Sources, Methods and Tools

In this chapter, we present the corpora chosen for the present study and the criteria for their compilation. In order to do so, we begin with an overview of the adopted periodization of the history of the North Germanic languages. We proceed then to a detailed presentation of each text included in the corpus. In the second part of the chapter, we describe the programme used in the annotation of the corpora as well as the statistical toolkit. We conclude with a discussion of some issues that arose in the course of annotation and the decisions that were made, which have a bearing on the final analysis.

3.1 Periodization of the North Germanic Languages

North Germanic languages or Nordic languages developed and became distinguishable from other Germanic languages most likely around 200 AD, but the Old Nordic language first took form around 500 AD (Bandle et al. 2002). Around 800 AD internal differences became more pronounced, as is attested in runic inscriptions from that period, and Old Nordic split into two distinct branches, western and eastern, as presented in Table 14.

The periodization of Old West Nordic seems quite straightforward (see Table 15). The Old Norwegian period begins conventionally around 1100, which is the end of the Viking Age. The oldest manuscripts of legal texts, however, date from the second half of the 12th century (Bandle et al. 2002:788). The year of the Great Plague, 1350, marks the delimitation between Old Norwegian and Middle Norwegian, as significant changes in the language took place after that time (Bandle et al. 2002:788).

The Old Icelandic period begins with the oldest manuscript of the legal codex *Grágás* (ca. 1170). Old Icelandic is usually treated as one long period between the years 1100 and 1550. While the term Middle Icelandic is not often used, some scholars point out the importance of distinguishing a middle period in the history of Icelandic, as several important sound changes took place in the 14th century (Bandle et al. 2002:1086).

In the tradition of Swedish and Danish language history, a three- or four-way periodization is employed (see Table 16). What is commonly regarded as Old Nordic, in the Swedish and Danish tradition is often called the *runsvenska* ‘Runic Swedish’ and *rundanska* ‘Runic Danish’ respectively. The begin-

TABLE 14 Periodization of the North Germanic languages before 1100

200–500	Proto-Nordic/Proto-Norse	
500–800	Ancient/Old Nordic	
800–1100	Old Nordic ca. 800 splits into:	
	Old West Nordic (Iceland and Norway)	Old East Nordic (Denmark and Sweden)

TABLE 15 Periodization of Old West Nordic after 1100

Old Nordic 800–1100			
Old Icelandic	1100–1350	Old Norwegian	1100–1350
Middle Icelandic	1350–1550	Middle Norwegian	1350–1530

ning of the Old Swedish period is conventionally set at the time the oldest legal manuscript was composed, i.e., *Äldre Västgötalagen* in 1225. The Old Swedish epoch is generally divided into two periods, the classical period (1225–1375) and the younger period (1375–1526). The delimitation year is usually 1375 for Swedish (cf. Wessén 1941), as again the Great Plague had an influence on language changes. It has, however, become customary to further divide Younger Old Swedish into two periods at 1450. A division into three periods better reflects the linguistic development (cf. Hirvonen 1987, Håkansson 2008, Skrzypek 2012). 1526, the year of the first complete translation of the New Testament into Swedish, conventionally marks the move towards New Swedish.

The Old Danish period begins around 1100, as the first Danish laws were written ca. 1170 (Bandle et al. 2002:788). This period is often called Middle Danish by Danish linguists and historians, but for the sake of consistency with other Nordic languages we use the denomination Old Danish. The Old Danish period is usually subdivided into two periods at 1350, i.e., Older Old Danish and Younger Old Danish. The year 1350 is conventionally used, as after that time a strong Low German influence on the vocabulary is attested (Bandle et al. 2002:788), which is also true of Norwegian. 1537, the year of the Reformation, is traditionally taken as the beginning of New Danish.

To create a uniform periodization for the Old Swedish, Old Danish and Old Icelandic texts used in the present study, we had to compromise between the delimitations of periods described above. Throughout the study we use the

TABLE 16 Periodization of Old East Nordic after 1100

Runic Swedish	800–1225	Runic Danish	800–1100
Old Swedish:		Old Danish:	
Period I	1225–1375	Period I	1100–1350
Period II	1375–1450	Period II	1350–1537
Period III	1450–1526		

TABLE 17 Periodization used in the present study

Old Swedish, Old Danish, and Old Icelandic	
Period I	1200–1350
Period II	1350–1450
Period III	1450–1550

periodization indicated in Table 17. As the oldest manuscripts used in the study were written in the early 13th century, the year 1200 is used as the beginning of Period I. The year 1350 is the beginning of Period II, as most Nordic languages use that year as the marker of language changes that took place due to rapid reductions in the morphological and syntactic systems (partly influenced by the Great Plague) and the influence of other languages (mostly Low German influence on Danish). In line with the Swedish tradition, we further distinguish Period III, beginning at 1450, in order to distinguish the most modern texts written after 1450, which demonstrate considerable differences compared with those from Period II.

At this point we should mention that Old Norwegian texts are not considered in the present study, for practical reasons. Firstly, Old West Nordic exhibits little variation between the two varieties (Icelandic and Norwegian) compared with Old East Nordic. Thus, in the literature the term Old Nordic or Old Norse is often synonymous with Old West Nordic (Bandle et al. 2002:34). Secondly, Old West Nordic is predominantly preserved in Icelandic texts, which are additionally much more diverse as regards text types than the text originating in Norway. It is thus not an easy task to find medieval texts from the 12th and 13th centuries that are unequivocally written in Old Norwegian and not in Old Icelandic, as they are often labelled as Old Nordic and often originate in Iceland. Further, it is assumed that there was a greater loss of manuscripts in Norway than in Ice-

land (Bandle et al. 2002:805). Naturally, there are available texts written in Old Norwegian, like the so-called *Landslög*, King Magnús Hákonarson's Norwegian Code from 1274, or the Norwegian didactic work from ca. 1250 *Konungs skuggsjá* ('The King's Mirror'). Old Norwegian texts from the period 1200–1350 are nonetheless few and are not easily available. Middle Norwegian, on the other hand, is well represented by diplomas, i.e., legal documents and letters. These, however, are not comparable to other texts included in the corpus, as Swedish, Danish and Icelandic diplomas were not included in the study (see section 3.2).

3.2 The Corpus

To make the present project possible, we decided to compile a new corpus of Old North Germanic texts from existing sources. Historic texts written in Nordic languages are well-documented and easily accessible thanks to such projects as *Fornsvenska textbanken* by Lars-Olof Delsing, which is the largest repository of Old Swedish texts,¹ and *Middelalder og renæssance* by Det Danske Sprog- og Litteraturselskab, which is the largest available corpus of Old Danish texts.² Another project that merits mentioning is *Medieval Nordic Text Archive* (Menota) led by Einar Haugen, which has the aim of preserving and publishing medieval texts and manuscripts written in Nordic languages.³ Some of the Icelandic sagas were also obtained via *The Icelandic Parsed Historical Corpus* (IcePaHC, Wallenberg et al. 2011).⁴ Thanks to these projects we were able to choose the texts and fragments of texts that we deemed most suitable for our project's objectives.

