SSL *Scandinavian Studies in Language, 12(1), 2021 - https://tidsskrift.dk/sss*

# FROM LINGUISTIC FEATURES TO CULTURAL PATTERNS: PERSPECTIVES FROM LARGE-SCALE TEXT STUDIES OF DANISH SERMONS

Anne Agersnap
Aarhus University
anag@cas.au.dk

Kirstine Helboe Johansen
Aarhus University
kp@cas.au.dk

**Abstract**
This article discusses the concept of reading and presents a method that combines distant and close reading, while drawing on insights from computational humanities. Focusing on basic features in language, distant reading allows for the construction of new types of text. By close reading these texts, it is possible to analyse cultural patterns across individual texts. This method of reading is illustrated by two cases stemming from a project based on a corpus of 11,955 Danish sermons. The first case begins with a distant reading of gendered pronouns in the corpus. The second case begins with a distant reading of named agents.[*]

**Keywords:** Sermons, collective text productions, large-scale text analysis, distant reading, close reading, semantic structures

## 1. Introduction

A sermon is an oral event. People gathered for Sunday service sit in the pews listening to the preacher's words. In a Danish context, a sermon is typically also a written manuscript and a text file on the preacher's computer. The preacher may revise the manuscript several times, reuse it or come back to it later for inspiration. It might also be published on the church website or in a booklet. Like any other speech, a sermon is the words of the orator at the spoken event and in the written

[*] We owe special thanks for the engaging comments following the review of our article. Moreover, we are grateful for our engaging and fruitful partnership with Ross Deans Kristensen-McLachlan, Uffe Schjødt and Kristoffer Laigaard Nielbo. The case studies we present in this article have been conducted in collaboration with all project partners in the sermon project.

manuscript. However, a sermon is also a genre and a shared practice. As a genre, it is bound to its setting and occasion, and it is an established practice for which pastors are trained, leaning on the long tradition of preaching preceding them. Whereas preaching as an oral event is bound to the moment in which it takes place, sermon manuscripts manifest the sermons as concrete objects and enable them to be archived as resources for later. As objects, sermon manuscripts document preaching as ongoing cultural productions – as a collective practice of preaching. Literary scholar Franco Moretti famously brought attention to text in terms of quantity, which allows text productions to be understood as collective systems (Moretti 2007:3-4). We argue that also sermons should be studied as a collective system. To study texts as collective systems entails new understandings of what reading is and demands distinct methodological tools to interpret the cultural significance of these practices. However, before we can extract cultural information from them, we need to step back and think of texts as 'language'.

In an ongoing collaborative sermon project investigating a corpus of 11,955 Danish sermons, we have strived to study sermons from this perspective with the overall objective to uncover how pastors in their sermons navigate between religious tradition and contemporary culture. Neither of us are trained linguists. Nevertheless, to achieve our goal of approaching sermons as a collective practice, we have explored new ways of reading our material, and the strategy that proved most insightful was to focus on basic linguistic features. In our case, this means paying attention to words and texts based on linguistic coherence structures, such as syntactic function of terms as well as relations between certain parts of speech in language. Thus, underlying the overall objective of the sermon project was an implicit and more methodological question: *How can we through basic linguistic features access culturally significant information in large text collections?*

In this article, we demonstrate how this underlying question has guided the sermon project and allowed us to gain culturally significant information, thereby helping us to achieve the overall objective. The two case studies presented in the following show how we by attending to basic features such as gendered pronouns and named agents have gained insight into gender and religion and into relations between cultural and religious domains.

## 2. The sermon project: A collection and large-scale text study of 11,955 Danish sermons

The sermon project is a collaborative project with expertise from 'the sociology of religion', 'practical theology' and 'computational humanities'.[1] The cases included in this article were initiated within the framework of Anne Agersnap's PhD project (Agersnap 2021). As part of the PhD project, Agersnap assembled a text corpus of 11,955 sermons including basic socio-demographic variables from 95 pastors in the Evangelical-Lutheran Church in Denmark (ELCD). All sermons were received directly from the pastors via e-mail and carefully cleaned, annotated and archived as a digital corpus (Agersnap et al. 2020). Due to the size and quality of this new data set, the corpus offers previously inaccessible knowledge about how pastors in the ELCD practice religion and culture.

The sermon project aims to study sermons as a collective production of texts conceived in the intersection of tradition and contemporary culture. The main research question is: how do pastors' collective encounters with Christianity and cultural fields unfold in sermons? Sunday after Sunday, year after year, pastors in the ELCD are obliged to expound prescribed and recurring biblical passages in their sermons as part of church services. While the biblical texts are prescribed by a cyclical church calendar, the momentary situation in which the biblical passages are interpreted is always new. Sermons are therefore deeply intertextual and entail a unique meeting between biblical and contemporary time (Agersnap 2021).

The ELCD employs roughly 2,000 pastors, and most pastors write down their sermons as a manuscript prior to the church service. This means that pastors collectively and continuously contribute to a comprehensive production of texts– even though the sermons remain the property of the individual pastors after the service. With the newly established sermon corpus, we have studied how pastors collectively navigate between Christian tradition and contemporary culture. The case studies included in this article were originally designed to address this overall question. The first study explores the role of cultural discourses in sermons with a focus on gender constructions, while the

[1] The sermon project is based at Aarhus University, and the project partners include associate professor in practical theology Kirstine Helboe Johansen, PhD in the study of religion Anne Agersnap, associate professor in the study of religion Uffe Schjødt, software engineer at the *Centre for Computational Humanities Aarhus* Ross Deans Kristensen-McLachlan and associate professor and leader of the *Centre for Computational Humanities Aarhus* Kristoffer Laigaard Nielbo.

second study investigates how cultural and biblical domains are encountered in sermons. The main purpose of the cases in this article is, however, not to provide sociological and theological answers, but to demonstrate and discuss how automated linguistic tools have enabled us to pursue these questions.

