Orthography-based dating and localisation of Middle Dutch charters

Thesis,
written by Dieter Van Uytvanck,
student of Linguistics at the Radboud University Nijmegen,
majoring in Language & Speech Technology,
to obtain the degree of Master of Arts

December 2007

Thesis supervisor:
Hans van Halteren (Radboud University Nijmegen)

Preface

The last months have been a very intensive period. I think that I genuinely experienced all of the things that most scientists encounter during their research: from curiosity, through the despair when all things seem to go wrong, to the thrill of doing some discoveries. All in all, it was a fascinating time. However, let it be clear that this was only possible because of the extensive support by a lot of people.

First, of course, my sincere thank goes to Hans who suggested this subject in the first place, turned out the be a supportive supervisor and managed to guide me flawlessly through the whole process — even when I sometimes did not know anymore which direction to choose.

Margit gave indispensable support on the historic and linguistic aspects of my thesis. When I had trouble with finding suitable data, she pointed me in the right direction.

With regards to the preliminary research on the corpus data I would also like to thank Griet Coupé and Judith Kessler for granting access to their research material.

Hans Adamse, Jort Florent Gemmeke and Bert Cranen deserve praise for helping me sorting out some technical problems and winning the battle against the mosix cluster. More generally, the whole free software community made it possible to do this research using a broad range of excellent software.

Some final warm words go to my parents and Katrijn and Thomas. They unconditionally supported me, even when I decided to take up a second study and helped me to deal with my once in a while hectic life of studying, working an travelling.

To conclude in style, I would like to add a typical Middle Dutch charter element, although it is slightly altered[1]:

dit was ghedaen jnt jaer ons heeren als men screef MM ende vii den xiv dach van december

[1] It is left to the reader as an exercise to discover the anachronism.

Abstract

In this study we build models for the localisation and dating of Middle Dutch charters. First, we extract character trigrams and use these to train a machine learner (K Nearest Neighbours) and an author verification algorithm (Linguistic Profiling). Both approaches work quite well, especially for the localisation task. Afterwards, an attempt is made to derive features that capture the orthographic variation between the charters more precisely. These are then used as input for the earlier tested classification algorithms. Again good results (at least as good as using the trigrams) are attained, even though proper nouns were ignored during the feature extraction. We can conclude that the localisation, and to a lesser extent the dating, is feasible. Moreover, the orthographic features we derive from the charters are an efficient basis for such a classification task.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter will introduce the reader into the broad context in which this research project is situated. First we will discuss some important concepts and the historic background. In the next section we will give an overview of relevant research in this field. We conclude with the basic idea behind and motivation for our specific research goals.

1.1 The context

Old relics have always fascinated people. Archaeologists study a potsherd, and upon close investigation a world of information opens to them. Not only can they determine which materials were used, but the same piece can reveal something on the daily life, habits, eating culture and the society as a whole. The object of the study becomes a rich and captivating source of historic information — something no one probably thought of at the time it was made.

Within historic linguistic research, the situation is very similar. An old text can deliver far more information than one could think of at first sight, and definitely more than the original author intended. As with the piece of pottery, there are many layers of information stored in a historic manuscript. Next to the actual content, it can provide information on the historic context, the author's identity (directly or through e.g. the handwriting) and a broad range of linguistic factors, from discourse style through grammar to orthography.

Often studying linguistic phenomena within large collections of historic documents is a matter of sheer drudgery and we cannot but respect the erudition that has been devoted to the work in this area. However, since the rise of the computer this field has gone through a revolution. As it became possible to store large amounts of information and to make even larger numbers of calculations a whole wave of new approaches has been introduced in the historic language research.

It is exactly in this new paradigm of using modern technology to analyse historic documents that this study will be situated. We chose to concentrate on the orthography of historic Dutch. After learning more on the available data (as explained in chapter 3), the focus went to Middle Dutch. This is not a real language on its own but merely a collective term for the dialects as used in the Dutch language area between roughly 1200 and 1550. This inherent variety enhances the potential for attempts at deriving where and when a text has been written, based on the spelling particularities.

When performing experiments on the Middle Dutch documents, we need metadata, like the place and time of writing. A specific subset of the existing Middle Dutch material that has been digitised, fits exceptionally well with this requirement. Separate juridical declarations of someone's rights or obligations contain explicit information on the place and time of writing. Subjects that are being handled in such documents are e.g. selling of ground to someone or inheritances. All of them are more or less in line with a formal structure and thus are very apt for mass processing. Such an official document in the Medieval context is called a charter. For an extensive historic background on charters, we refer to Rem (2003).

We have now introduced all elements that form the basis of this thesis: the Middle Dutch charters, the fact that there is a large variation within Middle Dutch and the aim to use language technology to analyse orthographic differences, relating them to time and place. In the next sections we will place these elements within the context of related research, explaining in what aspects our design differs.

1.2 Research overview

Since some corpora of Middle Dutch charters have been published in electronic form in the last decade of the 20th century, a variety of research based on language technology has arisen.

First of all there is the approach by Mooijaart & Van der Heijden (1992) to use correspondence analysis on a set of 160 dichotomous dialect properties, trying to relate these to the 56 locations where Middle Dutch 13th century charters are found. The properties were selected from the Atlas of Early Middle Dutch linguistic variants (Mooijaart, 1992). It is demonstrated that there is a clear relation between the 2-dimensional feature space that remains after the correspondence analysis and the geographical reality. This is explained both by dialectic similarities between neighbouring locations and the occurrence of fixed expressions that are typical for certain places.

