[1473] North Frisian dialects: A quantitative investigation using a parallel corpus of translations

Magnus Breder Birkenes

Abstract. *North Frisian is well-known for its small-scale variation and the traditional classification found in Århammar (1968) assumes as many as ten dialect groups within a small area. Until this day, however, a dialect classification based on quantitative methods is lacking and the criteria for the traditional classification are also far from clear. In order to address this problem, the paper uses parallel text material (the questionnaires from Georg Wenker's "Sprachatlas des Deutschen Reichs") and character n-grams (trigrams). Applying cosine distance to the trigram inventories of 55 North Frisian questionnaires, the paper employs several dimension reduction techniques, e. g. multidimensional scaling, Neighbor-Net and hierarchical cluster analysis and compares the results with the traditional classification. While the latter can be confirmed to a large extent, the distinctions within Southern Mainland North Frisian seem to be less clear. Using an association measure (log-likelihood), prominent features are extracted for the six main dialect groups that emerge on basis of the aggregated data. Finally, the paper discusses the quality of the North Frisian Wenker questionnaires in the light of these findings.*

Keywords. *n-grams, North Frisian dialects, classification, parallel corpus*

1 Introduction

Among the Germanic languages, North Frisian counts as a prime example of a small language with a substantial degree of linguistic variation: "[N]owhere in the Germanic speech community is there more rampant dialectal split in such a microcosm as there is in North Frisian." (Markey 1981, 211; similarly Århammar 1968, 295). Within a small area of roughly 2000 km^2, ten dialect groups are traditionally assumed (cf. Walker and Wilts 2001, 284-286). The classification of North Frisian dialects is primarily based on the work of Siebs (1889; 1901), and was further corroborated by Hofmann (1956) and Århammar (1968). The ten North Frisian dialects are separated into two main groups: Insular and Mainland North Frisian, a division that is mostly linked to two different waves of immigration into North Frisia (cf. Walker 2001,

266).The geological situation of parts of North Frisia in the Wadden Sea, with a landscape that has dramatically changed shape within the last few hundred years due to rising sea levels and human intervention, also did likely play an important role in the fragmentation of North Frisian dialects. Most of the small islets now known as *Halligen* are geologically young islands, and likewise, parts of today's mainland (i. e. in the Wiedingharde and the Bökingharde) were surrounded by water a few hundred years ago and first became part of the mainland due to land reclamation in recent centuries. The (alleged) limited intelligibility of many dialects had an interesting sociolinguistic consequence: Frisian is mostly limited to the local level, whereas outside Low and High German are used as a *lingua franca* of inter-Frisian communication (cf. Walker 1990, 3-4). Some centuries ago, the influence of Danish must have been more important than it is today (cf. Hofmann 1956, 79). Multilingualism is a defining feature of the area until this day.

But even though in the tradition of Frisian linguistics ten dialects are assumed, the criteria for doing so are far from clear. This was the motivation for Walker (1980) to look at the grouping of the Mainland North Frisian dialect of the Bökingharde into sub-dialects. According to Walker, the current classification into ten dialects lacks transparency. In his own words:

> Obwohl die Dialektzersplitterung als solche und auch die Mundarteinteilung allgemein bekannt sind, ist es schwierig herauszufinden, nach welchen Kriterien diese Einteilung vorgenommen wird. Es scheint überhaupt keine systematisch durchgeführte Übersicht der nordfriesischen Mundarten zu geben, die diese Einteilung rechtfertigt [...] Mit anderen Worten, man spricht von einer Dialektzersplitterung, aber die Kriterien für diese Behauptung und für die daraus folgende Mundarteinteilung sind weder systematisch erforscht noch erfasst (Walker 1980, 1).

Or as Markey (1981, 222) puts it:

> We even lack comprehensive maps of salient phonological and lexical features, and it is features from these levels of the grammar that serve to group areas into micro-dialectal units.

Until this day, a complete survey of North Frisian dialects is lacking, although relevant overviews can be found in Siebs (1901), Hofmann (1956), Århammar (1968), Walker (1990), Walker and Wilts (2001) and Århammar (2001). While compensating for this certainly cannot be the goal of this paper, it is my aim to shed some light on the classification of North Frisian by linking the results of a computational analysis to previous findings in traditional North Frisian dialectology. The material I am going to use is a corpus of

parallel texts, i. e. dialect translations from the 19th century completed under the supervision of Georg Wenker ("Sprachatlas des Deutschen Reichs"). The method that I use may be characterized as "linguistically naive": Other than a certain normalization of the transcriptions (documented in section 3.1), the material is not annotated in any way. Rather, I will attempt a dialect classification based on the frequencies of character sequences (so-called character *n*-grams) in the Wenker questionnaires and extract prominent features using a statistical association measure. I also want to assess the usefulness and pitfalls of the Wenker materials as a resource for North Frisian linguistics.

The approach probably lies somewhere between "atlas-based dialectmetry" and "corpus-based dialectometry" (Szmrecsanyi 2013, 4). Clearly, the Wenker questionnaires form a corpus of texts. These texts were, however, constructed with the specific purpose of producing a dialect atlas. While being standard in computational linguistics, character *n*-grams have to my knowledge hardly been used in the study of dialects of Germanic languages before (but see Hoppenbrouwers and Hoppenbrouwers 1988 to be discussed in section 3.2 and Dipper and Schrader 2008 using Middle High German parallel texts). In this paper, I will only be concerned with the (dis)similarities of North Frisian dialects as shown by the Wenker questionnaires and generally not consider the question of language contact and the complex relations between Frisian and other languages in this area, as this subject has been covered extensively by Lameli (2010).

2 The material

The material used for this study comprises the North Frisian questionnaires elicited by the "Sprachatlas des deutschen Reichs". Within the borders of the German Empire, 46,011 questionnaires were collected in the years 1879-1888 (cf. Wenker 2013, 2), including translations of 40 sentences into the local dialect – not only in Low or High German, but also in languages like Polish, Sorbian, East and North Frisian or Danish spoken in the German Empire at that time (cf. generally Fleischer 2017 and for North Frisian Fleischer 2012 and Bosse to appear).

The translations – hand-written lay transcriptions – were carried out by schoolteachers, sometimes supported by their pupils. The questionnaires form the basis of a colorful hand-drawn atlas with the same name (consisting of slightly more than 500 maps), which due to technical limitations could not be published until the beginning of the 21st century (in digital form), although a

simplified version, "Deutscher Sprachatlas" (DSA), appeared in the years 1927-1956 (cf. Herrgen 2001, 1522). All in all, there exist 66 questionnaires for North Frisian in this collection: 61 questionnaires from the first survey 1879-1880 (cf. Wenker 2013, 3), and five from later surveys (cf. Bosse to appear), with a good geographical spread. Thus, the North Frisian area is all in all well represented. The project director, George Wenker, seems to have had a special interest in North Frisian and the language situation of North Frisia, as documented by his map of the language situation in this area (cf. Wenker 2013, 5; see also Lameli 2008) and his frequent remarks on North Frisian throughout his commentary, which was only recently published (cf. Wenker 2013 and Fleischer 2017, 40).

Interestingly, the maps and questionnaires of the "Sprachatlas" have not been much used within Frisian philology and linguistics (cf. Bosse to appear). For example, the current reference work on Frisian linguistics, *Handbook of Frisian studies* (Munske 2001), only mentions Georg Wenker and his "Sprachatlas" once, and only very indirectly (cf. Bosse to appear). One prominent exception is Selmer (1926) on the complex article systems of North Frisian, where the Wenker questionnaires and maps are used together with religious and literary texts. The skepticism towards the material may be illustrated by the following quote from Dietrich Hofmann:

> Es ist verständlich, daß der Deutsche Sprachatlas bei solchen Verhältnissen manchmal versagt, vor allem, wenn es sich um lautliche Unterschiede handelt. (Hofmann 1956, 86-87)

According to Hofmann, the complex vowel and consonant inventories of the North Frisian dialects are not captured in a satisfactory way by the "*Sprachatlas"*.

Criticism of the "Sprachatlas" was indeed not limited to the Frisian materials. Especially the use of lay transcriptions started a long controversy at the end of the 19th century (cf. Herrgen 2001, 1524). Interestingly, Wenker's most vocal critic, the phonetician Otto Bremer (cf. Bremer 1895, and, in retrospective, Haas 1995) was himself active on the field of Frisian linguistics (cf. Bremer 1887/1888). There is no doubt that the indirect method used in the "Sprachatlas" is a two-edged sword: On the one hand, delegating the translation task to local teachers made the whole enterprise possible, considering its extremely dense grid of locations. On the other hand, the experts had little control over the informants and transcriptions and relevant biographical information like year of birth is not found on the questionnaire. In the case of North Frisian, Hofmann (1956) laments that many of the

teachers were not familiar with Frisian or came from other parts of North Frisia. But as Bosse (to appear) shows, at least for 45 questionnaires it is plausible to assume that people who had competence in the local dialect were involved in the translation.

One further reason why the Wenker materials have been neglected in Frisian linguistics may have been that the sentences were assumed not to provide enough interesting material for a study of Frisian. From a qualitative point of view this would make sense: The 40 sentences were carefully constructed in order to capture relevant phenomena in, first and foremost, German dialects (cf. Wenker 2013, 1), among others words affected by the High German consonant shift, various monophtongization and diphthongization processes and diminution. They do thus not represent "natural language", but rather a mixture of phenomena known to follow a certain areal distribution. Seen from a quantitative perspective, this may however be a strength in the case of North Frisian in that the material is not biased towards this language (cf. Lameli 2010, 25). Thus if we still find clear areal differences, this could potentially provide independent, strong support for a classification of North Frisian dialects. In the following, I will leave the methodological doubts aside and carry out a quantitative dialectometrical analysis on the material, but I will return to some of the problems in the discussion.

When talking about the "Wenker materials", I mean both the maps of the "*Sprachatlas"* and the questionnaires upon which these are based. But in my quantitative analysis, I will resort to machine-readable versions of the questionnaires, transliterated from hand-writing (mostly in the "Deutsche Kurrent") and I will only occasionally use the maps for qualitative exploration and validation. When referring to the questionnaires only, I will use the term "Wenker data". The questionnaires were digitized at the Forschungszentrum *Deutscher Sprachatlas*[1], the transliteration and correction of the 66 North Frisian versions were carried out by various researchers and assistants in Kiel, Marburg and latest by Temmo Bosse in Flensburg. Temmo Bosse, who currently plans an edition of the North Frisian questionnaires, kindly provided me with digital access to his latest version.

Of the 66 questionnaires available, I decided to leave out 11 for various reasons to be discussed below. The 55 questionnaires that I use in this study are listed in table 1 and plotted on a map in figure 1. In accordance with

1. Metadata and scans from: https://regionalsprache.de/Wenkerbogen/Katalog.aspx

Fleischer (2017, 8-10) I cite the questionnaires with their respective questionnaire number followed by the location name, i.e. 46824 Norddörfer. In tables and graphics, I omit the location name and append the abbreviations for the various dialect groups for convenience, in the case of Norddörfer on Sylt: 46824_SY.[2]

Dialect	Questionnaires
Syltring (SY)	(6): 46824 Norddörfer, 46885 Westerland, 46886 Tinnum, 46887 Keitum, 46888 Archsum, 46889 Morsum
Föhring-Amring (FA)	(10): 46572 Nebel, 46747 Norddorf, 46748 Utersum, 46749 Oldsum, 46750 Toftum, 46751 Borgsum, 46753 Alkersum, 46754 Midlum, 46756 Wrixum, 46757 Boldixum-Föhr
Wiedingharde Frisian (WI)	(6): 46699 Horsbüll, 46700 Emmelsbüll, 46891 Uphusum, 46890 Norderdeich, 46892 Klanxbüll, 46893 Rodenäs
Bökingharde Frisian (BÖ)	(7): 46701 Marienkoog, 46702 Deezbülleck, 46703 Niebüll, 46707 Risum-Lindholm, 46708 Nordlindholm, 46760 Dagebüll-Kirche, 46761 Fahretoft
Karrharde Frisian (KA)	(7): 46709 Wester Schnatebüll, 46711 Klintum, 46764 Stedesand, 46765 Sande, 46766 Enge, 46767 Schardebüll, 46779 Soholm
Nordergoesharde Frisian (NG)	(9): 46575 Süd-Ockholm, 46576 Büttjebüll, 46577 Sterdebüll, 46578 West-Bordelum, 46579 Dörpum, 46763 Nord-Ockholm, 46771 West-Langenhorn, 46772 Loheide, 46773 Mönkebüll
Mittelgoesharde Frisian (MG)	(3): 46586 Drelsdorf, 46587 Almdorf, 46588 Bohmstedt
Südergoesharde Frisian (SG)	(3): 46638 Altendeich, 46640 Wobbenbüll, 46646 Hattstedt
Hallig Frisian (HA)	(4): 46636 Hooge, 46759 Oland, 46573 Langeness, 52969 Gröde

Table 1: Data set

2. The names for the dialect groups are taken from Walker and Wilts (2001, 285), based on Århammar (1968, 296), and were slightly adapted to English. I will use them throughout the paper.

Figure 1: Locations of the questionnaires

All in all, we see that the ten dialects of North Frisian are quite well and evenly represented, with the prominent exception of Helgoland. Unfortunately, the only questionnaire from Helgoland contains a very idiosyncratic transcription system with a plethora of diacritics that are difficult to interpret. It is also a rare example of a "Sprachatlas" questionnaire that was elicited directly on-site.[3] For these reasons it had to be excluded from the current study, thus leaving one of ten groups out. We also note that there are only three questionnaires per group for the two southernmost dialects in the Mittel- and Südergoesharde. This shows that already in the end of the 19th century it was difficult finding speakers of North Frisian in these areas (the last speaker of Südergoesharde North Frisian died in 1981, cf. Walker and Wilts 2001, 284). As a matter of fact, Bosse (to appear, 13) upon closer inspection found that none of the teachers in questionnaires from the Südergoesharde originated from this area. But contrary to Hofmann's (1956) criticism this is the exception rather than the rule in the North Frisian Wenker data.