## 3. What is reading and what do we read?

The magnitude of 11,955 sermons naturally complicates an initial in-depth reading of the material before commencing analysis. Thus, upon archiving the sermons, we had not read any of the manuscripts in full, meaning that we did not know the content of the individual sermons. While we have no prior knowledge of the specific subject matter in the sermons, we can presume that all pastors rely on fundamental linguistic principles in order to construct coherent texts. All of the included case studies are therefore based on a fundamental understanding of the corpus as a collection of language. However, uncovering linguistic structures is not our immediate purpose. Rather, our aim is to identify semantic patterns as collective representations in the text collection through information carried by linguistic features. Therefore, we need to develop pipelines that allow us to uncover cultural information in the form of semantic structures from naturally occurring linguistic features in the text collection. This includes methods to automatically juxtapose and contrast content in each document with content in all other documents in the corpus. These pipelines include a process of 'distant reading' and 'close reading'.

### 3.1 Distant reading

The concept of 'distant reading' is associated with the works of Franco Moretti (Moretti 2000), and has since then been a common concept within the 'digital humanities'. Moretti used the term to explicate what it means to study texts through the aid of computational resources, when working with large literary corpora. This approach entails applying computational and quantitative visualisation techniques to represent large collections of texts in different ways – for example in the form of graphs, maps or trees (ibid. 2007). In a state-of-the-art report on common digital visualisation techniques, Stefan Jänicke and collegues (2015) review concrete tools and techniques used to perform distant reading by summarising large data sets – such as heat maps, geospatial maps and a range of graph visualisation tools.

We use the concept of distant reading in a slightly different way than Moretti and Jänicke and colleagues (2015), as distant reading in

our methodology defines the initial phase of the analytical process. This phase includes targeting relevant linguistic components and quantifying the relationship between them, which requires computational resources. In each of the case studies – whether they investigate gender constructions or cultural and religious domains – we need to be extremely clear as to how we observe those phenomena in the corpus. Furthermore, since we study sermons as a collective production of texts, we need to target text phenomena that can potentially occur in each of the texts in the corpus. Our overall approach is therefore to target generic linguistic text components that, in relation to other text components, implicitly carry information about pastors' handling of cultural and religious discourses. In both case studies, we apply automated tools developed based on linguistic principles that can target and retrieve specific text components individually or in relation to other components. With this strategy, our distant reading dissolves the natural coherence and structure of the corpus texts, as we remove specific text features from their original text context. What we are left with is a text structure that runs across the text collection and thus reveals a pattern of repetitive relations. The distant reading, however, leaves us with no interpretation of this pattern. Instead, it provides us with a new and reconstructed text that requires a 'close reading'. Thus, in our use of the concept, distant reading is a method that implies creating new types of texts for close reading. It is rarely a full reading in itself.

### 3.2 Close reading

Moretti originally used the term distant reading to challenge traditional text readings and perspectives within literary scholarship that required meticulous scrutiny of literary works individually. He argued that such close readings cannot target the vast majority of works from a given period, and scholars therefore tend to overlook 'the great unread' of literature (Moretti 2000). Consequently, the same 'few' works are studied over and over again, which leaves behind canonised understandings of literary epochs (ibid. 2007). Jänicke and colleagues (2015) demonstrate that close reading does not necessarily contrast with digitally informed text studies, as they present various tools used for visualising and annotating individual texts (i.e. not large-scale text collections) in creative and helpful ways.

In our methodology, close reading is not the practice of full text reading as problematised by Moretti, nor is it the process of computationally re-visualising individual texts as demonstrated by Jänicke and colleagues. Instead, we use the term to define the

interpretative phase of the analytical process, which entails closely reading and understanding the cultural and religious significance of the recomposed text that we retrieve computationally through our distant reading. This close reading requires no computational competency in itself, but implicates a traditional meeting between text and analyst – even though the text looks different from the form and structure that humanist scholars traditionally work with. When we insist that close reading is an inherent part of the analytical process of large-scale text studies, we emphasise that in the end, we are first and foremost interested in our study object as texts rather than as numbers. The automated analyses and quantitative measurements are not the goal, but the means to perform close readings of text phenomena in large-scale text collections. Therefore, the aim of the close readings in the case studies is to analyse and interpret the structures that have emerged from distant readings in order to understand what these patterns established tell us about sermons. Just like with distant reading, we consider close reading a generic analytical phase, and the concrete undertaking of close reading therefore depends on specific questions and design. In the specific cases that follow, the close readings entail top-down as well as bottom-up coding analysis known from text studies in the social sciences. In the same way as a range of different computational tools can be deployed for distant reading depending on the concrete questions and data, we believe that close reading as defined above can, in principle, be carried out through adaptations of text analysis from other and different humanist disciplines.

Overall, our use of distant and close reading implies a shift between first distancing ourselves from the individual sermons and observing patterns across the text collection, and then zooming in on these patterns to observe and interpret them closely. The distant reading analysis makes it possible to observe semantic structures in the sermons, while the close reading analysis makes it possible to interpret and contextualise the semantic content. Thus, our approach is a movement between complexity reductions (removing text abundance to get a clear view of the targeted text features) and complexity increases (interpreting the various dynamics that allow these patterns to emerge in sermons).[2]

[2] The methodology of distant and close reading was initially developed in Agersnap's PhD dissertation (see Agersnap 2021).