Rem (2003) uses a set of manually selected binary — mostly phonetic — features, the so-called locators. These locators are in fact significant attributes that have been mainly identified beforehand in the literature. Using these locators in combination with a self-developed classification algorithm, it proves possible to predict the provenance of the scribes[1] according to their language use. The classifier was trained on the data of the 14th century Middle Dutch corpus (see chapter 3).

In Van Dalen-Oskam et al. (2004), finally, the occurrence of a specific term (*beseghelen*, 'to seal') is studied. This is done both for the material of the 13th and 14th century. It gives a good overview of the differences between the corpora for the respective centuries.

1.3 Our approach

One of the commonalities throughout the approaches mentioned in the section above is the manual selection of the criteria used for the analysis or classification. They are mostly based on observations made in "classic" analyses, often phonetically inspired, like e.g. Mooijaart (1992).

In contrast with this frequent strategy, we will user large sets of automatically derived features and leave the decision whether and how to use these to the classification algorithms. We look how well the location and date of each charter can be predicted, and which machine learning algorithms perform best on this task, using n-fold cross-validation.

In the first instance we will rely on very simple attributes: character n-grams. At the same time we will setup a suite of benchmarks and analysis methods in order to measure the performance of the resulting classification.

During a later stage we try to capture the specific orthography of each charter, storing it as a separate feature vector. Using the earlier developed methodology we then run the classification again, this time with the new data, and compare its outcome to that of the classifier trained on the n-grams. Finally we will also assess the combined judgements of both methods.

[1] The clerks writing the charter. In Rem's work, the focus is on the chancery clerks of the Count of Holland.

Chapter 2

Problem definition and goals

After the overview we gave in the first chapter on the field of language technology-based research on historical Dutch texts, we now want to define clearly what we plan to do during this research project. Therefore we provide a clear problem definition, combined with the subquestions that we plan to answer in the course of this study.

2.1 Problem definition

Is it possible to determine inductively where and when a Middle Dutch charter was written based on its orthography? If so, to which degree of granularity is this possible?

We seek an initial solution to this problem in standard techniques for document classification. Subsequently, we will use the same techniques in combination with document features based on a more sophisticated representation of orthographic variation, instead of or in addition to features representing merely the presence of certain character strings. Finally, we will compare the outcome of both approaches.

2.2 Subquestions

During the cause of this study we will try to find an answer to the following questions:

- For those charters whose origin is well known, is it possible to recognize the location in which they were written on the basis of orthographic features, initially character n-grams?
- The same question as above for the dating of the charters.
- Can we develop a representation for orthographic variance which relates spelling variants to each other or to a canonical base form? If so, how do the classification results compare to the character n-gram approach and how does a combination of both methods perform?
- What is the relation between the granularity of the classes to predict (e.g. years versus decades and cities versus regions) and the accuracy of the classification? Does it pay off to use a multilevel classifier as an alternative to a single classifier?
- It is generally known that proper names and especially toponyms are less prone to diachronic changes than other word classes. From this perspective the question arises whether we should exclude proper names from the material that will be used by the orthographic temporal classifier. Additionally we ask ourselves what their influence on the classification results is – if they are included.

Chapter 3

The quest for corpus data

The research we intend to do largely depends on the existence of digital corpora that contain Middle Dutch texts. As creating such a corpus requires huge efforts, we almost directly decided to make an assessment of existing corpora that could be used for our study. In this chapter we describe the requirements that are put forward towards the corpus data. Then some potential sources are discussed, followed by a substantiation for the selection we made. Finally we give an overview of the conversion of the selected corpora towards a unified format that is ready to use for our research goals.

It should be noted that as a part of the work in this chapter has been done within the framework of a preparatory phase, we still considered the use of corpora that fall outside the scope of Middle Dutch charters. Therefore we will explain why we chose the latter as the basis for this project.

3.1 Requirements

To perform useful research on orthography the source material should fulfil several requirements:

Format The data has to be digitised, i.e. being available in a format that is computer-readable. Preferably a standardised structured format (e.g. SGML) is chosen for this purpose.

Authenticity The texts in the corpus have to be a trustworthy representation of the originals. Especially the orthography should not have been altered.

Size It goes without saying that a corpus should be big enough. Certainly if one wants to apply inductive learning techniques on it.

Variation If you want to classify texts based on certain parameters (location and time in this case) you need access to a collection of documents that is spread well over the classes that are to be distinguished.

Consistency For the metadata as well as for the annotations (e.g. how are abbreviations marked?) it is important that a consistent coding standard is used.

3.2 Corpus assessment

The fields of study related to Middle Dutch linguistics are quite diverse. This can be clearly observed by the number of related electronic databases, and even more by the way they are set up. Although it is obvious that research on poetry demands other data then a study on syntax, we have to come to a balanced decision with regards to the question which of the existing data sources is apt for our aim — orthography-based classification. For an extensive report on this issue, we refer to Van Uytvanck (2007), where this exercise was made. Here we will restrict ourselves to a short overview of the options we considered, and explain the decision about the inclusion or exclusion of each.

