Script and Text in Time and Space

Introduction

The project *Script and Text in Time and Space* will create a new foundation for working with medieval Danish texts by creating and deepening fundamental knowledge about the development of medieval script in Denmark as well as the dating and localization of relevant text bearing objects (manuscripts and other document types). The project is placed in the field of *Digital Humanities,* meaning that digital tools and general guidelines for editing and analysing primary sources play a central role.

Script – the written manifestation of language – is key to most analyses of medieval language and literature as well as history and culture in general. In fact, it is usually because of script and the language it preserves that a primary source can be placed in time and space. In recent years there has been an increased interest in the transmission and reception of historical texts and documents, which renews the need for and underlines the importance of detailed studies in the history and origin of these sources.

Despite dating and localization of a written text based on the language and script being central to philological studies, there are surprisingly few tools and fundamental reference points available when it comes to Danish script. Particularly compared to the substantial work on West Norse, little systematic analysis has been conducted on Danish medieval texts, including the circumstances of their production and the development of the script.

The project addresses this issue by conducting both qualitative and quantitative studies of the relevant features in medieval Danish charters (official documents) from the thirteenth to the sixteenth centuries. Unlike almost all other written

sources from the Middle Ages, charters usually provide very exact information on when, where and by/for whom they were written. Moreover, many charters are preserved in their original form (as opposed to a copy), meaning that they are excellent sources for linguistic and palaeographical analyses.

The project is divided into three sub-projects, which work with the same corpus material. The corpus consists of over 470 medieval Danish charters that once belonged to the St Clara monastery outside of Roskilde. Sub-project 1 produces a state-of-the-art digital scholarly edition of the charters and makes the richly annotated electronic corpus available online. Sub-project 2 conducts a systematic, qualitative analysis of both the script, the language and historical circumstances documented in the charters and feeds that information into the corpus in the form of mark-up. Sub-project 3 develops digital tools to support and facilitate the work of sub-projects 1 & 2 and to improve quantitative analysis of historical sources in general.

The way scholarly editions are prepared is in the process of changing, and we are in the midst of a transition from (traditional) printed editions to digital solutions for producing and publishing high-quality textual editions. Sub-project 1 will develop methods and technologies for improving digital scholarly editions of text bearing objects. The text preserved in charters is usually strongly linked to the "real word" through explicit references to the writing process itself as well as people and places involved. Therefore, these documents present a particularly good basis for re-thinking digital mediation and dissemination of historical texts.

The linguistic and palaeographical studies of sub-project 2 work towards developing new methods for automated dating, localization and grouping of scribes. With that aim, the examination of the corpus material is supplied by quantitative analyses of a large number of additional charters from both Denmark, Norway and Iceland. Furthermore, we believe strongly that the application of machine learning techniques and the knowledge created in sub-project 3 will help generate new knowledge and a better understanding of the relevant factors for establishing the date, place and scribe of primary sources from the Middle Ages. Such insight will be particularly useful and welcome with regard to Danish material, for which a solid foundation is still to be established.

The three sub-projects are very closely related and highly benefit from each other as they utilize the same material and create a common richly annotated corpus. Further, the different sub-projects share empirical, theoretical and method-

ological approaches, and the interaction between both qualitative and quantitative research promises to yield interesting results.

The project will be hosted at the Department of Nordic Research (University of Copenhagen), where it is related to both current and future activities within the department's core research areas such as palaeography, linguistics, onomastics, language technology, textual editing and digital humanities. Among others, the project will play an important role for the development of a digital edition series at the department.

The combination of systematic analysis of the charters with an open-source scholarly edition and development of customized digital tools will enable a new dimension of research opportunities. The results are expected to stimulate not only textual, linguistic, historical and cultural studies in Denmark but to have a major international impact.

Sub-project 1: Digital textual scholarship

Sub-project 1 focuses on theoretical, methodological and infrastructural questions related to producing a state-of-the-art digital edition of medieval primary sources. The main goals of the sub-project are:

1. Theoretical discussion of the defining criteria of a state-of-the-art digital edition of medieval sources.

2. Establishment of an advanced digital (dynamic, interactive) edition of all charters from the St Clara monastery.

3. Development of open-source software (editing tool) that enables analysis, publication and dissemination of handwritten medieval texts in both printed and digital form.

Traditionally, scholarly editing of charters has focused almost exclusively on the text that is preserved in these documents and its contents. Material and provenance-related aspects of these artefacts have received much less attention.

In line with recent developments, the edition will incorporate material-philological, socio-historical and other interdisciplinary approaches. As a result, the charters will be presented as complete physical objects that can be investigated

and researched from multiple angles. The edition will include facsimile images, transcriptions of the texts on several editorial levels, as well as rich meta data on historical, linguistic and other circumstances related to the act of writing the charters.