## 4. Case studies: Distant and close readings of cultural and religious patterns

Both of the case studies are motivated and framed within different disciplines – 'practical theology', 'the sociology of religion' and 'the digital humanities' – and consequently presented in publications for different audiences. In the following, our primary venture with the cases is methodological, as we present the application of automated and linguistically based tools and the shift between distant and close reading. The background, research context and findings of each case will therefore only be stated briefly.

### 4.1 Gender discourse in sermons

Unlike other church traditions today, gender theologies are not explicitly prominent in the Evangelical-Lutheran Church of Denmark. Meanwhile, discussions of gender equality have become increasingly conspicuous in Danish society over the last years. We therefore designed this study to explore implicitly how pastors represent gendered characters in their sermons in the contemporary discursive climate in society[3].

Our entry point is male and female personal pronouns in the nominative case (Danish *han* 'he' and *hun* 'she') and oblique cases (Danish *ham* 'him' and *hende* 'her') and their associated verbs. Gendered pronouns provide concrete text features that are thoroughly integrated in the Danish language to point out characters as gendered. With their associated verbs, we are able to observe on a very fundamental level the agency patterns that pastors ascribe to female and male agents, respectively. Before retrieving text components, we segment the full sermon corpus into two separate sub-corpora, one including sermons from female pastors and one from male pastors. This allows us to study gender on two levels: how gender 'in' texts is constructed, and whether text constructions differ depending on the gender of the authors 'behind' the texts. As we design the study to include gender categories thoroughly integrated in Danish language use, we are not able to consider more diverse understandings of gender in this study. Alternative designations of gender would not be particularly widespread – if at all – in the corpus. With our approach, we are able to observe gender categories in the full collection.

As a first step in our distant reading pipeline, we use a part-of-speech tagger (POS tagger) that automatically recognises and annotates

[3] The full study appears in Agersnap and colleagues (2020).

word classes in texts. From the annotated corpora, we retrieve every instance of one of the four types of gendered pronouns and their associated verbs. This gives us a list of 3,314 unique verbs in sermons by women and 3,765 unique verbs in sermons by men. We then measure the relationship between each verb and each of the four pronouns by calculating 'pointwise mutual information' scores (PMI scores). PMI scores measure how strongly a verb is associated with each pronoun in comparison to the remaining pronouns. If the score is above 0, there is a significant and systematic pronoun-verb relationship; if the score is below 0, there is a systematic non-relationship between pronoun and verb; and if the score is 0, there is no systematic relationship, meaning that it is arbitrary whether verb and pronoun co-occur. This measuring system devalues verbs that occur frequently and are commonly associated with all four pronouns – such as the Danish *siger* 'says', *er* 'is' and *kommer* 'comes'. Therefore, verbs that do not indicate gender dependency are automatically discarded. However, this also means that verbs that are used rarely in the corpora are more likely to obtain high PMI scores. To avoid these outliers, we discard verbs that have a frequency score below 10 in sermons by male and female pastors, respectively.

This distant reading pipeline – annotating, retrieving and measuring data – outputs a table for each sub-corpora with the verbs as rows, pronouns as columns and PMI scores as the cells designating the associative strength of the relationship between pronouns and verbs. With these steps, we have recomposed the corpora into a new form of text. Figure 1 is an illustration of the content and structure of this text type.

| | word | frequency | han | ham | hun | hende |
|---|---|---|---|---|---|---|
| 0 | accepterede | 32 | 2.478988105224 9 | 0.0 | 0.0 | 5.622482465089852 |
| 1 | accepterer | 64 | 1.123465402771763 | 0.0 | 4.223102332944738 | 0.0 |
| 2 | accepteret | 63 | 0.4460665791799567 | 0.0 | 0.0 | 0.0 |
| 3 | adlyde | 37 | 0.0 | 3.949399422066509 | 0.0 | 0.0 |
| 4 | adlyder | 19 | 0.0 | 3.5172660668761835 | 0.0 | 0.0 |
| 5 | adresseret | 1 | 0.0 | 5.768557865482679 | 0.0 | 0.0 |
| 6 | adskille | 14 | 0.0 | 0.0 | 0.0 | 4.870494784506972 |
| 7 | adskiller | 83 | 1.2689729864430013 | 1.3497172576860803 | 0.0 | 0.0 |
| 8 | adskilt | 47 | 0.7390537038614309 | 0.0 | 0.0 | 0.0 |
| 9 | adskilte | 24 | 1.4111474752235438 | 0.0 | 0.0 | 0.0 |
| 10 | advarede | 18 | 3.3082674601094 25 | 2.8781861075865 14 | 0.0 | 4.619180356226066 5 |
| 11 | advarer | 80 | 2.6050699436959786 | 0.0 | 0.0 | 3.1275254794483494 |
| 12 | advaret | 23 | 0.0 | 3.2836512156946784 | 0.0 | 0.0 |
| 13 | afbilde | 4 | 0.0 | 4.382263504362788 | 0.0 | 0.0 |
| 14 | afbillede | 4 | 0.0 | 4.382263504362788 | 0.0 | 0.0 |
| 15 | afbilledet | 38 | 0.9516151458451036 | 2.1309717057562927 | 0.0 | 0.0 |