3.2.1 Dutch in Transition

Within the framework of the project "Variation and standardisation: the influence of language contact on the emerging Dutch standard language"[1] , the *Dutch in Transition* (DiT) corpus was created with Dutch texts from the period 1400-1700. However, the main focus of this project is grammar. A part of the data in this corpus is based on so called critical editions - texts whose spelling has been normalised, often during the 19th century. This does not pose problems for syntactic research, but for ours it does, so we had to exclude this corpus[2].

3.2.2 DBNL

The DNBL (Digitale Bibliotheek voor de Nederlandse Letteren, 'Digital Library for the Dutch Literature') is a large digitalised library of Dutch literature from the Middle Ages until now[3]. However the collection suffers from the same problem as the DiT: the orthography has not been maintained consistently. Moreover, literature is a prime example of language use which is heavily influenced by copying. As such, it is very difficult to relate a piece of historic Dutch literature to a single location or year. All these factors gave cause to the exclusion of the DBNL material.

3.2.3 Anna Bijns

Another potential source we considered was the collection of poems by Anna Bijns, written between 1520 and 1550 (Kessler & Oosterman, 2007). Because of a lack of geographical and temporal spreading, we discarded this source as well.

3.2.4 Corpus Gysseling 1

The largest part of Middle Dutch texts from before 1301 is included in a series of 15 books, known as the Corpus Gysseling, named after the work of Gysseling & Pijnenburg (1987). In turn, it exists of two parts: the first 9 books (abbreviated as CG1) contain 1937 charters, the last 6 (CG2) hold literary works.

The books of the Corpus Gysseling were published in 1998 as a part of the cd-rom with the dictionary of Middle Dutch (Instituut voor Nederlandse Lexicologie, 1998). Just as the books the cd-rom is a diplomatic edition and thus it provides the genuine spelling. This, together with the fact that charters are rewarding research material because they can mostly be well localised and dated, pleads for the use of the CG1.

[1] `http://oase.uci.ru.nl/~varstan/`

[2] At the moment of writing, the DiT corpus is being extended with metadata on the trustworthiness of the orthography of the included documents. This means that in the future it might be used for research on spelling nevertheless.

[3] `http://www.dbnl.org/`

The CG2 does not contain this information (or in a lower proportion). Because its literary nature — which means that it is often copied and thus multiple language layers can be found in it — we decided not to include the second part of the Corpus Gysseling.

3.2.5 Corpus Van Reenen-Mulder

Chronologically following on the CG1, there the corpus with 14th century Middle Dutch charters, also known as the Corpus Van Reenen-Mulder (CRM), after Van Reenen & Mulder (1993). It contains 2720 charters, from 345 location. 149 of these are not cities or villages, but regions, because a more exact localisation is lacking. The localisation was done using non-linguistic criteria, as described in Rem (2003, p. 24). This avoids circular reasoning, like the induction of dialect properties on the basis of texts that have been classified by those very dialect characteristics.

All charters were diplomatically transcribed and there is a pretty good temporal and geographical spread because relatively many 14th century charters have been preserved. So apart from the CG1 we have good reasons to use this corpus as well for our research.

3.3 Gathering, editing and analysing the data

Now that we have made a choice between several available sources for historic Dutch, we will have a closer look at the qualitative and quantitative properties of the two chosen corpora. Besides we explain how we convert them into a convenient uniform format that is usable for further computer-assisted research.

3.3.1 Corpus Gysseling 1

3.3.1.1 Properties

When the CG1 was created, an attempt was made to set up an as large as possible collection of charters. This indicates in advance that the amount of authentic documents from the thirteenth century that still exist is not that big. It can also be noted that this corpus hardly holds any material from outside Flanders (Van Dalen-Oskam et al., 2004). These are factors one can live with, but nonetheless it should be brought into account, e.g. when creating new computer models. As a whole the corpus contains 1937 charters. Together they hold 41,325 types (unique word forms, ignoring capitals).

3.3.1.2 Conversion

On the Middle Dutch cd-rom (Instituut voor Nederlandse Lexicologie, 1998) the CG1 corpus is only accessible through a tailored windows application. Although this suits the needs for the consultation of individual texts, it does not offer an opportunity to process all data together. Therefore we decided to export all charters to a format that is easier to manipulate, as described below.

Export to HTML Using the export to HTML option and a macro package we stored all of the charter texts to HTML files. We chose this format above pure text, as some of the annotations are —- similarly to the printed edition of the CG1 — indicated by text markup, like italics for abbreviations that have been expanded.

Extraction of the metadata The resulting HTML files are unfortunately not semantically structured (as XHTML is), but exists mainly of representational codes. Luckily the typesetting is very consistent and therefore we can automatically convert these codes into their semantic equivalents. In that way we can distinguish the following elements for each charter:

- the identification number
- the dating
- the localisation
- the introduction text, written by the editor of the corpus
- the text of the charter itself

Using a script and regular expressions we have extracted and built an XML-structure (based on the the parts mentioned above) as follows:

```
<charter>
  <number>...</number>
  <place>...</place>
  <date>...</date>
  <description>...</description>
  <text>...</text>
</charter>
```

Enrichment of the metadata Although we have now added some structure to the data, some parts of the metadata are still not fully suitable for automatic processing. For some charters the exact dating or localisation lacks. This often is indicated by a question mark or the note "place X or place Y". Yet for most of the texts the exact place or date is known. To take into account this situation we have chosen to add some extra fields to each file: a normalised date and place. This holds:

- <localisation>, containing the city or village name if known and otherwise the statement "unknown"
- <year>, with the year in which the charter was written or again the "unknown" label

For some of the texts a dating of the type "year *n* or year *n+1*" was used. In order not to loose this information we have (arbitrarily) chosen to use *n* as <year>-component. When the deviation was more than one year we marked the <year> as "unknown". These simplifications can always be detected by comparing the <year> and the <date> fields. The latter has been left unchanged.