A prerequisite for producing such an edition is a thorough theoretical discussion of what a state-of-the-art scholarly digital edition is compared to its traditional counterpart. While the field is in the midst of a transition from printed editions to digital editions, most digital solutions are still modelled on the static prototype provided by traditional editions.[1] In order to move on to digital editions that take full advantage of the dynamic opportunities of digital solutions, theories and reflections need to go hand in hand with actual implementations. Therefore, sub-project 1 and 3 will collaborate closely in order to not only rethink what a state-of-the-art, dynamic and interactive edition should be, but also realize and implement such a new form of a digital edition, using the charters from St Clara as example material.

One of the main characteristics of digital editions is their dynamic structure. Other than a printed edition, a digital edition may provide customizable display options for, among others, the transcription level(s). Being based on an electronic, richly annotated corpus, digital editions also open up for dynamic search and sampling features as well as highly advanced export options. That means a well-executed digital edition with dynamic text and data display is an active research tool. It can provide excellent built-in features for users with various interests and greatly facilitate their research. Additionally, a good digital edition with such sample features frees up valuable time for actual analysis that would otherwise be used for sampling and excerpting texts. On-the-fly excerpting moreover enables reproducibility, which make research more transparent and ensures better quality. Finally, it allows one to test all kinds of hypotheses, also concerning superficially less-important seeming details, which would otherwise be too time-consuming to investigate.[2]

[1] Characteristics of possibilities of digital editions are frequently discussed in the field of scholarly editing. See e.g. various articles in Burnard et al. (2006), Deegan et al. (2009) and Ambrosio et al. (2014); the latter has a particular focus on digital diplomatics.

[2] The value of on-the-fly excerpting of highly annotated material from handwritten sources is discussed, among others, in Kjeldsen (2015).

The specifics of the digital edition are highly dependent on the outcome of the theoretical discussion led by the sub-project. Therefore, many of the details and features are not yet known. Nonetheless, the edition will certainly include an advanced user interface that supports customized searches for time, place, people, scribes, palaeographical features, various actor's roles, keywords as well as many different linguistic traits. Additionally, each word will have a hyperlink to either an online dictionary or, in the case of place names, a digital map. Together with the multi-level display of the text, this ensures the usability of the edition by a wide audience and will serve as an unsurpassed dissemination and teaching resource.

The technical realization of the edition will be based on a system developed by Alex Speed Kjeldsen. He designed a dynamic online application for digital publishing of the oldest Icelandic charters, which will serve as a starting point for the new edition.[3] The system is implemented as a Single Page Application in JavaScript and is highly adjustable, which is why it is also expected to form the basis for a planned digital edition series for manuscripts, *Editiones Arnamagnæanæ Electronicæ*. Among the system's most important characteristics are: **1)** Simple architecture that is independent of external programs and visualization tools. This allows it to function as an off-line tool. **2)** Usage of commonly employed open-source technologies such as HTML5 and JavaScript. **3)** Data structures with clearly laid-out mapping from and to XML, ensuring straight-forward import and export of texts. **4)** Openness to various adjustments and extensions due to its self-contained infrastructure.

Examples of the functionality of the system designed by Kjeldsen are:

- Multi-level text display, e.g. the three levels recommended by Menota (Medieval Nordic Text Archive), *facsimile*, *diplomatic* and *normalized*[4]:
  - **facs:** Wær arne m; guȝz naȝh ƀp
  - **dipl:** Wær arne m*ed* gudz nadh b*ysco*p
  - **norm:** Vér Árni með Guðs náð byskup
- Complete linking between text and image on a word-by-word basis. Such linking can, among others, be used in searches.

[3] A beta version of Kjeldsen's charter edition is available online at `https://dl.dropboxusercontent.com/u/2327395/udgave/index1.html`.

[4] On levels of text representation see chapter 3 of *The Menota handbook* v2.0.

- Advanced searches that take advantage of the different text representations, information on lemmata, morpho-syntactic annotation, orthographic-phonological and palaeographical mark-up, identification and geo-tagging of proper names, as well as layout and structural information.
- Automatically generated lists of, for instance, the frequency of special characters, word forms, lemmata and grammatical categories or overviews of grapho-phonemic relations, alphabetically or grammatically ordered word lists, as well as extensive lists of people and places with links to digital maps.
- Multiple options for presenting extracted data. For instance, if a user searches for a phrase, the results can be shown both as a KWIC-concordance and using an n-gram viewer.

The majority of philologists and textual scholars perceive the lack of custom-made digital tools as a considerable challenge or barrier that prevents them from benefiting from digital solutions in their research. From experience we know that many find it ineffective to work with raw XML data, and it is a major issue that not only the individual scholars, but also many institutions do not have the necessary expertise of XML-related technologies. In particular, for institutions involved in editing that want to make the transition to digital work practices, this is highly problematic. Therefore, the sub-project will develop a customizable specialised tool for editing, analysing and disseminating handwritten texts in both printed and digital form.

The software (editing tool) will be made available under an open-source license and allow users to customize it according to their individual editorial and scholarly needs. The system can be configured through an easy-to-handle user interface, which allows for selection and de-selection of various options. Based on the chosen functionality, the program generates the individually relevant HTML, CSS and JavaScript files, which – together with corpus files – form the entire edition. This set-up minimizes the need for technical support for projects using the system. At the same time, it ensures a high degree of re-usability of source code between various editorial projects.

