CLARIN

Common Language Resources and
Technology Infrastructure

Tour de CLARIN
DENMARK

Edited by **Darja Fišer** and **Jakob Lenardič**

DENMARK

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents Denmark and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

DENMARK

Introduction

Written by **Costanza Navarretta**

Denmark has been a member of CLARIN ERIC since February 2012 and is one of its founding members.[1] The Danish infrastructure CLARIN-DK was funded through two projects, the DK-CLARIN (2008-2010), and the DIGHUMLAB project (2011-2017). Since 2018, CLARIN-DK has been funded by the Faculty of Humanities and the Department of Nordic Studies and Linguistics, at the University of Copenhagen. The Danish national coordinator is Costanza Navarretta and the leading institution is the Centre for Language Technology, which is part of the Department of Nordic Studies and Linguistics.

CLARIN-DK involves the following institutions:

- The University of Copenhagen
- The Royal Danish Library

CLARIN-DK is a stable national research infrastructure where researchers can deposit, share and download language resources such as domain-specific corpora (e.g., The Danish Parliament Corpus 2009–2017 and the Johannes V. Jensen Corpus, which is a literary corpus collecting the works of the famous modernist poet Johannes Jensen from the early 20th century), as well as lexicons, word lists, speech transcriptions, and audio/video files in a secure way. CLARIN-DK also offers on-line language technology tools comprising e.g. a tokenizer, PoS tagger, a lemmatizer for Danish and English, a named entity recognizer for Danish, a keyword extractor, a TEI-to-text converter and a pipeline to linguistic annotation. Tools for performing basic frequency counts of words in textual data are also included, as well as visualization and corpus linguistics tools developed by other research groups, such as Korp and Voyant. Aside from being a certified B Centre, CLARIN-DK also runs a Knowledge Centre called DANSK, which provides expertise and help with using the language resources and technologies offered by the Danish consortium together with the Danish Language Council.

CLARIN-DK is involved in various Danish research projects and networks. For example, it is part of the Danish collaboration initiative DIGHUMLAB that involves various research communities, such as NetLAB, which is aimed at the cross-disciplinary study of internet materials, and LARM.fm, which is an online platform used for automatically locating the missing metadata of broadcast radio programmes. CLARIN-DK is also partner in an external funded research project Infrastrukturalisme with PI Henrik Jørgensen, of Aarhus University. The consortium is also involved in a research network, Multimodal Child Language Acquisition, with the University of Hong Kong and the Chinese Hong Kong University, (PI Costanza Navarretta), and contributes tools and guidance in a number of research activities comprising the linguistic annotation of medieval documents and TEI encoding of literary corpora, mainly at the University of Copenhagen. CLARIN-DK is also involved in research data management and the promotion of FAIR data in the Humanities.

The CLARIN-DK team participates in the following CLARIN committees: Standing Committee for CLARIN Technical Centres (Lene Offersgaard, Bart Jongejan), Legal and Ethical Issues Committee: Sussi Olsen, Assessment Committee (Lene Offersgaard as Chair).

The Clarin-DK group at the University of Copenhagen: Mitchell John Seaton, Costanza Navarretta, Dorte Haltrup Olsen, Bart Jongejan, Sussi Olsen and Lene Offersgaard

[1] https://clarin.dk/

Tool | **CST Lemmatizer**

Written by **Bart Jongejan** and **Costanza Navarretta**

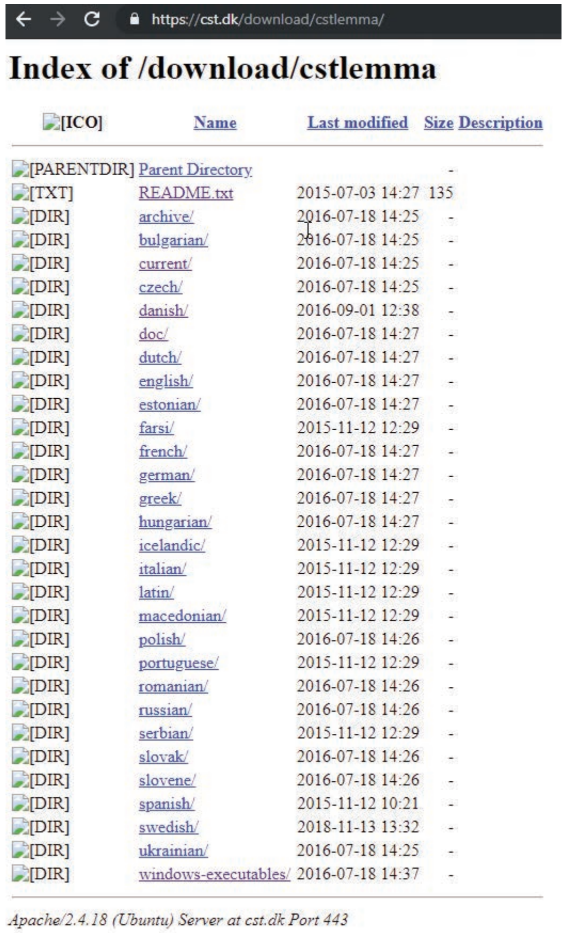
Lemmatizers generalize over the different forms of a word used in free text and provide its lemma, which is the base or dictionary look-up form. They are therefore one of the basic NLP tools which are not only important for NLP, but also for lexicographic work and all text-based studies. They are especially indispensable in morphologically rich languages that have a large number of word forms for the same lemma, which severely hinders querying or processing all of them in running text.

The CST lemmatizer has been developed over many years and as part of various projects, especially the Danish STO (Jongejan and Haltrup 2005) and the Nordic Tvärsök (Jongejan and Dalianis 2009).[2] While it was initially used as a tool to support Danish lexicographic work, it has gradually been extended with a dynamic self-learning algorithm which learns new lemmatization rules from morphological lexica that contain the relations between word forms and their corresponding lemmas. The lemmatization rules are organized in a decision tree.

In comparison to other state-of-the-art stemmers and rule-based lemmatizers, the current version of the CST lemmatizer not only learns lemmatization rules from word endings, but also recognizes a wide variety of derivational patterns; e.g., prefixation, infixation, suffixation. Therefore, it can deal with languages with different morphological systems. Currently, the CST lemmatizer has been trained on 25 languages. The list of these language-trained versions of the CST lemmatizer available from the Center for Language Technology is shown in Figure 1.