For the small number of charters for which a date could not be derived automatically — for these the <year> value "unknown" was assigned by default — we manually checked whether it was possible to determine the date of writing. If possible, we added the correct <year> data.

The addition of annotations Some of the notations of the CG1 could not be simply transferred into a standard text format. This is the case for expanded abbreviations (which are marked in italics), reconstruction of fragments (between square brackets) or even text parts that have been added later on (marked with a red text colour). To be sure that we have access to both the diplomatic text and the version with extra information we converted the existing annotations into a code that is easily manipulable. It exists – for each word – of the original text, an @-sign, and an indication of the annotation type and finally the text critical information or addition. More concretely we used:

A@guess:B The text A is has not been retained or is badly readable, presumably B was written.

A_@full:B The part of A indicated with _ was an abbreviation, the full form is B.

A@note:B B contains a note about the word A.

<nl> Indication of a new line.

3.3.1.3 Selection

Now that we have a workable format for the editing and selection of individual charters within the corpus, the next important step is to exclude those texts that are not suitable for automatic classification. More concretely we can distinguish the following cases:

Special charters The first text in the CG1 consists of a wide selection of Latin charters, mixed up with Middle Dutch. Because this differs significantly from all other charters and it is widely spread over several years (1210-1240) we decided to exclude it from our material. Charters 92, 93, 94, 307 and 1253 are each written by multiple authors (at different times, sometimes a successive part was authored after twenty years). Therefore these have been excluded as well.

Forgeries and copies Some charters are copies (2*, 14*) or forgeries (1505*, which is doubtfully situated in Kortrijk) — we will therefore not consider them as research material.

Documents from chancelleries Rem (2003) concludes that charters that have been written at the ducal or count's chancelleries are often more difficult to determine on the basis of dialect properties. Therefore no charters of this type have been included in the Van Reenen-Mulder corpus. For this very reason we decide not to include the thirteenth century charters written at "higher instances" either. This means that we removed 216 documents from the CG1. A list of these can be found in appendix A.

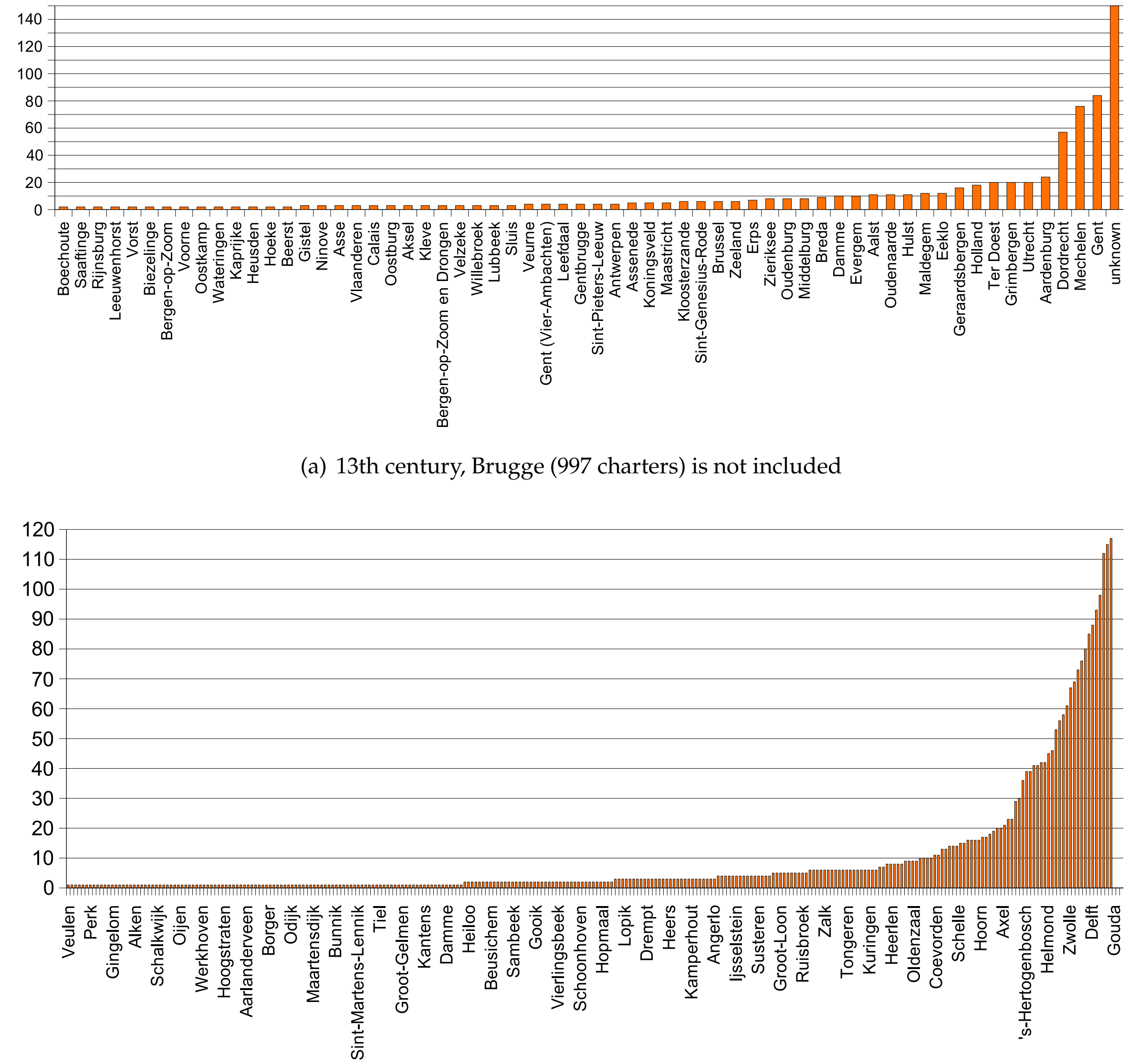
Copies Not all texts from the CG1 are based on original documents. Sometimes they contain copies of the originals. Because this increases the risk of multiple language layers — the person creating the copy possibly adds some of his own language background — we removed these. A full enumeration of the copies (48 charters) that were removed is added to appendix A as well.

