Making Sense of Normalization

A Corpus-driven Approach to Place-name Norms and Norm Negotiations in East Norse

Alexandra Petrulevich | ORCID: 0000-0002-8478-2040
Department of Scandinavian Languages, Uppsala University,
Uppsala, Schweden
alexandra.petrulevich@nordiska.uu.se

Abstract

This article examines current practices of normalization of names in Norse philology and computational linguistics that to a large extent build on deductive reasoning and external authoritative sources such as grammars, dictionaries and gazetteers. Instead, a survey of manuscript evidence and quantification of name forms at several levels of abstraction is proposed as an alternative inductive principle of normalization. A case study of name-form distributions in a dataset of 6,633 spatial attestations in East Norse literature from the Norse World resource serves as a point of departure for a discussion of the advantages and disadvantages of the approach. The comparison between attestations linked to the five most frequent place-names in Old Swedish and Old Danish shows the existence of typical spellings. However, there are still examples of norm negotiations and competitive distributions. Thus, the first inductive step of normalization can be complemented by further processing based on correspondences between phonology and spelling. Finally, stratified normalization of place-names pioneered by Norse World is seen as more versatile compared to traditional methods; the approach has a potential to facilitate both more nuanced philological and linguistic research as well as the further development of named-entity recognition tools.

Keywords

normalization – East Norse – Old Norse – philology – place-names – named entity recognition
1 Introduction*

Modern place-names – Uppsala, Heidelberg or the Murray River – have standardized spellings; these spellings are declared a national and/or international standard by expert authorities such as map agencies, language institutions and/or the United Nations Group of Experts on Geographical Names. Viewed diachronically, the existence of standardized names is a very recent phenomenon. Present-day conventional norms with respect to name orthographies should thus be seen as an exception rather than the rule (cf. Gammeltoft, forthcoming). In medieval sources, names of any kind are subjected to spelling variation along with the rest of the lexicon. This variation is treated somewhat differently – either retained or “neutralized” by subsequent normalization – depending on the field of study and the purpose of the academic enterprise. In the tradition of textual criticism, for instance in handbooks in philology or historical linguistics, names are seldom paid any attention and name variation and normalization is hardly ever explicitly discussed. There are in practice two main directions for processing name stocks in Scandinavian editions of Norse texts; names are as a rule normalized, e.g. in the Íslensk fornrit series publishing works in Old Norse, but not, e.g., in the East Norse tradition. Similarly, computational linguists working with natural language processing and/or named entity recognition (NER) choose either to normalize place-names in accordance with modern spelling conventions (Pettersson 2016, 50; Bollmann 2018, 20) or to retain variation (e.g. Archer, Kytö, Baron, & Rayson 2015, 12). Finally, spatial humanities seem to favour normalized forms at the expense of place-name variation, which limits applicability of such datasets for literary, philological and linguistic inquiry (cf. Petrulevich, forthcoming).

In digital infrastructural resources, for instance digital editions or databases of spatial references, there is no need to choose between the two options, because there is a practical demand and a technological possibility to do both, at the same time. Place-name variation is utterly important as the empirical foundation of any study based on the material, be they etymological inquiries, examination of names in multiple textual witnesses, discussion of textual relations or genre etc. However, in order to enable adequate processing of a large amount of variants, either by quantifying them or examining them

* This work was supported by the Swedish Foundation for Humanities and Social Sciences Riksbankens Jubileumsfond Infrastructure under Grant IN16-00932. The author would like to express her gratitude for the input of project participants Simon Skovgaard Boeck and Jonathan Adams, the Society for Danish Language and Literature, Denmark.

1 Variation in spelling can of course indicate language change, see Empirical Place-name Norms and Norm Negotiations below.
qualitatively in a comparative, contextual perspective, they have to be linked to their respective “tokens” or normalized forms. Traditionally, normalization principles regarding names are the result of deductive reasoning rather than of empirical study of attested material; the names are normalized in accordance with a pre-defined standard, for instance an influential dictionary or present-day conventional forms. This article is instead concerned with both deductive and emerging inductive approaches to normalization. The latter requires large quantities of name data from heterogeneous sources, in other words suitable research infrastructure that has not existed until very recently. In addition to analysis of editorial perceptions of norms and editorial practice, I will explore empirical foundations of place-name norms and possible norm negotiations in medieval East Norse, i.e. Old Swedish and Old Danish, material. The overarching aim of the article is to give an evidence-based empirical background to current deductive editorial practice and to offer an alternative principle of place-name normalization.