Alongside the problematic questionnaire from Helgoland, I decided to exclude ten further questionnaires: Five questionnaires show an interesting pattern where one half is translated into North Frisian, the other half into another language spoken within the village (Low German, Danish), reflecting the complex language situation in North Frisia (this applies to 46755

3. As Fleischer (2017, 42-49) reveals, questionnaire 47862 Helgoland stems from Georg Wenker himself.

Oevenum, 46769 West-Bargum, 46770 Ost-Bargum, 46897 Neukirchen, 46898 Hörn; cf. Bosse to appear, 7-8). Three questionnaires were collected at a later stage, and could thus possibly reflect language change (200000 Nebel, 200001 Oevenum, 200002 Oldsum). Last, for two locations, Westerland (Sylt) and Hallig Gröde, there exist multiple questionnaires: 52963 Westerland shows a transcription system similar to that of 47862 Helgoland[4], whereas 46574 Gröde displays unexpected influence from Amrum (i.e. <a> in unstressed syllables: *wesan* 'been').[5]

Since duplicates only exist for these two locations, I settled on the presumably better questionnaires 46885 Westerland and 52969 Gröde. Other than in these two cases, I did not exclude any questionnaires on subjective grounds. Hofmann (1956, 87) explicitly mentions 46761 Fahretoft and 46709 Wester Schnatebüll as inadequate questionnaires showing "foreign" dialect features. The following study will show whether this holds true from a global perspective.

3 Methods

In the following, I will describe the methods used to normalize and process the transliterations of the North Frisian questionnaires. The transliterations of the 55 questionnaires listed in table 1 are stored in a central SQL database, where primary text and metadata are strictly separated. Each questionnaire is assigned to a location, whose coordinates are used in drawing maps. All further steps are carried out using the programming language and statistical package R (R Core Team 2018), the corpus model and most corpus processing functions are taken from the package `quanteda` (Benoit 2018).

3.1 Text cleanup and normalization

Since the translations were provided by individual teachers without any professional phonetic experience and without any rules on transcription, the principles used for rendering the dialect are also individual to a certain

4. As with 47862 Helgoland, this questionnaire is also from Wenker's hand (cf. footnote 3).
5. This probably relates to the fact that both 46572 Nebel and 46574 Gröde were translated by the very same teacher, working on Amrum but originating from Gröde (cf. Bosse to appear, 7). The translations from Amrum and Gröde were handed in on one questionnaire and later copied to two questionnaires. It is not impossible that some of the similarities are related to the copy process, unfortunately this is impossible to prove since the original questionnaire appears to be lost.

degree.[6] Thus there is a risk that we get writing profiles of individual authors/translators rather than true dialectal differences. There is also no single coherent orthography system of North Frisian, but there are indeed certain regional traditions, e. g. on Sylt with a long-standing continuity (cf. Wilts 2001, 305), that may have been used by some of the teachers. In order to deal with this graphematic variation, I decided to apply some normalization to the texts for the sake of comparability, but at the cost of accuracy. There were also metadata such as comments on pronunciation and variants in the translations (mostly enclosed in parenthesis), that needed to be left out before doing a quantitative analysis.

This cleanup and normalization was done as integral part of the processing and without physically changing the transliterations. In the following, I will give a simplified description of the individual steps and comment on my choices. Most importantly, I applied the methods described in Moran and Cysouw (2018), implemented in the R package `qlcData` (Cysouw 2018). This involves creating an "orthography profile" for the corpus where all so-called Unicode code points or characters[7] are listed, which can then be grouped together as graphemes: The Unicode system contains a wealth of optically similar characters separated into so-called blocks, which are, however, treated as different characters computationally. For example, a schwa character is found both in the IPA block and in the Cyrrilic block and both look very similar, but they are treated as two completely different characters. The picture gets more complex when diacritics are involved. Obviously, this is very important for the transliterations used here, having been created by different persons, using different applications. By creating an orthography profile of the North Frisian transliterations, I managed to reduce the number of characters in the texts from 122 to 29 and thereby drastically reduced the amount of characters only appearing in a few texts.

The following cleanup and normalization rules were applied to the data set in table 1 (no manual correction was done) in the order specified here, using the R package `qlcData` and regular expressions:

---

6. This is the case for the 1879/80 and the 1887 survey, from which all questionnaires stem. In the 1887/88 survey in Southern Germany, Wenker suggested the use of certain diacritics (cf. Fleischer 2017, 29), but this is irrelevant for the North Frisian data, with the exception of Wenker's directly elicited questionnaires from Helgoland and Sylt (which are not part of this study for that very reason).
7. I will ignore the rather intricate distinction between glyph and character, because it is not relevant for the discussion.

1. Convert all letters to lowercase
2. Strip all punctuation
3. Remove all text within parenthesis (comments, optionality)
4. Normalize certain character sequences depending on context:
   a) remove double vowels and consonants if the same character is repeated
   b) remove <h> if preceded by any vowel
   c) convert capital <I> or <J> to <j> before vowels, to <i> before consonants[8]
5. Normalize certain characters regardless of context:

look for	replace with
ck	k
th, dt	t
ſ	s
ß	ss
q	k
x	ks
ə, ë, ä, æ	e
œ	ö
å	o
“all diacritics except umlaut”	“base character without diacritics”

Of the rules above, only the rules 3-5 demand a further explanation. In rule 3, all text in parenthesis is removed. In the Wenker data, parenthesis may indicate e g. variation (for example Ø vs. schwa or a synonym) or a general comment (mostly on pronunciation) made by the teacher. Additionally, in the machine-readable versions used in this study, parenthesis may also be used by the transliterators to indicate uncertainties in interpreting the hand-writing. Due to these factors, I chose to omit text in parenthesis in the first place. Once an edition of the North Frisian questionnaires exists, this step could possibly be reconsidered.

Rule set 4 deals first and foremost with vowel length marking, which is variable in North Frisian, as in all other Germanic orthographies. There is,

8. Obviously, capital <I> and <J> were marked accordingly before all letters were converted to lowercase (step 1).

however, a certain North Frisian tradition for vowel doubling, which is not at all consistent (cf. Wilts 2001, 308). Both double vowels and double consonants were therefore simplified, as were vowel + *h* sequences. There is obviously a risk that relevant linguistic information other than length was lost here. For example, on Sylt <aa> can represent [ɔ], as in the older Danish orthographic tradition (cf. Wilts 2001, 307). I did, however, not choose to apply regionally different normalization rules. A remaining problem is the <ie> spellings that may, in the German tradition, represent a long [i:]. These spellings were not normalized due to the potential complexity of the rules. The last context-dependent rule should be uncontroversial, on the other hand: since capital <I> or <J> are identical in the "Deutsche Kurrent" – the type of handwriting used in most questionnaires – the transliterations, however, show variation between <J> and <I>, I decided to normalize capital <I> and <J> to <j> before vowels and to <i> before consonants. This is also the usual practice when editing documents in the "Deutsche Kurrent".

In rule set 5, certain consonants and vowels were normalized independently of the context. Some choices are unproblematic, for example the treatment of the purely orthographic long <ſ> or the Eszett <ß>. With the other rules, we sacrifice potential linguistic information for the sake of comparability, and the most radical rule is probably the last: Since the usage of accent acute, grave or breve is subject to variation and mostly undocumented by the teachers, I decided to leave all of them out and only to consider the base character. Umlaut diacritics, on the other hand, were kept, except for <ä>, which was normalized to <e>. The realization of open-mid short [ɛ] and close-mid [e] is prone to variation in the writing systems, and not all of the North Frisian dialects show a phonemic distinction here, especially in the short vowels and among the insular dialects (cf. Walker and Wilts 2001, 288; cf. Löfstedt 1928, XXII for parts of southern Mainland North Frisian). According to Wilts (2001, 308) <ä> is mostly used in Mainland North Frisian written tradition, <e> in Insular North Frisian (both representing [ɛ]), but in the Wenker questionnaires, we do find variation between <ä> and <e> within individual dialects in both areas. Given this unclear situation, we probably gain more than we loose by lumping these graphemes together for the purpose of this article.

An example of the normalization process is given in table 2, showing Wenker sentence (henceforth: WS) number 1 in all the questionnaires from the Bökingharde (the original to the left, the normalized version to the right):

	Transliteration	Normalization
46701_BÖ	Ön é Wónter flieé dé dröge Bleé dör é Luft ámbei.	ön e wonter flie de dröge ble dör e luft ambei
46702_BÖ	Am Wanteren fliee da drüge Bleſe dör ä Luft ambei.	am wanteren flie da drüge blese dör e luft ambei
46703_BÖ	Am wunterm flîe da drôge bléthe dœr-e luft ambâi.	am wunterm flie da dröge blete döre luft ambai
46707_BÖ	Önj n‘ wunter fliee dă dröge bleese ŏnj n‘ lŭft ămbei.	önj n wunter flie da dröge blese önj n luft ambei
46708_BÖ	Am Wuntermen flië da dröge Blese dör ä Luft ămbei.	am wuntermen flie da dröge blese dör e luft ambei
46760_BÖ	Äunje Wunter fleije dê dröge Bleese döre Luft embei.	eunje wunter fleije de dröge blese döre luft embei
46761_BÖ	Di Wonter fläie da drögge Blese aun ă Locht ambai.	di wonter fleie da dröge blese aun e locht ambai

Table 2: Transliterated and normalized version of WS 1 in all the questionnaires from the Bökingharde. Original German version: *Im Winter fliegen die trocknen Blätter durch die Luft herum*. ‘In winter the dry leaves fly around in the air’.

One of the questionnaires, 46707 Risum, is described by Bosse (to appear, 16) as one of the more problematic cases involving strange transcriptions, i.e. with a breve above the umlaut. A scan of the first three sentences in the original questionnaire 46707 Risum is included in figure 2, to give an impression of the amount of work needed to get from transcription to a machine readable normalized transliteration (the red underlinings and blue symbols were probably added later and are not relevant here):

Figure 2: Scan of the first three sentences of 46707 Risum in the Bökingharde

After normalization, this questionnaire can be rather neatly compared to the other questionnaires from the Bökingharde. Notice for example the variation in vowel length marking and the usage of diacritics in all questionnaires in table 2: In terms of the finite verb, *flie* ‘fly’, the number of word types is reduced from 6 to 3 and in the definite plural article *da* ‘the’, from 4 to 2 forms, hopefully without loosing too much phonetic information.

3.2 N*-grams, document-feature matrices and feature weighting*

N-grams are sequences of *n* items in running text. These items can for example be words or characters, depending on the task. *N*-grams are frequently used in computational linguistics and information retrieval for processing and classifying text, e.g. in statistical machine translation (cf. Manning and Schütze 1999, 191-202), language classification and identification (cf. Cavnar and Trenkle 1994), authorship attribution and topic modeling (cf. Jockers 2014 for an accessible introduction). Because the Wenker questionnaires contain heterogeneous and sparse dialect material, it is sensible to go beyond the level of the word and look at sequences of characters (although in this parallel corpus, one would come a long way even when looking at the word). This way, we are also able to capture phonological patterns or morphological endings, which are highly relevant in dialect classification.

Accordingly, this paper aims at a frequency-based approach to dialectometry (cf. Heeringa and Nerbonne 2013, 628). In the Groningen school of dialectometry, average Levenshtein string distance is used (cf. Heeringa 2004) which allows for a more fine-grained comparison of corresponding words. The application of Levenshtein distance, however, relies on aligned words/strings, a process that is non-trivial even in parallel texts (e.g. deviations in translation, syntactic variation, clitization). Using a frequency-based *n*-gram method, we compare documents (here: questionnaires) rather than words without needing to align them. Although *n*-grams are widely used in computational linguistics, they have not been much used in dialectometry to my knowledge. But Hoppenbrouwers and Hoppenbrouwers (1988) used phone frequencies, i.e. unigrams, in a comparison of Dutch dialects.[9] In the following paper, which will work with higher order *n*-grams, additionally a significance test (log-likelihood) will be carried in order to check whether asymmetries in the frequency distributions can be attributed to mere chance.

Creating *n*-grams is simple: One only has to loop over text and extract sequences of *n* characters, moving one character to the right for each iteration. The result is a list of such sequences as shown in table 3. In my notation of *n*-grams, underscore represents space. This makes it easier to identify word boundaries and visually separate prefixes from suffixes:

9. I would like to thank one of the anonymous reviewers who referred me to the paper by Hoppenbrouwers and Hoppenbrouwers (1988).

Character unigrams:	w,i,_,s,a,n,_,t,r,e,t,_,a,n,_,h,a,_,t,a,s,t
Character bigrams:	wi,i_,_s,sa,an,n_,_t,tr,re,et,t_,_a,an,n_,_h,ha, a_,_t,ta,as,st
Character trigrams:	wi_,i_s,_sa,san,an_,n_t,_tr,tre,ret,et_,t_a,_an, an_,n_h,_ha,ha_,a_t,_ta,tas,ast

Table 3: Table of *n*-grams for WS 23 *wi san tret an ha tast* 'We are tired and thirsty' (46753 Alkersum, after normalization). Original German version: *Wir sind müde und haben Durst.*

In this paper, I chose to use sequences of three characters (so called trigrams). Obviously, relying upon only one-character sequences (unigrams) leaves out a whole lot of phenomena like diphthongization, consonant clusters and endings. Therefore, bigrams or trigrams are mostly employed in classification tasks. With trigrams, we capture many of the phenomena above, possibly even with some positional information like differences in distribution between a word-initial <sk>, in my notation <_sk>, and a word-final <sk>, that is <sk_>, which is important in certain Mainland North Frisian dialects (cf. Hofmann 1956, 105), where certain dialects show assimilation of the consonant cluster word-initially, but not word-finally. When creating trigrams, I decided to restrict these to the level of the word, i.e a trigram like <n_t> in table 3 spanning two words was not considered. Leading and trailing spaces were, however, kept for the reasons discussed above. Doing so means that we rely solely on phonological, morphological and lexical information – leaving syntax out. But the syntactic information potential of short character *n*-grams is limited anyway.