The lemmatizer is available for download via GiTHub. Figure 2 shows the CLARIN-DK web service for the CST-lemmatizer, while Figure 3 shows a Danish example sentence that was lemmatized with the tool.

[2] https://cst.dk/online/lemmatiser/uk/



https://cst.dk/download/cstlemma/

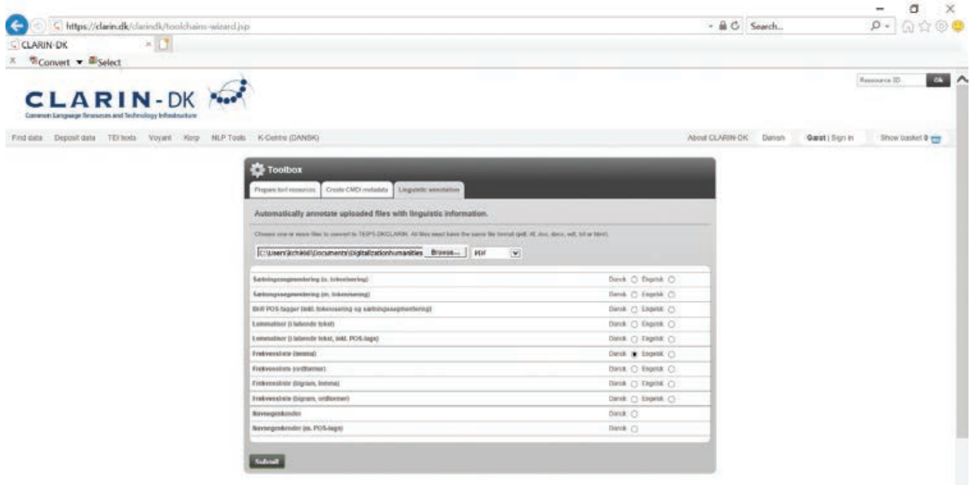
Index of /download/cstlemma

```
[ICO]        Name                  Last modified     Size Description

[PARENTDIR]  Parent Directory                          -
[TXT]        README.txt            2015-07-03 14:27  135
[DIR]        archive/              2016-07-18 14:25    -
[DIR]        bulgarian/            2016-07-18 14:25    -
[DIR]        current/              2016-07-18 14:25    -
[DIR]        czech/                2016-07-18 14:25    -
[DIR]        danish/               2016-09-01 12:38    -
[DIR]        doc/                  2016-07-18 14:27    -
[DIR]        dutch/                2016-07-18 14:27    -
[DIR]        english/              2016-07-18 14:27    -
[DIR]        estonian/             2016-07-18 14:27    -
[DIR]        farsi/                2015-11-12 12:29    -
[DIR]        french/               2016-07-18 14:27    -
[DIR]        german/               2016-07-18 14:27    -
[DIR]        greek/                2016-07-18 14:27    -
[DIR]        hungarian/            2016-07-18 14:27    -
[DIR]        icelandic/            2015-11-12 12:29    -
[DIR]        italian/              2015-11-12 12:29    -
[DIR]        latin/                2015-11-12 12:29    -
[DIR]        macedonian/           2015-11-12 12:29    -
[DIR]        polish/               2016-07-18 14:26    -
[DIR]        portuguese/           2015-11-12 12:29    -
[DIR]        romanian/             2016-07-18 14:26    -
[DIR]        russian/              2016-07-18 14:26    -
[DIR]        serbian/              2015-11-12 12:29    -
[DIR]        slovak/               2016-07-18 14:26    -
[DIR]        slovene/              2016-07-18 14:26    -
[DIR]        spanish/              2015-11-12 10:21    -
[DIR]        swedish/              2018-11-13 13:32    -
[DIR]        ukrainian/            2016-07-18 14:25    -
[DIR]        windows-executables/  2016-07-18 14:37    -
```

Apache/2.4.18 (Ubuntu) Server at cst.dk Port 443

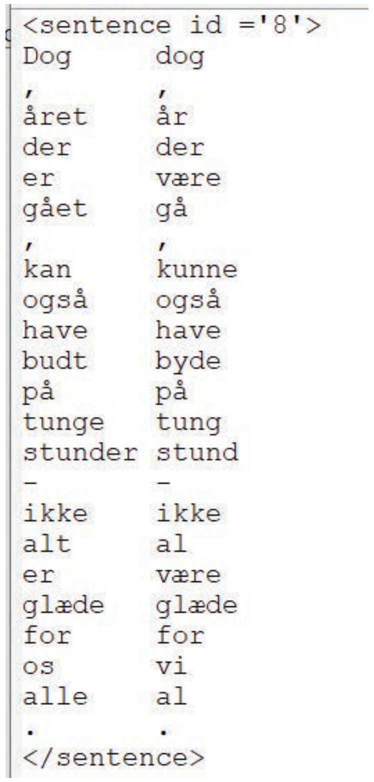
Figure 1: *The languages for which the trained CST-lemmatizer is available. Danish and English texts can be lemmatized online with the CST lemmatizer.*



https://clarin.dk/clarindk/toolchains-wizard.jsp

Automatically annotate uploaded files with linguistic information.

Figure 2: *The online CST lemmatizer on CLARIN-DK*



```
<sentence id ='8'>
Dog     dog
,       ,
året    år
der     der
er      være
gået    gå
,       ,
kan     kunne
også    også
have    have
budt    byde
på      på
tunge   tung
stunder stund
-       -
ikke    ikke
alt     al
er      være
glæde   glæde
for     for
os      vi
alle    al
.       .
</sentence>
```

Figure 3: *Lemmatization of the Danish sentence* Dog, året der er gået, kan også have budt på tunge stunder – ikke alt er glæde for os alle *("However, the past year can also have provided sad moments – not everything can give happiness to all of us "), which is taken from the 2017 New Year's Eve speech by the Danish Queen*

The CST lemmatizer trained for Danish has been used in many NLP projects, but also outside the NLP community. Frederik Hjorth, who is a political science researcher at the Department of Political Science, at the University of Copenhagen, has applied the CST lemmatizer to political speeches as one of the preprocessing steps in order to investigate how members of the existing political parties have addressed right-wing populists who have been challenging the order of the established political system (Hjorth 2018). The results of the study indicate that young politicians are often willing to engage with the populists as well as with other politicians across the political spectrum in the name of democratic freedom (which Hjorth calls the *strategy of engagement*), while older politicians often describe the populist challengers as morally illegitimate (which Hjorth calls the strategy of *disparagement*) and refuse to enter into discussions with them.