2 Deductive Approaches to Normalization of Names and Current Editorial Practice

The term normalization is often used as an umbrella concept for any type of editorial intervention into the form and/or the substance – to loan one of Saussurean dichotomies – of a textual witness. In Norse philology, the notion of normalization seems to be self-explaining, because even research articles, the sole purpose of which is to examine the phenomenon, do not include explicit definitions (Williams 2012; Berg 2014; Williams 2017; Frederiksen 2013). Those that do, most often equate normalization with homogenized orthography (e.g. Skovgaard Boeck 2015, 79; Haugen 2019, 163), although the broadest definitions (cf. Skovgaard Boeck 2015, 79) include for instance change of medium and type-setting of the editorial output. Furthermore, there is a general awareness that normalized orthography most often implies normalized morphology and even syntax (Frederiksen 2013, 3; Berg 2014, 37, 43; Skovgaard Boeck 2015, 79; Haugen 2019; cf. Zeevaert 2014, 986). The purpose of the final product, be it an edition, a collation of textual witnesses or a collection of examples for a language grammar or a textbook, guides the editors’ choices with respect to normalization (Skovgaard Boeck 2015, 79–84; Williams 2017, 56–57; Haugen 2019, 164–167). An editor can for instance choose to homogenize the orthography throughout one and the same textual witness with the scribe’s practical “norm” as point of departure or to normalize the text transmitted in several manuscripts to either a pre-defined external language norm – premodern or modern – or to
Making Sense of Normalization

Empirically anchored internal norm e.g. by introducing the oldest among the attested language forms (Haugen 2019, 164–167; Berg 2014, 36–37 with references; cf. Frederiksen 2013, 3).

Deductive reasoning with respect to normalization and the choice of norm dominates the Norse tradition of textual criticism. The norm is seldom constructed on the basis of empirical studies of manuscript witnesses; rather, the constructed forms of authoritative grammars and/or dictionaries set the standard (cf. Berg 2014; Frederiksen 2013, 3–4; Haugen 2019, 174).2 Interestingly, the East Norse normalization debate of the last decade builds at least partly on the idea of a uniform West Norse normalization (Williams 2012; 2017) – and/or the lack of possibility to introduce the same type of standard in East Norse (Skovgaard Boeck 2015, 81). One of the arguments against normalization in East Norse, for instance, is that it is impossible to construct a single norm for all the surviving texts because of the rapid language change – and consequently immense amount of variation and competing forms – in above all Old Danish (Frederiksen 2013, 3–4; Skovgaard Boeck 2015, 81). However, as Ivar Berg (2014, 39–41) points out, several edition series of West Norse texts, most importantly the immensely popular Íslenzk fornrit, take the work’s time of composition as the point of departure for the external norm, not the surviving copies. There are in practice several sets of external normalization standards even in the West Norse editing tradition (cf. also Paulsen 2017, 2–3). Alternative internal options of normalization have not been explored in detail in the debate.3

Rendering of names or proper nouns has to a large extent been overlooked in the scholarly dialogue on normalization.4 In handbooks of philology, names either appear in the context of etymological examination (e.g. Særheim 2013) or are treated as factual information to be corrected by the editor in case of errors (e.g. Kondrup 2011, 140; see Petrulevich 2016, chap. 2:2 for a more comprehensive summary) even though name materials and variation in names are of much value for textual criticism (see e.g. McDonald Werronen 2016, 31, 48–51; Zeevaert et al.: 21; overview in Petrulevich, forthcoming). Treatment of name stocks in edited Norse texts has in practice two main directions corresponding to the standpoints of the normalization debate outlined above. The editing tradition of Old Norse is characterized by normalization of names in accordance with an external norm, although some name variants can be included in

2 There are of course exceptions; for instance, internal normalization is often employed in textbooks (Skovgaard Boeck 2015; Haugen 2019 with references).
3 See however a short discussion in Skovgaard Boeck 2015, 83 and Haugen 2019, 165, 172–173 that I will come back to in Empirical Place-name Norms and Norm Negotiations.
4 Treatment of names in translation of Norse texts has received scholarly attention, e.g. Hollander 1954; Novotná & Starý 2014 with references.
footnotes or endnotes. For instance, *Síþan gripu vikingar sveina þessa báþa ok hoffu þá út of haf til lórsalalands. Þeir seldu þá í siáborg þeiri es Cesarea heitir húsreyiia aubígrí, ok hét sú lusta ok var Gýpinga kyns*, ‘Afterwards the pirates seized both these boys and carried them off over the sea to Palestine. They then sold them in the sea town which is called Cesarea to a rich lady, and she was called Justa and was of Jewish family’ (Carron 2005, 6, 7), where the proper nouns lórsalaland, Cesarea, lusta and Gýpinga are given normalized spellings, not least with respect to capitalization of the initial letter. The external norm is at least partially encoded in authoritative dictionaries, e.g. Cleasby & Vigfusson and Dictionary of Old Norse Prose (ONP). The problems associated with normalization of names that are not included into authoritative publications is seldom discussed; Zeevaert et al. (21) is the only piece I am aware of that explicitly but briefly addresses the issue in the context of editing multiple textual witnesses of Njáls saga.