*Figure 1: Excerpt of PMI data for close reading*

The data used for close readings is based on these new tabular texts (one from each sub-corpora). It provides an overview of the verbs most strongly associated with the pronouns individually, which allow us to define and interpret the agency pattern for the male and female subjects (*han* and *hun*). However, we also sort the table to show the verbs most related to a pronoun in the nominative case (such as *hun*) and to a pronoun of opposite gender in the oblique case (such as *ham*). This means that we can detect the type of action most likely to be performed by a female agent towards a male agent and vice versa. We thus study four types of agency patterns in each sub-corpora – one pattern for male subjects, one for female subjects and two for interacting agents (*he-her* and *she-him*). We use a qualitative coding strategy to categorise agency types bottom-up by closely reading and evaluating the verbs assigned to the agent classes. This strategy discloses four agency categories that were dispersed differently in the agent classes and in the sub-corpora: 'active–competent agency' (verbs that express acts of vigour); 'cogitative–perceptive agency' (verbs that express acts of sensing and interpreting); 'communicative–expressive agency' (verbs that express extroverted or contact-oriented acts); and 'religious agency' (verbs that express acts connected to a religious context in particular). See table 1 for an example of the categories represented for male and female subjects individually (*han* and *hun*) in sermons by female pastors.

| | **Single case: Male subject** | | | **Single case: Female subject** | | |
|---|---|---|---|---|---|---|
| **Category** | **Verb** | **Freq.** | **PMI** | **Verb** | **Freq.** | **PMI** |
| Active-Competent agency | Undertook | 21 | 3.2 | Gave birth | 72 | 5.2 |
| | Promise | 388 | 2.9 | Involves | 20 | 4.0 |
| | Overcame | 57 | 29 | Managed | 37 | 3.4 |
| | Direct | 82 | 2.8 | Achieve | 39 | 3.3 |
| Cogitative-Perceptive agency | Sensed | 29 | 3.0 | Misunderstand | 28 | 4.1 |
| | Regrets | 63 | 2.8 | Dreamt | 82 | 3.7 |
| | Decides | 87 | 2.9 | Imagined | 42 | 3.6 |
| | % | % | % | Recognises | 201 | 3.3 |
| Communicative-Expressive agency | Discussed | 32 | 3.1 | Greeted | 22 | 4.3 |
| | Propagate | 32 | 2.9 | Kisses | 72 | 4.3 |
| | Declare | 32 | 3.1 | Cries | 366 | 3.4 |
| | Asserted | 27 | 3.2 | Expressed | 83 | 3.3 |
| | Sermonised | 129 | 2.8 | Anointed | 110 | 3.2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Religious agency | Heals | 213 | 2.8 | % | % | % |
| | Preached | 101 | 2.8 | % | % | % |
| | Revealed | 101 | 2.8 | % | % | % |

*Table 1: Distribution of analytical categories in sermons by female pastors. The table presents verbs representative of one of the four categories.*

The analyses show that male and female pastors ascribe the categories differently to female and male agents. Male pastors seem to reserve the 'active-competent agency' to male subjects and 'cogitative-perceptive agency' and 'communicative-expressive agency' to female subjects. Comparatively, female pastors distribute these three agency traits to male and female agents alike. 'Religious agency' is a trait mostly found in sermons by women, who in particular ascribe it to male subjects or to female subjects (*hun*) interacting with a male object (*ham*).

These findings suggest that male pastors represent social worlds with a clear difference between the agential spaces of male and female characters, while female pastors represent social worlds indicating that female and male characters hold more and similar agential spaces. Our interpretation of these observations is that pastors reassert general social patterns pertaining to men and women. In sociolinguistic theory (Wood 2014:102-122, Tannen 1992:23-95), male speech communities are characterised by language used to assert differences between conversational partners and language used instrumentally to provide solutions for concrete objectives. Comparatively, female speech communities are characterised by a language that seeks to establish connections between communicating partners and language used to display support and responsiveness. Male speech communities are thus concerned with asymmetric relationships and female speech communities with symmetric relationships. The findings thus suggest that sermons are not gender-neutral documents, but influenced by the socialisation processes that tend to pertain to men and women. From these observations, we cannot say whether these traits affect the theologies that male and female pastors represent. Instead, we can observe that discursive strategies in society at large also exist in sermons and that the social worlds in sermons are affected by the social reality that pastors themselves are part of.

### 4.2 Cultural and religious domains in sermons

The purpose of the second case study is to explore which aspects of culture and Christian tradition pastors tend to interact with most

distinctly in sermons. Similar to the first case study, we want to investigate these tendencies based on features that are common in language and therefore bound to be widely disseminated in sermons. Following this principle, we choose named characters as features for the study, since language tends to revolve around subjects doing something. In contrast to for example literature, sermons are not fictional texts, and pastors therefore do not construct characters in sermons freely. Instead, pastors invite characters in from a range of different contexts, whether biblical, literary, political, popular cultural, contemporary or historical. The characters in sermons consequently provide unique insight into the aspects of culture and Christian tradition that pastors represent the most and most distinctly[4]. We develop a pipeline to extract and label characters from the sermon corpus and to study character relationships in order to disclose how characters and thereby cultural and religious domains interact in sermons. The pipeline includes more than one shift between distant and close reading.