Linguistically localised documents In the corpus Gysseling certain charters have been localised according to dialect properties found in the texts. Rem (2003, p. 30) remarks that this kind of localisation should not be trusted. In order to prevent circular reasoning (using linguistically localisations for the benchmarking of a linguistic localisation method) we decided to remove the location label for these charters. To be more precise: those mentioned in Mooijaart (1992), number 660, 661, 1193, 1222 and 1268, will not be used for localisation purposes.

3.3.1.4 Analysis

After this selection procedure we keep a corpus of 1748 charters, that contains more than 664,000 tokens. The distribution relative to dating and localisation can be found in respectively figure 3.1 and 3.2.

Timewise there is a reasonable spread. Almost obviously there are more "younger" charters represented. As noticed in Van Dalen-Oskam et al. (2004), the material of the thirteenth century is scarce, and this certainly is the case for the begin of that century.



(a) 13th century, Brugge (997 charters) is not included

(b) 14th century, only some examples are shown on the X-axis

Figure 3.1: Location distribution (number of charters) for the 13th and 14th century corpus.

As for the locations[4] we can conclude that most of the texts (998) are from Brugge. Furthermore there are only fifteen places with ten or more texts. The other places are mainly represented by one or two charters.

3.3.2 CRM

3.3.2.1 Properties

In contrast with the CG1, a larger number of sources was available for the realisation of the 14th century corpus. This had repercussions for the composition: it became possible to spread the included charters more or less equally over the locations and years. As such we have about 2500 relatively well spread charters at our disposal. They contain 40,508 types.

[4] We will further on use the Dutch place names, even when an English variant (Bruges, Ghent, Brussels, The Hague, etc.) exists, as most of the locations we are dealing with do not have an English toponym.

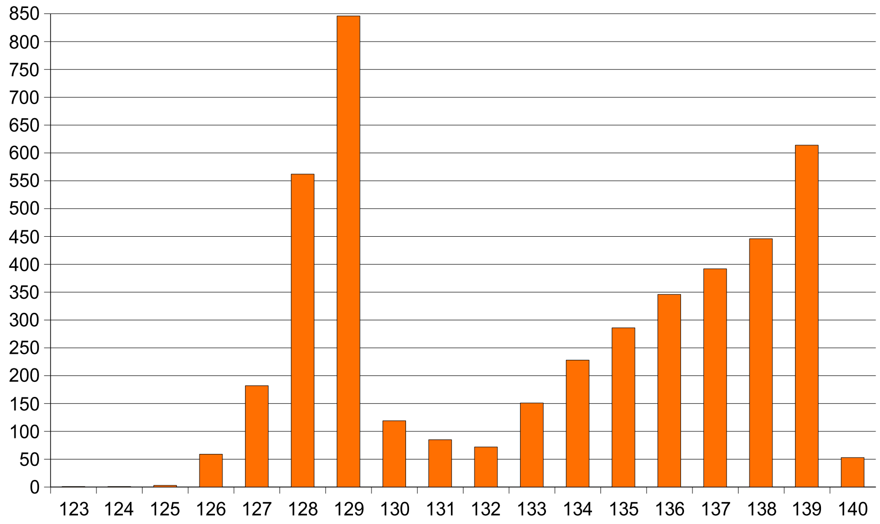


Figure 3.2: Decade distribution (number of charters) for the 13th and 14th century corpus. The first 3 numbers on the X-axis indicate the decade.

3.3.2.2 Conversion

Compared to the original CG1 this corpus has already been incorporated into a form that is more directed towards computer-based processing. The starting point on the way to the final format as described in section 3.3.1.2 is a file with one charter on each line, preceded by a code as described in Rem (2003, p. 22). An example of such a code is _o:E109p39509 and exists of:

_o:	a reference to a charter ("oorkonde" in Dutch)
E109	the code (Kloeke number, see Kloeke, G.G. (1927)) for Amsterdam
p	this is an exact place indication (while r stands for a region)
395	year indication, this text dates from 1395
09	the serial number within all charters from Amsterdam of 1395

3.3.2.3 Extraction of metadata

All necessary metadata is already contained in the charter code. We mapped the Kloeke number to the place name (prefixing it with REGION if it is a regional place indication), This metadata, together with the text of the charter, is saved as an XML-file.