The corpus comprises Old Swedish, Old Danish and Old Icelandic texts written between 1200 and 1550. To make our analyses more nuanced and to make the texts more comparable across the three languages, we divided the studied epoch into three periods. Period I encompasses the texts written between 1200 and 1350, Period II the texts written between 1350 and 1450, and Period III the texts written between 1450 and 1550. The corpus consists of 280,024 words in total. The total length of the texts for each language and period is given in Table 18.

¹ *Fornsvenska textbanken*, <https://project2.sol.lu.se/fornsvenska/>.

² *Middelalder og renæssance*, <https://dsl.dk/website?id=32>.

³ *Medieval Nordic Text Archive*, <https://menota.org/forside.xhtml>.

⁴ *The Icelandic Parsed Historical Corpus*, [https://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_\(IcePaHC\)](https://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)).

TABLE 18 Length of the texts in the corpus (total word counts)

Period	Old Swedish	Old Danish	Old Icelandic
Period I (1200–1350)	47,434	13,231	56,761
Period II (1350–1450)	27,700	12,078	73,946
Period III (1450–1550)	12,027	7,813	29,034
Total	87,161	33,122	159,741

TABLE 19 Number of NPs tagged in the corpus

Period	Old Swedish	Old Danish	Old Icelandic
Period I (1200–1350)	1,194	1,097	1,536
Period II (1350–1450)	1,093	1,016	1,037
Period III (1450–1550)	635	787	648
Total	2,922	2,900	3,221

Admittedly, the length of the texts differs greatly between languages. Old Icelandic is represented by the longest texts, whereas in Old Danish much shorter texts are chosen for the corpus. The length of the texts, however, does not have a substantial bearing on the data analysed, as the entirety of texts was not analysed, but rather fragments chosen evenly for each language. Specifically, 9043 noun phrases were extracted from the corpus. In Table 19, we give the number of noun phrases tagged and analysed in each language and period.

The number of NPs analysed is quite evenly distributed, as we aimed to make the data commensurate, in particular so that each period would be comparable across the three languages. Thus, in Period I, the mean number of tagged NPs per language is 1276 (SD = 231), in Period II the mean number is 1049 (SD = 40), and in Period III the mean number is 690 (SD = 84).

The texts chosen for the corpus represent four major genres: legal prose, religious prose, profane prose and sagas. Sagas are specific to the Icelandic literary tradition. As they are prose narratives describing historical or local events or the lives and legends of saints, they are comparable to the other profane and religious prose texts available for the Continental Scandinavian languages of that time. Texts representing these four genres were chosen because they are examples of the oldest extant texts written in the respective languages (legal texts for Swedish and Danish, sagas for Icelandic), and because they include

longer descriptive and narrative passages (especially profane prose and religious prose), which offer an abundance of different types of reference of NPs, e.g., specific, non-specific, unique or generic referents. Legal texts are the only ones written without foreign models, while prose texts are usually free translations from German or Latin. There might thus be some influence from the source languages. The only remaining genres of texts available from the Nordic countries during the studied period (from the 13th to the 16th century) are poetry and diplomas. The latter is a specific set of legal documents, mostly letters, which pertain to the history of a given Nordic country or, more locally, to a given province or settlement. We chose not to use diplomas in our corpus, as the documents and letters are usually short and do not contain many narratives. The genre is thus very distinct and not comparable to others. We also decided to omit poetry, as the form of these texts is often governed by metrics and rhyme. Previous studies have also indicated that poetry is not a reliable source material for research on definiteness. Poetry, usually including translations from German, exhibits a significantly higher proportion of preposed determinatives and a lower proportion of postposed articles (Haskå 1972). The fragments of texts in the corpus were for the most part chosen randomly. The only ground rule was to make sure that the excerpts chosen included longer narrative passages and be sufficiently self-contained that any potential referents introduced in the omitted fragments would not be missed. Thus, fragments containing many dialogues were often excluded. In the case of shorter texts, especially in the Old Danish corpus, the texts were studied in their entirety.

As regards the uniformity of the corpus, we strove to make the texts as comparable as possible, especially within one period. When working with medieval texts, however, it is not always possible to choose freely the texts to work with. Thus, the texts in our corpus are not entirely comparable across the three periods studied. In the first period, the majority of the texts are examples of hermetic legal prose. These are the oldest extant texts written in Nordic languages in Latin script, and for that reason they must be included in the study, even though they may be of a very specific genre. In the later periods there are no legal texts, but exclusively profane and religious prose, as these were the types of texts that were regularly translated and adapted from foreign models at that time, with the exception of the native Icelandic sagas. Whenever possible we included the same text in different variants or translations. For instance, the *Karl Magnus* saga is represented by fragments in both Old Swedish and Old Danish; there are also Old Swedish and Old Icelandic versions of *Didrikssagan*, as well as Old Swedish and Old Danish versions of the religious text *Själens tröst*. The sections below describe the texts used. The corpus texts are classified here first by period, then by language, and finally by genre.

3.2.1 *Period I (1200–1350)*

3.2.1.1 Swedish

Legal texts: *Äldre Västgötalagen* (AVL), *Östgötalagen* (OgL), *Dalalagen* (DL)

Äldre Västgötalagen is the oldest extant text written in Swedish using Latin script. The oldest fragment is dated to 1225, and the whole text comes from a later manuscript (Holm B 59) written around 1280. It is a legal codex of the province of Västergötland. It is divided into smaller parts dealing with different areas of law, e.g., criminal offences, marital and inheritance law, regulations on religious matters, etc. The fragments chosen for the study are *Af mandrapi*, § 1 ('On manslaughter') and *Ärvdabalken*, § 1–6 ('The inheritance codex'). The language in AVL is very archaic, even compared with other legal texts from the same period (Holmbäck and Wessén 1979, Skrzypek 2012:17).

Östgötalagen is a legal codex for the province of Östergötland written around 1280. The earliest preserved copy of the manuscript (Holm B 50) is dated to ca. 1350. Similarly to AVL it includes several codices; the fragments excerpted for the study include *Drapa balken*, § 1–2 ('The manslaughter codex') and *Gipta balken*, § 1–7 ('The marriage codex'). The language of OgL is considered among the most modern found in legal texts (Stähle 1967).

Dalalagen, also known as *Äldre Västmannalagen*, was written at some time between 1200 and 1320, and the only preserved manuscript, Holm B 54, is from around 1350. Whether this legal codex was intended for the province of Dalarna or Västmanland is unclear. The fragment used for this study comes from *Kristnu balkar*, § 6–9 ('The church codex').