East Norse philology advocates diplomatic transcription rather than normalization of names. A prototypical East Norse edition is a so called best-text edition that follows the orthography and other features of a textual witness considered to be the oldest or otherwise the closest to the archetype of the work. For instance, the edition of Erikskrönikan (The Chronicle of Duke Erik; Pipping 1963) diplomatically transcribes its principal witness, Stockholm, National Library of Sweden, D 2 (1400–1523): *fiempte jwla i oslo | Til varfrw kirkio the han baro* (Pipping 1963: 109), ‘on the fifth day of Christmas in Oslo, they carried him to the Church of Our Lady’. The quotation mentions the Norwegian city of Oslo and the city church Mariakirken where Vitslav II of Rügen was buried; the name forms of the source are not manipulated in any way except for expansion of abbreviations indicated by use of italics. Every variant reading from the rest of the relevant medieval and post-medieval manuscripts is collected in the critical apparatus that includes for instance the place-name variant *Oxlo* from Stockholm, National Library of Sweden, D 5 dated 1500–1525. In other words, the East Norse tradition of textual criticism lacks both a conventional

---

5 There is little consistency with respect to critical apparatus in editions of Old Norse texts. Many of Íslenzk fornrit books include only a selection of variants, e.g. Gúðnason 1982, while e.g. the Samfund til Udgivelse af Gammel Nordisk Literatur (STUAGNL) edition of Knýtlinga saga in af Petersens & Olson 1919–1925 has a very detailed collection of variant readings.

6 Initial capitalization is the only aspect of normalization of proper nouns specified by Haugen (2019, 174).

7 It should be noted that ONP does not capitalize the first letter in proper nouns.

8 There are of course exceptions to this general rule. For instance, Roger Andersson’s editions of the Saint Birgit of Sweden (e.g. Andersson 2017) where a synthetic approach is preferred, or Henrik Williams’ edition of the Euphemia’s lays (2020) with modernized spelling.
practice of, and a practical need of, name normalization and for this reason sets of established normalized name forms.\(^9\) A complementary reason underlying this state of affairs is the almost total lack of lookup forms of names in the authoritative dictionaries of East Norse, Söderwalls Ordbok öfver svenska medeltids-språket (Söderwall; 1884–1973) for Old Swedish and Gammeldansk Ordbog for Old Danish.\(^10\) In name indices that accompany some editions, the reader can find both diplomatically transcribed (e.g. in Henning 1954) and normalized name forms (e.g. in Wiktorsson 2020); in the latter case, normalization follows modern dictionary forms of names where possible, although main variants can be included as secondary lookup forms.

Here, an important although both rough and to a certain extent anachronistic distinction between domestic and foreign names should be introduced. Domestic names are understood as names of places, people and other named entities situated, originated or otherwise linked to a country within its present-day borders, while foreign names cover named people and entities outside that particular country. The former category is considered cultural heritage by most states today and as such is subject to archival preservation and archive research. In the Nordic countries, name archives such as Namnarkivet at the Swedish Institute for Language and Folklore, Stednavne- og Personnavnesamlinger (Arkiv for Navneforskning) at the University of Copenhagen, Denmark, or Språksamlingane at the University of Bergen, Norway, are responsible for maintaining name records from premodern sources as well as editing and publishing name materials for further use by researchers and the general public. The publications, e.g. Devine 2021 or SMP 17, contain both present-day standard forms of names as well as a selection of older variants; therefore, most of the domestic name materials are easy to normalize in accordance with these standard forms. For instance, editions of Swedish charters include personal names and place-names in diplomatic transcription – in accordance with the general principle of the East Norse tradition, e.g. Thord and Lwthingø, but normalized modern forms of the names are provided in the indices (Svenskt diplomatarium 121, vi), e.g. Tord and Lidingö. Foreign names, especially less frequent ones, have not been of much interest in the context of construction of national cultural heritage and identity; thus, the point of departure for normalization of foreign names is completely different: the normalized forms have to be coined from scratch.

\(^{9}\) Kalinke 1999 is one example of an attempt to normalize Old Swedish including names; for a discussion and further details, see Williams 2012 and 2017.

\(^{10}\) For a more comprehensive discussion, see Place-name Normalization in Norse World.
Normalization of premodern texts is an established operation or a set of operations in computational linguistics used to facilitate the implementation of natural-language processing (NLP) tools on historical corpora. The concept of normalization has been defined in various ways in the field; according to the mainstream definition used here, normalization is seen as standardization of word forms in alignment with a particular norm (Bollman 2018, 17 with references). The norm in question can be either internal, for instance based on the most frequent or the most consistent variants in a text or in multiple texts, or external, i.e. based on an authoritative source such as a dictionary encoding either a historical or contemporary standardized language (ibid.).