A large part of the specialized functionality will draw on language technology (see sub-project 3), such that the system will support semi-automated orthographical standardization (normalization), lemmatization and morpho-syntactical

annotation. Recent developments in natural language processing will moreover come to use in the implementation of advanced searching mechanisms, linking between images and transcriptions, as well as integrating external resources such as online dictionaries and digital maps. Thus, besides being a state-of-the-art edition with many advantages for the user, the system also functions as an excellent internal editing tool that facilitates the work of editors inserting and correcting data. It moreover supports exports of the annotated text to various relevant formats such as TEI-compatible XML (for data exchange), LaTeX (for printed publications) and the edition's own corpus format.

Like the user-end for the digital edition, the editorial side of the system is to a large part based on an application developed by Alex Speed Kjeldsen. The internal editing tool, currently with the working title Menotab, uses the free and open-source editor GNU Emacs and its built-in org-mode, that provides an excellent working and encoding environment.[5] Emacs provides an exceptionally high degree of programmability, since the editor at its core is an interpreter for Emacs Lisp, a dialect of the Lisp programming language with extensions to support text editing. That means each user is in principle able to extend and adjust the editor and its system based on his or her specific requirements. This ensures the system's usability across projects working on very diverse kinds of texts and with various languages. The high degree of programmability also makes it possible to use the same system throughout the different phases of a project ranging from the initial transcription to the publication of the digital edition. Similarly, when writing a scholarly introduction to an edition or working on articles based on the corpus material the editorial users benefit from the system's analytical tools. Not least, the system enables and supports reproducible research.[6]

Finally, sub-project 1 will provide a free and open-source monospaced font that includes all the relevant letters and glyphs for Old Danish. It will be based on an already existing free font, DejaVu Sans Mono, which will be expanded with a large number of characters defined by MUFI character recommendation v4.0.

[5] The name Menotab is a play on words referring to Menota's transcription levels (which are directly mapped and automatically exported) as well as the tabular set-up of the system.

[6] Among others, to use the built-in analytical functions and integrate them in scientific writing based on empirical studies has proven extremely fruitful and time-saving, as it eliminates the need for double bookkeeping (e.g. Kjeldsen [forthcoming]).

Sub-project 2: Studies in writing. The archive of St Clara monastery

Sub-project 2 focuses on palaeography, the history of language and script as well as orthography. Furthermore, it engages in diplomatics and studies in the monastery's history. This sub-project follows a decidedly interdisciplinary approach for investigating writing and text production in the middle ages, but does so through the clearly defined corpus of charters from St Clara monastery.

The Arnamagnæan Collection, which since 2009 has held the status of Cultural World Heritage, accommodates, among others, the archive from St Clara monastery in Roskilde. The former monastery archive consists of over 470 original charters attesting to possessions and privileges of the convent, which makes it the largest of its kind in Denmark. The wide range of documents stretches from a papal letter of protection issued in 1253 to a court judgement from 1551, and the charters were composed in various languages, most notably Latin, Danish, Low German and Swedish. The archive of St Clara monastery thus forms a unique corpus that is both varied and coherent at the same time, providing excellent research opportunities.[7]

The Franciscan convent St Clara in Roskilde was founded in 1256 by Ingerd af Regenstein (ca. 1200–1258), who was part of the very influential Danish noble family Hvide. After the reformation all catholic monasteries in Denmark were closed down and the properties of St Clara came into the possession of the Danish Crown. In a property exchange in 1561, king Frederick II gave the former estate and buildings of St Clara to the University of Copenhagen. The monastery archive then became part of the university's archives, and according to a registry from 1633, all its charters were stored in a separate drawer in the Academic Council's building. As a member of the Academic Council, Árni Magnússon was tasked to organize the university's archives in 1716. On that occasion he borrowed the entire drawer and brought it home in order to go through the charters and copy them. The documents were never returned and form now part of the charter-section, the so-called Diplomata Danica, of the Arnamagnæan Collection (see Hansen 2015a).

Sub-project 2 uses this archive as its primary empirical source for studying writing and text production in a Danish medieval monastery. The analyses can

[7] The historian and archivist Thelma Jexlev has catalogued the archive (Jexlev 1973) and described it from a monastic and social-historical point of view (Jexlev 1976, 1977, 1994).

be grouped into three main fields, which are closely related and will complement each other in favour of a wider and more general understanding. The three fields are:

1. Studies in palaeography and linguistics
2. Studies in monastic history and diplomatics
3. Studies in conservation science

Studies in palaeography and linguistics

The project will examine the development of script in Denmark from the middle of the thirteenth century to the end of the sixteenth century, and the charters from St Clara will form the natural core of the analysis. However, other related material will be considered for comparison. In order to conduct studies across different languages in the best possible way, the palaeographic analysis will play a major role. Nonetheless, a number of linguistic features with regards to medieval Danish shall be investigated, among others, by means of case studies focusing on orthographic-phonological relationships as well as morphological and lexical evidence. Special attention will also be paid to the writing of scribes who penned documents both in Danish and in Latin.