The CST lemmatizer was also used for many other languages in different linguistic projects. For example, it was trained on Russian (Sharoff and Nivre 2011) and then used e.g. for event identification (Solovyev and Ivanov 2016), and for anaphora and co-reference resolution (Toldova et al. 2014).

References:

Jongejan, B. and Haltrup, D. 2005. The CST *Lemmatiser*. Center for Sprogteknologi, University of Copenhagen version 2.7. http://cst.dk/online/lemmatiser/cstlemma.pdf.

Jongejan, B. and Dalianis, H. 2009. Automatic Training of Lemmatization Rules That Handle Morphological Changes in Pre-, in- and Suffixes Alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1* - ACL-IJCNLP '09. Vol. 1 Suntec, Singapore: Association for Computational Linguistics, 145.

Hjorth, F. 2018. Establishment Responses to Populist Challenges: Evidence from Legislative Speech. 2018 *Annual Meeting of the Danish Political Science Association*. http://fghjorth.github.io/papers/responses.pdf.

Sharoff, S. and Nivre, J. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proceedings of Computational Linguistics and Intelligent Technologies DIALOGUE2011*, Bekasovo, 591–604. https://pdfs.semanticscholar.org/36df/5fbe04f425e9b089437e979581d1f5375a94.pdf.

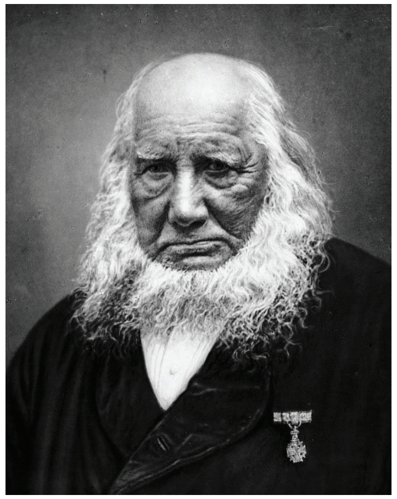
Solovyev, V. and Ivanov, V. 2016. Knowledge-driven event extraction in Russian: corpus-based linguistic resources. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155%2f2016%2f4183760.

Toldova, S. et al. 2014. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. *Computational Linguistics and Intellectual Technologies* 13 (20): 681–694.

Resource | **Grundtvig's Work Corpus**

Written by **Dorte Haltrup Hansen** and **Costanza Navarretta**

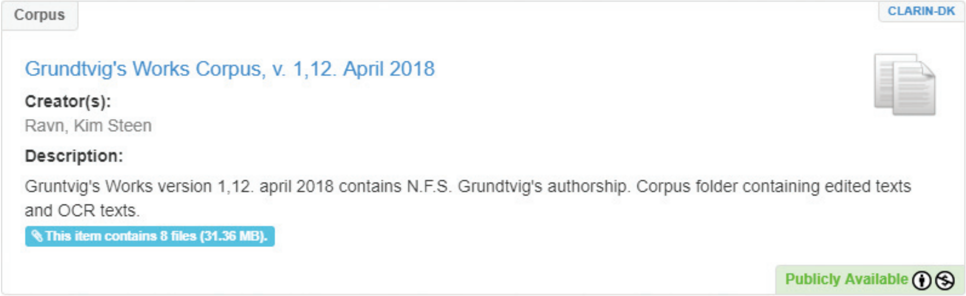
Nikolai Frederik Severin Grundtvig was a theologian, a priest, a philosopher, a poet, a writer, a teacher and a politician (member of the Rigsdagen, one of the two parts of the Parliament), who lived in Denmark between 1783 and 1872. He was a contemporary of Hans Christian Andersen and Søren Kierkegaard. Grundtvig's ideas have had a lasting impact on many areas of Danish culture such as education, politics and the church. For example, Grundtvig advocated for a reform of the school system, which also included educating adults to participate actively in society and cultural life. Therefore, Grundtvig is considered to be the mind behind the folk high school. He was part of the national romantic movement, and contributed to the development of the Danish national awareness. Grundtvig's written works are thus an important key to the understanding of Danish culture and mentality.



Nikolai Frederik Severin Grundtvig

The collection Grundtvig's Works are published by the Grundtvig Centre at the University of Aarhus and will contain 1,000 critically annotated texts written by N.F.S. Grundtvig when finalized in 2030.[3] The works are available to the public through a searchable interface, including registers of persons, places and Bible citations. The researchers at the Grundtvig Center wanted to a reliable and consistent way to cite the publication and a sustainable and interoperable environment in which they could share the work among other scholars and the public in general. Since the Grundtvig Centre itself does not offer the possibility for downloading the underlying files, CLARIN-DK was approached as a repository provider.

[3] http://www.grundtvigsvaerker.dk/



Corpus

CLARIN-DK

Grundtvig's Works Corpus, v. 1,12. April 2018

Creator(s):
Ravn, Kim Steen

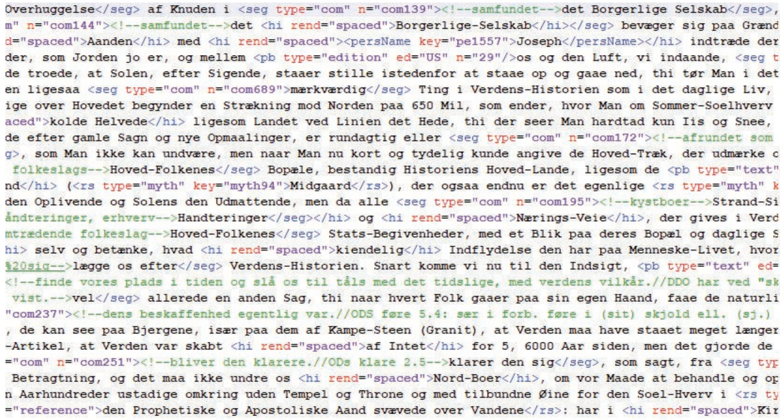
Description:
Gruntvig's Works version 1,12. april 2018 contains N.F.S. Grundtvig's authorship. Corpus folder containing edited texts and OCR texts.