Multiple automatic tools have so far been either tested or developed to normalize historical texts in different languages (for an overview see Pettersson 2016, chap. 3; Bollmann 2018, chap. 4). Similar to the state-of-the-art in philology presented above, the treatment of names varies; names, especially place-names, are either normalized, most often in accordance with present-day standard (cf. Bollmann 2018, 20), or left as they are (cf. treatment of personal names in Pettersson 2016, 50–51).

Among NLP approaches, there is a set of specific text-mining techniques for automatic extraction of proper nouns from unstructured text, NER tools. Studies of NER implementation in historical corpora are still not very common, but this subfield is developing rapidly (overview in Won, Murrieta-Flores, & Martins 2018; Humbel, Nyhan, Vlachidis, Sloan, & Ortolja-Baird 2021). Interestingly, two recent studies comparing multiple NER systems both individually and in combination, Kogkitsidou & Gambette 2020 and Won, Murrieta-Flores, & Martins 2018, come to contradictory conclusions with respect to potential correlation between prior normalization of text and the tools’ performance. Eleni Kogkitsidou & Philippe Gambette (2020) apply six NER tools from CasEN, CoreNLP, Perdido, SEM, spaCy and CasEN+R, to the sixteenth- and seventeenth-century French literary texts both in original and normalized spelling. The multiple normalization steps included neutralization of orthographic inconsistencies and capitalization of initial letters in proper

---

11 Cf. definitions used in philology. A complete survey of available terminology accounting for nuances of meaning, for instance the difference between regularization and normalization, is not of interest here; see Bollman 2018, chap. 2 for a comprehensive discussion.

12 Here, I use the term normalization in the general sense outlined above. Normalization of names thus has nothing to do with named-entity normalization (NEN), a technique of assigning unique identifiers to named entities (see e.g. Weston, Tshitoyan, Dagdelen, Kononova, Trewartha, Persson, Ceder, & Jain 2019).

13 CasEN+R stands for CasEN output further manipulated by GeoNER_repair script (see Kogkitsidou & Gambette 2020, 4 for more details).
nouns in accordance with a named entity dictionary. Normalization was further performed both automatically and manually. NER tools applied to manually normalized texts in most cases showed much better performance compared to both original and automatically normalized texts (Kogkitsidou & Gambette 2020, 2, 5). Contrary to this conclusion, Miguel Won, Patricia Murrieta-Flores, & Bruno Martins (2018, 10) do not see any correlation between normalization and performance of NER tools\footnote{The authors have tested Stanford NER, NER-Tagger, the Edinburgh Geoparser, spaCy, and Polyglot-NER.} in their study of English correspondence from the seventeenth and eighteenth centuries. However, they admit that spelling variation is a problem; for that reason, some initial pre-processing including normalization procedures such as expansion of abbreviations and removing word hyphenations were performed (4). The extracted items were likewise matched against an external authoritative resource, a gazetteer. The authors’ conclusion should be viewed against the design of the study; the normalization or modernization of spelling was performed automatically with two different tools, not manually (Won, Murrieta-Flores, & Martins 2018, 8). NER tools have also been used to link place-names from fourteen Swedish medieval charters and spatial data from seventeenth-century land survey maps of Sweden (Karsvall & Borin 2018). However, name variation was not considered – or constituted a problem – because only normalized place-name forms from editorial charter summaries were used.

3 Place-name Normalization in Norse World

Norse World is a digital interactive platform that aggregates foreign\footnote{Defined in accordance with the discussion in the previous section, more precisely as referring to places outside the current, modern-day borders of Sweden and Denmark. The corpus includes texts of a variety of genres from roughly 1100 until 1530. For a comprehensive presentation of the Norse World resource, please see Petrulevich, Backman, & Adams 2018; Petrulevich & Skovgaard Boeck, forthcoming.} spatial references from medieval East Norse literary texts. It is one of several gazetteers with pre-modern focus that has been built in recent years; however, unlike the comparable research infrastructures Pleiades Gazetteer of the Ancient World, World Historical Gazetteer or The Icelandic Saga Map, Norse World has an ambition to provide every single attestation of spatial references in diplomatic transcription linked to several levels of normalized data.\footnote{For a comprehensive presentation of the Norse World approach in comparative perspective, please see Petrulevich, forthcoming.} The