The palaeographic analysis relies on methods developed within digital palaeography, i.e. a systematic, detailed annotation of selected characters' manifestations based on a (sub-graphemic) feature analysis combined with a registration of various features regarding the individual components of the letter. The sub-project will follow the formal description model introduced by DigiPal (`http://www.digipal.eu/`; see Stokes 2014), but will adjust the system in such a way that the annotated material can directly be incorporated into the digital edition (sub-project 1). The palaeographic meta-data collected in this sub-project could in the future become a cornerstone in a larger Danish/Nordic palaeographic database.

The benefits of studying the specific letterforms with regards to identifying scribal hands can be demonstrated on the material from St Clara monastery. For example, one of the scribes that was associated with St Clara and wrote several Latin charters there may be characterized by a special shape of the letters

'ø' and 'æ'. Among others since the same features also occur in the writing of the manuscript AM 187 8vo, a Danish medical book, the scribe of the charters was shown to also have written the medieval book.[8]

The linguistic analyses of the charters will be based on the thoroughly annotated transcriptions of the corpus material. All words will be lemmatized, morpho-syntactically and grapho-phonemically annotated allowing for broad excerpting of the material.[9] The approach follows to a large extent the principles of *reproducible research*, and will go hand-in-hand with the customized functions of the editorial system implemented by sub-project 1.

The palaeographical investigation will be conducted in close collaboration with the machine learning part of sub-project 3, among others for identifying central features for the dating and localization of written material as well as for scribal attribution and grouping of hands.

Studies in monastic history and diplomatics

The archive from St Clara monastery provides unique material for the study of medieval convents. Even by European standards, the charters offer an exceptionally good record of not only the monastery's possessions but also its relations to the outside world – both religious and secular. Moreover, the material allows for a better understanding of the day-to-day life of the nuns living in the convent, a subject that has hardly been studied so far. Therefore, the linguistic and palaeographic analysis of the charters will be complemented by a more general examination of the historical circumstances, under which the text bearing artefacts were produced. The charters contents, the deeds and actors described, as well as the charter type (i.e. original vs. certified copy and the narrower sub-category) account for the principle data for these investigations.

Even though the archive features charters from various issuers, it reflects to a high degree the writing and reading practices at the receiving institution. Many of the charters were written by request of the monastery's management or nuns, and naturally, only documents that were relevant to the convent and its interests

[8] Kroman (1944) has described the scribe's special way of writing 'ø', Hansen (2015b) of the ligature 'æ'.

[9] In terms of methodology, the work will be highly inspired by the approaches taken in Kjeldsen (forthcoming).

were kept in the archive. In addition, a considerable part of the charters preserved are certified copies (rather than the original), which were produced on behalf of the convent.

An example of an issuer that features repeatedly in the corpus is the Danish noble woman Anna Pedersdatter Jernskæg (died after 1432). She issued three charters that are preserved in the St Clara archive and reveal interesting facts about the monastery's acquisition of real estate. The oldest charter is dated 24 June 1408 and was composed on the occasion of Anna's daughter joining the monastery. The charter records that Anna gave her daughter two estates, which after the nun's death would fall to the monastery. Another charter from 31 October 1412 is a rental contract, which attests to Anna renting one of the monastery's properties in Roskilde for the price of one mark of silver. Finally, in the third charter, issued 28 October 1432, Anna presents the monastery with a deserted farm in Dalby in exchange for requiem masses for both her and her sons.

As a whole, the collection of charters from St Clara has the potential to provide an unprecedented insight into the writing practices and diplomatic activities of a medieval Danish monastery. It allows for research questions such as: what kinds of facts were considered necessary to be written down; who issued the relevant charters, to whom, when and in which context(s); and to which degree were certain scribes associated with the monastery and/or its social network? The analysis of the Danish charters will be supplemented with a comparison of related material from other northern European convents based on ongoing studies.

Places mentioned in the charters will be analysed in particular detail using 'geotagging'. Firstly, the location named as the place of issue is registered in order to better understand the spatial provenance and, thus, the circumstances of production and scribes involved. Secondly, potential other places mentioned in the charters are worth including, since they witness where certain people lived at a time as well as what the geographic frame of the monastery's network was. Also, various spellings of recurring proper names reveal valuable clues to the individual scribe's linguistic background and indirectly his or her identity. This kind of 'geotagging' will make the project and Danish research internationally leading in this currently up-coming field.