This item contains 8 files (31.36 MB).

Publicly Available

Figure 4: *The corpus in the CLARIN-DK repository*

The corpus, now deposited in CLARIN-DK's D-Space repository[4], consists of around 1,300 TEI encoded XML-files, of which approximately 450 are critical editions manually annotated with person names, place names, mythological names, Bible citations and comments. When new versions of the works are released, they will be uploaded as new versions of the corpus in the CLARIN-DK repository.



```
Overhuggelse</seg> af Knuden i <seg type="com" n="com139"><!--samfundet-->det Borgerlige Selskab</seg>,
m" n="com144"><!--samfundet-->det <hi rend="spaced">Borgerlige-Selskab</hi></seg> bevæger sig paa Grænd
d="spaced">Aanden</hi> med <hi rend="spaced"><persName key="pe1557">Joseph</persName></hi> indtræde der
der, som Jorden jo er, og mellem <pb type="edition" ed="US" n="29"/>os og den Luft, vi indaande, <seg t
de troede, at Solen, efter Sigende, staaer stille istedenfor at staae op og gaae ned, thi tør Man i det
en ligesaa <seg type="com" n="com689">mærkværdig</seg> Ting i Verdens-Historien som i det daglige Liv,
ige over Hovedet begynder en Strækning mod Norden paa 650 Mil, som ender, hvor Man om Sommer-Soelhverv
aced">kolde Helvede</hi> ligesom Landet ved Linien det Hede, thi der seer Man hardtad kun Iis og Snee,
de efter gamle Sagn og nye Opmaalinger, er rundagtig eller <seg type="com" n="com172"><!--afrundet som
g>, som Man ikke kan undvære, men naar Man nu kort og tydelig kunde angive de Hoved-Træk, der udmærke c
 folkeslags-->Hoved-Folkenes</seg> Bopæle, bestandig Historiens Hoved-Lande, ligesom de <pb type="text"
nd</hi> (<rs type="myth" key="myth94">Midgaard</rs>), der ogsaa endnu er det egenlige <rs type="myth" k
den Oplivende og Solens den Udmattende, men da alle <seg type="com" n="com195"><!--kystboer-->Strand-Si
åndteringer, erhverv-->Handteringer</seg></hi> og <hi rend="spaced">Nærings-Veie</hi>, der gives i Verd
mtrædende folkeslag-->Hoved-Folkenes</seg> Stats-Begivenheder, med et Blik paa deres Bopæl og daglige S
hi> selv og betænke, hvad <hi rend="spaced">kiendelig</hi> Indflydelse den har paa Menneske-Livet, hvor
%20sig-->lægge os efter</seg> Verdens-Historien. Snart komme vi nu til den Indsigt, <pb type="text" ed=
<!--finde vores plads i tiden og slå os til tåls med det tidslige, med verdens vilkår.//DDO har ved "sk
 vist.-->vel</seg> allerede en anden Sag, thi naar hvert Folk gaaer paa sin egen Haand, faae de naturli
"com237"><!--dens beskaffenhed egentlig var.//ODS føre 5.4: sær i forb. føre i (sit) skjold ell. (sj.)
, de kan see paa Bjergene, især paa dem af Kampe-Steen (Granit), at Verden maa have staaet meget længer
-Artikel, at Verden var skabt <hi rend="spaced">af Intet</hi> for 5, 6000 Aar siden, men det gjorde de
="com" n="com251"><!--bliver den klarere.//ODs klare 2.5-->klarer den sig</seg>, som sagt, fra <seg typ
 Betragtning, og det maa ikke undre os <hi rend="spaced">Nord-Boer</hi>, om vor Maade at behandle og op
n Aarhundreder ustadige omkring uden Tempel og Throne og med tilbundne Øine for den Soel-Hverv i <rs ty
="reference">den Prophetiske og Apostoliske Aand svævede over Vandene</rs>: har i <hi rend="spaced">Kir
```

Figure 5: *A look into* Haandbog i Verdens-Historien (Handbook in World History) *from 1833*

The language excerpt in Figure 5 shows the old orthography from before the Danish language revision in 1948, e.g.:

Original	… som Man i det attende Aarhundrede troede, at Solen, efter Sigende, staaer stille istedenfor at staae op …
Normalized Danish	som man troede i det 18. århundrede, at solen efter sigende står stille i stedet for at stå op
Literal English translation	… as thought in the 18th century, that the sun after what they said, is staying still instead of rising …

[4] http://hdl.handle.net/20.500.12115/31

Furthermore, the excerpt shows the manual mark-up of the corpus, done by philologists at the Grundtvig Centre. There are references to, for example, person names (Joseph), mythological places (Midgaard) and actual places (Europe) and comments on parts of the text (*Overhuggelse af Knuden* <com139>, literal English translation: the cut of the knot). The actual comment is not shown in the text.

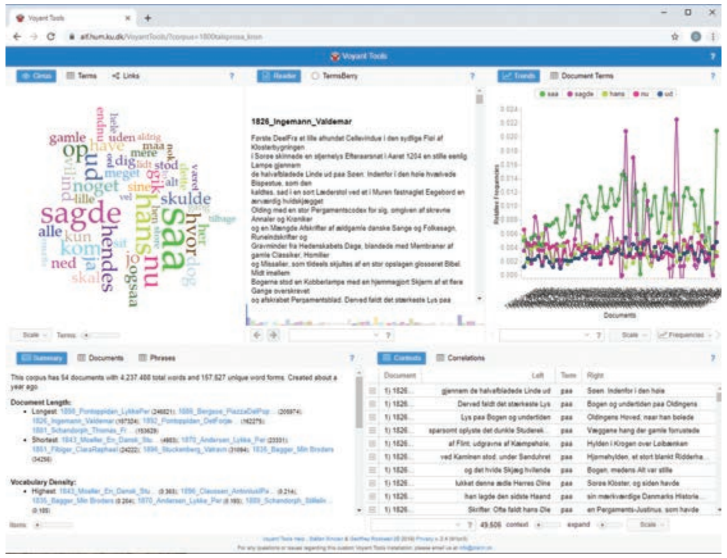
The corpus is an excellent resource for researchers who wish to apply digital methods to investigate various aspects of Grundtvig and his epoch. For example, researchers might want to investigate Grundtvig as a historical person, address the 19th century's literary language or orthography, or dig into his work when studying the theoretical background of the Danish folk high school tradition. The corpus is also important for scholars applying Linked Data in order to investigate the 19th century, since the corpus contains the annotations of people, places and events.