For the following paper, *n*-grams for all 55 questionnaires were created in the way illustrated above and then counted, thereby creating types: For example, a high proportion of trigrams with the substring *an* like in <san> or <an_> above could point at a tendency of vowel lowering in the relevant dialect (e g. OFr *sin(d|t)/sen(d|t) > sen > san* 'are'). The end-product is a so-called document-feature-matrix containing each trigram type in the corpus (as columns) and the number of times it occurs in a questionnaire (as rows). Without cleanup and normalization, this first led to a total of 9,102 trigram types in the document-feature-matrix, with a very high sparsity degree of 88.4%., i.e. the proportion of cells that contain 0, that is, features that only appear in a few or only in one questionnaire. This is hardly surprising for this kind of data: Not only are we dealing with a heterogeneous area, we also have heterogeneous spellings with a subtle amount of individual variation. For the following study, it was of value to reduce the amount of idiosyncratic types. After cleanup and normalization, I decided to leave out all trigrams only

occurring within one document and trigrams occurring less than 15 times in the whole corpus.[10] This lead to a substantial reduction from 9,102 to 984 features with a sparsity degree of 46.8%.

Table 4 shows a small subset of the document-feature matrix used, with the top 10 trigrams in the whole corpus taken from a randomly chosen sample of nine documents (the ordering is not random: it is grouped by the two main branches of North Frisian and ordered from north to south):

	er_	an_	en_	at_	_he	_da	_be	_de	_en	de_
46889_SY	37	9	66	7	8	8	9	5	19	3
46750_FA	33	17	56	5	9	6	9	25	17	12
46891_WI	21	4	53	14	15	18	14	12	16	12
46701_BÖ	30	16	42	16	13	29	8	10	20	4
46707_BÖ	23	21	42	10	14	25	8	12	19	11
46766_KA	24	4	62	12	21	11	14	33	22	28
46578_NG	28	4	41	16	18	13	16	14	8	4
46640_SG	31	5	51	21	17	18	12	15	5	12
46759_HA	25	3	65	14	16	12	15	10	19	8

Table 4: Document-feature matrix of the overall top features in a subset of nine documents

The top three trigrams <er_>, <an_> and <en_> mostly represent verbal and nominal endings. Their distribution is quite telling. All in all, <er_> seems to be more common in Insular North Frisian than in Mainland North Frisian (which might be related to a certain morphological preference, cf. section 4.3.2). We also notice an asymmetry between <an_> and <en_> in that the former is very common in the Bökingharde and in Föhring-Amring, whereas the latter is more frequent elsewhere. Such distribution differences will play a very important role in the following.

We could use the document-feature-matrix computed so far for the computation of distances between the questionnaires. Doing so would, however, give a substantial weight to the very frequent trigrams shown in table 4, whereas less frequent, but highly characteristic trigrams would contribute less to the end-result. One common way of correcting this is to

10. The chosen threshold is arbitrary: Generally, by setting document frequency to a minimum of 2, we single out certain idiosyncratic spellings. By additionally setting a term frequency threshold, we remove further potential noise and data sparsity (i .e. variables where most of the observations are 0).

weight the document-feature-matrix. There are many ways of doing this: I settled on logarithmic weighting (with base 2), as shown in (1):

(1) $log2(tf + 1)$

The formula in (1) takes the logarithm of the term frequency increased by one. This way, a feature with a term frequency of 0 also gets a weighted frequency of 0 (increasing the term frequency by one circumvents the zero-division problem), a feature with a frequency of 1 also gets a weighted frequency of 1, but differences in higher frequency terms are given lower weight. For example, the frequency of <an_> in 46707_BÖ and 46766_KA (as shown in table 4), which is 21 : 4, is only 4.46 : 2.32 after weighting.

There are also various other weighting schemes with which I experimented in the course of the investigation: One common scheme is the so-called tf-idf (term frequency / inverse document frequency). With tf-idf, the frequency of a term (i.e. a trigram) is seen in inverse relation to the number of documents (document frequency) in which it occurs. The idea is that a trigram that occurs in all documents is not very useful when it comes to classification, whereas one that is common only in one or a smaller number of documents is. The problem here is that in excluding certain endings such as *-en* or *-er* (that occur in all documents) we also loose important isoglosses between Insular North Frisian and Mainland North Frisian, blowing up the distances between the dialects severely. Log-weighting, then, seems to be a reasonable compromise.

3.3 Similarity and distance

The final step is to compare the (weighted) trigram inventories of all data points or questionnaires in terms of similarity/distance. There exists a vast amount of similarity measures one could use: One of the simplest methods is the so-called Jaccard index: Here, the similarity of two documents is defined as the ratio of trigrams shared by two dialects in relation to the total amount of trigrams in both dialects, i. e. the size of intersection divided by the size of the union of the two trigram sets to compare. Two identical sets have the index 1, two disparate sets have 0. Although this in many cases may suffice, it is very simplistic in that the frequency of the various trigrams is not considered, giving each trigram the same weight. In order to account for frequency, we resort to another similarity metric: the so-called Cosine similarity. The idea behind cosine similarity is to compare two vectors, i.e. the (weighted) frequencies of all trigrams in two documents and compute the cosine of the angle between them, where each feature represents one dimension in the vector. A cosine similarity of 0 indicates that the documents

are completely different (90° angle), a cosine similarity of 1 (0° angle) means that they are identical. Because term frequencies cannot be negative, 90° is the maximum of the angle, and therefore the boundaries of cosine similarity in text mining are always positive [0,1].[11]

When comparing the trigram inventories in this way we get the similarity between two questionnaires. We are, however, more interested in the distances between them, so that we can project these onto a map. This is done simply by subtracting the cosine similarity from 1. I will give a practical example on how to compute cosine distance in the following, using WS 23 *Wir sind müde und haben Durst* 'we are tired and thirsty' and the same sample of questionnaires as in section 3.2. The sentences to be compared are shown in table 5:

Questionnaire	normalized text (WS 23)
46889_SY	*wü sen tred en ha töst*
46750_FA	*wi san tred en ha tast*
46891_WI	*wi sen trart en hewe tast*
46701_BÖ	*wi san trat en hewe torst*
46707_BÖ	*we san trat en hewe torst*
46766_KA	*we sen trad en heve torst*
46578_NG	*wi sen trat un hewe törst*
46640_SG	*wi sen trat un het torst*
46759_HA	*wi sen troet en hef torst*

Table 5: Sentences (documents) for comparison

The so-called distance-matrix returned by R is shown in table 6 below. The matrix is symmetric: because the distance between x and y is the same as the distance between y and x, repeated values above the diagonal can be omitted. The distance between x and x is 0. The closer the distance to 0, the more similar are the documents.

11. Hoppenbrouwers and Hoppenbrouwers (1988) use Pearson correlations to compare the feature vectors. This is actually very similar to cosine similarity. The only real difference is that Pearson correlations are invariant to adding a constant (for example: Laplace smoothing) to the frequencies because the means are subtracted (centered values). Cosine similarity, on the other hand, would lead to slightly different results here (cf. Moisl 2015, 99). I settled on cosine similarity since it is more used in text mining and information retrieval (cf. Manning and Schütze 1999, 299).

	46889_SY	46750_FA	46891_WI	46701_BÖ	46707_BÖ	46766_KA	46578_NG	46640_SG	46759_HA
46889_SY	0								
46750_FA	0.44	0							
46891_WI	0.56	0.48	0						
46701_BÖ	0.74	0.51	0.43	0					
46707_BÖ	0.76	0.64	0.5	0.09	0				
46766_KA	0.56	0.74	0.45	0.43	0.36	0			
46578_NG	0.64	0.73	0.38	0.35	0.43	0.52	0		
46640_SG	0.68	0.72	0.51	0.33	0.46	0.41	0.28	0	
46759_HA	0.56	0.64	0.45	0.43	0.55	0.36	0.52	0.32	0

Table 6: Distance matrix

In table 6 we notice that the two questionnaires from the Bökingharde (46701_BÖ, 46707_BÖ) have a cosine distance of only 0.09, which indicates a close relationship. When inspecting the sentences in table 6, we see that they only differ in terms of the vowel in the 1st person plural of the personal pronoun: *we* vs. *wi*. Table 7 lists all the features in both documents and their distribution (I use raw frequencies here for simplicity).

	_wi	wi_	_sa	san	an_	_tr	tra	rat	at_	_en	en_	_he	hew	ewe	we_	_to	tor	ors	_we	rst	st_
46701_BÖ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
46707_BÖ	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1

Table 7: Vectors (calculation of cosine)

The cosine similarity is calculated as follows:

(2) $\cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\,|\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$ (Manning and Schütze 1999, 300)

In cosine similarity, the angle between two vectors ($\vec{x}$, $\vec{y}$) is measured. These vectors are the two rows in table 7 containing term frequencies for each document. This involves calculating the so-called dot product between the two vectors (3) and then for each of them individually (4) and (5):

(3)
$$(1*0)+(1*0)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*2)+(1*1)+(1*1)+(1*1)+(0*1)+(1*1)+(1*1)=19$$

(4)
$$(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(0*0)+(1*1)+(1*1)=20$$

(5)
$$(0*0)+(0*0)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(2*2)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)=22$$

Now, we can calculate the cosine between the two documents by dividing the dot product of $(\vec{x},\ \vec{y})$ by the product of the squared dot products of each individual vector $\vec{x}$ and $\vec{y}$:

(6) $\frac{19}{\sqrt{20}\sqrt{21}} = 0.9057895$

The cosine distance is now simply:

(7) $1 - 0.9057895 = 0.0942105$

It does not sound unreasonable that these two locations are close considering the fact that they belong to the same dialect group. The furthest distance, on the other hand, is between one of the two questionnaires from the Böking-harde (46707_BÖ) and the questionnaire from Sylt (46889_SY): Here, the distance is 0.76. Bear in mind that this chosen example relies on one very short sentence only. For this paper, I computed a distance matrix for all questionnaires (= 55) with vectors of length 984 (= all trigrams considered).

4 Results

In this main section, I will look for structure in the data by employing methods of multidimensional scaling, cluster analysis and association measures. The normalized text corpus, the document-feature matrices and the distance matrices described in section 3 form the basis of this quantitative and qualitative exploration. Before we do that a short report on corpus statistics is in order.

4.1 Corpus summary

The total size of the corpus is 25,692 word tokens and 1,262 word types, i.e. unique word forms (after normalization has been applied, see 3.1) in 55

documents, the average per document is 468 word tokens (the original German questionnaire has 474 tokens), where tokens are simply delimited by whitespace. Although this may sound like a very small amount of text, it is larger than many corpora used in comparable studies. For instance, Heeringa (2004) achieved significant results for Norwegian dialects only using the fable "The North Wind and the Sun", where he looked at 58 word types (cf. Heeringa 2004, 199).

When breaking up the word tokens into trigrams, we get an increase in token and type size due to the fact that word boundaries are counted as well: Without any frequency cuts, the corpus then consists of 91,880 trigram tokens and 2,952 types. After the frequency cut (minimum document frequency of 2, and minimum token frequency of 15 on corpus level) described in section 3.2, we end up with 84,283 trigram tokens and 984 trigram types, with an average per document of 1,532 trigram tokens and 525 trigram types. In total, then, we are analyzing 55 x 984 = 54,120 data points (feature frequencies).

point pairings	1485
mean	0.30
standard deviation	0.08
minimum	0.01
median	0.32
maximum	0.53
skewness	-.1
kurtosis	-.24

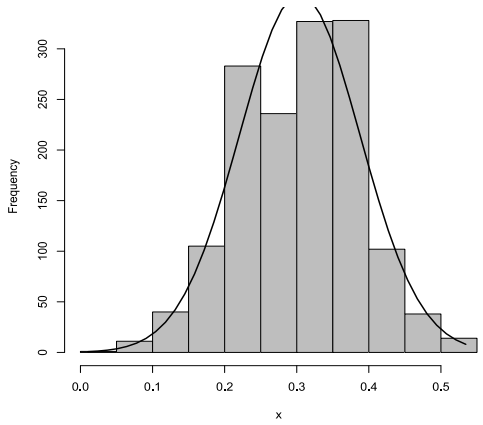


Figure 3: Histogram with aggregate cosine distances

Since we will be working directly on the distance matrices, some key statistics on them is in order (shown in table 3 below) as well. The number of point pairings, i.e. comparisons between two questionnaires, is 55*54/2 = 1485. The mean distance is 0.30, with a standard deviation of 0.08. The minimum distance between two data points is 0.01: This happens to be between the two questionnaires from Amrum (46572 Nebel and 46747 Norddorf), which indicates that they are nearly identical. The furthest distance is 0.53: This is between Amrum (46572 Nebel) and one questionnaire from the Wiedingharde (46700 Emmelsbüll). All in all, the data seem fit for a quantitative

analysis in that the distances are fairly normally distributed. The skewness is –.1, which means that there are a few more pairing points showing bigger distances than smaller. But all in all, skewness and kurtosis values within ±1.0 are accepted within a fairly normal distribution (cf. Szmrecsanyi 2013, 72).

4.2 Dialect continua and dialect areas

4.2.1 Multidimensional scaling and Neighbor-Net

We will start by exploring the distances between the questionnaires using two methods: Multidimensional scaling and Neighbor-Net. Both of these are good for a first visual exploration of the distance data. They can also help us answer the question whether the dialects are more continuum- or area-like. Based on the literature on North Frisian, one would expect dialect areas with sharper borders rather than a smooth dialect continuum.