Studies in conservation science

In order to achieve a deeper understanding of the writing habits and the particular circumstances in the convent's scriptorium, sub-project 2 will also involve conservation science. The text bearing parchment will be analysed visually, among others using microscopes and multi-spectral imaging. The aim is to chart the used writing support with regards to potential geographical and chronological variation. The conservation workshop at the Arnamagnæan Collection, that has specialized knowledge on the examination and treatment of parchment (see Fazlic 2009), will be responsible for this part of the sub-project. The analyses will include:

- Micro-analysis of hair follicle patterns that allows for the identification of the animal species from which the parchment was made (sheep, goat or calf). This analysis is carried out using digital microscopy.
- Surface examination with regards to mechanical marks left by tools used during the production and treatment of the parchment. This survey is carried out using transparent light.
- Systematic study of the parchment's quality and condition on a general level (colour, thickness, transparency). The results from this visual survey may be incorporated into the database of the international program Improved Damage Assessment of Parchment (IDAP) (`https://www.idap-parchment.dk/`), of which the institute is a member.
- Qualitative analysis of ink by means of multi-spectral imaging.[10] Earlier tests have shown that this technology may be used for identification of ink types and has given positive results with regards to recognizing inks with the same surface and reflectance patterns across documents.

[10] Multi-spectral scanning takes a variety of images at different wavelengths of light in order to enhance features not visible in (regular) white light. The Department of Nordic Research holds multi-spectral equipment (a VideometerLab 2), which has 19 LED-light sources and takes images in the spectrum from 375 to 970 nm.

Sub-project 3: Language Technology and historical source material

Sub-project 3 is closely tied into the work carried out in the other two sub-projects. Firstly, this sub-project will provide advanced tools for the quantitative philological studies of the other sub-projects. The software solutions will further be integrated into the editorial tools created by sub-project 1. Secondly, sub-project 3 will develop methods for automated dating, localization and grouping of scribal hands based on machine learning techniques. These will play an important role for, among others, the linguistic and palaeographic investigations of sub-project 2.

To work with historical source material poses many interesting and engaging challenges to the techniques commonly used in language technology. For instance, variant spelling forms of words do not only occur across primary sources from different periods, but also to a high degree within individual documents. This is due to the less clear writing standards at the time as well as the particular circumstances of the act of writing such as potential impact from an exemplar.

Semi-automated lemmatization and morpho-syntactical analysis

To lemmatize and morpho-syntactically analyse unnormalized text from handwritten sources is a large and time-consuming task when done by hand. Therefore, digital tools can support and speed up both the annotation work as well as the qualitative philological analysis that is based on the linguistically annotated corpus material. In particular for working with unnormalized source material such as medieval Nordic texts, the introduction of semi-automated lemmatization and morpho-syntactical annotation will have a considerable impact. Existing tools for other, mostly modern, languages will be taken as a starting point for designing customized tools for the specific needs of the project. Currently available annotation tools include IceNLP (see `http://www.malfong.is/index.php?lang=en&pg=icenlp`), which comprises a number of language technological tools for Modern Icelandic and have been successfully adjusted for Old Icelandic (see Kjeldsen 2014, pp. 46–49), as well as tools developed at the Centre for Language Technology (part of the Department for Nordic Research) such as a lemmatizer that has been trained and evaluated on various languages including Danish, Polish, German, Dutch, Greek and Icelandic (Jongejan & Dalianis, 2009).

The results of the developed annotation software will be compatible with the guidelines for lemmatization and morpho-syntactically analysis of medieval Nordic languages as presented by *The Menota Handbook* (ch. 8). As a consequence, the linguistically annotated texts produced by the project can be immediately incorporated into the digital archive for Medieval Nordic Texts, *Menota*.

Automated dating, localization and grouping of hands

The majority of medieval text sources do not have any explicit indication of a date or place of writing. Yet, information about the age and area of production is crucial to further interpretation of the material. Therefore, the dating and localization of medieval sources is often solely based on detailed philological investigation of the language and script. Such analyses are both difficult and time consuming, and for medieval Danish texts the situation is even more challenging, because of the lack of reliable reference points regarding linguistic and palaeographic changes.

The project attempts to correct this situation by means of developing digital tools for automated dating, localization and grouping of scribes. These tools will be based on methods and techniques commonly used in the field of Machine Learning.

Machine Learning Techniques (MLT) are highly common in modern language technology, for instance for developing new programs or adjusting existing software to handle additional languages and domains. The Centre for Language Technology has a long-standing history and ample experience with MLT, among others with regards to semantic annotation and analysis of texts (Johannsen et al. 2014), automated tagging (Plank et al. 2014), as well as analysis of video data (Paggio and Navarretta 2013).

The basic principle of MLT is to feed the system enough textual data – potentially annotated data – that it can detect patterns which then enable qualified analyses of new, unseen material. Such techniques are for example employed in automated classification of junk mail, since the email program learns to recognize the characteristics of that kind of mail after having been fed with both wanted and unwanted mails. Another example of the usage of MLT is automated classification of user-generated contents in social media with regards to positive, negative or neutral comments (so-called *sentiment analysis*).

The classification of medieval texts with regards to time, place and scribe (or scribal milieu) conducted in sub-project 3 will primarily be approached as a supervised learning task. In other words, the system will be trained on annotated and classified data before facing unseen, not annotated data. However, since some of the corpus material is not sufficiently well annotated, the sub-project will also experiment with semi-supervised methods, i.e. mixing annotated and non-annotated material in order to increase the degree of precision in the learning process.