Event | **Teach the Teachers – the Voyant Tools**

Written by **Lene Offersgaard** and **Dorte Haltrup Hansen**

Digital methods are only slowly gaining ground in the teaching of literary studies in Denmark. While many lecturers are interested in introducing digital methods to their students, they often lack the knowledge of existing tools. From previous workshops, CLARIN-DK learned that neither traditional NLP tools like lemmatizers, POS-taggers, and named entity recognizers, nor simple command line scripting, were suitable in such teaching scenarios. This is why CLARIN-DK started to explore other technologies, such as data visualization tools that could serve as a better and easier entry point to the use of digital methodologies for non-computational researchers and teachers.

We opted for Voyant Tools,[4] introduced to us by information specialists from HUMlab – a datalab at the Copenhagen University Library. Voyant Tools is an online environment that performs automatic text analysis with functionalities such as word frequency lists, frequency distribution plots, and KWIC displays (Figure 6). CLARIN-DK and HUMlab have organized several interactive workshops presenting the use of this environment to lecturers and researchers at the Faculty of Humanities at the University of Copenhagen. CLARIN-DK hosted a dedicated event at the Department of Nordic Studies and Linguistics on 21 November 2018, which was attended by 12 teachers and researchers.



1826_Ingemann_Valdemar

Første DeelFra et lille afrundet Cellevindue i den sydlige Fløi af Klosterbygningen
i Soroe skinnede en stjernelys Efteraarsnat i Aaret 1204 en stille eenlig Lampe gjennem
de halvafbladede Linde ud paa Søen. Indenfor i den hele hvælvede Bispestue, som den
kaldtes, sad i en sort Læderstol ved et i Muren fastnaglet Egebord en ærværdig hvidskjægget
Olding med en stor Pergamentscodex for sig, omgiven af skrevne Annaler og Krøniker
og en Mængde Afskrifter af ældgamle danske Sange og Folkesagn, Runeindskrifter og
Gravminder fra Hedenskabets Dage, blandede med Membraner af gamle Classiker, Homilier
og Missaler, som tildeels skjultes af en stor opslagen glosseret Bibel. Midt imellem
Bogerne stod en Kobberlampe med en hjemmegjort Skjærm af et flere Gange overskrevet
og afskrabet Pergamentsblad. Derved faldt det stærkeste Lys paa

This corpus has 54 documents with 4.237.488 total words and 157.827 unique word forms. Created about a year ago.

Document Length:

- Longest: 1856_Pontoppidan_LykkePer (248821); 1888_Bergsøe_PiazzaDelPop... (205974); 1826_Ingemann_Valdemar (187524); 1892_Pontoppidan_DetForje... (162275); 1881_Schandorph_Thomas_Fr... (153629)
- Shortest: 1843_Moeller_En_Dansk_Stu... (4883); 1870_Andersen_Lykke_Per (23301); 1851_Fibiger_ClaraRaphael (24222); 1896_Stuckenberg_Vatravn (31084); 1835_Bagger_Min Broders (34256)

Vocabulary Density:

- Highest: 1843_Moeller_En_Dansk_Stu... (0.363); 1896_Claussen_AntoniusIPa... (0.214); 1835_Bagger_Min Broders (0.204); 1870_Andersen_Lykke_Per (0.193); 1889_Schandorph_Stille... (0.193)

Document	Left	Term	Right
1) 1826...	gjennem de halvafbladede Linde ud	paa	Søen. Indenfor i den hele
1) 1826...	Derved faldt det stærkeste Lys	paa	Bogen og undertiden paa Oldingens
1) 1826...	Lys paa Bogen og undertiden	paa	Oldingens Hoved, naar han bøiede
1) 1826...	sparsomt oplyste det dunkle Studerek...	paa	Væggene hang der gamle forrustede
1) 1826...	af Flint, udgravne af Kæmpehøie,	paa	Hylden i Krogen over Løibænken
1) 1826...	ved Kaminen stod, under Sanduhret	paa	Hjørnehylden, et stort blankt Ridderha...
1) 1826...	og det hvide Skjæg hvilende	paa	Bogen, medens Alt var stille
1) 1826...	lukket denne ædle Herres Øine	paa	Soroe Kloster, og siden havde
1) 1826...	han lagde den sidste Haand	paa	sin mærkværdige Danmarks Historie...
1) 1826...	Skrifter. Ofte faldt hans Øie	paa	en Pergaments-Justinus, som havde

49.505

Figure 6: *Voyant Tools*

[4] https://alf.hum.ku.dk/VoyantTools/

In order to tailor the events to the needs of the participants, CLARIN-DK asked some of them in advance which literary works were most relevant to be showcased and which research questions could be investigated and discussed during the events. They opted for novels written around the Modern Breakthrough period, an era in the Scandinavian literature which started at the end of the 19th century and in which Naturalism replaced Romanticism. The Archive of Danish Literature (http://adl.dk) provided a collection of 54 novels. The novels were preprocessed and uploaded to a local instance of Voyant Tools by the CLARIN-DK team and information specialists from HUMlab.

A research question addressed the use of terms before and after the Modern Breakthrough (1870–1890). If it was possible to visualize changes in the use of, for example, terms for emotions (like *love*) which are typical for the Romanticism period compared to the use of more concrete terms (like *work*) which should be more common in the Naturalism novels. Using the *Trends* tool in Voyant (Figure 7), it was found that the term for love is used relatively more often before 1875 than after 1888. Moreover, the term for *work* is not used before 1875 in the novels, while it was used after then. Therefore, the use of these terms indicates that there is a shift in the use of common themes around the Modern Breakthrough. However, by using this simplistic method, it is impossible to differentiate novels representing the Modern Breakthrough.