Multidimensional scaling (MDS) tries to visualize the distances between data points in low-dimensional space, i. e. mostly in two or three dimensions (cf. Levshina 2015, 336). Building upon this method, Heeringa (2004) suggested taking the first three dimensions of a MDS solution and project these to a two-dimensional map where each dimension represents its own base color (red, green, blue). The method has since been widely adopted within linguistic geography, since it makes it possible to visualize three dimensions (ideally explaining more than 90% of the data) of variation on a traditional map. Instead of using Voronoi polygons, however, I chose to use simple points for reasons of exactness (i.e. the individual nature of each questionnaire) and due to the special geography of North Frisia.

The MDS/Heeringa map[12] is pictured in 4 (plotted on the dialect map of Århammar 1968) and shows a clear division between Insular and Mainland North Frisian with distinct areas within Insular North Frisian and certain, albeit less clear areas in Mainland North Frisian. In Insular North Frisian, we clearly recognize Syltring and Föhring-Amring. Within Mainland North Frisian, one could make the case for a threefold division: Wiedingharde and Bökingharde show up as distinct areas, whereas the situation in southern Mainland North Frisian (Karrharde, Nordergoesharde, Mittelgoesharde, Südergoesharde and the Halligen) is less clear in that we do not immediately recognize any clear tendency other than that the area seems to be different from the Wiedingharde and Bökingharde.

12. In the print version of this article, the map is drawn in grayscale, reducing its usefulness. Please consult the digital copy for the color version.

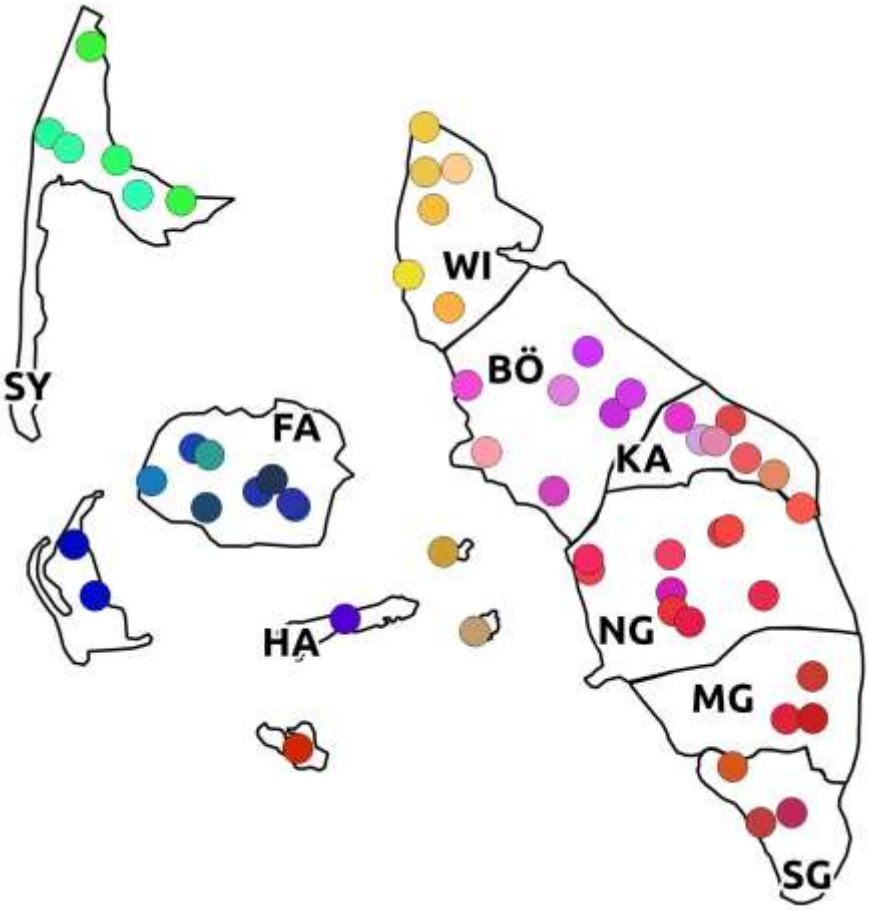


Figure 4: MDS plot (Heeringa style; type: Kruskal non-metric, r2 = 0.918)

The Karrharde and the Halligen appear as transitional zones: In the case of the Halligen, Langeness shows clear connections to Föhring-Amring, whereas Hooge, Gröde and Oland conform with (southern) Mainland North Frisian, as one would expect. The transitional status of the Karrharde is possibly due to quality issues with multiple Wenker questionnaires from this area (see section 5.2).

How reliable is this MDS solution, which is based on Kruskal's non-metric MDS? I computed the so-called R^2 score (cf. Heeringa 2004, 160), to measure the amount of variation in the data that the three-dimensional MDS solution can account for, i. e. the squared Pearson's correlation between the original distance matrix and the Euclidean distance between the points of the MDS solution. A value above 0.6 (or 60%) is the smallest possible according to

consensus, but too high values are also problematic (cf. Spruit, Heeringa, and Nerbonne 2009, 1632). My data led to a R^2 value of 0.918, which means that the solution attributes for ca. 92% of the variance in the distance matrix. This is very good considering the fact that we work with rather heterogeneous data. I additionally tried two other MDS methods: Classical (Metric) MDS and Sammon's Non-Linear Mapping. Both methods produced a similar structure as above with good, but lower R^2 values (Classical MDS = 0.854, Sammon's Non-Linear Mapping = 0.805). Heeringa (2004, 161) also found that Kruskal performed better with his data, without being able to explain this.

Another way of visualizing distances which has gained popularity in recent years is Neighbor-Net (cf. Cysouw 2007), an algorithm for computing phylogenetic networks originally developed within bioinformatics (cf. Bryant and Moulton 2004). Neighbor-Net, as implemented in the SplitsTree4 package (Huson and Bryant 2006), produces an unrooted tree with paths and splits at various points. Distances between nodes (i.e. documents) are represented as paths here that are split at certain points, and one gets the distance between two points by following the paths and adding the lengths. In this way, we can visualize the distances between many elements using only two dimensions. The unrooted tree produced on the basis of the cosine distance matrix of the trigram Wenker corpus is shown in 5:

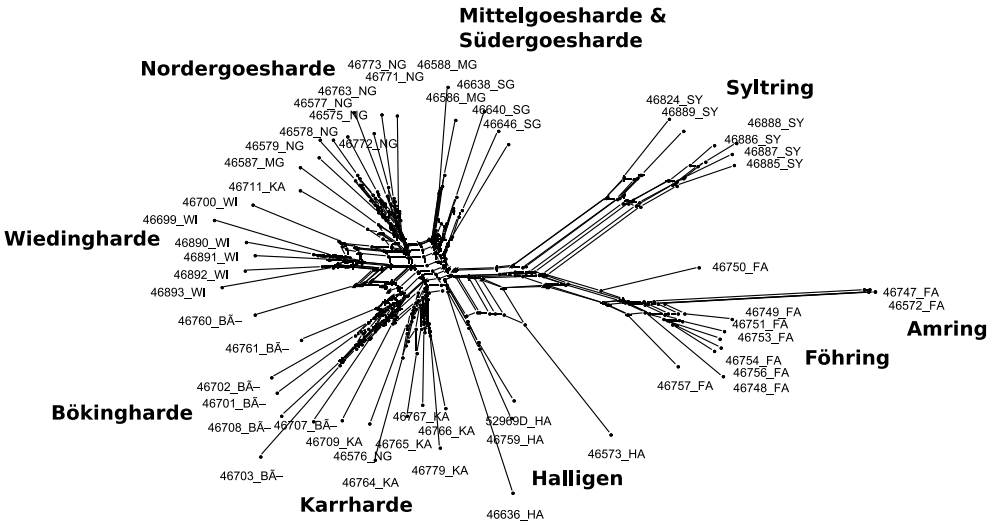


Figure 5: Neighbor-Net

The Neighbor-Net tree shows a clear division between Insular and Mainland North Frisian, similar to the MDS map in 4, where the Halligen occupy an

intermediate position between Syltring/Föhring-Amring and Mainland North Frisian proper. Note that the Halligen, especially 46636 Hooge and 46573 Langeness, have very long branches indicating rather idiosyncratic questionnaires. Within Insular Frisian, there is a clear-cut distinction between Syltring and Föhring-Amring, where Amring is also separate from Föhring. The other divisions within Syltring and Föhring do not seem to follow any areal pattern. Especially, we find no indication for a clear distinction between the three main dialects of Föhr (i. e. Westerlandföhring, Osterlandföhring and Südföhring according to Walker and Wilts 2001, 284) in the material.

On the mainland, while certain outliers can be found, we see a fairly clear picture, which confirms the traditional classification to a large extent, with the exception that we do not see a clear separation of Oster- and Westermooring, which are sometimes postulated as sub-groups within the Bökingharde (cf. Walker and Wilts 2001, 284), from the rest of the Bökingharde. Südergoesharde and Mittelgoesharde form one unit, and they appear to be closest to the Halligen, which – as far as Südergoesharde is concerned – is perfectly in line with the traditional assumption (cf. Löfstedt 1928, VIII). Then we see clusters for the Nordergoesharde, Wiedingharde, Bökingharde and the Karrharde. The outliers are 46587 Almdorf and 46711 Klintum (which appear to be closer to each other), 46709 Wester Schnatebüll (cf. 1 and Hofmann 1956, 87) and 46576 Büttjebüll. I will return to these in section 5.2.

4.2.2 Hierarchical clustering

Whereas MDS and Neighbor-Net are useful for a general overview over the data and for investigating dialect continua, they are not so good when it comes to separating clear groups. For this, we will resort to two different algorithms of hierarchical clustering: Ward and UPGMA (Unweighted Pair Group Method using Arithmetic Averages) as offered by the R package hclust. These are agglomerative or "bottom up" approaches where each document starts out as an own cluster and is merged subsequently into larger and larger clusters until we reach the top of the tree. There are various ways of doing this, and the models are not very easy to interpret. Whereas Ward produces compact clusters of similar size, UPGMA tries to reduce the distance to a computed average (this method is more sensitive to outliers than Ward). Ward is one of the most popular clustering algorithms within corpus linguistics and dialectometry (cf. Szmrecsanyi 2011, 61), but not without problems since it is biased towards clusters of equal size even if these are not found clearly in the data by other methods. Therefore UPGMA will be used as well as a

corrective. For a good overview of various hierarchical clustering approaches, see Heeringa (2004, 146-156) and Levshina (2015, 309-311). These agglomerative clustering methods produce hierarchical so-called dendrograms or trees showing the relationships between the clusters and the distances between them.

The major problem with all hierarchical clustering models is to find the "best cut": since, in the end, each document may be regarded as one cluster, finding the "right" number is not trivial. There is no consensus regarding what is the best method to use here. To quote one recent textbook on the subject: "There have been attempts to formalize selection of a best cut [...], but the results have been mixed, and the current position is that the best cut is the one that makes most sense to experts in the subject from which the data comes." (cf. Moisl 2015, 215). In the following, I will draw maps with 2, 6 and 9 clusters for each of the two methods. We expect two major groups (Insular and Mainland North Frisian) with a total of nine dialects in our area of investigation (since Helgoland is not part of this study). The results are shown in figures 6-11. The comparison reveals certain differences between the two algorithms although the big picture is relatively similar in both. In the 2-cluster solution, the algorithms show complete agreement: We find Insular and Mainland North Frisian as one would expect. Within these two groups, however, we see divergence, as revealed by the 6-cluster solution. Ward shows a picture which

Figure 6: 2 clusters (Ward)

Figure 7: 2 clusters (UPGMA)

Figure 8: 6 clusters (Ward)

Figure 9: 6 clusters (UPGMA)

Figure 10: 9 clusters (Ward)

Figure 11: 9 clusters (UPGMA)

is fairly close to that given by the MDS (cf. figure 4), whereas UPGMA suggests an own cluster for the Hallig Hooge and for the Bökingharde vs. the rest of the mainland. Remember that Ward favors even-sized clusters, whereas UPGMA looks for clusters that are furthest away from a computed average. Thus, the Bökingharde seems to be the most untypical of the Mainland North Frisian dialects. When looking at the Neighbor-Net

visualization of the distances (cf. figure 5), Hooge is very far away from the mainland (this also makes sense geographically), but as noted earlier, there are certain problems with this questionnaire.

It is further interesting to note that 46760 Dagebüll clusters with the Wiedingharde in both clustering algorithms (the proximity to the Wiedingharde can be seen in the Neighbor-Net model in figure 5 as well). There seems to be a certain linguistic justification for this as we will see in the discussion in section 5.1. In the Ward 9-cluster solution, a picture emerges that confirms the traditional classification to a fair degree, but with certain differences in that the distance between Föhring and Amring appears to be greater than that between Mittelgoesharde and Südergoesharde. We also recognize some of the outliers shown by the Neighbor-Net model. In the UPGMA 9 cluster solution, we see no clear differences at all in the south of Mainland North Frisian. This is also supported by the MDS analysis in figure 4.

Finally, I will test for the validity of the cluster solutions using the cophenetic correlation coefficient (cf. Heeringa 2004, 150-151): this is a test for correlations between the distances in the clustering solution and the distances in the original distance matrix. The question is to what degree the clustering solutions represent the distances found in the original distance matrix. When comparing the results of the two algorithms, we see that UPGMA clustering has an edge over Ward in this corpus: the UPGMA solution explains 78%, whereas Ward accounts for 71% of the variation. The UPGMA solution, however, seems to be sensitive to certain outliers in the material (to be discussed in section 5.2), as can be seen in figure 11.