The benefit of using MLT in the project is that it allows us to take advantage of the dated charter material for building a reference and training corpus. Statistical models can be trained on this reference corpus to suggest dates, places and scribal groupings based on linguistic, palaeographic and other characteristics of the source material. Whereas digital philology has mostly focussed on automated analysis of visual data (image processing) (see e.g. Wahlberg et al. 2014, Mårtensson et al. 2015, and He et al. 2016), this project will primarily work from textual data that is enriched in other ways than by means of images (as mentioned previously, the semi-automated linguistic analysis and palaeographic annotation will result in highly enriched data).

Experience with medieval Icelandic sources indicates that – when dealing with material written in the vernacular – new insights into criteria for dating, localization and scribal attribution can be gained by including linguistic features in the analysis.[11] Special attention will be paid to character variants (including specifically annotated palaeographic features as analysed by sub-project 2), character frequency, graphotactics, phoneme-grapheme-relationships, manifestation of particularly frequent word forms, morphology, morpho-syntax as well as lexical features. By means of studying textual data instead of image processing data (which has already been done successfully), we expect to contribute with new important insight that will complement the results produced from visual material. As a consequence, it will be possible to reach even better results from combined linguistic and visual analysis, once we have gained a deeper understanding of the role of the various linguistic features.

[11] On the usefulness of orthographic vs. palaeographic criteria for identifying scribes see e.g. Kjeldsen (2013, pp. 407–12) and the references within. See further Mårtensson (2013).

To use linguistic criteria for the analysis moreover enables comparisons across different script types. For instance, it will be possible to compare a charter that is written in *semi-cursive* with a manuscript in *textualis*, which is highly relevant for the attribution of date, place and hands, since professional scribes of many time periods used different script types for the various kinds of media they wrote.

For the discussion of digital editorial practices (sub-project 1) it will be important to explore the possibilities of analysing patterns of interest based on various forms (or degrees) of annotation. Does palaeographic analysis, for example, contribute with relevant novel aspects for dating compared to the more conventional analysis of a text based on a facsimile transcription? How much better does a facsimile transcription support the grouping of scribal hands compared to a diplomatic transcription (and which effect does it have to include, for instance, grapheme-phoneme-relationships)? In this context, the results from an analysis of the vernacular material is expected to be particularly interesting, because the high degree of orthographic variation in these sources might lead to good results even if the text is hardly annotated.

Apart from the corpus material from the St Clara archive, the empirical studies of sub-project 3 are based on a larger amount of Icelandic, Norwegian and Danish charters that have already been digitized and partially annotated. Accordingly, the developing work of the tools and models may start even before the main corpus of the project, the material from St Clara, is entirely transcribed and annotated. Particularly in the beginning phase of the project, a sub-corpus of ca. 330 already highly annotated Icelandic charters from the period ca. 1300–1450 will be of great use. In addition, a fully lemmatized and morpho-syntactically annotated sub-corpus of about the same size consists of Norwegian charters and was created by the so-called Menotec project. Finally, sub-project 3 will have access to selected charters from the online edition Diplomatarium Norvegicum as well as the corpus of previously digitized Danish charters from the first half of the fourteenth century.

In order to develop, train and test the classification software, sub-project 3 will employ the *open source* software WEKA (Hall et al., 2009). That system provides a number of algorithms for both supervised and semi-supervised machine learning tasks.

An important aspect of the studies conducted in sub-project 3 is the meta-analysis of statistical models. Among others, this allows us to inspect the pro-

cesses involved in dating, localization and scribal hand attribution and identify the features that are most relevant for correct predictions. This is especially interesting from a philological perspective, because it may lead to new knowledge and provide valuable guidance regarding which aspects to focus on in future philological studies of (historical) primary sources.

Research dissemination and access to digital tools

The project's research results will be published in printed as well as digital media and will be presented at international conferences. The software solutions provided by the project will be made available on the internet under an *open source* license. The project's deliverables include, but are not limited to:

- A digital edition of all charters from the archive of St Clara monastery, also including an edition of its oldest known register from 1586.
- An anthology about the monastic and writing history of St Clara.
- Articles about topics falling into the fields of the three sub-projects, which will be published in respective relevant academic journals.
- Presentations of research progress and results at international conferences, such as the *Annual conference of European Society for Textual Scholarship* (textualscholarship.eu) in fall 2017, the *International Medieval Congress* (IMC) in Leeds in July 2017 and 2019, *Care and conservation of manuscripts* 17 (`http://nfi.ku.dk/cc/`) in April 2018, the annual conference organized by Digital Humanities in the Nordic Countries (dig-hum-nord.eu) in spring 2019 and the conference *Monastic Europe* 3 (place and date t.b.a.).
- An editorial tool for the study, publication and dissemination of handwritten texts in digital and printed form. The software will be made available under an *open source* license at the latest one year after the funding period of the project.
- A freely accessible monospaced font with a large selection of characters from MUFI character recommendation v. 4.0.