Religious prose: *Codex Bureanus* (Bur), *Pentateukparafrasen* (Pent)

Codex Bureanus is a collection of hagiographic legends, and is a free translation of *Legenda aurea* by Jacobus de Voragine. It was written between 1276 and 1307 (Jansson 1934:4); the manuscript used here, Holm A 34, is dated to ca. 1350. Apart from the legal texts described above, *Codex Bureanus* is the oldest text written in Old Swedish. The fragment chosen for the analysis is the first legend in the collection, *Af ioakim* ('On Joachim') on pages 3–7.

Pentateukparafrasen is a free translation of the five books of Moses. It was most likely written in Vadstena around 1330 (Klemming 1848:577). There are two extant manuscripts: the A manuscript is dated to 1430–1450, and the B manuscript to 1526. The latter manuscript, also called *Codex Thott*, is presumed to be more faithful to the original, as its text is less modernized than that of the earlier copy. The fragment chosen for the study comprises pages 156–159 and tells the story of Abraham.

3.2.1.2 Danish

Legal texts: *Eriks Lov* (ErL), *Valdemars Lov* (VL), *Skånske Kirkelov* (SKL), *Skånelagen* (SL)

Eriks Lov, also called *Eriks Sjællandske Lov*, is a legal codex for the province of Zealand. It consists of three parts: the inheritance law, the criminal law, and a supplement in which various laws are laid down, e.g., laws concerning the king. The fragments chosen for the study include chapters 1:13, 1:47, 2:4, 2:69, 3:1 and 3:23. *Valdemars Lov* is also a codex for the province of Zealand; it is an older law than *Eriks Lov*, and thus ErL is considered to be a continuation of VL. There are three extant editions of VL, the earliest of which is assumed to have been written before 1300. The language of VL is considered to be quite modern for its times, with properties of an oral text rather than formal legal prose. Chapters 12, 13, 27, 47 and 66 are included in the present study. The manuscript that contains both *Eriks Lov* and *Valdemars Lov* is coded as AM 455 12mo. It was written ca. 1300, most likely in the Sorø Kloster monastery on the Danish island of Zealand (VL, Skovgaard Boeck 2015).

Skånske Kirkelov (Scanian Church Law) is, despite its name, not an actual law but rather a legal agreement between the archdiocese of Lund and its inhabitants. The text was composed ca. 1200 and is preserved in around 30 manuscripts, the earliest of which is dated to 1300 (manuscript AM 28, 8vo). For the present study chapters 3, 5 and 11 were chosen; these include the codex on breaches of the law in church.

Skånelagen is another codex from the province of Scania (covering also the provinces of Halland, Blekinge and Bornholm). It has been preserved in many manuscripts both in Denmark and Sweden. The first Old Danish version of the law, written entirely in runes, is dated to 1203–1212, although the earliest manuscript, *Codex Runicus*, is from ca. 1350. In this study we make use of the Holm B 76 manuscript in Old Swedish and Latin script, which is dated to 1300. The paragraphs chosen are §1–5.

Religious prose: *Marialegende* (ML), *Skriftemålsbøn* (SMB), *Mariaklagen* (MK)

Marialegende and *Skriftemålsbøn* are religious texts found in one of the oldest manuscript fragments, dated to ca. 1300. The manuscript K 48 is fragmentary; it was split into two separate volumes early in its history, and a part of it is known as *Cambridgefragmentet* ('The Cambridge fragment'), as it was found and stored in Cambridge. *Marialegende* consists of two short stories relating miracles of St. Mary; the text is studied in its entirety. Following ML in the manuscript is the prayer *Skriftemålsbøn*. It is assumed to have been written later

by a different author, as it has no connection to the preceding legends or the herb book that follows it in the manuscript. The relatively long prayer is studied in its entirety.

Mariaklagen is a short religious text relating the prayers and lament of St. Mary witnessing the crucifixion of her son. The text was first written in its entirety in runes; the text in Latin script (manuscript Cod. Holm. A120) is dated to 1325 (MK, Brøndum-Nielsen & Rohmann 1929). The text is studied in its entirety.

3.2.1.3 Icelandic Legal texts: *Grágás* (Gragas)

Grágás is the only legal codex written in Old Icelandic included in our corpus. It is one of the longest Nordic legal codices from the Middle Ages. It is preserved in two major manuscripts: the so-called *Konungsbók* ('King's book', GKS 1157 fol.) dated to 1258–1262, and *Staðarhólsbók* ('Book of Stadarholm', AM 334 fol.) dated to 1262–1271. For the analysis only a short fragment was chosen, namely § 9, which describes the punishment for murder.

Sagas: *Þiðreks saga af Bern* (Did), *Geirmundar þáttur heljarsskinns* (Ge), *Þorsteins saga hvíta* (To), *Þetubrot Egils Sögu* (Egil), *Grettis saga Ásmundarsonar* (Gret)

Þiðreks saga af Bern is based on the Old German story of Didrik of Bern. The Old Norse version was written ca. 1250 in Norway. It is preserved in several manuscripts, the earliest of which, Perg. fol. nr. 4 at the Royal Library in Stockholm, is dated to the late thirteenth century. For the analysis we chose chapters 82–86.

Geirmundar þáttur heljarsskinns is a short saga from ca. 1300 describing the lives of Geirmund and his twin brother Håmund. They were the sons of a king, but were abandoned as infants due to their horrific appearance. The saga tells the story of their coming of age and successful Viking raids. It is preserved in the manuscript of the so-called *Sturlunga saga* (manuscript AM 122 a fol.), a collection of Icelandic sagas from the 13th and 14th centuries. Chapters 1–7 were selected for the corpus.

Þorsteins saga hvíta belongs to sagas of the Icelanders from the 13th century. It tells the story of the settlement of Iceland by Þorstein the white and his family. Dated to ca. 1230, it is only preserved in one manuscript, namely AM 496 qu from 1629.

Þetubrot Egils Sögu or simply *Egils saga* is a family saga by Snorri Sturluson tracing the lives of the clan of Egil Skallagrímsson in a period spanning the

years 800–1000. The saga begins with Egil's grandfather settling in Iceland with his two sons following a conflict with the royal family. The text is preserved in several manuscripts. The oldest manuscript (AM 162 A 0 fol.), often called the Theta manuscript, is dated to ca. 1250. The fragment of approximately 1300 words chosen for the study comes from the beginning of the Theta manuscript.

Grettis saga Ásmundarsonar portrays the life and family of an outlaw, Grettir Ásmundarson. The fragment chosen for the study, Chapters 14 and 15, describes Grettir's childhood and his immediate family. The saga is dated to ca. 1300–1320. The various fragments of the text are preserved in more than 50 manuscripts, most of them written in the 17th and 18th centuries. The oldest manuscripts, on which most printed editions are based, are dated to ca. 1400–1500 (AM 152 fol., AM 551a, AM 556a).