In the first phase of distant reading, we use a tool for 'named-entity recognition' (NER) to automatically detect and retrieve all characters in the corpus. The NER tagger is trained to identify entities such as persons, locations and organisations based on their syntactical position and function in the texts. We only include characters mentioned 10 times or more in the corpus, and based on this information, we then obtain a tabular overview of the most frequent characters in the corpus. In the first phase of close reading, we label each character in order to categorise them. First, we distinguish between four main types: biblical characters (*Jesus, God*), non-biblical but known characters (*Astrid Lindgren, Hitler*), non-specific characters holding a proper noun (*Anders, Sophie*) and characters named according to a role or function (Danish *kongen* 'the king'*, gartneren* 'the gardener'). For the purpose of the study, we are most interested in the biblical characters, which we sub-label as New Testament or Old Testament characters, and the non-biblical characters, which we sub-label according to the cultural domain they represent – such as literary authors or politicians. All characters are sub-labelled in regard to gender and geographical origin as far as possible. See the coding manual below in table 2.

[4] The full study appears in Agersnap and colleagues (2022).

| CATEGORY | CODE | DEFINITION |
|---|---|---|
| Type | B | Biblical character |
| | B_a | Ambiguous biblical character |
| | N | Non-biblical character |
| | P | Proper nouns |
| | R | Social role |
| Gender | 1 | Male character |
| | 2 | Female character |
| | 0 | Character unidentifiable as male or female |
| Time | OT | Old testament character |
| | NT | New testament character |
| | ONT | Character or name in the old and new testament |
| | CON | Character who was alive in 2011-2016 |
| | PAST | Character who was not alive from 2011-2016 |
| Nationality | DK | Danish character |
| | NOR | Icelandic, Finnish, Swedish, Norwegian, Faroese or Greenlandic character |
| | EUR | European character |
| | NAM | US or Canadian character |
| | GLO | Non-European and non-North American character |
| Domain | POL | Politician |
| | LIT | Fictional writer |
| | MUS | Musician |
| | ART | Artist within the visual arts |
| | THE | Theologian |
| | ACT | Actor |
| | ROY | Royal character |
| | REA | Character from reality show |
| | TV | TV personality (show host, anchor etc.) |
| | RAD | Radio personality (show host, anchor etc.) |
| | SOC | Social media personality (bloggers, youtubers etc.) |
| | SPO | Athlete |
| | BUS | Personality associated with business communities |
| | ACA | Academic |

| | |
|---|---|
| DEB | Public voices or debaters |
| FIC | Fictional character ex. from literature, movies, tv shows etc. |
| LEG | Legendary character |
| OUT | Unidentifiable character |

*Table 2: Coding manual for named entities*

This process demonstrates that our use of close reading does not necessarily equate to a qualitative reading, since the labelling procedure provides our data with quantitative variables. Most importantly, close reading in our terminology entails observing and evaluating retrieved data meticulously through a process that presupposes automated text processing, but cannot be automated itself. With this first phase of close reading, we therefore manage to correct some of the flaws and insufficiencies made by the NER tagger – such as different spellings of the same character or text components wrongly tagged as personal entities. By correcting these mistakes, we end up with a data set of 600 unique characters. We thus use the first analytical phase of close and distant reading to get an overview of the frequencies and types of characters in the corpus.

In the second analytical phase, we perform a 'social network analysis' (SNA) to investigate how these characters encounter each other in sermons. SNA measures and describes how components in any type of system, called 'nodes', interact or relate to each other, which is visually demonstrated by an 'edge' between nodes. As a tool for distant reading, we use the open-access program Gephi to construct and statistically evaluate the character network (Bastian et al. 2009). In our network, the retrieved characters are the nodes, and the edges – the relationship between characters – are defined as the co-occurrence of characters in a sermon. We use Gephi to detect sub-groups of characters in the full network with the program's community detection algorithm. The algorithm detects 10 modularity classes (i.e. sub-groups) of characters that tend to co-occur systematically or to co-occur with similar characters. This procedure provides us with a rematerialised text in tabular form with characters as rows and all network parameters – including modularity classes – and previously assigned labels as columns. See figure 2 below.