3.3.2.4 Enrichment of the metadata

We simply copied the content of the just extracted <date> and <place> to <year> and <location>. This ensures that we can access the elements <year> and <location> in a consistent way for all charters (from both the 13th and 14th century).

3.3.2.5 Selection

Charters that were written at a "higher" instance, as a chancellery, are not included in this corpus, as described in Rem (2003, p. 25). This means that we do not have to exclude them explicitly — as was the case for the CG1.

3.3.2.6 Analysis

As a whole this corpus consists of 2720 charters, that contain in turn about 794,000 tokens (including new line indications). Again we have created some histograms to give an impression about their distribution over year and location, as can be seen in figures 3.2 and 3.1.

It becomes immediately clear that there are more documents towards the end of the 14th century, but unlike the CG1 every year is represented. For 62 locations (i.e. cities and regions) at least 10 charters are available.

Chapter 4

Classification feasibility

Now that we have gathered a charter set, we will first have a look at a basic classification approach. This involves choosing meaningful classes, extracting features from the documents and testing several classification algorithms.

4.1 Feature selection

The outcome of machine learning classification depends for a large part on the feature set that is used. Therefore it is important to choose a well-reasoned procedure to extract features from the documents we are considering. We are on the other hand initially dealing with a basic approach to get an idea about the quality of models that are based on simple features. Remember that we want to assess such a basic strategy as a reference point for further models that will use more sophisticated attributes.

4.1.1 Character trigram frequency

In the first instance one would think about using the widely spread naive bayes classification, which relies on the occurrence of certain words in a document. However this would most probably bring along with it the risk that we base the classification on some content words (e.g. city names and years) instead of orthography.

To counter this issue we decided to use character trigrams instead of whole words. This still holds some risk of content recognition in the feature vectors — think about parts of cities, e.g. *Dam* is probably derived from *Damme*. However the risk is largely reduced. Word boundaries were accounted by adding spaces at the begin and end of every word when extracting the trigrams. The word "scepene" e.g. was converted into the following trigrams: " sc", "sce", "cep", "epe", "pen", "ene" and "ne ".

By counting the frequency of all encountered character trigrams, we create a feature vector for each document from the corpus. In order to neutralise the influence of the charter's size we normalise the frequency count by dividing it through the number of trigrams in the document. Then we multiply the outcome by 1000, in order to get a result which is easier to grasp.

4.1.2 Normalising the frequencies

4.1.2.1 Procedure

Now each document corresponds to a feature vector containing the relative frequency counts for all character trigrams. As an extra normalisation we finally calculated the Z-score for every attribute. This was done using the following formula:

$$f_i' = \frac{f_i - \overline{f_i}}{\sigma(f_i)} \tag{4.1}$$

In other words: for every feature f_i we calculated the number of standard deviations it is positioned from the mean value for this feature. If this measure for a feature is 0, this means it corresponds to the mean value for this feature. The more the absolute value of f_i' differs from 1, the more remarkable this feature is. If it has a positive value, it occurs more than on average, and the other way around for a negative value.

Arguably this operation could be regarded as a source of overfitting when splitting these feature vectors into separate training and test data sets. However, as we do not use the class value in order to perform this normalisation, it could at most be seen as preprocessing the data with a non-supervised method.

4.1.2.2 Afterthought

In retrospect we came to the conclusion that the way of calibrating the relative frequency count we mentioned above (the calculation of f_i') might suffer from some theoretical imperfections. The main problem is caused by the calculation of the standard deviation. We used the following following estimator for the variance, relying on the assumption that the population of relative frequency counts for each character trigram (x_i) is normally distributed over all N documents:

$$\sigma(f_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

However, the relative frequency counts are not normally distributed. In fact, we cannot even guarantee that even if there was one distribution that fits all trigram relative frequency counts, that distribution would have a variance.

To explain this reasoning, we first need to focus on the distribution of the character trigrams among all documents. The distribution of words in corpora has been investigated extensively in Baayen (2001). One important conclusion is the fact that words are not normally distributed. Instead, as Baroni (2007) states:

> The distribution of words is akin to finding a population made of few giants, rather few people in the medium height range and an army of dwarves.

The randomness assumption is invalid for the distribution of words, as some elements are used extremely more often than others. According to Zipf's law, most words only occur once, even in large corpora. Cavnar & Trenkle (1994) showed that the same reasoning is valid for character n-grams as well. The great difference in the frequency for all trigrams forces us to consider different cases:

4.1.2.2.1 Hapaxes The largest part of all trigrams only appears once, i.e. in a single charter. This means that the maximal information that can be derived from the feature in question is a binary hint: does the trigram occur or not? It does not make any sense to make any further calculation, as only inter-feature comparisons will be made further on.

4.1.2.2.2 Non-hapax trigrams In every case where there is more than one occurrence of a trigram, it theoretically could pay off to use more than just a binary variable. With the induction techniques in mind, one of the prerequisites of the features to be created would be the comparability. In other words: does a n-gram occur more often than another one? At this stage, such comparisons can already be made as we have calculated the relative frequency count earlier on.