issue of name normalization, or more specifically place-name normalization, has thus been of pivotal importance for the project. As outlined in the previous section, the normalization tradition in East Norse philology is largely non-existent. Moreover, the principle dictionaries of Old Swedish and Old Danish include very few names. If names are included, the principle of normalization either varies, e.g. compound names can be written both as one or two words, cf. Old Swedish *ryzaland* (Russia; Söderwall 4 1884–1973, 660) and *romara stadher* (Rome; Söderwall 2 1884–1973, 264), or is rather archaic, cf. the Old Danish dictionary form *Northmannia* (Normandy; taken from citation-slip collection of names in Gammeldansk ordbog) and most of the attestations in the database reading *Normandy*. As of February 1, 2022, the Norse World database includes 6,636 attestations (4,139 in Old Swedish and 2,497 in Old Danish) excerpted from 45 manuscripts and early books (20 in Old Swedish and 25 in Old Danish) as well as three medieval runic inscriptions. The ambition of the project is to excerpt spatial references from nearly 200 manuscripts and early books; new data are continuously being entered into the database.

The Norse World project proposed a pioneer place-name data structure to provide access to and facilitate quantitative analyses of names in original spelling, i.e. name forms at attestation level, and in normalized format.\textsuperscript{17} The normalization is carried out manually in two different ways yielding so-called variant and lemma forms; the approach is motivated by the difference in the objective of and the criterion behind the two normalization methods. Variant forms give users an overview of spelling variation in the material, while lemma forms showcase variance in name formation. In practice, both deductive and inductive approaches to normalization are employed. The general framework is of a deductive nature and relies largely on the principles of normalization underlying the external authoritative sources for East Norse, Söderwall and Gammeldansk Ordbog.\textsuperscript{18} Normalization of variant forms is empirically anchored and rather superficial, because it only includes minor interventions such as capitalization of the initial letter. To the contrary, the principles of normalization at the lemma level follow the national dictionary tradition; lemma forms can thus be labelled somewhat archaic constructs or abstractions. For instance, the spelling of the generics or specifics in lemmas is harmonized with the spelling of corresponding words attested in the dictionaries, cf. the

\textsuperscript{17} For a comprehensive presentation of the Norse World approach to name variation and name normalization, see Petrulevich, forthcoming.

variant form *Didimibergh* and its normalized version *Didimibiærgh* incorporating *biærgh* “mountain” (Söderwall 1884–1973 1, 117). However, there are cases where lemma forms go back to original spellings rather than dictionary forms. If a place-name attestation is interpreted as a re-analysis of the exemplar’s form, it is kept as a new lemma. For instance, the Old Swedish lemma forms denoting the Neva River include *Nyn* and *Nynan* attested in the Linköping, Linköping Diocesan Library, H 131 (1500–1525) version of Erikskrönikan (The Chronicle of Duke Erik; Pipping 1963, 83–84). Both forms are likely to have originated as scribal mistakes leading to reinterpretations of the exemplar’s *Nya; Nyn* might be explained as ending with the definite morpheme *-n*, while *Nynan* can be seen as a compound ending in the generic *-an* “river” in the definite form. Every new name formation thus gives a new lemma, i.e. normalization at the lemma level is not based on a named-entity dictionary where each referent only has one official name or a couple of official names or name forms. One and the same location can have multiple lemma forms, e.g. there are three lemmas in Old Swedish and six lemmas in Old Danish linked to the referent Egypt. I will illustrate the Norse World data structure and data processing with another example. The attestation *tha haffuith frøss i aalandh || the tha komma til fynlandh* (Klemming 1867–1868, 127) ‘when the sea froze in Åland, they then came to Finland’ from the Old Swedish Sturekrönikorna (Sture’s Chronicles) in the manuscript D 5 contains the original forms *aalandh* and *fynlandh*. These are first normalized through capitalization of the first letter to the variant forms *Aalandh* and *Fynlandh* respectively; the variants in their turn are further adjusted in accordance with the general normalization framework to *Aland* and *Finland*.

Finally, the original, variant and lemma forms of place-names are linked to spatial metadata such as coordinates and type of locality by so-called standard forms defined as “the most commonly used form[s] of (…) place name[s] in the English language”.19 In practice, however, standard forms include forms in English retrieved from authoritative sources such as GeoNames.org and other digital gazetteers, alternative forms such as historical names of places, e.g. *Reval* in addition to the present-day official name *Tallinn*, and even lemma forms in East Norse in case of unidentified spatial referents that lack common names in English.20 The principle behind the choice of standard forms can be

compared to the use of normalized name forms in East Norse editions as well as to one of normalization procedures employed in NER, the verification of extracted entities against external authoritative resources. Assigning – both choosing and coining – standard forms to further categorize and enrich processed spatial references in Norse World can thus be seen as another form of normalization, a procedure of adapting complex linguistic data to the standard gazetteer model.\textsuperscript{21} The purpose of such normalization is to make it possible to explore the dataset on geographical grounds, i.e. with the geographic location as point of departure.