| | Id | modular | Domain | Gender | Type | Degree | Weighted | Eccen | closeness | harmonic | clo | betweenness | centr | Authority | Hub | pageranks | clustering | triangles | eigencentrality | Frequenc | Type | Gender | Domi | Time | Nationa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Id | | | | | | | | | | | | | | | | | | | 1 | | | | | |
| 2 | Jesus/Jesus Kristus | 0 | B | 1 | NT | 405 | 9859 | 3 | 0,7506 | 0,836672 | | 39839452724 | 0,19717 | 0,19717 | 0,021025 | 0,08844 | 7235 | | 0,463246 | 1754 | B | 1 | NT | NT | % |
| 3 | Josef | 0 | B | 1 | ONT | 101 | 1391 | 4 | 0,5292 | 0,574708 | | 1107087581 | 0,09464 | 0,09464 | 0,004537 | 0,36436 | 1840 | | 0,463246 | 1754 | B | 1 | ONT | ONT | % |
| 4 | Kristus | 0 | B | 1 | NT | 204 | 2892 | 4 | 0,5907 | 0,664858 | | 4408255909 | 0,14445 | 0,14445 | 0,009247 | 0,19477 | 4033 | | 0,714983 | 2384 | B | 1 | NT | NT | % |
| 5 | Paulus/Saulus | 0 | B | 1 | NT | 182 | 3486 | 4 | 0,5776 | 0,646216 | | 2850001301 | 0,13998 | 0,13998 | 0,008123 | 0,23381 | 3851 | | 0,689483 | 3802 | B | 1 | NT | NT | % |
| 6 | Jakob | 0 | B_a | 1 | ONT | 116 | 1177 | 4 | 0,5392 | 0,588898 | | 1210819952 | 0,1026 | 0,1026 | 0,005204 | 0,31934 | 2130 | | 0,503115 | 1669 | B_a | 1 | ONT | ONT | % |
| 7 | Johannes | 0 | B_a | 1 | NT | 149 | 3823 | 4 | 0,5567 | 0,617001 | | 2019452323 | 0,12013 | 0,12013 | 0,006668 | 0,26102 | 2878 | | 0,591119 | 6398 | B_a | 1 | NT | NT | % |
| 8 | Maria | 0 | B_a | 2 | NT | 81 | 2765 | 4 | 0,5195 | 0,557735 | | 2700115152 | 0,08777 | 0,08777 | 0,003542 | 0,48827 | 1582 | | 0,427601 | 6003 | B_a | 2 | NT | NT | % |
| 9 | Peter/Simon Peter | 0 | B_a | 1 | NT | 131 | 2509 | 4 | 0,5445 | 0,600306 | | 2629781649 | 0,10851 | 0,10851 | 0,00615 | 0,27704 | 2359 | | 0,533491 | 6066 | B_a | 1 | NT | NT | % |
| 10 | Stéphane Hessel | 0 | N | 1 | DEB | 1 | 2 | 4 | 0,4291 | 0,44616 | | | 0,00265 | 0,00265 | 0,000294 | | 0 | | 0,013887 | 18 | N | 1 | DEB | CON | EUR |
| 11 | Adriana | 0 | P | 2 | % | 1 | 1 | 4 | 0,4291 | 0,44616 | | 0 | 0,00265 | 0,00265 | 0,000294 | | 0 | | 0,013887 | 12 | P | 2 | % | % | |
| 12 | Herren | 1 | B_a | 1 | ONT | 293 | 1492 | 3 | 0,6561 | 0,742348 | | 1441036763 | 0,17404 | 0,17403 | 0,013926 | 0,13486 | 5769 | | 0,873172 | 1249 | B_a | 1 | ONT | ONT | % |
| 13 | Martin Luther | 1 | N | 1 | THE | 214 | 1228 | 4 | 0,599 | 0,674318 | | 6293308131 | 0,14735 | 0,14735 | 0,009974 | 0,18446 | 4204 | | 0,73138 | 1481 | N | 1 | THE | PAST | EUR |
| 14 | N.F.S. Grundtvig | 1 | N | 1 | THE | 206 | 1081 | 4 | 0,5925 | 0,666806 | | 5506629742 | 0,14863 | 0,14863 | 0,009568 | 0,20549 | 4339 | | 0,735544 | 1154 | N | 1 | THE | PAST | DK |
| 15 | Søren Kierkegaard | 1 | N | 1 | THE | 158 | 541 | 4 | 0,5598 | 0,623678 | | 2678520567 | 0,12372 | 0,12372 | 0,007131 | 0,24752 | 3070 | | 0,609682 | 614 | N | 1 | THE | PAST | DK |
| 16 | Esajas | 1 | B | 1 | OT | 133 | 1119 | 4 | 0,5485 | 0,603645 | | 1117874641 | 0,1173 | 0,1173 | 0,005818 | 0,31397 | 2756 | | 0,575045 | 989 | B | 1 | OT | OT | % |
| 17 | Gud | 1 | B | 0 | ONT | 126 | 651 | 4 | 0,5445 | 0,597524 | | 1329211153 | 0,11214 | 0,11214 | 0,005607 | 0,32343 | 2547 | | 0,549375 | 357 | B | 0 | ONT | ONT | % |
| 18 | David/Kong David | 1 | B_a | 1 | OT | 122 | 885 | 4 | 0,5411 | 0,59335 | | 957880752 | 0,1099 | 0,1099 | 0,005368 | 0,33261 | 2455 | | 0,538623 | 1060 | B_a | 1 | OT | OT | % |
| 19 | Herodes | 1 | B | 1 | NT | 120 | 721 | 4 | 0,5401 | 0,591681 | | 1093652657 | 0,10844 | 0,10844 | 0,005276 | 0,33978 | 2426 | | 0,530484 | 847 | B | 1 | NT | NT | % |
| 20 | Faderen/Fader | 1 | B_a | 1 | NT | 109 | 568 | 4 | 0,5348 | 0,582499 | | 1344412206 | 0,10261 | 0,10261 | 0,004959 | 0,36731 | 2162 | | 0,501743 | 429 | B_a | 1 | NT | NT | % |
| 21 | Augustus/Kejser Augu | 1 | B | 1 | NT | 95 | 265 | 3 | 0,5373 | 0,570117 | | 6261076 | 0,09025 | 0,09025 | 0,004276 | 0,3794 | 1694 | | 0,441784 | 177 | B | 1 | NT | NT | % |
| 22 | Pontius Pilatus | 1 | B | 1 | NT | 94 | 338 | 4 | 0,5231 | 0,567195 | | 31800628 | 0,09769 | 0,09769 | 0,00405 | 0,45733 | 1999 | | 0,475873 | 267 | B | 1 | NT | NT | % |
| 23 | Menneskesøn/Menne | 1 | B | 1 | ONT | 86 | 385 | 4 | 0,52 | 0,560796 | | 605684681 | 0,08915 | 0,08915 | 0,003842 | 0,44679 | 1633 | | 0,434789 | 258 | B | 1 | ONT | ONT | % |
| 24 | Satan | 1 | B | 0 | ONT | 86 | 314 | 4 | 0,5159 | 0,558292 | | 284620052 | 0,08743 | 0,08743 | 0,00374 | 0,43529 | 1591 | | 0,426479 | 279 | B | 0 | ONT | ONT | % |
| 25 | Jeremias | 1 | B | 1 | OT | 84 | 290 | 4 | 0,52 | 0,559683 | | 439219629 | 0,08359 | 0,08359 | 0,00376 | 0,40333 | 1406 | | 0,40952 | 280 | B | 1 | OT | OT | % |
| 26 | Elias | 1 | B_a | 1 | OT | 82 | 376 | 4 | 0,5124 | 0,553979 | | 236421947 | 0,08743 | 0,08743 | 0,003561 | 0,48359 | 1606 | | 0,424939 | 448 | B_a | 1 | OT | OT | % |
| 27 | Kai Munk | 1 | N | 1 | THE | 81 | 164 | 4 | 0,5182 | 0,5569 | | 338714385 | 0,08224 | 0,08224 | 0,003581 | 0,42315 | 1371 | | 0,40209 | 172 | N | 1 | THE | PAST | DK |