One could think of rescaling such a feature for every vector, using the Z-score (formula 4.1). However that would require us first to find out with which distribution the frequencies for the trigram in question could be modelled. Only then we can determine the variance and thus the standard deviation.

Furthermore, even if we found a distribution that matches the frequency counts, it is most probable that for each trigram the distribution parameters should be estimated separately. Even worse: it is very well possible that we should use different distribution models for the trigrams at the tail and those at the top of the Zipf curve.

All in all we can conclude that rescaling the relative frequency counts using the Z-score in a theoretically correct way would be utterly complex if not practically impossible[1]. For our research aims, the use of the relative frequency should have been sufficient.

4.1.2.3 Why does it work nevertheless?

Now, if the rescaling method we applied to the relative frequency counts erroneously assumes a normal distribution, why do we still get sensible results when using the rescaled data for machine learning? This can be explained by the fact that we apply a mapping to the feature vectors that does not destroy the inherent information.

For all trigrams that only appear once, the mean will be just a little bit above 0. The same goes for the standard deviation, though the latter will be a fraction higher than the mean. This means that for a hapax trigram the following formula applies:

$$f'_i \quad = \quad \begin{cases} \frac{-\overline{f_i}}{\sigma(f_i)} \in [-1, 0[& \text{if } f_i = 0 \\ \frac{f_i - \overline{f_i}}{\sigma(f_i)} \in]0, +\infty[& \text{if } f_i \neq 0 \end{cases}$$

So instead of 0 (if the trigram does not occur) and positive numbers (if it occurs once) the features contain respectively a negative number above -1 and a positive number.

For non-hapax features, the situation is slightly different:

- A rescaled feature can then be lower than -1 if in other vectors the same feature has a much higher count. Then $|f_i - \overline{f_i}|$ will be larger than the standard deviation. This situation is relatively seldom encountered.
- If a trigram occurs more than average in a document, the rescaled result will again be larger than 0.
- If a feature is encountered equally frequently in all charters (which is very unlikely), the standard deviation becomes 0. In this case we left the relative frequency count as is, to avoid the division by zero. In the end this feature will not be able to deliver any information during the further classification stages anyway.
- In case the frequency count exactly equals the mean count and the standard deviation differs from 0, f'_i also becomes 0. Note that where this would be the most frequent case for a population that matches the normal distribution, it is rather rare for our data.

From this analysis we can learn that the applied standardising, although it is superfluous, does not harm the information held by the feature vectors.

[1] It might even be theoretically impossible. Some distributions simply do not have a variance. Examples are the Zipf distribution itself and its continuous equal, the Pareto distribution for certain parameter settings (k<2, which would certainly be the case for trigrams with at least one low count, even when taking smoothing strategies into account).

Table 4.1: Century classifier accuracy using the absolute frequencies of character trigrams (ABS) and the Z-scores of relative trigram frequencies (REL).

learner algorithm	ABS	REL
majority	0.6116	
naive bayes	0.9773	0.9802
KNN	0.8601	0.9462
decision tree	0.9906	0.9885
SVM	0.9849	0.9795

4.2 A model for each corpus?

Now that we have selected two corpora that stem from a different century it is time to make an important choice. Are we going merge the data of these sources when creating our classification models or is it better to create separate models for the 13th and 14th century? To answer this question, it needs to be sorted out in what way the corpora differ. If there are considerable differences, this pleads for a separate treatment.

Moreover, the geographical spread of the locations differs quite a lot for the two corpora: the 13th century features more southern places, while the northern locations make up the largest part of the 14th century material. This encompasses the risk that instead of creating location models, we are in fact (partially) modelling the originating century.

Therefore we assess the mutual measure of the correspondence by setting up an automatic classifier for predicting the century in which a charter was written (and thus to which corpus it belongs). If such a classifier can distinguish both centuries reasonably well, we can assume it makes sense to create separate models further on.

4.2.1 A century recognizer

As features we used the absolute frequency of all character trigrams that appear in one of both corpora. This resulted in vectors of 8407 features. We ignored the difference between capitals and non-capital characters, as in preliminary research we found only a marginal influence of case sensitivity. Furthermore, ignoring capitals was also a very welcome technique to reduce the number of attributes. As an alternative feature set we used the Z-score of the relative frequency counts, as explained in section 4.1.2.

The results for predicting the century are shown in table 4.1. The baseline to compare with is the majority classifier that simply selects the largest class, i.e. the 14th century.

4.2.2 Results, discussion and consequences

The results in table 4.1 make it very clear that we can almost flawlessly derive a document's originating century. When using the absolute trigram frequency count in combination with a decision tree an accuracy rate of about 99% is reached. However, we cannot know for sure that we are actually recognizing the century itself. It might very well be the case that we are modelling some specific properties of the corpora instead[2].

For the rest of this research the actual reason for the easy discrimination between both corpora is not relevant. We can distinguish both centuries, so from here on we will create separate models. This approach eliminates the risk that we are modelling somehow the century or the originating corpus instead of another parameter. Should an unknown document be

[2] There is a minor indication this is not the case. The CRM corpus contains two 13th century documents (P176p-300-01 and P176p-300-01 , both written in 1300). When we added them to the class for the 14th century (which corresponds to the corpus they belong to), the classification accuracy slightly decreased. This does not support the hypothesis that we built models for the corpora instead of the centuries.

classified, then we can first run the century recognizer, followed by the century-dependent classifier.