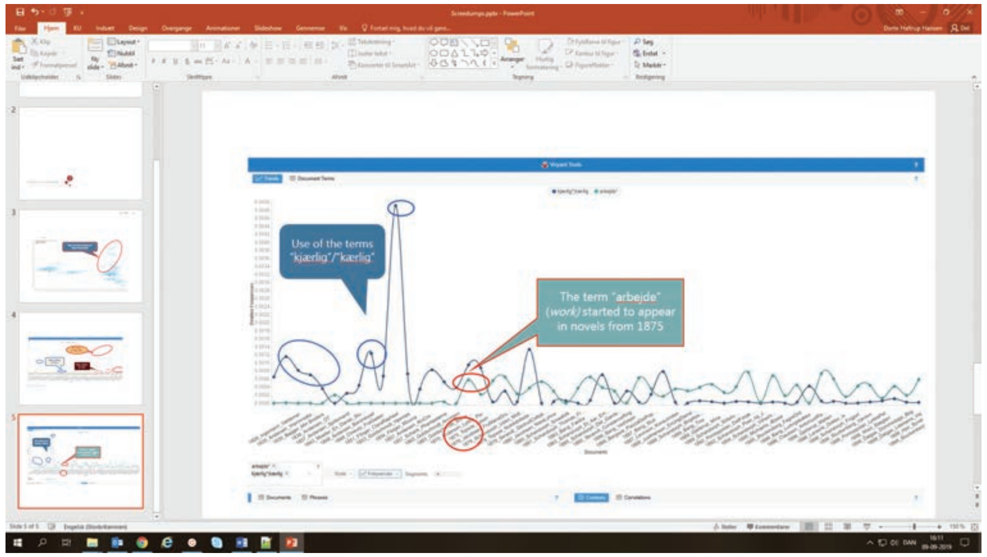

Use of the terms "kjærlig"/"kærlig"
The term "arbejde" (work) started to appear in novels from 1875

Figure 7: *The chronological distribution of the terms love vs. work for the period between 1826 and 1899 with regard to 54 novels*

We therefore investigated if other tools in Voyant could also confirm the differences between the two literary periods. In the *ScatterPlot tool* it is, among other things, possible to visualize the results of document similarity analysis. Figure 8 shows the document similarity using the TF-IDF frequency count for all novels in the corpus. In the figure, the novels by Herman Bang and a few novels by Sophus Schandorph are clearly separated from the other works. The novels from the late 19th century of these two writers are considered representatives of the Modern Breakthrough. It was now up to the researchers to interpret the similarities in the other groups of the scatter plot and from there to pose more research questions.

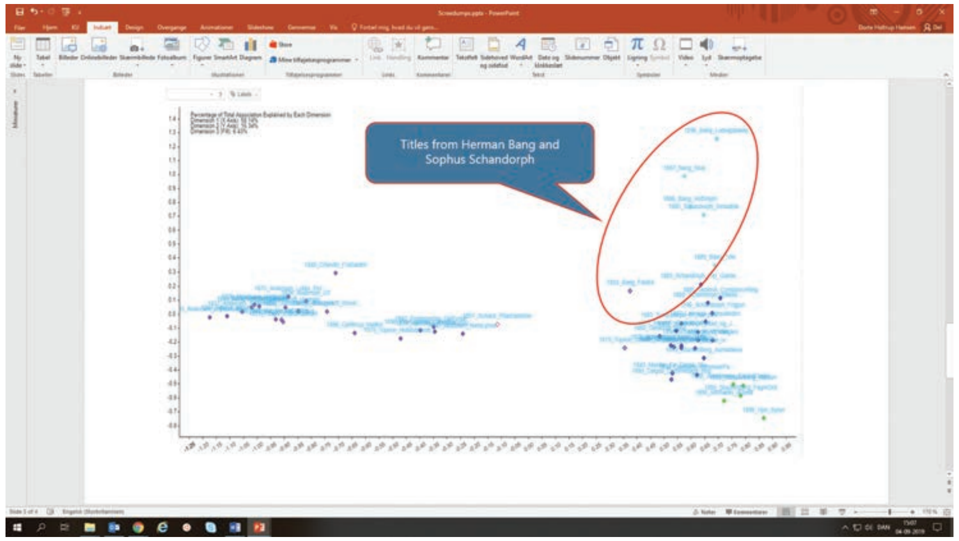


Figure 8: *Novel similarity based on TF-IDF counts*

In this and other workshops, the participants soon realized that studying texts through isolated words (word forms) was limiting, and there was a clear need for lemmatization. Moreover, the need for PoS-tagged texts became evident since some researchers were interested in investigating adjectives showing emotions, while others were interested in analysing events, requiring the automatic extraction of verbs. Despite this, Voyant Tools has proved to be very illustrative and useful to get a first quantitative overview of a collection of novels, and it allowed the comparison of two or more novels.

As a follow up to this event, the CLARIN-DK team will organize a workshop introducing corpus tools and corpus querying techniques in linguistically annotated texts for Literary Studies. The event will also showcase how automatic linguistic annotations are performed on texts from before and after the Danish orthographic reform of 1948, and discuss how it is possible to circumvent problems encountered when applying NLP tools developed for contemporary texts to older texts.

Interview | **Klaus Nielsen**

Klaus Nielsen is the chief editor at the Grundtvig Study Centre.

What is your scholarly background and your current academic position?

<

I obtained my PhD from the University of Copenhagen in 2012 and my thesis was a combination of traditional literary theory and book history, a philological field that focuses on a more mechanical-analytical study of the publication process of literary works. I focused on *Gittes monologer*, a famous collection of satirical poems by the Danish poet Per Højholt published in different versions between 1980 and 1984. I was able to observe crucial textual differences between their various published versions, which allowed me to arrive at a much richer interpretation of the poems that wouldn't be possible with the final, best-known 1984 version alone. This showed me how important it is to combine traditional qualitative literary analysis with analytical methods that also take into consideration non-textual information such as publication history.

I now work as chief editor at Grundtvig Study Centre,[5] where we are preparing a critical edition of the collected works of N.F.S. Grundtvig, a very prolific and multidisciplinary Danish author who published around 37,000 pages of text from 1804 to his death in 1872. We are making this corpus available in an online environment, with manual annotations that follow the scholarly standards of textual criticism. In a sense, my PhD was an important methodological stepping stone for my current work related to the Grundtvig's Works Corpus, which also involves a close study of the differences between the various published editions.

>

The Grundtvig's work corpus has been published through the CLARIN-DK repository. How did this collaboration start? How do you benefit from this collaboration?