Religious sagas: *Þorláks saga helga* (Þorlak), *Árna saga biskups* (Arna), *Jarteinabók* (Jart)

Þorláks saga Helga is a saga about the life of Saint Þorlákr Þórhallsson, the patron saint of Iceland. There are at least five major manuscripts of the texts. The earliest one is in Latin (manuscript AM 386 4to) and is dated to ca. 1200. The oldest Old Norse manuscript is dated at ca. 1350, and the remaining three manuscripts were written in the 14th to the 17th centuries. For the study, pages 111–117 are chosen.

Árna saga biskups is considered the primary source for the political and ecclesiastical history of Iceland during the episcopacy of Árni Þorláksson of Skálholt, i.e., in the years 1269–1298. The saga is dated to ca. 1325. The text is preserved in several manuscripts, the oldest of which, AM 220 VI fol., is dated to 1340–1360. For the study, the first 1100 words of the saga are chosen.

Jarteinabók is a religious saga which, together with *Þorláks* and *Árna*, makes up the so-called Bishop sagas (*Biskupa sögur*). It is a miracle collection composed in honour of Saint Þorlák, dated to ca. 1200–1220. *Jartein* is preserved in several manuscripts, the oldest being the so-called *Heilagramanna sögur* ('Legends of Saints', AM 645 4to) dated to 1220–1249. For the present study, chapters 1–6 of *Jart* are selected.

The Bishop sagas are in general the closest to the religious prose genre in Old Swedish and Old Danish, as they also relate miracles and legends of saints. They are, however, very much grounded in the Icelandic saga literary tradition.

3.2.2 *Period II (1350–1450)*

3.2.2.1 Swedish

Religious prose: *Järteckensboken* (Järt), *Helga manna leverne* (HML), *Själens tröst* (ST)

Järteckensboken is a religious text that comprises relatively short texts relating miracles connected with the host (the altar bread). It is dated to 1385, as is the manuscript from which the text is taken, the so-called *Codex Oxenstiernianus* (Holm A 110). The fragments used for the study are pages 3–12.

Helga manna leverne is a collection of short hagiographic texts; it is a translation of the Latin *Vitae Patrum*. It comes from the same manuscript as Järt; however, it is believed to have been written later than 1385 as the language and spelling are somewhat more modern (Mattsson 1957:232). For this study we chose the fragment on pages 185–189.

Själens tröst is a translation of the Low German *Seelenstrost*. It is a collection of legends written for the purpose of explaining the Ten Commandments. There is only one manuscript with the text preserved in Old Swedish, Holm A 108, written in Vadstena and dated to 1430–1450. The fragment on pages 326–330 was chosen for the study.

Profane prose: *Sju vise mästare* (SVM), *Karl Mangus* (KM)

Sju vise mästare is a collection of short stories with a Chinese box structure. The protagonists tell each other stories, each with a moral lesson, to convince the king either to save or to hang his son for his crimes. The text thus has an overarching main story and a number of shorter novellas woven into it. SVM is a translation from Latin or German, depending on which of the three manuscripts one considers. The most comprehensive and complete is manuscript A (Holm D 4), which is a translation from Latin, dated to 1430–1450. Pages 117–121 were chosen for the analysis.

The *Karl Magnus* saga is a Swedish translation of two tales of Charlemagne based on an earlier Norwegian version *Karlamagnús Saga*, which was based on a number of French poems (Skrzypek 2012:19). It was most likely translated at the end of the 14th and the beginning of the 15th century, and it is preserved in the same manuscript as SVM, dated to 1430–1450. The pages studied here are 265–270.

3.2.2.2 Danish

Religious prose: *Aff Sancte Kerstine hennis pyne* (Kerst), *Aff Sancta Marina* (Mar), *Huoel Sancte Pouel vort pint* (Pouel), *Sjalens trost* (ST)

Aff Sancte Kerstine hennis pyne is the Danish version of the Legend of St. Christina, a martyr who refused to worship pagan gods and suffered tortures at the hand of her father. The story focuses on various cruel tortures inflicted on her. There are two manuscripts in which fragments of the legend are preserved. The earlier manuscript dated to ca. 1300 contained only a short fragment of the legend (manuscript Cod. Upps. C871). The later manuscript, Cod. Holm. K4, often called *De hellige Kvinder* 'The holy women', is dated to ca. 1450. In the study we use the latter manuscript as it contains a longer, more comprehensive text. The pages used are 38–41.

Aff Sancta Marina is another legend about the life of a woman saint, Marina the Monk, who lived as a monk in an all-male monastery. *Huol Sancte Pouel vort pint* is a gruesome tale of St. Paul descending into hell and witnessing the grotesque punishments and suffering of lost souls. It is based on the apocryphal story *Visio Pauli*, which was translated into many languages of Europe and beyond during the classical period and the Middle Ages (Bullitta 2017:1). Both texts are preserved in the same manuscript as Kerst, i.e., Cod. Holm K4, dated to ca. 1450. The text of Mar is studied in its entirety, while for Pouel pages 24–26 are chosen.

Sjalens trost is the Old Danish version of the Old Swedish ST, based on the same German source text. The Danish version is split into two parts; the first part is preserved in one manuscript (Cod. Ups. C 529) and the second in another manuscript (Cod. Holm A 109). Both manuscripts are dated to 1425. The latter manuscript is used in the present study, and the fragment on pages 68–72 was chosen.

Profane prose: *Gesta danorum* (GD)

Gesta danorum is one of the first attempts to record the history of Denmark in the mother tongue. It is a collection of brief descriptions of individual events in the history of the country, as well as a list and description of its kings. There is a particular focus on the information about the church, particularly the archdiocese in Lund. That fact, and the presence of linguistic traits typical of the Scanian dialect, lead researchers to conclude that the text might have originated in Lund (GD, Akhøj Nielsen 2015). The text is preserved in two manuscripts, both from ca. 1450. Manuscript C67 is used here. The text itself is dated to around 1380–1400. Chapters 1 (Humble), 4 (Guttorm), 6 (Halfdan) and 7 (Rolf Krake) were chosen for the study.

3.2.2.3 Icelandic

Religious sagas: *Mörtu saga og Maríu Magdalenu* (Marta)

Mörtu saga og Maríu Magdalenu is the Old Norse version of the apocryphal legend of the sisters of Lazarus, Martha and Mary of Bethany. The character of Mary of Bethany was later conflated with Mary Magdalen. The legends tell stories of the sisters' lives and miracles they performed or witnessed. The text was translated from Latin ca. 1350. It is preserved in five manuscripts. The oldest manuscript (NoRA 79 fragm.) is dated to 1350 and is considered to be the original compilation, on which other manuscripts are based (Van Deusen 2012:201). In the NoRA manuscript only a short fragment of the legend survived; for this reason, the edition of the saga used in the present study is based on manuscript AM 233a fol., dated to 1350–1375. The fragment chosen for the study is the first 1133 words of the saga text.