*Figure 2: Excerpt from network data including statistical parameters and manually assigned labels*

We use the table for a close reading of the modularity classes, focusing on the thematic contexts that the characters, based on their original domain, represent now through their connections to other characters. The first phase of close reading thus identifies the broader cultural or religious context that sermon characters represent, while the second phase of close reading considers the context which the characters enter into 'in' sermons.

The network analysis unveils three interesting types of sub-groupings. Most of the sub-groupings constitute narrative clusters highlighting different stories and roles of Jesus: 'the saviour', based on connections with New and Old Testament characters representative of the nativity narrative in the bible; 'the teacher', based on connections with New Testament characters from gospel texts associated to the Trinity season of the church calendar[5]; and 'the compassionate', based on connections with biblical outcasts, but also literary authors, writing about moral and the breaking of norms (*Fyodor Dostoyevsky, Astrid Lindgren)*.

Another type of sub-group represents a theme connected to this-worldly strife and evil. The most central figure in the cluster is Adolf Hitler, who, along with other political figures, comprises a World War II motif. Furthermore, the sub-group includes a group of terrorists, such as the Norwegian perpetrator Anders Breivik. This sub-group of historical and contemporary figures constitutes a problematic domain in the sermons revolving around this-worldly atrocities and evil.

The final and largest sub-group in the network data is characterised by containing characters with low frequency scores who do not co-occur with many other characters in the corpus. This means that what links them as a group is that they are all anonymous characters in the sermons, unable to connect to coherent narratives, whether biblical or historical. These figures are in particular contemporary figures and Old Testament characters.

The sub-groups demonstrate which narratives and motifs pastors in the corpus agree the most about disseminating. Our findings show that pastors agree about enhancing some gospel narratives rather than others. For every church service in the church calendar, pastors in the ELCD are obliged to read aloud a prescribed gospel passage before giving their sermon, and they are supposed to base their sermon on this

[5] Trinity refers to a specific part of the liturgical year in the ELCD. Trinity runs from the second Sunday after Pentecost until Advent. The main topic of this part of the liturgical year is the holy Trinity.

passage. Since the nativity narrative and trinity stories appear as coherent structures in the network data, we hypothesise that pastors at large paraphrase and deal with these narratives in particular as compared to other gospel passages in the church year. Meanwhile, in sermons about biblical outcasts, pastors tend to include characters from a literary domain as well to emphasise certain themes of the gospel passages. In comparison to the biblical clusters, the cluster centred on a World War II motif and worldly atrocities illustrates a consensus among pastors that this particular historical period is important in a church context. In contrast, the large group of anonymous figures is made up of the characters whose roles and relevance pastors are not in agreement about yet, but may be negotiating. We hypothesise that these 'new' figures over time may obtain a more central status or disappear altogether.

### 4.3 From basic features of language to complex cultural patterns

The study on gender and the study on biblical and cultural domains in sermons demonstrate how we can study rather complex semantic and cultural patterns in large-scale text corpora through very basic assumptions of how language works: constructing sentences revolving around agents as well as applying gender categories or names to these agents are integral features of Danish language. The taggers applied in the studies (POS and NER) testify in themselves to these assumptions: since these features are integral to language, the process of identifying agents in texts can be automated. The gendered as well as named entities provide linguistic components that can be traced in the corpus in full. While we acknowledge that the individual representations of the named entities or gendered pronouns take part in many concrete and diverse contexts, we argue that the POS tagger and NER tagger allow us to investigate the larger structures that these components occur in and their implicit connotations in the corpus. These connotations provide insights into culturally significant information.

## 5. Zooming out and zooming in

In his book on practice theory, work and organisation, practice theorist Davide Nicolini describes the basic method of studying practice as a constant process of zooming in and zooming out (Nicolini 2012). Nicolini is not interested in text studies; he is concerned with field work and careful participant observation. However, as we have tried to describe above, studying a shared practice from textual resources entails the same basic methodology – though the first step is zooming