4 Empirical Place-name Norms and Norm Negotiations

This section is concerned with what an empirical name norm can look like in an East Norse context. The concept is operationalized as occurrences of frequency-wise typical spellings in the distribution patterns of name forms at variant or lemma level that imply conscious or unconscious standardization or supralocalization\textsuperscript{22} of names. The approach draws on interpretation of manuscript evidence as opposed to deductive approaches to normalization building largely on authoritative sources such as dictionaries and grammars. Despite multiple socio-linguistic studies showing the contrary, the perception of pre-modern spelling variation – especially in name materials – as free persists.\textsuperscript{23} An inductive case study of distribution patterns of name forms of the type presented below was not possible to conduct before due to lack of appropriate infrastructural resources as well as relevant theoretical and methodological tools to pre-process the data. Through comparison of the Old Swedish and the Old Danish material of the Norse World dataset, I would like to answer two research questions: What can distribution patterns of name forms reveal about emerging or established empirical norms and norm negotiations with respect to place-names? What circumstances can explain the patterns of variation occurring in the material? The analyses build on the dataset of 6,633 attestations, i.e. all attestations from manuscripts and early books, that was downloaded on February 2, 2022. Here, it is important to note that Old Swedish data (4,136

\textsuperscript{21} On the consequences of employing the model and for some alternative solutions, see Gammeltoft, forthcoming; Petrulevich, forthcoming.

\textsuperscript{22} The term supralocalization is preferred in historical sociolinguistics, e.g. Gordon, Oudesluïjs, & Auer 2020 with references. In the article, I prefer standardization, because geographical provenance of most of East Norse manuscripts is unknown, i.e. geography cannot be used as a variable, cf. discussion in Petrulevich & Skovgaard Boeck, forthcoming.

\textsuperscript{23} Cf. e.g. Gammeltoft, forthcoming and Gordon, Oudesluïjs, & Auer 2020.
attestations) is almost twice as big as the Old Danish one (2,497 attestations). Names are as is well known very infrequent in most contexts; furthermore, the majority of place-names in the Norse World database appear only one to three times. To ensure an appropriate case study material and comparable results, I have chosen to examine the attestations linked to the five most frequent standard forms attested in both languages, i.e. Norway, Rome, Jerusalem, Egypt, and France. However, the analysis is complemented by a discussion of how hapax spellings of place-names can be handled. Descriptive statistics is employed to explore patterns of name variation in the dataset.  

Frequency distributions indicate that empirical norms with respect to place-names were in practical use in medieval literature in both languages. At lemma level, see figure 1 for Old Swedish and figure 2 for Old Danish, there existed a preferred or standard name formation to denote most of the locations: Rom, Iherusalem, Egyptaland, and Frankarie to refer to Rome, Jerusalem, Egypt, and France in Old Swedish, and Norghe, Rom, Jerusalem, and Frankerie for Norway, Rome, Jerusalem, and France in Old Danish. Furthermore, the material includes less frequent lemmas as well as hapax forms; these unique name formations are essential for an overview of the different ways of referring to the same place such as loans from classical languages or possible vernacularizations or re-interpretations of exemplar forms, e.g. Egiptus and Romestath in Old Danish. There are however exceptions from the general rule: two competing lemma forms of Norway in Old Swedish and Egypt in Old Danish show near-equal frequencies. In the first case, language change and genre-specific preferences lie behind the distribution pattern. Most of the Swedish mentions of Norway come from the rich chronicle material that employs the etymologically more original Noreg as well as the contracted East Norse form Norge to refer to the Western neighbour of Sweden. The contracted competitor form is first attested in Norwegian charters in the fourteenth century and becomes increasingly popular thereafter (Sandnes & Stemshaug 2007, 236). Prose sources such as Olav den heliges saga (The Saga of Saint Olaf) and Prosaiska krönikan (Prosaic Chronicle), both composed and written in the fifteenth century, favour the contracted form in accordance with the
language development outlined above. However, *Noreg* is by far the most frequent lemma and variant form in the fifteenth- and even sixteenth-century manuscript witnesses of the oldest chronicle in verse, The Chronicle of Duke Erik (1320–1330);31 a combination of factors such as dating of composition and genre preference of three-syllable variants in rhymed works seems to be a likely explanation.32 The versed Karlskrönikan (The Chronicle of Karl; 1430–1452) occupies an intermediate position both with respect to dating and to the distribution patterns of the Norway forms.33 The youngest of the rhymed chronicles, Lilla rimkrönikan (The Small Rhymed Chronicle; 1448–1453) and Sturekrönikorna (Sture’s Chronicles; after 1470) use only the contracted form and thus language development finally trumps the genre-specific dynamics.34 Interestingly, the Danish material demonstrates a clear preference of the East Norse contracted form *Norghe*. Once again, a temporal factor might be at play. The Old Danish Rimkröniken (The Rhymed Chronicle) that contains most of the Norway attestations35 was composed around 1450, more than a century later than the Chronicle of Duke Erik.36 In the case of Egypt in Old Danish, the choice of either *Egipten* or *Egipteland* seems to be a genre-specific or possibly even a work-specific feature. Devotional, encyclopaedic and didactic texts favour the compound form,37 while the secular travel guide Mandevilles Rejse (The Travels of Sir John Mandeville) makes use of the simplex form in most cases.38