Additionally, the project's progress and results will be communicated continuously on the project's website (diplom.ku.dk). A web blog will be integrated into the website that facilitates the communication and dialogue with other interested parties. Furthermore, each month one charter will be featured as the Charter of the Month on the communicational and educational website of the department haandskrift.ku.dk.

Project participants, organization and network

The core group of researchers consists of four associate professors and senior researchers, a tenure-track assistant professor and three post-docs. By name, these are Associate Professor Peder Gammeltoft (PG), Associate Professor Anne Mette Hansen (AMH), Associate Professor Johnny G.G. Jakobsen (JGGJ), Senior Researcher Patrizia Paggio (PP), tenure-track Assistant Professor Alex Speed Kjeldsen (ASK), post-doc Seán Douglas Vrieland (SDV), post-doc Beeke Stegmann (BS) and one post-doc who is still to be announced. AMH (project-holder), PS and ASK will lead the project. Furthermore, photographer Suzanne Reitz (SR), conservator Natasha Fazlic (NF) and software developer Bart Jongejan (BJ), all part of the Department for Nordic research, will be associated with the project. Finally, three student helpers (SH1, SH2, SH3) will be employed. Their primary task will be to support the empirical work. The overview on the following page shows how the different tasks are divided in between the three sub-projects (digital photographing is not included).

The three post-docs will undertake the majority of the research, and they will each be associated with a different sub-project. BS has considerable experience with various aspects of working with digital texts (including digital editions) and will be the key figure behind preparing the digital edition of sub-project 1 (in collaboration with AMH and ASK). Similarly, SDV who is experienced in the mark-up and linguistic annotation of East Norse texts, will be mainly responsible for the palaeographical and linguistic analyses of sub-project 2 (in collaboration with AMH and ASK). The third post-doc, who is yet to be announced, will play a central role in the work with machine learning techniques of sub-project 3 (in collaboration with PP). Apart from the support by the more senior staff, the student helpers will assist the post-docs (mostly with regards to the empirical work).

The post-docs will start six months after the official beginning of the project in order to secure that the necessary philological tools and guidelines for annotation etc. are in place. Moreover, the delayed starting point of the post-docs gives a head start to the process of gathering empirical material (mostly carried out by the student helpers). The student helpers will be able to start their work at once, since a basic version of the editing tool – though not in its final state of development – is available and caters to their expected needs.

Good collaboration between the three sub-projects is essential for the project's success. To secure active communication, there will be monthly meetings for all participants of the project. Collaboration is further fostered by some of the project members, namely AMH and ASK, being involved in more than one sub-project. Thanks to her experience in editing medieval Danish texts, AMH will play a central role in both sub-project 1 and 2. ASK, who has expert knowledge on digital editions, diplomatics and history of language and script, will equally contribute to these two sub-projects. Additionally, ASK will assist BJ in developing the language technological tools that fall under sub-project 3, as well as being engaged in the machine learning part of that sub-project, since he is mainly responsible for developing the editorial tools of sub-project 1.

	Tasks/research areas	Primary actors
Sub-project 1: Digital textual scholarship	Transcription, non-linguistic annotation, correction	BS, AMH, SH1
	Development of philological tools	ASK
	Digital edition	BS, AMH, SH1, SH3
Sub-project 2: Studies in Writing. The archive of St Clara monastery	Linguistic and palaeographical annotation	SDV, SH2
	Analyses of text and script	SDV, ASK, SH2
	Monastic history and diplomatics	JGGJ, AMH
	Onomastics and geotagging	PG, SH3
	Conservation science	NF
Sub-project 3: Language Technology and historical source material	Lemmatization and POS-tagging (tools)	BJ, ASK, SH3
	Automated dating, localization and attribution of scribal hands (MLT)	NN, PP, SH3
Project management etc.	Project management, participation in workshops and conferences, publishing, establishing and maintenance of website	AMH, PG, SH1, SH2

The project will benefit from a wide academic network. Within Denmark, the project will be in close collaboration with the Diplomatarium Danicum, that has published a larger number of charters from the relevant time period in digital form. Internationally, an Icelandic group centred around the Orðabók Háskólans (in particular Kristín Bjarnadóttir has expressed interest in the project in regards to her own work on automated normalization of Old Icelandic) and the project Språkbanken in Gothenburg (in particular Gerlof Bouma and Yvonne Adesam who have worked with automated lemmatization and morpho-syntactical analysis of Old Swedish texts MAÞIR and will make available their tools under an *open source* license) may be named as collaborators. Finally, the project will work closely together with Christian-Emil Ore (Universitetet i Oslo), who has many years of experience with digital editions of Norwegian charters. Thanks to this connection, the project will also have access to the Norwegian sub-corpora.