<

We released the first version of our corpus through the CLARIN-DK repository in 2018 at the suggestion of Lene Offersgaard, with whom we were collaborating on a related project at the time. This was a great opportunity for us because we had been receiving feedback from some of our more devoted users who said they wanted the corpus in a downloadable format. We've also made an agreement with CLARIN-DK that as soon as we publish a new version of the corpus through our online environment, we'll also update the version deposited in the repository with the newest, more richly annotated one.

>

How is Grundtvig's corpus structured? What are some of the challenges you come across when annotating the corpus?

<

The corpus is extremely varied in terms of content, since Grundtvig was a polihistorian who wrote on a variety of different subjects. Perhaps most prominently, he wrote books on Danish history and Nordic mythology, carried out linguistic studies of Old Icelandic and Old English, translated from Latin, wrote political and philosophical texts, and composed around 1,500 hymns, many of which are still sung today in Denmark. For this reason, Grundtvig's views are representative of the intellectual and cultural zeitgeist of Denmark in the 19th century.

There's a downside to his varied repertoire, in that annotation is still manually intensive. We do use a database for place and person names that we feed into a named-entity recognizer, but even in this case, we often have to manually verify the results. For example, Grundtvig often refers to the philosopher Søren Kierkegaard, who was a contemporary of his, and our software is generally successful in identifying this particular named entity. However, Grundtvig often refers to him by his last name only, but since Søren Kierkegaard had a brother who was also a published author in the same period, we have to manually check the automatic recognition to make sure that the software made a link to the correct referent. In addition to this, we often come across obsolete words, in which case we manually add their possible historical meaning. This can only be done by closely reading and interpreting the surrounding text. Nevertheless, we will use the parts that have already been annotated as a baseline for a semi-automated processing of the remaining two-thirds of the corpus in the future.

[5] http://www.grundtvigsvaerker.dk/

One of the greatest challenges in terms of mark-up pertains to identifying Biblical references, especially in cases where Grundtvig doesn't use direct quotes taken from the Bible but his own modified variants, or where he makes indirect references to the more obscure motifs and quotes. Although we have theologians both internal and external who closely read the texts and manually identify such references, it would be invaluable if we could also make use of a language tool that would help automatize this process of identification. I don't think that such a tool exists yet, but it would be a very welcome addition to the CLARIN infrastructure in my opinion. Similarly, it would be great to have a tool that can automatically recognize proverbs and sayings, which abound in Grundtvig's works, given that his work is a major part of the Danish cultural heritage. Although I'm not an expert in digital technologies, it seems that developing such a tool wouldn't be too hard a task, as there already exist readymade digital collections of Danish proverbs that could be used as a baseline for training the tool.

>

Has the corpus been successfully used by an external research project?

<

Yes, Baunvig and Nielbo (2017)[6] have used our corpus in a case study to determine how digital methods can benefit the analysis of very large collections of written text, and to uncover new perspectives and interpretations. Grundtvig Studies is a popular subfield in literary history in Denmark, and many studies on Grundtvig have been published in the past fifty years. However, previous researchers weren't able to use digital methods and tools, which means that their claims were influenced by the limitations inherent to a purely manual approach to analysis. As I've said, Grundtvig produced around 37,000 pages in his lifetime, which is simply too much text for an individual researcher to read and then be able to recollect the finer details. For instance, there is an older study in which it is claimed that Grundtvig started suffering from a series of psychological problems in the 1830s, which was reflected in the texts he wrote in this decade. However, Baunvig and Nielbo (2017) were able to show, by using quantitative methods such as measuring the amount of information entropy in the corpus, that his psychological turmoil actually started earlier than was previously claimed, which is of course an important finding from a purely historical viewpoint. There has also been a follow-up study of our corpus conducted by Nielbo et al. (2018).[7]

>

What makes this corpus particularly valuable for the CLARIN infrastructure?

<

I think that our rather thorough manual approach to the corpus is an important contribution for a more accurate understanding of the historical developments of the Danish language, especially its orthography. What is important in this respect is that there were no orthographic rules in Grundtvig's time, only tendencies, which means that spelling was quite liberal in comparison to contemporary Danish. Consequently, we're often in doubt whether the way Grundtvig spelled a certain word is an instance of spelling variation that was attested at the time or if it is just a spelling mistake on his part. This is particularly problematic in cases where Grundtvig's idiosyncratic spelling can't be found in the historical dictionaries of 19th century Danish, since this intuitively makes you think that the spelling variant was a mistake. However, such dictionaries weren't compiled on the basis of the original edition but often used later published editions that had gone through the editing process, where spelling variation was normalized. This means that if a researcher wanted to study the vocabulary of 19th century Danish just on the basis of such dictionaries, he or she would miss the attested variations and consequently get a warped view of how people actually wrote at the time. By contrast, we spend a lot of time closely analysing and proofreading the materials, so we are able to present a resource that serves as a much more complex, as well as accurate, presentation of the linguistic situation at the time.

>

Could you give an example of such orthographic variation? How did you resolve it?

<

I actually came across a fairly interesting orthographic problem just recently when I was annotating Grundtvig's *History of the Northmen*, which is one of the few texts he wrote in English. In this text, Grundtvig used the word kempion in the sense of "champion" or "hero"; however, this spelling variant isn't listed in the *Oxford English Dictionary*, which only includes the variant campion with an a instead of an e. Because my colleagues and I weren't sure how to solve this issue, we consulted a Professor of Middle English, and he believed it to be a spelling mistake that should be corrected in the edited corpus, given that the Oxford *English Dictionary* is extremely comprehensive and thorough in its account of English etymology. However, when I searched for the variant kempion on Google, I found out that it was actually attested at the time, and it was for instance used by Sir Walter Scott in his 1822 novel *The Pirate*, which Grundtvig was alluding to.

>

[6] https://knielbo.github.io/files/valider_selvopgoer_kln.pdf

[7] https://doi.org/10.1093/llc/fqy054

Are there any other aspects of the CLARIN-DK infrastructure that are important for your work at the centre?

<

Yes, especially in relation to how proactively they reach out as part of their user-involvement initiative. Last year, CLARIN-DK organized a tutorial for the philologists at our centre where they demonstrated how Voyant Tools can simplify our annotation process. Using Voyant has turned out to be extremely helpful when we come across obsolete phrases the meaning of which we don't know and can't find in the historical dictionaries. By using Voyant's extended search capabilities and visualization tools, we are now able to easily chart the occurrences of this unknown phrase in the entire corpus, and then extract only those texts where this phrase seems to occur in a similar context, which then helps us determine its actual meaning.