Sagas: *Víglundar saga* (Vig), *Bandamanna saga* (Band), *Gunnars saga Keldugnúpsfífls* (Gun), *Finnboga saga ramma* (Finn)

Víglundar saga tells the love story of Víglund and Ketilríð and their desire to marry despite the objections of her family. The saga is dated to ca. 1400 and is preserved in many manuscripts. One of the oldest manuscripts is AM 551 4to, dated to 1500. The fragment chosen for the corpus consists of the first 1462 words of the saga.

Bandamanna saga tells the story of Grettir's nephew Ospakr and his friend and protector Oddr. Oddr wants to help his less fortunate friend and invites Ospakr to live with him and lead his household, from which many family disputes arise. The saga was composed around 1350, although some scholars argue that it was created at the end of the 13th century (cf. Magerøy 1981). It is preserved in two major manuscripts: the so-called *Möðruvallabók* (AM 132 fol.) dated to ca. 1350, and the so-called *Konungsbók* (GKS 2845 4to) dated to ca. 1425.⁵ In the present study chapters 1 and 2 are considered.

Gunnars saga Keldugnúpsfífls and *Finnboga saga ramma* are considered to be post-classical sagas, as they were composed later than the sagas previously described. Both sagas are dated to the late 14th century, Finn being a somewhat older saga than Gunnar. Finn is preserved, among others, in the manuscript *Möðruvallabók* (AM 132 fol.) and Gunnar survives only in quite late manuscripts

5 Note that this is a different manuscript than the earlier *Konungsbók* (GKS 1157 fol.) which contains a copy of *Grágás*.

from the 17th century (e.g., AM 572 c 4to). The fragments excerpted for the study are the first 1865 words of Gunnar and the first 2930 words of Finn.

3.2.3 *Period III (1450–1550)*

3.2.3.1 Swedish

Religious prose: *Linköpinglegendariet*, *Legenden om Sankta Amalberga* (Linc)

Linköpinglegendariet is a collection of hagiographic legends composed around 1520. In the first part of the work there are fifteen legends translated by Nils Ragvaldi, including the legend of St. Amalberga. The legend is preserved in the manuscript Linc B 70a, dated to the first half of the 16th century. For the present study pages 278–281 were chosen.

Profane prose: *Didrikssagan* (Did), *Historia Trojana* (Troja)

Didrikssagan is the Swedish translation of the Old German story of Didrik of Bern. The Swedish translation is dated to 1450 (Henning 1970:28) and is based on an earlier Old Norse version from ca. 1250 (Skrzypek 2012:21). The manuscript containing the text, Skokloster 115, 116, is dated to ca. 1500. For the analysis chapters 79–87 were chosen.

Historia Trojana is the Swedish translation from Latin of the story of the Trojan War. There is only one extant manuscript of the text, translated in 1529 and published in 1892 (Holm D 3a). Troja was chosen for the study as the last Old Swedish text, in line with many previous studies (Stähle 1967:121, Hirvonen 1987:63, Delsing 1999, Håkansson 2008:21). Pages 1–6 are included in the analysis.

3.2.3.2 Danish

Religious prose: *Af Jeronimi levned* (Jer), *Af Katherine legende* (Kat), *Jesu Barndoms Bog* (Jesu)

Af Jeronimi levned and *Af Katherine legende* are legends about Saint Jerome and Saint Catherine of Siena respectively. Both texts are preserved in the manuscript called *Mariager Legende-Håndskrift*, dated to 1488. It was composed at the Saint Bridget monastery in Mariager in Denmark. For the present study chapters 2–4 (pages 341–345) of Jer and chapters 20–22 (pages 346–347) of Kat are chosen.

Jesu Barndoms Bog is an apocryphal story about the childhood of Jesus Christ and the miracles he performed. Stories about St. Mary, the birth of Christ, etc. also make up an important part of the book. The manuscript is based on a ver-

sion by Gotfred af Ghemen from 1508. The stories are based on an apocryphal gospel by an unknown author. The Danish version is largely based on a German poem, *Marienleben* by a monk named Philip (Jesu, Skovgaard Boeck 2015). The chapters chosen for the present study are 4–7 (4: Abrathar, 5: Maria og Josef, 6: Bebudelsen, 7: Jesu fødsel).

Profane prose: *Karl Magnus Krønike* (KM)

Karl Magnus Krønike is the Danish version of the tales of Charlemagne and, like the earlier Swedish version, was based on the Old Norwegian *Karlamagnus Saga*. There are three Danish manuscripts; the oldest version, which we use in the present study, originated in 1480. The fragments chosen are pages 138–149.

3.2.3.3 Icelandic

Sagas: *Gísla saga Súrssonar* (Gis), *Vilhjálm's saga Sjóðs* (Vil), *Jarlmanns saga og Hermanns* (Jarl)

Gísla saga Súrssonar is one of the sagas of the Icelanders. It tells the story of Gisli who is forced to kill his brother-in-law as a revenge for another brother-in-law's death. Gisli is then outlawed resulting in a thirteen years' long exile, until he is finally hunted down and killed. Although dated to ca. 1200, it is only available in manuscripts from the beginning of the 1400s.

Vilhjálm's saga Sjóðs is a long chivalric saga. While probably inspired by translations of continental romances, it is an original Icelandic knightly saga. Vilhjálms' genre is quite distinct from that of the earlier classical sagas of Icelanders, as it does not focus so much on the history of Iceland, but rather on more exotic and romantic adventures of knights in distant countries. It is dated to the first half of the 15th century, and is preserved in many manuscripts from the 17th century. The earliest known manuscript is AM 548 4to, dated to 1543. The fragment chosen for the study encompasses the first 1039 words of the text.

Jarlmanns saga og Hermanns is another romance saga about the lives of two foster brothers, Hermann and Jarlmann, and their difficulties in finding wives. It is dated to the second half of the 15th century. The text of Jarlmann is preserved in numerous manuscripts, the earliest being the so-called *Eggertsbók* ('Book of Egert', AM 556b 4to) composed in the late 15th century. The fragment selected for the present study spans the first 1570 words of the text.

3.3 The DiaDef Tool and Its Benefits for the Study

3.3.1 *Tool Overview*

The process of collecting linguistic data was facilitated by computer software, the manual tagging system DiaDef (Diachrony of Definiteness), based on the open-source Tagger framework, available on the GitHub platform (<https://github.com/rjawor/tagging>). More information about systems of the Tagger family can be obtained at <http://rjawor.vm.wmi.amu.edu.pl/wiki>.

Similar systems are commonly used in contemporary linguistic research; see for instance Haig and Schnell (2015). Features of DiaDef include:

- import of texts in textual format (.txt or Microsoft Word) into documents;
- storing of information about the language of the document and the epoch in which it was written;
- automatic splitting of the text into sentences and individual words;
- manual assignment of tags to individual words.

The key functionality of the DiaDef tool is multi-level annotation of words and sentences in larger documents. The tool provides several features that improve the efficiency of use. For most annotation levels the system displays a context-sensitive list of prompts of available annotation tags. For a word under annotation the system displays a “prompt cloud”, which consists of a set of tag suggestions.

