ID.

Structure and Components of the COR

Unique Identification Numbers: Other Resources

Other COR resources should adhere to a similar syntax:

1. The resource abbreviation, starting with `COR`.
2. The lemma ID.
3. Any required subdivisions, specific to each resource.

Structure and Components of the COR

Unique Identification Numbers: Other Resources

Other COR resources should adhere to a similar syntax:

1. The resource abbreviation, starting with `COR`.
2. The lemma ID.
3. Any required subdivisions, specific to each resource.

Details for subdivisions must be provided on `ordregister.dk`.

Structure and Components of the COR

Relations

Relations establish connections between lemmas and word forms.

Structure and Components of the COR

Relations

Relations establish connections between lemmas and word forms.

They facilitate the organisation and search for data within COR.

Structure and Components of the COR

Relations

Relations establish connections between lemmas and word forms.

They facilitate the organisation and search for data within COR.

Various types of relations can be defined, each resource can define its own. e.g.:

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
hyr	hypernym for two or more COR indexes
hyp	hyponym for another COR index
rim	rhyme (for rhyming dictionaries)

Structure and Components of the COR

The COR Resource Landscape

1. Level 1: Corresponding to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.

Structure and Components of the COR

The COR Resource Landscape

1. Level 1: Corresponding to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
2. Level 2: Contains resources from professional language environments in Denmark. Includes a resource of supplementary lemmas from the Danish Dictionary, and a semantic extension to the basic register. Prefix: `COR.NAME`.

Structure and Components of the COR

The COR Resource Landscape

1. Level 1: Corresponding to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
2. Level 2: Contains resources from professional language environments in Denmark. Includes a resource of supplementary lemmas from the Danish Dictionary, and a semantic extension to the basic register. Prefix: `COR.NAME`.
3. Level 3: Encompasses all other resources without restrictions. Any project can be assigned a prefix and an ID range. Prefix: `COR.OPEN.NAME`.

Structure and Components of the COR

The COR Resource Landscape

1. Level 1: Corresponding to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
2. Level 2: Contains resources from professional language environments in Denmark. Includes a resource of supplementary lemmas from the Danish Dictionary, and a semantic extension to the basic register. Prefix: `COR.NAME`.
3. Level 3: Encompasses all other resources without restrictions. Any project can be assigned a prefix and an ID range. Prefix: `COR.OPEN.NAME`.

Each resource is allocated a series of unique ID numbers.

Structure and Components of the COR

The COR Resource Landscape

1. Level 1: Corresponding to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
2. Level 2: Contains resources from professional language environments in Denmark. Includes a resource of supplementary lemmas from the Danish Dictionary, and a semantic extension to the basic register. Prefix: `COR.NAME`.
3. Level 3: Encompasses all other resources without restrictions. Any project can be assigned a prefix and an ID range. Prefix: `COR.OPEN.NAME`.

Each resource is allocated a series of unique ID numbers. These should be used in combination with existing ones in other resources on the same or lower levels.

Structure and Components of the COR

The COR Resource Landscape

1. Level 1: Corresponding to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
2. Level 2: Contains resources from professional language environments in Denmark. Includes a resource of supplementary lemmas from the Danish Dictionary, and a semantic extension to the basic register. Prefix: `COR.NAME`.
3. Level 3: Encompasses all other resources without restrictions. Any project can be assigned a prefix and an ID range. Prefix: `COR.OPEN.NAME`.

Each resource is allocated a series of unique ID numbers. These should be used in combination with existing ones in other resources on the same or lower levels.

New numbers should primarily be used for non-existing lemmas and those that do not correspond one-to-one with an existing entry.

Structure and Components of the COR

COR-S

The Society for Danish Language and Literature (DSL, Det Danske Sprog- og Litteraturselskab) and the Centre for Language Technology (CST, Center for Sprogteknologi, University of Copenhagen) are currently working on a semantic extension, COR-S.

Structure and Components of the COR

COR-S

The Society for Danish Language and Literature (DSL, Det Danske Sprog- og Litteraturselskab) and the Centre for Language Technology (CST, Center for Sprogteknologi, University of Copenhagen) are currently working on a semantic extension, COR-S.

It will be made available in the usual place, `ordregister.dk`.

Accessing and utilising the COR

COR Linkers

COR Linkers are programs assigning the correct COR id to each word in a text, crucial for computational linguistics.

Accessing and utilising the COR

COR Linkers

COR Linkers are programs assigning the correct COR id to each word in a text, crucial for computational linguistics.

The CLINK project, developed by the Danish Language Council, is an example of a COR linker. It is currently undergoing beta testing.

Accessing and utilising the COR

COR Linkers

COR Linkers are programs assigning the correct COR id to each word in a text, crucial for computational linguistics.

The CLINK project, developed by the Danish Language Council, is an example of a COR linker. It is currently undergoing beta testing.