I am also pleased to say that CLARIN-DK has already made the first version of our corpus available through their installation of the Voyant Tools. We plan on updating this test version with newer ones with regularity. In the long run, I believe the availability of the corpus through CLARIN-DK's Voyant Tools will significantly streamline user assistance.

>

Your professional website says that you're also interested in audio literature. Is this something that you're still actively researching?

<

No, my research on audio literature was mostly confined to my PhD project, because Per Højholt, who is the author of the poems that I was analysing, had read them aloud on Danish radio in the 1980s. By using an audio-analysis software called PRAAT I measured prosodic features such as the author's pitch and reading speed, and I was able to see how he deliberately changed his voice in accordance with the way the point-of-view character developed through the course of the poems' narrative. This was a rather small but important finding, since it hadn't been previously acknowledged in the relevant literature on Gittes Monologer how the author's spoken performance of his own work added new dimensions to the understanding of the poems themselves.

>

What kind of new research questions does audio literature offer in the context of Digital Humanities? Do you think that CLARIN could contribute to this field?

<

When I was writing my thesis, research on audio literature was still a very new field, but nowadays it is more readily agreed upon that audio recordings can serve as crucial material for textual analysis. Literary theorists are now conducting important research on the link between the reader of the audio text and the content of the text itself, and this opens up many interesting questions. Let's say, for instance, that we are dealing with a novel written in the first person, and that the narrator is a woman. Should the reader of the audio version then also be a woman, or conversely, what interpretative repercussions would arise if the reader were actually a man? That is, the person's voice crucially affects the way people perceive the text, much in the same way that the sort of typography of an old book can evoke various pre-conceptions in the reader about the book's content.

Given how audio literature opens up interesting questions relevant for the emerging Digital Humanities, I think that new digital tools for analysing recorded literary works would serve as very welcomes additions to the CLARIN infrastructure.

>

What are your hopes for CLARIN-DK in the future?

<

I think that one of the future challenges for Digital Humanities in Denmark is to find a common platform where our whole research community can have a more unified and interoperable access to as many carefully annotated resources as possible. I believe that CLARIN-DK is an excellent candidate in the country for this, because our experience with releasing the Grundtvig's Work Corpus has proven to us that their repository is a stable environment through which corpora can be released in a sustainable fashion and with well-presented metadata. On top of that, the repository also allows us to integrate our corpora with other services in the consortium. For this reason, it can only be a good thing if more Digital Humanities scholars in Denmark decide to deposit their resources in the CLARIN-DK repository.

>

COLOPHON

This brochure is part of the ‘Tour de CLARIN’ volume II
(publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands

www.clarin.eu

CLARIN
Common Language Resources and
Technology Infrastructure

Overhuggelse</seg> af Knuden i <seg type="com" n="com139"><!--samfundet-->det Borgerlige Selskab</seg>,
m" n="com144"><!--samfundet-->det <hi rend="spaced">Borgerlige-Selskab</hi></seg> bevæger sig paa Grænd
d="spaced">Aanden</hi> med <hi rend="spaced"><persName key="pe1557">Joseph</persName></hi> indtræde der
der, som Jorden jo er, og mellem <pb type="edition" ed="US" n="29"/>os og den Luft, vi indaande, <seg t
de troede, at Solen, efter Sigende, staaer stille istedenfor at staae op og gaae ned, thi tør Man i det
en ligesaa <seg type="com" n="com689">mærkværdig</seg> Ting i Verdens-Historien som i det daglige Liv,
ige over Hovedet begynder en Strækning mod Norden paa 650 Mil, som ender, hvor Man om Sommer-Soelhverv
aced">kolde Helvede</hi> ligesom Landet ved Linien det Hede, thi der seer Man hardtad kun Iis og Snee,
de efter gamle Sagn og nye Opmaalinger, er rundagtig eller <seg type="com" n="com172"><!--afrundet som
g>, som Man ikke kan undvære, men naar Man nu kort og tydelig kunde angive de Hoved-Træk, der udmærke d
folkeslags-->Hoved-Folkenes</seg> Bopæle, bestandig Historiens Hoved-Lande, ligesom de <pb type="text"
nd</hi> (<rs type="myth" key="myth94">Midgaard</rs>), der ogsaa endnu er det egenlige <rs type="myth" k
den Oplivende og Solens den Udmattende, men da alle <seg type="com" n="com195"><!--kystboer-->Strand-Si
åndteringer, erhverv-->Handteringer</seg></hi> og <hi rend="spaced">Nærings-Veie</hi>, der gives i Verd
mtrædende folkeslag-->Hoved-Folkenes</seg> Stats-Begivenheder, med et Blik paa deres Bopæl og daglige S
hi> selv og betænke, hvad <hi rend="spaced">kiendelig</hi> Indflydelse den har paa Menneske-Livet, hvor
%20sig-->lægge os efter</seg> Verdens-Historien. Snart komme vi nu til den Indsigt, <pb type="text" ed=
<!--finde vores plads i tiden og slå os til tåls med det tidslige, med verdens vilkår.//DDO har ved "sk
vist.-->vel</seg> allerede en anden Sag, thi naar hvert Folk gaaer paa sin egen Haand, faae de naturli
"com237"><!--dens beskaffenhed egentlig var.//ODS føre 5.4: sær i forb. føre i (sit) skjold ell. (sj.)
, de kan see paa Bjergene, især paa dem af Kampe-Steen (Granit), at Verden maa have staaet meget længer
-Artikel, at Verden var skabt <hi rend="spaced">af Intet</hi> for 5, 6000 Aar siden, men det gjorde de
="com" n="com251"><!--bliver den klarere.//ODs klare 2.5-->klarer den sig</seg>, som sagt, fra <seg typ
Betragtning, og det maa ikke undre os <hi rend="spaced">Nord-Boer</hi>, om vor Maade at behandle og op
n Aarhundreder ustadige omkring uden Tempel og Throne og med tilbundne Øine for den Soel-Hverv i <rs ty
="reference">den Prophetiske og Apostoliske Aand svævede over Vandene</rs>: har i <hi rend="spaced">Kir

CLARIN 2019