DiaDef minimizes the cost of usage errors or system failure. Each annotation decision is saved automatically in a periodically backed-up database. There is no save button. This solution ensures protection against the loss of valuable annotations. The wide variety of configuration settings ensures the flexibility of the tagger, allowing it to be used in various scenarios. On request, DiaDef generates statistics concerning occurrences of specific classes of words and word collocations in a specified document or collection of documents.

3.3.2 *Multi-Level Tagging*

To begin the tagging process, the user uploads a text document into the system. Upon upload, the document is automatically split into sentences (see Figure 1).

The user can easily override the automatic sentence split (using “scissors” or “glue”; Fig. 1). The document is then annotated in sentence-by-sentence mode.

Each sentence is automatically split into words. A split sentence is presented in Figure 2. The user may also override the word split, for example to divide a word into a stem and a suffix. DiaDef may be configured to serve a variety of annotation tasks. The “configuration” option allows one to manage the languages of tagged documents as well as to configure annotation levels. Annotation levels may be freely ordered, added, deleted or edited. Editing of an

1. Thet var en timæ en værliz man oc han hæffde en dotter
2. Han fech then hugh At han vildhe sigh i closter giffue

FIGURE 1 Sentence splitting in DiaDef

1.	Thet	var	en	timæ		en	værliz	man
Lexeme	There	was	one	time		a	worldly	man
POS				N				N
ART				EN				EN ADJ-ST
REF				NON-SPEC				NEW SPEC
Anchor								
Grammar				SG M NO-END				SG M NO-END
Syntax				ADV				SBJ
Semantics				INANIM COUNT CONCR				ANIM COUNT CONCR P-OR

FIGURE 2 An example of an annotated sentence in DiaDef

annotation level consists in defining admissible values of respective tags. The DiaDef tool is thus very flexible, as it enables non-restricted editing and adding of tags on different levels of annotation, so that the information entered is well-suited for the objectives of the project at hand.

Words are annotated at eight levels: Lexeme (where the closest English lexical equivalent is given), POS (Parts of Speech), ART (article), REF (reference), Anchor (related “anchor” word), Grammar, Syntax, and Semantics. The annotation levels are presented in Table 20 together with the tags within these levels.

As indicated in Figure 2, each utterance is presented in the form of a table in DiaDef, where each lexeme appears as a separate column, and the abovementioned levels of tags appear as rows in the table. To facilitate and simplify the statistics generated in DiaDef, almost all of the information is annotated on the noun of the NP. We also occasionally tag possessive pronouns, but generally it is the noun that carries the full information. All of the tag annotation levels (all except for *Lexeme*, *Anchor* and *Add info*, which do not operate on tags) have a multiple-choice structure, i.e., several tags may be marked within one level, as long as the tags are not mutually exclusive. Clicking on the row of a given

TABLE 20 The annotation tags defined for the present project

No.	Annotation levels	Tags
1.	Lexeme (English translation)	–
2.	POS (Part of speech)	N (noun), PRON (pronoun), ART (article), NUM (numeral), P (preposition), V (verb)
3.	ART (articles, lack of articles, other elements)	BN (bare noun), BN-PL (plural bare noun), -IN (postposed definite article), EN (indefinite article), DEN (preposed definite article), DEM (demonstrative), ADJ-ST (strong adjective), ADJ-WK (weak adjective), REL (relative clause), NUM (numeral), POSS (possessive), Poss-refl (reflexive possessive pronoun), Poss-gen (genitive), Poss-pro (possessive pronoun)
4.	REF (type of discourse reference)	DIR-A (direct anaphora), INDIR-A (indirect anaphora), U (unique reference), G (generic reference), NEW (new referent), SPEC (specific reference), NON-SPEC (non-specific reference), NON-REF (non-referential)
5.	Anchor (of the discourse reference)	–
6.	Grammar	SG (singular), PL (plural), NOM (nominative case), ACC (accusative case), DAT (dative case), GEN (genitive case),

TABLE 20 The annotation tags defined for the present project (*cont.*)

No.	Annotation levels	Tags
7.	Syntax	OBL (oblique case),
		NO-END (no case ending),
		M (masculine gender),
		F (feminine gender),
		N (neuter gender)
		SBJ (subject),
		OBJ (object),
		ADV (adverbial),
		MOD (modifier),
		NEG (includes negation),
8.	Semantics	P-OBJ (prepositional object),
		PRED (predicative)
		ANIM (animate),
		INANIM (inanimate),
		COUNT (countable),
		MASS (mass NP),
		ABSTR (abstract),
		CONCR (concrete),
9.	Add info (additional information)	P-OR (possessor),
		P-UM (possessum)
		–

level highlights the tags that one can choose from. Simple keyboard shortcuts displayed in brackets are assigned to the tags, which facilitates the tagging process.

The first annotation level, called *Lexeme*, is a free write-in row where we enter the English translation of each lexeme. The second level defines the part of speech (POS) of a given lexeme. Since in this project we deal exclusively with nominal phrases, the vast majority of lexemes are tagged as N (noun) or PRON (pronoun). The third level is the article level (ART), where we annotate information about the types of articles (or the lack thereof) that appear in an NP. On this level we include not exclusively articles, but also other elements that appear with articles or in their place (e.g., possessive pronouns, adjectives, numerals) or elements that may have some influence on the articles used in an NP (e.g., a following relative clause). On the fourth level, REF, we tag

FIGURE 3
Automatically generated tagging suggestions in DiaDef

the discourse reference type, i.e., whether the referent is new (specific or non-specific), unique, generic, anaphoric (direct or indirect), or non-referential. The fifth level is called *Anchor*, and enables us to mark in the previous discourse the referent in which the reference of a given noun is anchored. The sixth level covers grammatical information, i.e., the number, case and gender of a given noun. On the seventh level information about syntactic roles is entered. The eighth level pertains to semantic information. Three pairs of mutually exclusive features are defined here, i.e., animate–inanimate, countable–mass, and abstract–concrete, as well as the role in a possessive construction, i.e., possessor–possessum. Next, there is an additional annotation level where we tag the types of indirect anaphora in a more detailed way. The last level is again a free write-in option where in the course of the tagging process we enter any additional information, such as notes on problematic cases.

The corpus texts were tagged in DiaDef manually by the authors with the help of a master’s degree student who assisted in the project for six months. The texts were tagged primarily from February 2017 to February 2018, although after that period the tagged texts were continually checked and corrected, and minor changes were also frequently made when necessary.

3.3.3 *Automatically Generated Suggestions*

To improve tagging efficiency, the system suggests hints whenever possible, i.e., when a word has already been tagged or when the tagging could be deduced automatically. Tag suggestions appear in a “cloud” above the word (Figure 3).

Figure 3 presents tag suggestions for the word *thet* ‘that’. The suggestions come from previous annotations of the word. The user can accept the set of suggestions by clicking the ‘check’ symbol in the leftmost column.