One problem with hierarchical clustering is instability, i. e. small changes in the data matrix can lead to big changes in clustering. There exist various solutions to validate cluster solutions, e. g. bootstrapping or noisy clustering (cf. Nerbonne et al. 2008). I applied noisy clustering using Peter Kleiweg's RuG/L04 package[13] to the difference matrix, using the optimal UPGMA solution. The result is that the divisions between Insular North Frisian and Mainland North Frisian and within Insular between Syltring and Föhring-Amring are absolutely stable (they appear in 100 % of the runs). Within Mainland North Frisian, the clusters are less stable, but still appear in the majority of the cases. Wiedingharde was found in 86 % of the runs, Bökingharde in 69 %, Karrharde + Nordergoesharde + Mittelgoesharde +

13. RuG/L04 can be downloaded at Peter Kleiweg's personal homepage: http://www.let.rug.nl/~kleiweg/L04/. I used the following parameters. Number of runs: 1000, noise = 0.5 (= standard deviation * 0.5). The noise value is the default, but I increased the number of runs from 100 to 1000.

Südergoesharde (without outliers) in 90 % and the Halligen (without Hooge) in 71 %. On the other hand, the outlier clusters Hooge and the the outlier within Karrharde and Niedergoesharde were found in 100 % and 93 % of the runs. These outliers, then, seem to be real, but they appear to be more of a data problem than a method problem (see section 5.2).

As a result of this, I chose to rely on the six-cluster Ward solution in the following, which shows a very similar picture to the MDS solution in figure 4 when leaving out the Halligen and the nine-cluster UPGMA solution in figure 11 when leaving out the outlier clusters. Thus, in the following, the Halligen are treated as a group for itself although it appears to be rather heterogeneous. Accordingly, the results for the Halligen (cf. section 4.3.4) should probably be treated with a certain degree of suspicion.

4.3 Prominent dialect features

So far, the article has been concerned with rather abstract areas, without discussing the features responsible for the various clusterings. As stated in the introduction, the literature on North Frisian dialectology is not rich on criteria for distinguishing the various dialects. In the following, I will use the Wenker trigram corpus and extract prominent features from the six most important dialect groups found in the previous section by using an association measure, the so-called log-likelihood ratio test.

4.3.1 Log-likelihood ratio test

The log-likelihood ratio test is a statistical test that compares the frequency of a word in a target group (i.e. the dialect to investigate) with its frequency in a reference group (i.e. all other dialects). In our case, if a trigram is particularly frequent within the target group (considering the total frequency of all trigrams in both the target group and the reference group) this will lead to a high log-likelihood ratio. Log-likelihood is calculated by using a so-called contingency table containing frequency information. In the following example, we will look at the distribution of the trigram <jem> in Mainland North Frisian and Insular North Frisian: We would expect <jem> to be more common in Mainland North Frisian due to the personal pronoun *jem* ‘you’ (2pl), where Insular North Frisian has *jam* or *i*.[14]

For the calculation we need four numbers: the frequency of <jem> in the target corpus (Mainland North Frisian) and the reference corpus (Insular

14. Of course, *jam* is also found in Mainland North Frisian in the areas showing lowering of OFr *i* > *a* (first and foremost: Bökingharde). Insular Frisian does not have *jem*, however.

North Frisian) and the total frequency of all trigrams in both corpora respectively. The contingency table is shown in table 8 below:

	target (Mainland)	reference (Insular)	total
frequency of trigram *jem*	185	5	190
frequency of other trigrams	60,528	23,565	84,093
total	60,713	23,570	84,283

Table 8: 2x2 contingency table with observed frequencies

Not surprisingly, we notice a clear asymmetry between Insular and Mainland North Frisian here: <jem> is only attested five times in Insular North Frisian (and the trigram does not refer to the personal pronoun but to participles like *kjemmen* 'come' in Syltring). Is this difference also statistically significant? We notice that the reference corpus is considerably smaller than the target corpus. In the following, I rely on Rayson and Garside (2000) for the calculation of the likelihood ratio.[15]

First we compute the expected values E1 (target) and E2 (reference) based on the frequency distribution of the observed values in table 8 using the formula in (8), where *n* represents the total size of the two corpora, S_i the total of row *i* and S_j the total of column *j*:

(8) $E_{ij} = \frac{S_i S_j}{n}$ (Levshina 2015, 211)

This leads us to the two expected values E1 and E2 in (9) and (10):

(9) E1: $\frac{190 * 60713}{84283} = 136.87$

(10) E2: $\frac{190 * 23570}{84283} = 53.13$

We notice that there is a discrepancy in the observed and expected values, *jem* is more common than expected in Mainland North Frisian given the sub-corpus size, and far less common than expected in Insular North Frisian. Now we compute the log-likelihood, using the formula in (11), to see if this value is significant:

15. Note that the R package `quanteda` (Benoit 2018) that I used to calculate the G^2 values in the following section, relies on a marginally different algorithm by Dunning (1993). The paper by Rayson and Garside (2000) is however much easier to follow for the average linguist reader.

(11) $-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$ (Rayson and Garside 2000, 3)

When applied to the numbers above, this results in the following G2 value:

(12) $G^2 = 2 * ((185 * ln(185/136.87)) + (5 * ln(5/53.13))) = 87.86$

The higher the G^2 value, the more significant is the difference between observed and expected frequencies. For 2x2 contingency tables we use one degree of freedom (two observations − 1): a value of >= 3.84 is $p < 0.05$ and >= 15.13 is $p < 0.0001$.[16] The difference is thus highly significant.

I applied this association measure to all trigrams in the various groupings in the following sections and sorted the frequency lists by the G^2 value. For reasons of space, I will limit the discussion to the top 25 trigrams for the comparison of the two main groups and within Insular North Frisian, and the top 10 trigrams in Mainland North Frisian, where we find less significant differences, all of which are at least statistically significant at the $p < 0.05$ level. This does of course not give a complete picture, but still a fairly good impression of what can be done with this method. I will start by looking at prominent features of Insular North Frisian and Mainland North Frisian. This is traditionally assumed to be the primary division in North Frisian, and is found in all methods used so far. Subseqeuntly, I will investigate the differences within Insular North Frisian and Mainland North Frisian.

4.3.2 Insular North Frisian vs. Mainland North Frisian

In the literature on North Frisian, mainly two criteria are used in order to distinguish Insular North Frisian (INF) from Mainland North Frisian (MNF) (cf. Hofmann 1956, 81-86; Sjölin 1969, 41; Markey 1981, 212). The first criterion is a phonological one, i. e. the different reflexes of *i*-umlaut of Germanic *ū* and *ō* in the two varieties: In the mainland dialects, the umlaut products coalesced with the reflexes of germ. *e*[1], in the insular dialects, however, they did not. The second criterion is morphophonological in nature: Insular dialects display plurals in *-er* and *-en*, mainland dialects in *-e*. This is related to apocope: Whereas INF historically shows apocope in reflexes of OFr *a* and *e*, MNF retains the schwa reflexes of OFr *a* and only has apocope of OFr *e* (cf. Siebs 1901, 1244, Versloot 2001, 769-770 and Århammar 2001, 756). The idea is that INF dialects "restored" the plural endings lost in

16. Significance values taken from: http://ucrel.lancs.ac.uk/llwizard.html (last accessed: 14.04.2019)

apocope through morphological innovation (perhaps under a certain influence of Danish, although the endings were independently available in North Frisian). In general, then, we would expect MNF varieties to show more schwa endings than INF dialects.

Figures 12 and 13 show the 25 most prominent trigrams for INF and MNF as given by the log-likelihood test:

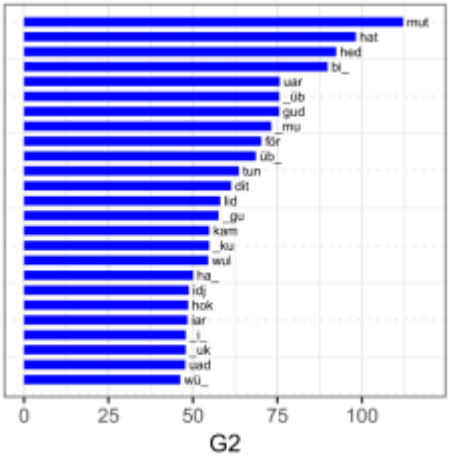


Figure 12: INF top 25 trigrams
(G2: max = 112.23, min = 46.27)

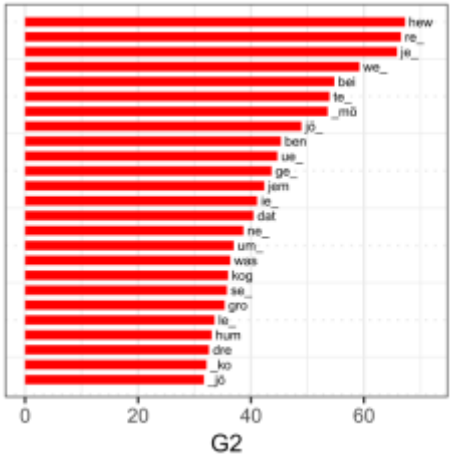


Figure 13: MNF top 25 trigrams
(G2: max = 67.28, min = 31.66)

The pervasiveness of schwa endings in MNF compared to INF is truly reflected here: in MNF, 10 of the trigrams display a <e> at the end of a word (e.g. *-re*, *-ne*, *-ge*, *-le*), in INF there is none. These are mostly verbs, adjectives or nouns whose insular counterparts take *Ø* or other endings, e.g. INF *beg* (46887 Keitum) vs. MNF *begge* (46893 Rodenäs) ‘build (inf.)’ (see also WA 471 ‘bauen’), INF *de brünn Hünj* (46753 Alkersum) vs. MNF *di brünne Hün* (46700 Emmelsbüll) ‘the brown dog’ (see also WA 531 ‘braune’) or INF *Üüffens Berger* (46749 Oldsum) vs. MNF *Ües Beerge* (46892 Klanxbüll) ‘our mountains’ (cf. WA 406 ‘Berge’). It is hardly surprising that this phenomenon gets a high score, given the fact that contexts with historical schwa endings are common in the Wenker sentences. It is worth noting that only the retained schwa endings show up as prominent, not the various innovations in the insular dialects. These are obviously more heterogeneous in nature, showing either the bare stem or some morphological replacement. The role of the

apocope, then, seems to be very important and is truly reflected by the material.

The first standard criterion, i.e. the reflexes of *i*-Umlaut, on the other hand, is not found among the top trigrams. Unfortunately, the Wenker materials do not provide all the relevant stimuli here: We do find examples for germ. *ō+i* in *Füße* in WS 8 (OFr *fēt,* cf. WA 107 'Füße'), and germ e^1 in *schlafen* in WS 24 (OFr *slēpa*, cf. WA 354 'schlafen'), we do not, however, find any adequate examples for germ. *ū+i*. The only candidate – *Kühe* in WS 37 (OFr *kī*, cf. WA 498 'Kühe') – is problematic, since /i/ is believed to be the historical form both for Insular and Mainland North Frisian here (cf. Siebs 1901, 1228).

What we do find, however, are what appears to be systematic differences in the reflexes of OFr *ō*: INF shows mostly *ū*/*u* (as illustrated by the many trigrams with <u>), whereas the situation in MNF is more complex, having dialects showing either old *ō/o* or various diphthongs (cf. Siebs 1901, 1222, Hofmann 1956, 86 and Århammar 2001, 751). In table 9 below, the realization of OFr *ō* in the lexemes OFr *gōd* 'good' and the 3rd person singular of OFr *mōta* 'must' is listed as they appear in the questionnaires (see also WA 243 'gut' and WA 325 'muss'):

	'good' (WS 17)	'must' (WS 22)
SY	6x <u>	5x <u>
FA	9x <u>	9x <u/ú>
WI	1x <ö> – 5x <oi>/<öi>/<eu>	4x <oi>/<ai>/<eu>
BÖ	2x <ő>/<ö> – 5x /<au>/<äü>/<äu>/<öi>/<eu>	1x <ö> – 6x <âü>/<äu>/<öi>/<au>/<ő̌j>/<ǒj>
KA	3x <ö> – 3x <äu>/<oe>/<eu>	3x <ö> – 2x <oe>/<öi>
NG	9x <öu>/<au>/<ou>/<öau>	9x <ou>/<öu>/<au>/<aou>/<öu>
MG	2x <o> – 1x <ou>	2x <o> – 1x <ou>
SG	3x <ö(ö)>	2x <o(o)>, 1x <ö>
HA	4x <ö> – 1x <oe>	4x <ö> – 1x <öe>

Table 9: Graphs representing OFr *ō* (as in questionnaire, not normalized)

Basically, this table confirms the description found in the literature (cf. the map in Hofmann 1956, 95) in that we see an opposition between INF <u> and MNF <o>/<ö> and diphthongs. We find a similar situation in certain lexemes reflecting OFr *ō* before *-rn*/*-rd*: OFr *korn* 'grain' and OFr **korf* 'basket' (see WA 553 'Korn' and WA 289 'Korb'). Here again, <u> is typical for INF

(with the exception of Sylt, cf. Århammar 2001, 751), <o> or diphthongs for MNF. We notice the trigram <_gu> in INF, with a fairly high rank: it reflects OFr *gunga/gonga* 'go' (germ. *a*). The *u* in this strong verb appears to be a special development in Old Frisian (cf. Siebs 1901, 1182, who suggests analogy): Generally, in the Wenker data INF has <u>, whereas in MNF <o> is prevalent (see also WA 239 'geh'). In the transmission of Old Frisian, <u> is typical for Eastern, <o> for Western Frisian (cf. Hofmann and Popkema 2008, 196-197).