out rather than zooming in. As Moretti points out, the same kind of texts are read continuously and in our case, sermons and preaching have been studied for ages. Computational tools such as those we have used and described are excellent for creating analytical distance to a highly studied field. Such tools allow scholars to read huge volumes of text material focusing on particular textual features. The output of the zooming out enabled by computational tools is new kinds of texts in strange formats that now have to be read for scholars to gain research insights. Close reading these texts entails zooming in on particular aspects of a text corpus, but for the reader to understand and interpret this, domain-specific competences and theoretical perspectives are needed as in any other interpretative process. These reflections resonate with ongoing developments in digital and computational humanities research. In 2009, Todd Presner and Jeffrey Schnapp argued that they could detect an increasing attention to "complexity, medium specificity, historical context, analytical depth, critique and interpretation" (Schnapp & Presner 2009:2). They described this development as the second wave of digital humanities in contrast to the first wave, which implied a more narrow understanding of computational power as primarily a helpful tool to ask new questions. With the second wave, there is an initiative to integrate tools to reconsider also epistemologically the main objective of study in humanities research, such as culture and human agency (Evans & Rees 2012). For example, Paul Baker and Tom McEnery demonstrate the convergences between corpus linguistic methods and discourse analysis (Baker & McEnery 2015). However, more theoretically encompassing attempts to define reading anew have also emerged. Literary scholar Andrew Piper uses the term 'topological reading' to define computational text analysis as a reading of horizontal networks of meaning inspired by Deleuzian theory (Piper 2013), and more recently, appellations such as *computational hermeneutics* and *thick reading* have appeared to integrate computational analysis of text within a more traditional hermeneutical framework in the humanities (Mohr et al. 2015, Fuenmayor & Benzmüller 2019). Comparatively, in the social sciences, Swedish sociologist Simon Lindgren has demonstrated how computationally informed methods align with theoretical horizons in social theory, arguing that relying on data science in other disciplines requires that one dismantles and reassembles in new ways existing theories and methods within one's field (Lindgren 2020). We find the same to be relevant and promising in humanities research.

It has commonly been argued that we are still lacking well performing computational linguistic tools for Danish texts. In our process of zooming out through distant reading, we have strived to use tools that are not technically complex. We are humanist scholars and we deploy digital tools to be better humanist scholars – not to transform into digital scholars. The data retrieved with the help of these rather basic tools are quite intuitive from a humanist perspective. The new types of texts that are created in the process of distant reading lend themselves almost seamlessly to the traditional analytical process of close reading a text. Thus, with this article, we hope to have demonstrated that we have been able to come a rather long way in the study of sermons through very basic tools. Selecting and paying attention to basic functions of language across large corpora of texts allows us to uncover core structures in the sermons; and close reading these core structures gives us new insights into different aspects of cultural and religious dynamics in sermons. The methodological pipelines that we deploy may easily be reproduced in other types of material. Close to the format that we have been working with – rather short texts meant for oral presentation – would be other types of public speeches in connection with Constitution Day, 1 May, New Year or the opening of parliament. Also, these types of speeches are both the speech of an individual person and a collective practice of public speaking on the specific occasion. However, because the digital tools that we have used are simple and because we focus on basic linguistic features, a similar approach may be used for a range of textual formats and genres. The revolution of text digitisation has indeed paved the way for even new understandings of texts and cultural practices fixed in texts. For us, the main revolution has been the creation of new challenging texts to read and interpret.

**References**

Agersnap, A., R. D. Kristensen-McLachlan, K. H. Johansen, U. S. & K. L. Nielbo (2020). Sermons as data: Introducing a corpus of 11,955 Danish sermons. In: *Journal of Cultural Analytics*, 5(2), 1-27. Online: https://doi.org/10.22148/001c.18238.

Agersnap, A. (2021). *Collective testimonies to Christianity and time - A collection and large-scale text study of 11,955 Danish sermons from 2011-2016*. (PhD dissertation). Aarhus University, Faculty of Arts.

Agersnap, A., K. H. Johansen, & R. D. Kristensen-McLachlan (2022). Unveiling the character gallery of sermons – A social network analysis of 11,955 Danish sermons. In: *Temenos – Nordic Journal of Comparative Religion*. [Forthcoming].

Baker, P. & T. McEnery (2015). Introduction. P. Baker and T. McEnery (eds.) *Corpora and discourse studies – Integrating discourse and corpora*, Palgrave Macmillan, 1-19.

Bastian, M., S. Heymann & M. Jacomy (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In: *International AAAI Conference on Weblogs and Social Media*.

Jänicke, S., G. Franzini, M. F. Cheema & G. Scheuermann (2015). On close and distant reading in digital humanities: A survey and future challenges. *Eurographics Conference on Visualization (EuroVis)*.

Evans, L. & S. Rees (2012). An interpretation of digital humanities. D. Berry (ed.) *Understanding Digital Humanities*. Palgrave MacMillan, 21-41.

Fuenmayor, D. & C. Benzmüller (2019). A computational-hermeneutic approach for conceptual explicitation. Á. Nepomuceno-Fernández, L. Magnani, F. J. Salguero-Lamillar, C. Barés-Gómez & M. Fontaine (eds.) *Model-based reasoning in science and technology: Inferential models for logic, language, cognition and computation*. Springer, 441-469.

Lindgren, S. (2020). *Data theory: Interpretive sociology and computational methods*. Cambridge: Polity Press.

Mohr, J. W., R. Wagner-Pacifici & R. L. Breiger. 2015. Toward a computational hermeneutics. *Big data & society*. Online: https://doi.org/10.1177/2053951715613809

Moretti, F. (2000). Conjectures on world literature.In: *New left review,* 1(1), 54-68.

Moretti, F. (2007). *Graphs, maps, trees – Abstract models for literary history*. London: Verso.

Nicolini, D. (2012). *Practice theory, work, and organization: An introduction*. Oxford: Oxford University Press.

Piper, A. (2013). Reading's Refrain: From Bibliography to Topology. *ELH* vol. 80(2), 373-399. John Hopkins University Press.

Schnapp, J. & T. Presner (2009). Digital Humanities Manifesto 2.0. Online: http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

Tannen, D. (1992). *You just don't understand: Women and men in conversation*. London: Virago.
Wood, J. T. (2014). *Gendered lives – Communication, gender and culture*. Cengage Learning.