TABLE 21 Overall numbers of annotated words in the DiaDef system

Language	Word count
Danish	15,788
Icelandic	20,308
Norwegian	989 ^a
Swedish	14,533
Total	51,618

a As it was difficult to obtain a comparable corpus of Norwegian texts, these annotated texts were not analyzed further.

3.3.4 Collected Data

The DiaDef system for Scandinavian languages has been in active use since early 2017. The current (as of 30 September 2019) counts of annotated words in four languages from the database are presented in Table 21.

The collected data have been used to compute various statistics.

3.3.5 Statistics Generator

The DiaDef tool has a Statistics Module, which is used to compute on demand statistics based on the data collected in the system. The module has two key functionalities: searching for words or phrases which meet specified criteria, and computing various statistics regarding those words. There are several functions that can be invoked from the Statistics Module, i.e., specific word form search, collocation search and proportional statistics.

Using the specific word form search option, the user can search for a specific word form with a single tag or various combinations of tags. The user can specify criteria for word search using the search window illustrated in Figure 4.

In this case the system will search for words tagged as nouns (N) in the singular (SG). The system enables preliminary sorting of the data by language and period (by using the filter panel at the top of the page), and it generates lists of cases displayed in larger context with direct links to the edition of a given case, as shown in Figure 5. By clicking on the “Edit” icon, the user is redirected to sentence edit mode, where the word in question is automatically highlighted. This feature greatly facilitates the control and correction of individual tags and examples within these tags. For instance, by searching for all occurrences of the dative case we can browse through the lists of words and correct possible mistakes.

Find annotated words

Add annotation criterion:

Selected criteria: (clicking on a criterion deletes it)

N (POS)

SG (Grammar)

Position in sentence

Any

Initial

Non-initial

Find words

FIGURE 4
The search window for simple statistics in DiaDef

No.	Document	Language	Epoque	Word	Context
1	DA_ErL.doc	da	classical	bonden	Takær bonden sinæ dottær man i flet til sich
2	DA_ErL.doc	da	classical	dottær	Takær bonden sinæ dottær man i flet til sich
3	DA_ErL.doc	da	classical	man	Takær bonden sinæ dottær man i flet til sich

FIGURE 5 An example of simple statistics generated in DiaDef—the dative case in Old Swedish

Using collocation search the user can also search for multiple words which appear together in a single sentence. These are referred to in the system as collocations. An example of a collocation search is given in Figure 6.

In this case the user is searching for sentences which contain both a noun and an article. It is possible to search for collocations of a maximum of five words. The number of criteria for each word is limited only by the number of tags available in the system. Results of collocations search are presented in the same way as in specific word forms search.

The last option is the proportional statistics search. This kind of search is based on the word form search, although it takes two sets of criteria: main search criteria and additional specific criteria. The system first takes the main search criteria and counts the number of words that meet those criteria. It then appends the additional specific criteria to the main criteria list and counts how many words meet all these criteria. Lastly, it computes the ratio of the number of words meeting all the criteria to the number of words meeting only the main criteria. For instance, we can search for the number of animate (ANIM) and singular (SG) referents within the main search of NPs with a postposed definite article (-IN). The result will be presented as in Figure 7. Note that the results can be refined with the use of filtering to a specified language, epoch and documents.

Find collocations

First word:

Selected criteria: (clicking on a criterion deletes it)

N (POS)

Second word:

Selected criteria: (clicking on a criterion deletes it)

EN (ART)

Find collocations

FIGURE 6
An example of collocation statistics search in DiaDef

Proportional statistics

Narrow by:

Language	Epoque
☐ Any	☒ Any
☒ Swedish	☐ middle
☒ Danish	☐ old
☐ Norwegian	☐ classical
☒ Icelandic	
☐ Modern Swedish	

Number of words tagged as: *Noun*, . Position in sentence: *Any*
1176

Number of words more specifically tagged as: *Noun*, , *singular*, . Position in sentence: *Any*
494 (42.01%)

FIGURE 7
An example of proportional statistics generated in DiaDef—animate singular referents within referents with postposed definite article

3.3.6 Regression Analysis

With the use of data collected in DiaDef, a set of statistics regarding the usage of articles was computed. The methodology of this analysis was the following. The problem of choosing the correct article was treated as multi-class classification in the set of all words tagged as nouns in the DiaDef system. The three classes (i.e., possible article choices) were:

- 1. -IN (the incipient definite article)
- 2. EN (the incipient indefinite article)
- 3. BN (bare nouns)

The independent variables were tags assigned to the nouns at the following levels: case, number, gender, function, animacy, countability, concrete and anaphora. The aim of the analysis was to measure the impact of these tags on the choice of article.

The analysis was performed using Vowpal Wabbit software, which is capable of performing efficient multi-regression for the needs of classification. It can also be used to output information about the relative impact of features on the choice of class. In this scenario, the features are the tags assigned to words, and the class, as mentioned above, is the choice of the article.

3.4 Annotation Issues

Anyone who has worked with authentic diachronic data will appreciate that even the best tool will not always help resolve all problems. Even though great care was taken to include only self-contained text fragments in the corpus, at times we had to struggle with the classification of a given form and its use. In this section we briefly present the most important issues that arose and how they were dealt with. The choices made were to some extent arbitrary, however, we strove to remain consistent in making similar choices in similar circumstances.

We encountered relatively few problems with tags concerning the form of the noun, such as information on number or gender. There are a number of excellent dictionaries available which may be consulted (e.g., the Swedish dictionaries at spraakbanken.gu.se), often the form of other constituents of the NP revealed the correct gender or number. However, tagging case was more of a challenge, since in the period studied the case system in Danish and Swedish is dismantled and by Period III can no longer be tagged. We have therefore introduced the tag NO-END to mark such nouns that had no ending, but had the system been viable, would have them.

The fact that the case system underwent such significant changes between 1200 and 1550 in both Danish and Swedish has consequences for the annotation of the corpus. While it is possible to annotate each NP for case in Icelandic in all three periods, the annotation presents more of a challenge in Danish as early as Period I, and in Swedish in both Period II and III. Apart from the tags NOM, ACC, DAT and GEN for distinct case forms, we also used OBL to annotate forms that were not unequivocally accusative, dative or genitive (most such forms were found among feminine nouns, but also quite a few among masculine nouns, such as *bondi* 'yeoman'). A certain degree of syncretism is a natural part of any inflectional system. However, we encountered problems when deal-

ing with certain nouns in Danish, and later in Swedish, which always appeared without endings, irrespective of their function in the clause. The lack of ending should not be confused with zero ending, which was a part of the paradigm in all languages studied. The focus of our study is the rise of definiteness and not the decline of case. For that reason, some nouns which had no case ending were still annotated with a given case if they were modified by adjectives whose form unequivocally displayed a case ending, even if, strictly speaking, the noun itself did not show case inflection at all, as in (118).