A survey of spelling variation in the dataset shows a similar picture. Orthography of most of the lemma forms is rather stable in both languages, see figure 3 and figure 4 for examples. Moreover, the most frequent spelling variants often correspond well with the normalized lemma forms, cf. e.g. the most frequent variant and lemma form *Frankarie* in Old Swedish and *Rom* in Old

---

31 105 of 123 attestations.
32 For a more detailed discussion, see Petrulevich, forthcoming.
33 Seven occurrences of *Noreg* and 16 of *Norge* in a dataset of 23 attestations in total.
34 Compared to its predecessors, Sture’s Chronicles demonstrates genre developments in other respects, see Petrulevich 2021.
35 68 of the total 73 attestations of Norway, of which 62 *Norghe* and six *Noreg*.
36 The datings of the manuscripts and early books containing the works in question can be summarized as a time period of 1400–1525; most of the material is dated to the fifteenth century. For further details, see the Norse World resource.
37 Sjælens Tröst (Consolation of the Soul) contains 21 attestations in total, of which 6 *Egiptus*, 1 *Egipterike* and 15 *Egipteland*; Lucidarius four attestations, all of *Egipteland*; Stenbog (Lapidary) and Vejleder for Pilgrimme (Pilgrims’ Guide to the Holy Land) one attestation of *Egipteland* each.
38 30 attestations in total, of which 1 *Mersen*, 2 *Kanopat*, 3 *Egiptus*, 1 *Egipteland* and 23 *Egipten*.
Figure 1  Frequency distribution of Old Swedish lemma forms in the dataset

Figure 2  Frequency distribution of Old Danish lemma forms in the dataset
Danish. Swedish material shows a remarkable orthographical stability: there is always a typical or standard spelling of a specific lemma form.\(^{39}\) In Old Danish, there are clear examples of norm negotiations as the situation is complicated by ongoing language change. Two most frequent variants linked to lemma forms *Norghe* (Norway) and *Frankerike* (France) are evenly distributed in the dataset; *Norghe* and *Norghe* appear 22 and 21 times respectively, while the variants *Franckerigy* and *Frankarige* are mentioned six times each. Weakening of unstressed vowels and the introduction of schwa as well as weakening and spirantization of stops in Danish resulted in spelling uncertainties and spelling variance in early and high medieval period (cf. Karker 2005, 1098; Riad 2002, 896–899, 904–905); for instance, the new schwa vowel is spelled both as ⟨æ⟩ and ⟨e⟩, while ⟨g⟩ and ⟨gh⟩ are used interchangeably. Additionally, purely orthographic conventions contribute to the pluralism of spelling options: the letters ⟨i⟩ and ⟨y⟩ are both used to indicate the high front vowel, for example.

![Frequency distribution of some Old Swedish variant forms in the dataset; the variants are linked to the most frequent lemma forms Norge, Rom, Iherusalem, Egyptaland, and Frankarike](image)

\(^{39}\) However, genre conventions and scribal practices can blur the general picture as well. For instance, most of the text and manuscript witnesses that mention Rome favour the variant *Rom*, but Själens tröst (Consolation of the Soul; 1420–1442), a devotional work excerpted from Stockholm, National Library of Sweden, A 108 (1400–1449), makes use of the spelling *Room* (63 attestations) in most cases, cf. *Rom* (8 attestations). The Rome distribution in this case is as close as we can get to norm negotiations in the Swedish context.
In the case study, I have examined distribution patterns of name forms at two different levels on normalization, the lemma level closest to the deductive approaches to normalization presented in the previous sections and the variant level firmly anchored in actual attestations of spatial references. Emerging and established empirical norms with respect to both lemma choices and variant spellings can be observed in the material; there are however some examples of norm negotiations where language change and stylistic variation seem to lie behind competitive distributions. It is important to note that East Norse place-name material as a whole shows much similarity, but the lemma forms normalized in accordance with external authoritative sources contribute to reinforce the image of seemingly divergent two systems, cf. the variants *Norge* (OSw.) and *Norge* (ODa.), *Iherusalem* (both OSw. and ODa.) lemmatized as *Norge* (OSw.) and *Norghe* (ODa.), *Jerusalem* (both OSw. and ODa.), *Egyptoland* (both OSw. and ODa.) lemmatized as *Egyptaland* (OSw.) and *Egipteland* (ODa.) in figure 1–4. Stratified normalization of names, place-names or other name types, is a more versatile alternative to the normalization principles discussed in the literature before. It is significant that the variant level closest to manuscript attestations confirms the existence of conscious or unconscious standardization of place-name spellings. Needless to say, the implementation of such a normalization model requires an appropriate medium, e.g. that of a digital interactive resource. However, typical name formations and name spellings could be and should be taken into consideration when names in normalized format are presented in above all
dictionaries and gazetteers. The typology of normalization employed here can be of much interest for the NER field; by studying and accounting for variation – instead of eliminating it – it should be possible to devise automatic tools that yield nested hierarchies of normalized name forms and thus give a more trustworthy picture of the studied material.