Another peculiarity of Insular Frisian as reflected by figure 12 above involves the diphthongs /ua/ and /ia/, found in the trigrams <uar>, <uad> and <iar>: This is also recorded in previous literature: "Wie das *ia* so ist auch das *ua* eine gemeinschaftliche Eigentümlichkeit sämtlicher Inselmaa." (Selmer 1921, 22, see also the overview in Walker and Wilts 2001, 288-289). /ua/ is a reflex of OFr *ā* (cf. Århammar 2001, 753), /ia/ is a reflex of OFr *ē* (cf. Århammar 2001, 752). According to Siebs (1901, 1161) and Århammar (2001, 753), diphthongization of OFr *ā* (= germ. *au*, *ai*, *a* before certain consonant clusters) historically lead to /ua/ in both groups. Later, monophthongization took place, but in Insular Frisian (and in the Süder-goesharde + Halligen) only before labials and velars. Thus we find e. g. *uaren* 'ears' (cf. WA 159 'Ohren') and *duad* 'dead' (cf. WA 192 'tot'). Equally, diphthongization of OFr *ē* to *ia* is also historically expected in both groups according to Siebs (1901, 1162), but here too, it is more pervasive in INF than in MNF due to later monophthongization in the latter group. However, monopthongization is found in Syltring as well (cf. Århammar 2001, 752). In INF, the diphthong is typical for Föhring-Amring, but in Syltring only before /r/ and /l/ (cf. Siebs 1901, 1404) and in certain words with "Akzentwechsel" (cf. Århammar 2001, 752). In INF, we find more /ia/ due to the merger of OFr *ē* with Ger. e^2 before diphthongization: Thus <iar> is a prominent trigram for INF, reflecting *wiar* 'was/were' and *diar* 'there' (Sylt/ Föhr), where MNF has monophthongs (see WA 78 'war' and WA 544 'da').

Also, diphthongization of OFr *ī* > *ai* word-finally and in hiatus (cf. Siebs 1901, 1220, Hofmann 1956, 87 and Århammar 2001, 749-750), reflected by the trigram <bei> in MNF, appears as typical for MNF (see <bi_> in INF). Generally, we would expect diphthongization in Föhring-Amring also; however in the frequent preposition *bi* 'by' the monophthong is retained on Föhr as well (see also WA 362 'bei' and Århammar 2001, 749-750), whereas MNF generally has *bei*. This preposition appears twice in the Wenker materials, in WS 9 and 25.

Within morphology, one important difference is found in the verb morphology of the verb 'have' (cf. Nübling 2000, 35-36): The most prominent trigram in MNF is <hew> reflecting *hewwe* (OFr *hewwe/hebbe*) 'have' (see also WA 338 'haben'), where INF shows the short form (without a labial consonant) and a different stem vowel: *ha(a)* (OFr *habbe*/*hawwe*). Both stems are found in Old Frisian, and according to Hofmann (1956, 84) the *a*-stems and *e*-stems are one of the oldest differences between INF and MNF. Another prominent morphological feature of MNF according to figure 13 is the neutral singular article and demonstrative pronoun *dat* (where Low German influence seems plausible, cf. Fleischer 2012, 70 on the relationship between *et* and *dat* in Northern Low German), which in this form is only found in MNF (INF has *det* (FA) or *dit* (SY)). In MNF, *dat* is also very often the expletive pronoun, whereas in Insular Frisian this is mostly *hat* (which again is found in figure 12 for INF).[17] Other important morphological divisions are found among the personal pronouns, which are probably over-represented in the Wenker materials (cf. Lameli 2010, 25): *jam* vs. *jem* 'you' or *i* in Syltring, or *jü* vs. *jö* 'she', where the latter forms are confined to MNF (see also WA 397 'ihr', WA 250 'sie'). In verb morphology, we find MNF *ben/ban* 'I am' (here reflected by the trigram <ben>), where INF has *san*/*sen*, which here is syncretic between 1st person singular and the whole plural.

4.3.3 Dialects of Insular North Frisian

As shown by the cluster analysis (figures 6-11) and the Neighbor-Net tree in figure 5, Syltring and Föhring-Amring belong to the same branch, however, they are still quite different from each other. At the same time, Föhring-Amring is somewhat closer to the mainland dialects than is Syltring. Figures 14 and 15 show the most prominent trigrams of Syltring and Föhring-Amring respectively. Remember that unfortunately Helgoland Frisian is not a part of this study due to the idiosyncratic transcription of the only questionnaire from this island.

17. An anonymous reviewer points out to me that while in the 19th century, *hat* can be used as an expletive pronoun, it already then was developing into a feminine singular pronoun. This is reflected by the Wenker data in the case of Oldsum on Föhr in that we find variation between *jü* (WS 9, referring to *Wüf* 'woman', hence showing semantic agreement) and *hat* (WS 17, referring to *Saster* 'sister'). Age seems to play a role in the retention of the old feminine form *jü* here: Ebert (1998, 269) notes that *jü* is only used "in respektvoller Rede für alte Frauen".

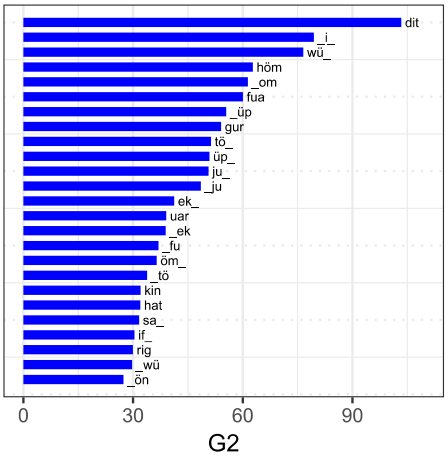


Figure 14: SY top 25 trigrams
(G2: max = 103.37, min = 27.39)

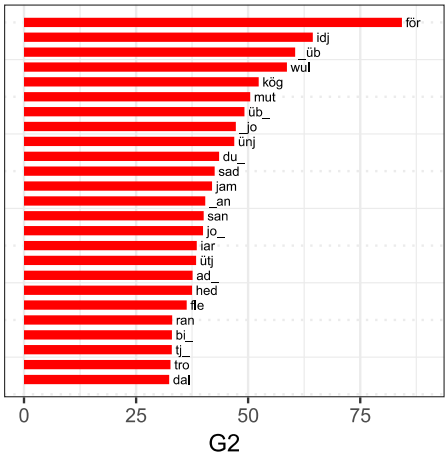


Figure 15: FA top 25 trigrams
(G2: max = 84.3, min = 32.4)

Phonologically, vowel lowering phenomena reflected by the verb and possessive pronoun *san* (masculine) show up as typical for Föhr/Amrum, whereas Sylt has *sin/sen*, which is also common in MNF, which is why the latter form does not show up among the prominent trigrams. Here, we find extreme lowering of OFr *i* through *e* to *a* (cf. Walker and Wilts 2001, 286), which is typical for Föhr/Amrum and Central MNF. In terms of the diphthongs /ia/ and /ua/ explored in section 4.3.2, we find that while being prominent in both dialects, /ua/ seems to be more common in Syltring, /ia/ more common in Föhring-Amring. In the case of /ua/, this seems to be more related to developments in single lexemes: /ua/ appears as particularly prominent in Syltring due to the form *fuar* ‘for’ (reflecting OFr *o*, probably with vocalized /r/), where Föhring-Amring has *för*. In Föhring-Amring, <ia> is more common than in Sylt Frisian (this is to be expected, see above), e.g. *siar* ‘weh’ (SY: *siir*), *iarst* ‘erst’ (SY: *jest*/*jen*), *iar* ‘als/before’ (found on both Föhr and Amrum according to Braren and Wilts 1986, 123, in the Wenker data only on Amrum, whereas Sylt and Föhr have *üs*).

A characteristic feature of Föhr is the palatalization of certain consonants, reflected by the trigrams <idj>, <ütj> and <tj_>, <ünj> in words like *letj* ‘little’ and *betj* ‘a little bit’ (see also WA 434 ‘bisschen’). According to Hofmann (1961, 8), the palatal consonant in *letj* can be traced back to a *k*-diminutive suffix **litik* with subsequent vowel loss, also found in words like *ētj* ‘Essig’ (**etik*/*edik*) or *pretji* ‘preach’ (**predikia*); the form *betj* can be

explained by the presence of a diminutive suffix *-ka* (**bit-i-ka*) (cf. Hofmann 1961, 78). In the latter case at least, Sylt has non-palatalized *bet*. But 'little' is not a Wenker lemma in this case: rather, it is used instead of the diminutive form of the original High German text, e.g. *Vögelchen, Mäuerchen*: Thus, in WS 36 – *Was sitzen da für Vögelchen oben auf dem Mäuerchen?* 'What are those (little) birds sitting up there on the (little) wall?' –, we find *Wat ſat dīär för letj Fögler bāwen ūb das letj Mūr?* (46748 Utersum), while on Sylt, we mostly find the diminutive ending *-ken* in this case, which is more common on Sylt than on Föhr/Amrum and a possible loan from Low German (cf. Hofmann 1961, 44).

In the morphology, we find mostly Syltring characteristics: The personal pronoun of the 1st person plural in Syltring is *wü* 'we', whereas Föhr/ Amrum together with MNF have *wi*. In the 2nd person plural, Sylt has *i* 'you' – the unique oblique form *juu* is found in figure 14 as well, whereas Föhring-Amring has *jam*. In the 3rd person singular, the Syltring masculine/ neuter oblique form *höm* 'him/it' turns up. In the article system, Sylt is radically different from Föhring-Amring and the MNF dialects in that the *a*/*e*-article completely lacks here (cf. Wilts 1995, 15). Further, we notice the most prominent trigram <dit> representing the neuter demonstrative pronoun and article *dit* where other dialects have *dat* or *a*/*e*-article. We also notice the negation particle *ek*, which is typical for Syltring, where the other dialects have *ei/ai* (cf. the discussion in Hofmann 1956, 95, where both forms are considered to be a loan of Danish *ikke*). In Syltring, we also find a lexical feature: The trigram *_ju* represents *jungen* 'child', where the Föhring-Amring questionnaires have *kind* or *bjarn*/*bjern*.

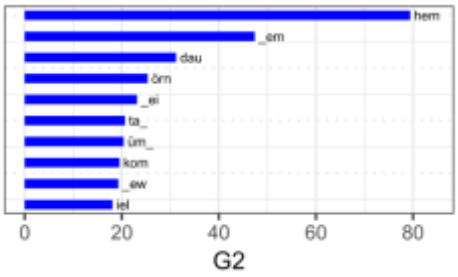
4.3.4 Dialects of Mainland North Frisian

The sections 4.2.1 and 4.2.2 showed that Mainland North Frisian seems to consist of four areas: Whereas Wiedingharde and Bökingharde appear as clear entities, the situation in southern Mainland North Frisian (without the Halligen) is less clear. In this section, I will therefore look at Wiedingharde, Bökingharde, southern Mainland Frisian proper and the Halligen and not investigate any further sub-divisions. While relevant features can be extracted from individual varieties within southern Mainland North Frisian as well, these are fewer and far less significant than features from Insular North Frisian and northern Mainland North Frisian.

Wiedingharde and Bökingharde

As shown in the figures 16 and 17, we notice that the amount of very

significant trigrams (= the length of the bars) is larger in the Bökingharde than in the Wiedingharde. Some of the trigrams in the Wiedingharde (e.g. <örn> representing *börn* 'child' and <iel> reflecting *hiil* 'wholly'[18] seem to be bound to single lexemes).



-Figure 16: WI top 10 trigrams
(G2: max = 79.42, min = 18.08)

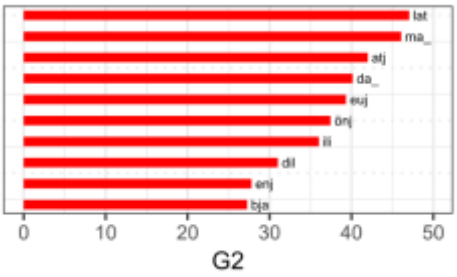


Figure 17: BÖ top 10 trigrams
(G2: max = 47.05, min = 27.28)

The most important difference between the Wiedingharde and the Bökingharde seems to be the lack of lowering of OFr *i* > *a* in the former dialect, where lowering stopped at *e*. This is reflected by the most prominent trigram *hem* (3sg personal pronoun 'him' where southern dialects have *ham*) and the second-most prominent <_em> reflecting the preposition and complementizer *em*, where southern dialects show *am* (cf. WA map 157 'um'). In the Bökingharde, on the other hand, we find extreme lowering in two of the most prominent trigrams <lat> and <ma_> reflecting e. g. the words *latj* 'little' and *ma* 'with'. According to Löfstedt (1933, 8) lowering of OFr *i* > *a* is most prominent in the Bökingharde of all MNF dialects.[19] A third trigram with a low vowel, <da_>, reflects the definite article plural *da* (OFr *thâ*, i e. no lowering here) and is found in all locations in the Bökingharde except Dagebüll, which has *dä/de* similar to Wiedingharde and most other MNF dialects.

18. Interestingly, *gans* is used elsewhere in the Wenker data, probably due to (Low) German influence. Wiedingharde thus seems to be the most resistant region with respect to this lexical item, perhaps with support from Danish *hel(t)*.
19. Regarding the reflexes of OFr *ō* before low vowels (cf. Siebs 1901, 1223) we find <au>, i.e. a falling diphthong like in *dau* 'do', to be prominent in the Wiedingharde (cf. WA 26 'tun'). This form appears in the Bökingharde as well, see for example <kau> representing *kaurv* 'bucket' or *kaul* 'coal' and is thus constitutive for Northern Mainland Frisian. In the Bökingharde, however, other lexemes are used in the relevant WS 3 *Thu Kohlen in den Ofen [...]* 'Put coal in the oven [...]' (*led, fu*), leading to a lower score here. Thus, this trigram may be less constitutive for the Wiedingharde.

A further important difference is found among the consonants: As in Föhring-Amring, graphemes with consonant/vowel + *j* are particularly frequent among the trigrams of the Bökingharde, e. g. <atj>, <önj>, <enj> representing inflected word forms like *latje* 'little', *mönje* 'must', *verstönj* 'understand', *hinje* 'bad' where the <j> stands for a palatalized consonant [c] and [ɲ] (cf. Walker 1990, 11-12). In the MNF dialects, palatalization is found in Wiedingharde, Bökingharde and Karrharde (cf. Århammar 2001, 759), but among these three, it appears to be most frequent in the Bökingharde.