- (118) *oc ropapo mz högh-o röst sighiande*
 and called with high-OBL voice saying
 'and called with a high voice saying' (SV_HML, Period II)

It is important to bear in mind that the tag NO-END was limited only to cases in which there was no declined form of the noun available; it was not used where a zero-ending contrasted with other forms, as in (119).

- (119) *Ok ginstan wiste diäfwl-en at thät war sand-ir gudh*
 and immediately knew devil-DEF that it was true-NOM god
som präst-in fördhe
 which priest-DEF carried
 'And the devil knew immediately that it was the true God which the priest carried' (SV_Järt, Period II)

In (119) above the noun is in the expected form, with a zero ending signifying nominative case (which is also visible on the adjective), and contrasting with *gudh-s* in the genitive and *gudh-i* in the dative (the accusative being syncretic with the nominative).

However, in a number of examples in Period II in both Danish and Swedish the nouns appeared without the expected endings, e.g., the genitive -s (the expected forms after the preposition *til* 'to' are *bords* 'table's', *lands* 'land's').

- (120) *Nu tha the sama quinna-n thiänte til bordh sin-om*
 now that DEF same woman-DEF served to table her.REFL-DAT
herra vm middaghin
 master.OBL about dinner
 'Now that the same woman served her master at the table at dinner-time.' (SV_Järt, Period II)

- (121) *The toko hona in j thera skip oc fōrdho hona mz*
 they took her in in their ship and led her with
sik til land
 themselves to land
 'They lifted her (out of the water) onboard their ship and carried her
 with them to the land.' (SV_ST, Period II)

In such cases the tag NO-END was used, to demonstrate that the syntactic position would call for the genitive ending, had the case system been intact.

While the annotation of formal features did cause some problems, it was the annotation of the function of a given NP that proved to be challenging, since at times it required making arbitrary choices. The functions could be tagged as: direct anaphor DIR-A, indirect anaphor INDIR-A, larger situation use (unique reference) U. A rule of thumb was that if an NP included a noun which had a co-referring NP with an identical lexeme in the preceding text, it was tagged as direct anaphor; if there was no connection between the NP and any other element (nominal or verbal) in the text it was tagged as larger situation use; all uses that could be considered anchored in the text were tagged as indirect anaphora. Any element that could serve as an anchor was linked via the Anchor tag with the indirect anaphor.

As it turned out, this rule of thumb worked well. Nevertheless, in several cases the choices were not easy to make. In (122) an example from Danish is presented. The noun marked in bold, *haff-it* 'sea-DEF', is definite and could be definite for a number of reasons. It could be treated as anaphora, since 'sea' is mentioned earlier in the text, albeit by a different term, *mær-it* 'sea-DEF'.⁶ The question remains, which anaphora is it? Is it a direct one, since obviously the same referent is presented by both NPs? Or indirect, anchored in the situation of traveling on a sailing ship? Or perhaps 'locally' unique, connected with the area in which the scene takes place? It could even be considered unique in the sense of Hawkins' larger situation use, if we assume that for the people of the 15th century the world consisted of one large piece of land, one large reservoir of water and the skies. In this interpretation 'the sea' would receive definite marking just as 'the heaven' does.

6 There were two terms which described large reservoir of water in Old Swedish, *hav* and the obsolete *mær*, which could be used synonymously.

- (122) *Tha the vpa mærit como saghdho the som hæenne fôrsho*
 when they upon sea-DEF came said they who her led
at hon skulde theras vilgha giôra ælla the vildo hæenne i
 that she should their will do or they will her in
haff-it casta
 sea-DEF throw
 ‘As they came out to sea, those who had kidnapped her said that she must abide by their will or they will throw her into the sea.’ (DA_ST, Period II)

In fact, the referent is presented as definite at the beginning of the story; the text opens with the passage in (123) below.

- (123) *budho sinum thiænarum at the skuldo hæenne mæth skip ofvir*
 asked their servants that they should her with ship over
mærit fôra
 sea-DEF take
 ‘(They) asked their servants that they should take her by ship over the sea.’ (DA_ST, Period II)

Mærit ‘sea-DEF’ is used in definite form from the first mention; the referent may be considered anchored in the phrase *mæth skip* ‘by ship’ and in this sense constitute an indirect anaphoric use.

In this particular case we opted for indirect anaphora. Even though the referent has been mentioned previously, the choice of a different lexeme suggests different qualities of the referent. It is also anchored in the situation of sea voyage presented in the text.

The choices that had to be made naturally imply that the quantitative results could be slightly different, had different choices been made. This is inevitable in any diachronic study. Therefore, we have also examined the data qualitatively and inspected each example separately. In this section it is the most troublesome examples that are of greatest interest and provide greatest insight (see Chapter 5).

3.5 Summary

In this chapter, we have presented our corpus, or rather a collection of three sub-corpora, comprising Danish, Icelandic and Swedish texts from 1200–1550. 1200 corresponds roughly to the beginnings of literacy based on the Roman

script rather than the runic alphabet futhark. The date marks the beginning of the so-called 'old' period in Danish, Swedish and Icelandic language history. 1550 is a discretionary date concluding the 'old' period, which coincides with a number of important social events, such as the divorce from the Catholic church, the introduction of print and the establishment of the national states. It is customary to subdivide the 1200–1550 period into older (or classical) and younger with, again, a discretionary date 1350, which coincides with the Black Death plague, which had a catastrophic effect on the population of the Scandinavian countries and profound results for their languages. We have additionally subdivided the younger period into two, 1350–1450 and 1450–1550, in order to study the ongoing linguistic change more precisely. In the book, we refer to these periods as Period I (1200–1350), Period II (1350–1450) and Period III (1450–1550) respectively.

From the extant corpora of texts, we have chosen whole texts or longer passages representing all prosaic genres of the time, i.e., legal, religious and profane prose. The total amount of tokens is about 280,000, from which we have annotated over 9000 NPs, ca. 3000 for each language in the sample.

For annotation, a programme called DiaDef was used, which is a manual tagging system allowing multi-level tagging and adding both new tags and new tagging levels. The tool also includes a statistics generator which can be used to obtain a number of statistical data, in particular proportional statistics, which were utilized to a great extent in the present study. It further allows export of data to other statistics generators, so that other types of analysis can be conducted. For the present study, we have used this function to perform regression analysis.

We have also discussed some issues arising when annotating historical texts, concerning both the form of a given noun, in particular the annotation of case, and its function, in particular the indirect anaphora as opposed to the direct anaphora, on the one hand, and the larger situation use, on the other. The case system is weakened or dismantled in Danish and Swedish within the period studied, which has consequences for the categorization of a given form. Also, the limited context does not always allow unequivocal categorization of the NP function and it is the decision of the annotator which function is chosen. The decisions made have consequences for the statistical analysis presented in the following chapter and the most troublesome examples are discussed at length in Chapter 5.