Bibliography

- Ambrosio, Antonella; Barret, Sébastien & Vogeler, Georg (eds., 2014). *Digital Diplomatics. The Computer as a Tool for the Diplomatist?* Köln/Wien.
- Burnard, Lou; O'Brien O'Keefe, Katherine & Unsworth, John (eds., 2006). *Electronic Textual Editing*. New York.
- Deegan, Marylin & Sutherland, Kathryn (eds., 2009). *Text Editing, Print and the Digital World*. Farnham.
- Fazlic, Natasha (2009). *Undersøgelser af effekten ved behandling af pergament med vand og alkohol* (specialeafhandling). København.
- Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter & Witten, Ian H. (2009). 'The WEKA Data Mining Software: An Update'. *SIGKDD Explorations*, Vol. 11, Issue 1.
- Hansen, Anne Mette (2015a). 'Adkomstbreve i Skt. Clara Klosters arkiv'. I M.J. Driscoll og Svanhildur Óskarsdóttir (red.), *66 håndskrifter fra Arne Magnussons samling*, pp. 138–139. København.
- – (2015b). 'Den Arnamagnæanske Lægebog'. I *66 håndskrifter fra Arne Magnussons samling*, pp. 204–205. København.
- He, Sheng; Samara, Petros; Burgers, Jan & Schomaker, Lambert (2016). 'Historical Document Dating using Unsupervised Attribute Learning'. 12th IAPR International Workshop on Document analysis Systems (DAS 2016).
- Jexlev, Thelma (1973). *Lokalarkiver til 1559. Gejstlige arkiver I. Ærkestiftet og Roskilde stift*. København.
- – (1976). 'Roskildenonnernes sociale og økonomiske forhold i senmiddelalderen'. I *Fra dansk senmiddelalder: Nogle kildestudier*. Odense.
- – (1977). 'Nonneklostrene i Roskilde'. I *Historisk årbog fra Roskilde amt 1977*, pp. 25–40. Roskilde.

- – (1994). 'Et nonneklosters godshistorie gennem 300 år'. I *Historisk årbog fra Roskilde amt 1994*, pp. 13–32. Roskilde.
- Johannsen, Anders; Hovy, Dirk; Martinez Alonso, Hector; Plank, Barbara; Søgaard, Anders (2014). 'More or less supervised super-sense tagging of Twitter'. 3rd Joint Conference on Lexical and Computational Semantics. Dublin.
- Jongejan, Bart, & Dalianis, Hercules (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*, Vol. 1, pp. 145–153. Singapore.
- Kjeldsen, Alex Speed (2013). *Filologiske studier i kongesagahåndskriftet Morkinskinna*. Bibliotheca Arnamagnæana. Supplementum, Vol. 8. København.
- – (2014). 'Middelalderdiplomer – i en digital tid: Præsentation af et forskningsprojekt'. I A. S. Kjeldsen (red.), *Arne Magnusson 350 år: Fem foredrag i anledning af 350-året for Arne Magnussons fødsel*, pp. 39–56. København.
- – (2015). 'Filologi og historisk lingvistik: Lidt om filologiens rolle i sproghistorisk forskning'. I *Studier i Nordisk 2010–2011*, pp. 81–98.
- – (forthcoming). *Notarius publicus Jón Egilsson*. Expected to be published in Bibliotheca Nordica, fall 2017.
- Kroman, Erik (1944). 'Dansk Palæografi'. I Johs. Brøndum-Nielsen (red.), *Palæografi A. Danmark og Sverige*, pp. 36–81. København.
- The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources. Ed. Odd Einar Haugen. Version 2.0. Bergen. Medieval Nordic Text Archive, 2008. `http://www.menota.org/HB2_index.xml`.
- Mårtensson, Lasse (2013). *Skrivaren och förlagan. Norm och normbrott i Codex Upsaliensis av Snorra Edda*. Bibliotheca Nordica 6. Oslo.

- Mårtensson, Lasse; Wahlberg, Fredrik & Brun, Anders (2015). 'Digital Palaeography and the Old Swedish Script. The Quill Feature Method as a Tool for Scribal Attribution'. I *Arkiv för nordisk filologi* 130, pp. 79–100.
- Paggio, Patrizia & Navarretta, Costanza (2013). 'Head movements, facial expressions and feedback in conversations – Empirical evidence from Danish multimodal data'. *Journal on Multimodal User Interfaces – Special Issue: Multimodal Corpora*. Berlin.
- Plank, Barbara; Hovy, Dirk; McDonald, Ryan; Søgaard, Anders (2014). 'Adapting taggers to Twitter using not-so-distant supervision'. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pp. 1783–1792. Dublin.
- Stokes, Peter A. (2014). 'Describing Handwriting, Part IV: Recapitulation and Formal Model'. DigiPal: Digital Resource and Database of Manuscripts, Palaeography and Diplomatic, acessible online at `http://goo.gl/3IzVMU`.
- Wahlberg, Fredrik; Brun, Anders & Mårtensson, Lasse (2014). 'Scribal Attribution using a Novel 3-D Quill-Curvature Feature Histogram'. In *Proceedings. International Conference on Frontiers in Handwriting Recognition*.