Southern Mainland North Frisian

For southern Mainland North Frisian, a vast amount of the questionnaires stem from the Nordergoesharde (which is also the largest area). Therefore this area probably dominates the results in figure 18:

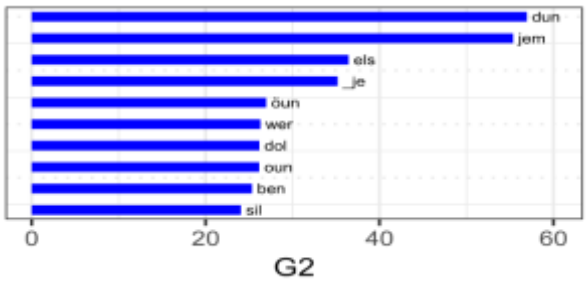


Figure 18: SMNF top 10 trigrams
(G2: max = 56.94, min = 24.05)

The top trigrams in figure 18 are <dun> and <jem>, which seem to capture this area fairly well, as shown by the original maps, reproduced in figure 19 and 20.

When looking at WA map 148 'tun', we see that the form *dun* is typically found in the Nordergoesharde and the Mittelgoesharde. This map actually yields a very similar picture as the six-cluster Ward solution in figure 8. <jem> is also found in the whole area. This appears to be a common feature of Southern Mainland Frisian as also shown by WA map 397 'ihr'.

In phonology, we notice the graphemes representing diphthongs in <oun>/<öun>, reflecting OFr *ō* (cf. Århammar 2001, 751-752): In the Wenker data, it seems like these are tied to Nordergoesharde and found in words like *oun*/*öun*/*aoun* 'in' (elsewhere: *an*/*en*) or *stounen*/*stöunen*/ *staounen*

Figure 19: WA map 148 ‘tun’ (fragment)

Figure 20: WA map 397 ‘ihr’ (fragment)

‘stand’. There is a geographical pattern here: diphthongs with <ö> as the first component are found in the north of Nordergoesharde (46771 West-Langenhorn, 46772 Loheide, 46773 Mönkebüll), diphthongs with <o> in the south (46576 Büttjebüll, 46577 Sterdebüll, 46578 West-Bordelum, 46579 Dörpum). This is confirmed by Brandt (1913, 44-45), according to whom <ö> is close in pronunciation to English *but*, i.e. an open-mid back unrounded vowel [ʌ]. The dipthong *öu* is according to Brandt (1913, 39) unique to the Nordergoesharde. This is indeed supported by the Wenker data: the diphthongic trigrams with <öu> are with one minor exception not found elsewhere. We also notice signs of rhotacism and lenition in the trigrams <wer>, <rer>: i.e. <wer> represents *pewer* ‘Pepper’, *awer* ‘but’, *wer* ‘weather; would’ (cond.). This fits well to the findings of Hofmann (1956, 98-99), according to whom lenition is an eastern innovation that had the biggest impact on the Central MNF dialects (see also Århammar 2001, 758).

One prominent morphological feature is the plural suffix -s after /l/ found in Nordergoesharde, Mittelgoesharde and Südergoesharde in the two words *apel* and *fogel*, e.g. *Wat fette der for letje Vögels bawen à di letje Mür?* (46646 Hattstedt).

Halligen

For the last group of Mainland North Frisian, the Halligen, the picture is not very clear, as shown in figure 21 (and by the global comparisons in section 4.2):

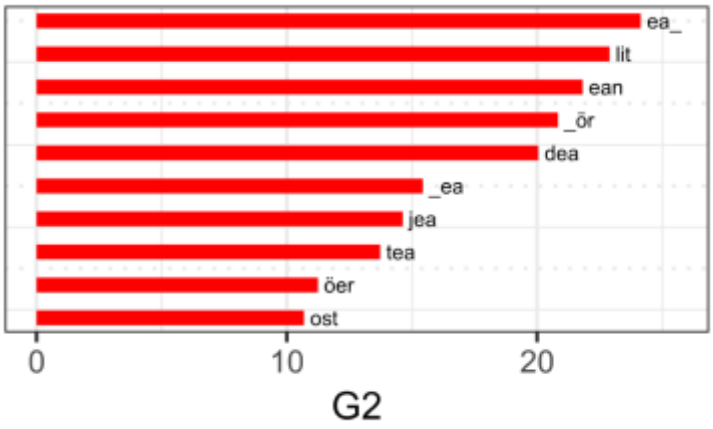


Figure 21: HA top 10 trigrams
(G2: max = 24.14, min = 10.69)

All in all, the Halligen appear to be more heterogeneous in nature. Certain idiosyncratic transcriptions represent a further challenge. What is most striking in figure 21 is the many trigrams with <ea>. These actually represent the spellings of one teacher from Hooge (46636 Hooge), which is why this questionnaire has a fairly large distance to all the others. Many of these forms are actually *d*-articles which are not known to have a diphthong pronunciation in Hallig Frisian (cf. Lorenzen 1982, 5). On the other hand, the trigrams <ör> and <öer> point at connections with the Südergoesharde, where we find centering of OFr *ō* > *ö* (cf. Århammar 2001, 751), e.g. *Brör* 'Bruder'). Lastly, we find an example of un-lowered OFr *i* in the trigrams which is to be expected according to Löfstedt (1928, 163), e. g. *litj* 'little'.

5 Discussion

In this last part of the paper, I will return to the question what the Wenker data can tell us about the situation of the North Frisian dialects at the end of the 19th century and discuss some important dialect borders explored in section 4. Finally I will look at some quality issues with the Wenker data.

5.1 Dialect dissimilarities and dialect borders

In earlier literature on North Frisian dialects, the fragmentation and dissimilarities of the North Frisian varieties have been a frequent topic. There appears to be some disagreement as to whether the varieties of Insular North Frisian or Mainland North Frisian are the most dissimilar. Siebs (1901) seems to consider the variation within Mainland North Frisian to be of greater importance:[20]

> Für die verschiedenen Dialekte, name
>
> ntlich des Festlandes, ist die ausserordentlich starke mundartliche Differenzierung charakteristisch. (Siebs 1901, 1164)

In modern handbooks, such as Markey (1981), however, we read that

> [g]enerally speaking, however, the number of similarities shared by the mainland dialects is greater than that for the island dialects. (Markey 1981, 209; similarly Sjölin 1969, 41)

At a later point, he also notes:

> [...] but Sylt normally presents fewer features that are shared with the mainland than do Föhr and Amrum. (Markey 1981, 233; similarly Århammar 1968, 295)

My data support the latter analysis: Whereas Mainland North Frisian appears to be more of a dialect continuum, Föhring-Amring and Syltring are definitely clearly separated dialect areas (Århammar 1968, 295 compares the difference between the latter varieties to that between Danish and Swedish). At the same time, especially Föhring is closer to the mainland (Bökingharde) than is Sylt.[21] A division between Westerland- and Osterland-Föhr is, however, not supported by the data as shown in figure 5. A distinction between Föhr and Amrum, on the other hand, appears in my data: In the two questionnaires from

20. The quote is ambiguous, though. An anonymous reviewer notes that Siebs could refer to quantity, not quality, here, i e. there are more dialect groups on the mainland than on the islands.
21. When using raw frequencies as input for the cosine distance, Föhring and Bökingharde appeared closer than shown in this paper. In the case of raw frequency comparisons, Bökingharde and Föhring-Amring build one cluster. The explanation seems to be that that certain words/trigrams like *jam* ‘they’ and *san* ‘am/are’ showing vowel lowering of OFr *i* > *a* shared by both the Bökingharde and Föhring-Amring carry important weight due to their high frequency. When using log-weighted frequencies, the similarities appear as somewhat lower, but still Föhring-Amring is closer to the mainland than is Sylt.

Amrum (46572 Nebel and 46747 Norddorf), which seem to be nearly identical expect for certain diacritics, the teachers systematically and correctly use <a> for vowels in unstressed syllables, a situation not found elsewhere in the North Frisian data, i.e. *bedar* 'better', *stælan* 'stolen', *botal* 'bottle'. This leads to a situation where Amrum is, although connected, fairly distant from Föhr due to many trigrams with *a*.

On the mainland, the borders between the Wiedingharde and the Bökingharde on the one hand and the Bökingharde and the Karrharde/ Nordergoesharde on the other appear to be the most significant ones in my analysis. Regarding the border between the Wiedingharde and the Bökingharde, Emmelsbüll is sometimes seen as taking an intermediate position (cf. Löfstedt 1933, 71 and the map in Århammar 1968), but this is in no way reflected in my data. Also Walker (1980, 220), on the basis of a quantitative analysis, considers the border between the Bökingharde and the Wiedingharde to be a sharp one ("Hauptmundartgrenze"), where Emmelsbüll clearly belongs to the Wiedingharde. All in all, the Bökingharde, in my data, appears to be the most diverse area of any single North Frisian dialect, something which is also partly reflected by the traditional classification by Århammar (1968), where we find a threefold division (Westermooring, Ostermooring and an unspecified third group), and by Walker (1980). Furthermore, the cluster analysis revealed that Dagebüll has a special status within the Bökingharde in that it actually seems to be closer to the Wiedingharde. One explanation for this could be the lack of vowel lowering from OFr *i* to *a* in this location, which used to be an islet of its own until the 18th century, and thus did not have its current status as a point of departure to the islands (cf. Hofmann 1956, 88).

Walker (1980, 220) also shows that the border between the Bökingharde and Karrharde/Nordergoesharde in the south is less clear than the northern border between Bökingharde and Wiedingharde. This is supported by my data (best shown in the MDS map in figure 4), e. g. the cosine distance between 46701 Marienkoog in northern Bökingharde and 46700 Emmelsbüll in southern Wiedingharde is 0.26, whereas the distance between 46761 Fahretoft in southern Bökingharde and 46763 Nord-Ockholm in northern Nordergoesharde is 0.20. The latter border runs parallel with a medieval political border, that between the so-called Uthlande (Wiedingharde/ Bökingharde + today's islands) and the Geestharden (southern Mainland North Frisian in my terminology). The former areas were directly ruled by

the Danish king until the 15th century, whereas the latter were under the duke of Schleswig (cf. Brandt 1913, 3).[22]

Between the Nordergoesharde and the Karrharde there is no clear border according to Löfstedt (1933, 70), something which is also supported by my analysis. According to Löfstedt, however, the border between the Nordergoesharde and the Mittelgoesharde is one of the most prominent in Mainland North Frisian, whereas the one between Mittelgoesharde and Südergoesharde appears as weaker (cf. Löfstedt 1933, 56 and 70). Borders south of the Nordergoesharde do not appear as very significant in my data. The border between the Nordergoesharde and the Mittelgoesharde first appears in the Ward 9 cluster solution (but not in the corresponding UPGMA solution, see figures 10 and 11). Interestingly, however, the border is the far most prominent in Mainland North Frisian according to the Neighbor-Net visualization in figure 5. Unfortunately, assessing the quality of Neigbor-Net models (similar to the R^2 score used in MDS) is non-trivial (cf. Bryant and Moulton 2004, 263), thus, this remains an open question. All in all, the southernmost varieties are less well represented in the Wenker data, something which probably also reflects the language situation of the area, where Low German has been on the rise for a long time. Alltogether, the distinctiveness of the Wiedingharde and Bökingharde compared to the rest of Mainland North Frisian may be related to the degree to which they retain North Frisian features: Lameli (2010, 36), on the basis of a quantitative analysis of feature variants from 104 randomly chosen Wenker maps, comes to the conclusion that Western Wiedingharde and Southern Bökingharde appear to be most independent (= show most Frisian features) among the mainland dialects, whereas especially the southern areas share more features with Low German due to language contact.[23]

The Halligen represent the biggest problem for this study: They clearly belong to Mainland North Frisian, but at the same time, they seem to be something for themselves. Here, a qualitative investigation of the Wenker data seems indispensable. When confronted with the findings of Löfstedt (1928), however, the situation of the Halligen seems at least somewhat less

22. Jabben (1931) says about the Geestharden that "Sie stehen seit alter Zeit sowohl in sprachlicher als auch in politischer Beziehung in einem deutlich erkennbaren Gegensatz zu den übrigen friesischen Gebieten an der Westküste Schleswig-Holsteins." (Jabben 1931, 8).
23. Note that Lameli (2010, 26) defines "Frisian features" *ex negativo*, i. e. as variants that are neither found in Danish or German questionnaires.

surprising: He initially expected North Frisian to be especially well preserved on the Halligen, but this turned out to be wrong, on the contrary he found that

> die Halligmaa. in lautlicher und lexikalischer Hinsicht dermassen destruiert waren, dass eine Klarlegung ihrer Lautentwicklungen mit erheblichen Schwierigkeiten verbunden waren. (Löfstedt 1928, VII)

The inconclusiveness of the data from the Halligen corresponds to Löfstedt's assessment, which is essentially diachronic.

5.2 Quality issues

In section 3 we saw that the Wenker materials were met with skepticism by many scholars of North Frisian in the 20th century. I hope to have shown that at least the Wenker questionnaires actually can be fruitfully used in analyzing North Frisian dialects. The classification into ten dialects for the area covered seems to be justified – at least for the 19th century – although the differences in the southernmost dialects of Mainland North Frisian do indeed appear less clear in the Wenker data.

One of Hofmann's (1956) main criticisms is the informants coming from other parts of North Frisian or from outside of this area. As Bosse (to appear) has shown, however, as many as 45 out of the 61 North Frisian questionnaires from Wenker's surveys were completed with the assistance of competent dialect speakers. A further point made by Hofmann (1956) is the inadequacies of the transcription systems used in the questionnaires. In my computer-assisted investigation, I was able to identify five really problematic questionnaires. I will not discuss Südergoesharde here, because the status of this area seems debatable even at the time of the Wenker survey. Three of these five problematic questionnaires are related to the teacher's origin: As shown by Hofmann (1956, 87), in 46709 Wester Schnatebüll in the Karrharde, the teacher, who stems from the Bökingharde, uses forms from his own dialect. Therefore, this questionnaire clusters with the Bökingharde. In two further questionnaires from the Karrharde (46711 Klintum) and the Mittelgoesharde (46587 Almdorf) both teachers originate from Bargum in the Nordergoesharde and both these locations cluster with the Nordergoesharde, which indicates that they use forms from their own dialect.