It uses modules can be swapped freely and there is potential for future developments, like an AI-based module.

Accessing and utilising the COR

Historical Lexicography

The Danish Language Council's RO^{hist} project (`rohist.dk`) is a search engine for comparing Danish orthographical dictionaries from 1872 to 2012.

Accessing and utilising the COR

Historical Lexicography

The Danish Language Council's RO^{hist} project (`rohist.dk`) is a search engine for comparing Danish orthographical dictionaries from 1872 to 2012.

Efforts are being made to expand RO^{hist} with all Danish historical orthographical dictionaries and other orthographic resources.

Accessing and utilising the COR

Historical Lexicography

The Danish Language Council's RO^{hist} project (`rohist.dk`) is a search engine for comparing Danish orthographical dictionaries from 1872 to 2012.

Efforts are being made to expand RO^{hist} with all Danish historical orthographical dictionaries and other orthographic resources.

Plans are in place to assign COR numbers to the historical dictionaries in RO^{hist}, serving as level 2 resources with their own prefix and ID number range.

Accessing and utilising the COR

Historical Lexicography

The Danish Language Council's RO^{hist} project (`rohist.dk`) is a search engine for comparing Danish orthographical dictionaries from 1872 to 2012.

Efforts are being made to expand RO^{hist} with all Danish historical orthographical dictionaries and other orthographic resources.

Plans are in place to assign COR numbers to the historical dictionaries in RO^{hist}, serving as level 2 resources with their own prefix and ID number range.

The same COR ID number is reused if the lemma is the same, even if the spelling changes across editions. For example, *fråse*, *frådse*, and *fraadse* ("gorge") in different editions will all share the COR number 37337.

Accessing and utilising the COR

Historical Lexicography

The Danish Language Council's RO^{hist} project (`rohist.dk`) is a search engine for comparing Danish orthographical dictionaries from 1872 to 2012.

Efforts are being made to expand RO^{hist} with all Danish historical orthographical dictionaries and other orthographic resources.

Plans are in place to assign COR numbers to the historical dictionaries in RO^{hist}, serving as level 2 resources with their own prefix and ID number range.

The same COR ID number is reused if the lemma is the same, even if the spelling changes across editions. For example, *fråse*, *frådse*, and *fraadse* ("gorge") in different editions will all share the COR number 37337.

This approach greatly simplifies the implementation of RO^{hist}, easing the process of searching for a lemma across dictionaries.

Accessing and utilising the COR

The COR API

The COR can be accessed in two ways:

1. Downloading the entire register as a CSV file from ordregister.dk for offline work and system integration.

Accessing and utilising the COR

The COR API

The COR can be accessed in two ways:

1. Downloading the entire register as a CSV file from ordregister.dk for offline work and system integration.
2. Using the online interface at ordregister.dk to search data and access lemma and word form information. This information can be displayed in HTML or accessed from a program in either CSV or JSON format.

Accessing and utilising the COR

The COR API

The COR can be accessed in two ways:

1. Downloading the entire register as a CSV file from ordregister.dk for offline work and system integration.
2. Using the online interface at ordregister.dk to search data and access lemma and word form information. This information can be displayed in HTML or accessed from a program in either CSV or JSON format.

Here is an example of how to use the API with Python to lookup a lemma given an ID number:

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

Accessing and utilising the COR

Crowdsourcing

Companies and individuals can contribute to the COR by applying for a unique prefix and number series, turning COR into a largely crowdsourced resource.

Accessing and utilising the COR

Crowdsourcing

Companies and individuals can contribute to the COR by applying for a unique prefix and number series, turning COR into a largely crowdsourced resource.

We encourage contributors to publish their lemma lists on `ordregister.dk`, aiding the process of finding relevant data.

Accessing and utilising the COR

Crowdsourcing

Companies and individuals can contribute to the COR by applying for a unique prefix and number series, turning COR into a largely crowdsourced resource.

We encourage contributors to publish their lemma lists on `ordregister.dk`, aiding the process of finding relevant data.

We hope many will release COR-linked corpora and lexical resources, enhancing the overall utility of COR.

Future Prospects and Conclusion

A semantic component currently under development will enrich the database, allowing for sophisticated linguistic analyses and applications.

Future Prospects and Conclusion

A semantic component currently under development will enrich the database, allowing for sophisticated linguistic analyses and applications.

New lexical resources, tools and applications will hopefully enhance the COR's utility and promote its adoption in language research and technology.

Acknowledgements

- The Society for Danish Language and Literature (DSL, Det Danske Sprog- og Litteraturselskab)
- Centre for Language Technology (CST, Center for Sprogteknologi, University of Copenhagen)
- The Digitalisation Agency (Digitaliseringsstyrelsen)
- My colleagues at the Danish Language Council (Dansk Sprognævn)

The End

Any questions?