Simon Skovgaard Boeck (2015: 83) describes two possible approaches to normalized East Norse – quantification of word spellings labelled “text-for-text normalization” and orthographic normalization based on correspondences between etymological phonological components and spelling as well as external norms encoded in dictionaries and grammars. In my opinion, the two approaches can complement each other; the former is a more attractive first step of normalization than the latter, not least because it builds upon actual manuscript evidence. However, there is no need to limit quantification of variant spellings to separate texts; on the contrary, it is of much interest to examine variance, empirical norms and norm negotiations at multiple levels – the East Norse corpus as a whole, a single manuscript or manuscript witnesses linked to a single work or a single genre. Skovgaard Boeck highlights that competitive distributions of spelling variants poses a concrete problem for text-internal normalization. It is most likely that this type of distribution will become most evident in Old Danish, cf. the case study results above and the Frankerike-spellings as the most striking example. What is the most appropriate normalization in such a case will as always depend on the task’s objectives; it is possible to use both of the competitive forms or include a subsequent deductive step of normalization based on the aforementioned correspondences between phonology and spelling. Still, I cannot help wondering how much trouble competitive distribution would cause. The answer is that we simply do not know. At the moment, the discussion of empirical foundations of normalization is conducted on deductive grounds. What we need in order to make further progress in developing inductive methods of normalization is an annotated corpus of East Norse at a manuscript level and possibly a referential framework of the kind developed for Old Norse (cf. Paulsen 2017). It might be the case we find that (some of) the East Norse we know is not that variant after all.

\[40\] The available corpora, e.g. Fornsvenska textbanken at Språkbanken, Sweden, a collection of Old Danish texts offered by the Society for Danish Language and Literature, Denmark, or electronic editions of some of Norse manuscripts at the Medieval Nordic Text Archive, are unfortunately not yet suitable for this kind of task for various reasons. Only a tiny fraction of the material is available in them; furthermore, editing principles and available formats vary.
5 Conclusions

Normalization is a concept frequently referred to in philology, historical linguistics and computational linguistics. However, its definitions as well as its operationalization and practical implementations differ greatly within the same field and between fields. Deductive approaches to normalization based on external sets of norms, e.g. authoritative dictionaries or gazetteers, have dominated the scholarly discussion, particularly in Norse philology. Inductive principles of normalization, be it text-internal normalization or normalization involving multiple texts or manuscript witnesses, have not received that much attention (see, however, Boeck 2015, 82–83; Haugen 2019, 165 and discussion above). Until very recently, there was in general very little infrastructural capability to account for and automatically process variation in Norse datasets including names and thus facilitate the development of empirically anchored methods of normalization. At the moment, Norse World is the only research infrastructure for geocoded humanities that provides a structured overview of raw and normalized data at multiple levels. The case study has shown that the repertoire of place-name forms in East Norse is limited and that their distribution fall into clear patterns; in the majority of cases, there are typical ways of spelling names in both Old Swedish and Old Danish. Theoretical implications and practical application of these findings extend beyond the field of philology; the dataset and the analyses can be significant for further development of NER tools in computational linguistics. Normalization will surely remain a deductive process in the future, but an examination or a survey of empirical norms across a given corpus should serve as a foundation for any decision on a suitable norm. I would like to encourage more empirically anchored normalization in the Norse tradition of textual criticism, but the first step towards the goal is as always to build an appropriate corpus.

References


---

41 ONP provides information on both unique name formations and spellings. However, the variants are not linked to each other or nested, which means that users have to know what forms they are looking for in order to look up each of them separately.


Linköping, Linköping Diocesan Library, H 131.


Stockholm, National Library of Sweden, D 2.

Stockholm, National Library of Sweden, D 5.