The questionnaire from Hooge (46636), on the other hand, uses idiosyncratic spellings, and thus appears as an outlier (see Neigbor-Net and UPGMA). In this questionnaire, the teacher, originating from Langenhorn in the Nordergoesharde, uses a wealth of <ea> spellings (the sequence appears 148 times in the whole document), where it is not exactly clear whether this

represents a diphthong or a monophthong, but where a monophthong interpretation seems plausible in most cases.

One last questionnaire from the Karrharde, 46764 Stedesand, is also problematic for this study, but for other reasons. In this case, the problem seems rather to be that the informant was a bit too fluent in his own dialect. As Bosse (to appear, 10) shows, this questionnaire was translated by Moritz Momme Nissen, a famous 19th century North Frisian scholar and poet, born in Stedesand. The questionnaire, however, being more of a free translation, contains many deviations from the original German version, leading to a different trigram inventory. Since the amount of text per location is rather limited, this has serious consequences. Accordingly, the Karrharde, of all the North Frisian dialect groups, has the largest share of truly problematic questionnaires from a quantitative and comparative point-of-view.

All in all, however, the Wenker questionnaires appear less problematic than suggested by Hofmann (1956). When we exclude the six questionnaires discussed above, 49 "good" questionnaires remain (when leaving out the problem of the Südergoesharde). This means that of all 58 complete North Frisian questionnaires (cf. section 3), 86% seem rather unproblematic from a global perspective.

6 Final remarks

This paper dealt with the classification of North Frisian dialects, an area known for its small-scale linguistic variation. Instead of inspecting phenomena qualitatively, as has been mostly done in previous research on North Frisian dialects, the paper followed a quantitative approach, looking at sequences of characters in a parallel corpus of dialect translations (the Wenker questionnaires from the 19th century) and then corroborated their distributions with traditional assumptions. Despite having not been used much by Frisian linguists until fairly recently, mainly due to the alleged bad quality, this paper has shown that the questionnaires may very well be used fruitfully, if certain problematic questionnaires (that form a true minority) are excluded.

Using statistical methods of multidimensional scaling and cluster analysis, it was shown that the traditional classification is to a large degree supported by a quantitative analysis of the Wenker data. Besides identifying the two main groups of North Frisian dialects, I was also able to single out two important borders (between the Wiedingharde and the Bökingharde on the one hand and the Bökingharde and the Nordergoesharde/Karrharde on the other) within Mainland North Frisian, one of which also runs parallel to an

older political border. It was, however, more difficult to find structure in the southern Mainland North Frisian material. This is probably related to the dialect situation in this area, where North Frisian has been loosing ground for decades or even centuries.

Acknowledgments

I would like to thank Temmo Bosse (University of Flensburg), Hanna Fischer, Jürg Fleischer, Sara K. Hayden (University of Marburg), Lars Johnsen (National Library of Norway, Oslo) and three anonymous reviewers for very helpful comments on the manuscript of this article and especially Temmo Bosse for providing me with machine-readable versions of the North Frisian transliterations. All remaining errors are of course my sole responsibility.

Universität Marburg

References

Århammar, Nils. 1968. Friesische Dialektologie. In Ludwig Erich Schmitt (ed.), *Germanische Dialektologie* (ZDL-Beihefte 5). Wiesbaden: Steiner, 264-317.

Århammar, Nils. 2001. Grundzüge nordfriesischer Sprachgeschichte. In Munske, Horst Haider (ed.), Handbuch des Friesischen = Handbook of Frisian studies. Tübingen: Niemeyer, 744-765.

Benoit, Kenneth. 2018. quanteda: Quantitative Analysis of Textual Data. R package version 1.3.0. https://doi.org/10.5281/zenodo.1447219

Bosse, Temmo. to appear. Die Wenkermaterialien in nord- und ostfriesischer Sprache. In Fleischer, Jürg, Alfred Lameli, Christiane Schiller, Luka Szucsich (eds.), *Minderheitensprachen und Sprachminderheiten. Deutsch und seine Kontaktsprachen in der Dokumentation der Wenker-Materialien* (Deutsche Dialektgeographie.). Hildesheim/Zürich/New York: Olms.

Brandt, Ernst. 1913. *Die nordfriesische Sprache der Goesharden*. Halle: Buchdruckerei des Waisenhauses.

Braren, Elene and Ommo Wilts (eds.). 1986. *Wurdenbuk för Feer an Oomram*. Amrum: Verlag Jens Quedens.

Bremer, Otto. 1887/1888. Einleitung zu einer amringisch-föhringischen Sprachlehre. *Niederdeutsches Jahrbuch* 13/14. 1-13, 155-157.

Bremer, Otto. 1895. *Beiträge zur Geographie der deutschen Mundarten in Form einer Kritik von Wenkers Sprachatlas des deutschen Reichs* (Sammlung kurzer Grammatiken deutscher Mundarten 3). Leipzig: Breitkopf & Härtel.

Bryant, David and Vincent Moulton. 2004. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Molecular Biology and Evolution 21(2). 255-265.

Cavnar, William B. and John M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161-175.

Cysouw, Michael. 2007. New approaches to cluster analysis of typological indices. In Grzybek, Peter and Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text*. Berlin: de Gruyter, 61-76.

Cysouw, Michael. 2018. *qlcData: Processing Data for Quantitative Language Comparison (QLC)*. R package version 0.2.1. https://CRAN.R-project.org /package=qlcData.

Dipper, Stefanie and Bettina Schrader. 2008. Computing Distance and Relatedness of Medieval Text Variants from German. In Storrer, Angelika, Alexander Geyken, Alexander Siebert and Kay-Michael Würzner (eds.), *Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing (KONVENS-08).* Berlin: de Gruyter, 39-51.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1). 61-74.

Ebert, Karen H. 1998. Genussynkretismus im Nordseeraum: die Resistenz des Fering. In Winfried Boeder et al. (eds.), *Sprache in Raum und Zeit 2*. Tübingen: Narr, 269-281.

Fleischer, Jürg. 2012. Pronominalsyntax im nordwestlichen Niederdeutsch: eine Auswertung des Wenker-Materials (mit Einbezug der friesischen und dänischen Formulare). *Niederdeutsches Jahrbuch* 135. 59-80.

Fleischer, Jürg. 2017. *Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen: Dokumentation, Entdeckungen und Neubewertungen* (Deutsche Dialektgeographie 123). Hildesheim: Olms.

Haas, Walter. 1995. Wenker contra Bremer oder: Empirie und Theorie des Dialekts. In José Carot, Ludger Kremer and Hermann Niebaum (eds.), *Lingua Theodisca. Beiträge zur Sprach- und Literaturwissenschaft. Jan Goossens zum 65. Geburtstag*. Münster / Hamburg: Lit, 331-340.

Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance* (Groningen dissertations in linguistics 46). Groningen: Rijksuniversiteit Groningen.

Heeringa, Wilbert and Nerbonne, John. 2013. Dialectometry. In Hinskens, Frans and Johan Taeldeman (eds.), *Language and Space. An International Handbook of Linguistic Variation. Volume 3: Dutch.* Berlin and Boston: de Gruyter, 624-645.

Herrgen, Joachim. 2001. Dialektologie des Deutschen. In Auroux, Sylvain and Herbert Ernst Wiegand (eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft: Geschichte der Sprachwissenschaften* (Handbücher zur Sprach- und Kommunikationswissenschaft 18.2). Berlin, New York: de Gruyter, 1513-1535.

Hofmann, Dietrich. 1956. Probleme der nordfriesischen Dialektforschung. *Zeitschrift für Mundartforschung* 24. 78--112.

Hofmann, Dietrich. 1961. *Die k-Diminutiva im Nordfriesischen und in verwandten Sprachen* (Niederdeutsche Studien 7). Graz: Bühlau Verlag.

Hofmann, Dietrich and Anne Tjerk Popkema. 2008. *Altfriesisches Handwörterbuch*. Heidelberg: Winter.

Hoppenbrouwers, Cor and Hoppenbrouwers, Ger. 1988. De featurefrequentiemethode en de classificatie van nederlandse dialecten. *Bulletin voor Taalwetenschap (Tabu)* 18. 51-92.

Huson, Daniel H. and David Bryant. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23(2). 254-267.

Jabben, Oltmann Tjardes. 1931. *Die friesische Sprache der Karrharde: Lautlehre*. (Veröffentlichungen der Schleswig-Holsteinischen Universitätsgesellschaft: 30; Schriften der Baltischen Kommission zu Kiel 19). Breslau: Hirt.

Jockers, Matthew L 2014. *Text Analysis with R for Students of Literature* (Quantitative Methods in the Humanities and Social Sciences). Cham: Springer.

Lameli, Alfred. 2008. Was Wenker noch zu sagen hatte... Die unbekannten Teile des `Sprachatlas des deutschen Reichs'. *Zeitschrift für Dialektologie und Linguistik* 75(3). 255-281.

Lameli, Alfred. 2010. Relationen im Sprachkontakt: Das Beispiel der nordfriesischen Mehrsprachigkeit. *Zeitschrift für Dialektologie und Linguistik* 77(1). 19-53.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: Benjamins.

Löfstedt, Ernst. 1928. *Die nordfriesische Mundart des Dorfes Ockholm und der Halligen*. Lund: Gleerupska-Univ.-Bokhandel.

Löfstedt, Ernst. 1933. *Beiträge zur nordfriesischen Mundartenforschung* (Lunds Universitets årsskrift. Avdelningen 1, Teologi, juridik och humanistiska ämnen 2). Lund: Ohlsson.

Lorenzen, Jens. 1982. *Halligfriesische Sprachlehre* (Nordfriisk Instituut 72). Bredstedt: Nordfriisk Instituut.

Manning, Chris and Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Markey, Thomas L. 1981. *Frisian* (Trends in linguistics. State of the art reports 13). The Hague: Mouton.

Moisl, Hermann. 2015. *Cluster Analysis for Corpus Linguistics* (Quantitative Linguistics). Berlin/Boston: de Gruyter.

Moran, Steven and Cysouw, Michael. 2018. *The Unicode cookbook for linguists: Managing writing systems using orthography profiles*. Language Science Press.

Nerbonne, John, Peter Kleiweg, Franz Manni and Wilbert Heeringa. 2008. Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering. In Preisach, Christine, Hans Burkhardt, Lars Schmidt-Thieme, Reinhold Decker (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the 31st Annual Conference ofthe Gesellschaft für Klassifikation e.V., Albert-Ludwigs Universität Freiburg, March 7-9, 2007.* Springer, 647-654.

Nübling, Damaris. 2000. *Prinzipien der Irregularisierung: eine kontrastive Analyse von zehn Verben in zehn germanischen Sprachen* (Reihe Linguistische Arbeiten 415). Tübingen: Niemeyer.

R Core Team, 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rayson, Paul and Garside, Roger. 2000. Comparing Corpora using Frequency Profiling. In *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, 1-6.

Selmer, Ernst W. 1921. *Sylterfriesische Studien*. Kristiania: Dybwad.

Selmer, Ernst W. 1926. *Über den Gebrauch des bestimmten Artikels im Nordfriesischen*. Oslo: Dybwad.

Siebs, Theodor. 1889. *Zur Geschichte der englisch-friesischen Sprache*. Halle: Niemeyer.

Siebs, Theodor. 1901. Geschichte der friesischen Sprache. In Paul, Hermann (ed.), *Grundriß der Germanischen Philologie* vol. Vol. 1. Straßburg: Trübner, 1152-1464.

Sjölin, Bo. 1969. *Einführung in das Friesische* (Sammlung Metzler 86). Stuttgart: Metzler.

Spruit, Marco René, Wilbert Heeringa and John Nerbonne. 2009. Associations among linguistic levels. *Lingua* 119(11). The Forests behind the Trees, 1624-1642.

Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6(1). 45-76.

Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects*. Cambridge: Cambridge University Press.

Versloot, Arjen P. 2001. Allgemeine und vergleichende Aspekte des Friesischen. In Munske, Horst Haider (ed.), *Handbuch des Friesischen = Handbook of Frisian studies.* Tübingen: Niemeyer, 767-775.

Walker, Alastair. 1980. *Die nordfriesische Mundart der Bökingharde: zu einer strukturell-dialektologischen Definition der Begriffe "Haupt"-, "Unter"- und "Dorfmundart"* (ZDL-Beihefte 33). Stuttgart: Steiner.

Walker, Alastair. 1990. Frisian. In Charles V. J. Russ (ed.), *The Dialects of Modern German*. London: Routledge, 1-30.

Walker, Alastair G. H. 2001. Extent and Position of North Frisian. In Munske, Horst Haider (ed.), Handbuch des Friesischen = Handbook of Frisian studies. Tübingen: Niemeyer, 263-284.

Walker, Alastair G. H. and Ommo Wilts. 2001. Die nordfriesischen Mundarten. In Munske, Horst Haider (ed.), Handbuch des Friesischen = Handbook of Frisian studies. Tübingen: Niemeyer, 284-304.

Wenker, Georg. 2013. *Gesamtausgabe der Schriften zum „Sprachatlas des Deutschen Reichs“. Herausgegeben von Alfred Lameli. Unter Mitarbeit von Johanna Heil und Constanze Wellendorf. Band 1: Handschriften: Allgemeine Texte, Kartenkommentare 1889-1897* (Deutsche Dialektgeographie 111). Hildesheim, New York, Zürich: Olms, 2013.

Wilts, Ommo. 1995. *Friesische Formenlehre in Tabellen IV: Sylt*. Husum: Matthiesen Verlag.

Wilts, Ommo. 2001. Die Verschriftung des Nordfriesischen. In Munske, Horst Haider ed.), Handbuch des Friesischen = Handbook of Frisian studies. Tübingen: Niemeyer, 305-313.