4.3 Selecting classification methods

An important step for the further experiments is the selection of a machine learning algorithm. We used the Orange data mining toolkit (Demsar & Zupan, 2004) to compare several methods. Note that these preliminary comparisons were made for both corpora at once because we only decided afterwards to create separate models for both centuries. However the relative scores of the classifiers — "which one has the highest accuracy?" — should be valid for century-specific models as well.

4.3.1 Classifiers: an overview

We investigated the following machine learning techniques:

- Naive Bayes, with the default parameter settings.
- Decision Trees, with the default parameter settings.
- K Nearest Neighbours (henceforth "KNN"), with k = 1 as a comparison learnt this gave the best results for our data set.
- Support Vector Machines (SVM), a module within Orange that acts as an interface to LIBSVM (Chang & Lin, 2001). The default Radial Basis Function (RBF) kernel was chosen as this turned out to be the optimal setting.

4.3.2 Classification assessment

4.3.2.1 Setup

As a first exploratory step we used all documents of the 13th and 14th century corpora and applied ten-fold cross-validation to measure the performance of the machine learning algorithms mentioned above. We tried to predict the following classes:

- The decade in which a charter was written.
- The location where a charter was written.

As features we used again the absolute frequency and the the relative frequency count Z-score. Both vector versions exist of 8407 features.

4.3.2.2 Quantitative analysis

The results of this first attempt at basic classification are shown in table 4.2. The baseline to compare with is the majority classifier that simply selects the largest class. We can conclude that the location is better predicted than the decade. Still both tasks are performed on a level that clearly exceeds the baseline, so this gives a firm substantiation for the claim that it is possible to induce the originating time and location for a charter from our collection.

For both the decade and location prediction the combination of KNN and the relative feature variants outperformed all of the other methods. Besides the KNN algorithm is relatively fast and reasonable with regards to the memory usage, so all in all it seems to be a good choice for further classifications on our data set.

Table 4.2: Classifier accuracy using the absolute frequencies of character trigrams (ABS) and the Z-score of the relative trigram frequencies (REL).

	decade (ABS)	decade (REL)	location (ABS)	location (REL)
majority	0.1903		0.2574	
naive bayes	0.5456	0.5583	0.6014	0.6053
KNN	0.4431	**0.5668**	0.5222	**0.7009**
decision tree	0.5209	0.4253	0.2644	0.2579
SVM	0.5043	0.5378	0.6285	0.5678

Table 4.3: Confusion matrix for the decade (ranging from 1230 to 1400, indicated by the first 3 numbers) classifier, using KNN and the Z-score of the relative trigram frequencies. The leftmost column indicates the real class, the header row shows the predicted class.

	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
123	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
124	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
125	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
126	0	0	0	15	17	16	7	0	0	0	0	0	2	1	0	1	0	0
127	1	0	0	4	75	64	33	1	1	0	0	0	0	1	1	1	0	0
128	1	0	1	0	19	392	137	3	4	0	0	1	1	1	0	0	1	1
129	2	0	0	2	21	124	645	9	18	1	5	4	5	2	5	2	1	0
130	1	0	0	0	2	13	32	42	13	2	3	3	4	2	1	0	1	0
131	0	0	0	0	1	2	14	4	35	5	5	8	5	2	1	2	1	0
132	0	0	0	0	0	2	14	1	16	180	3	4	3	5	4	0	2	0
133	0	0	0	0	0	3	15	2	13	4	59	21	13	9	5	3	4	0
134	1	0	0	0	2	7	14	0	16	2	12	104	20	22	10	9	9	0
135	0	0	0	0	1	4	11	0	7	2	3	25	148	33	17	17	16	2
136	1	0	0	0	1	5	9	0	9	1	4	17	31	166	35	34	33	0
137	0	0	1	0	0	2	8	2	8	0	0	9	17	40	176	71	57	1
138	1	0	0	0	2	3	4	0	10	2	1	9	18	16	57	210	112	1
139	1	0	0	0	2	10	14	1	4	0	1	9	13	16	28	82	419	14
140	0	0	0	0	0	0	2	0	4	0	0	2	0	2	0	12	20	11

4.3.2.3 Qualitative analysis

For a qualitative analysis of the optimal feature set and classifier combination we take a closer look at the confusion matrix for the decade prediction in table 4.3. Note that we did the same for the location prediction, but the resulting confusion matrix (for 351 classes) would be too big to include here. However, the observed tendencies are largely the same.

In spite of the relatively low overall accuracy we can see that most predictions are more or less pointing in the right direction, though there is often confusion with the neighbouring decades.

Another important observation can be done for the classes that contain more documents than average: they tend to attract a lot of charters from other classes. Examples of this can be found at the end of both centuries, for which both corpora contain proportionally more documents: 1290-1299 and 1390-1399. The same phenomenon was spotted for the locations: out of 351 classes and 4321 documents, there are 172 documents that are incorrectly classified as belonging to Brugge, which is by far the location class containing the most (exactly 1111) charters.

Furthermore, the classification model has to make a choice between a broad range of classes. This implies that if a wrong decision is made, we cannot know for sure what exactly went wrong. It might be that it is impossible after all to model the characteristics of the real class, or another somehow similar class could have been chosen, or maybe the classifier ended up in a tie between two classes and randomly chose the wrong one.

To counter the latter issue we could take a look behind the scenes of the classification algorithm to find out where things went wrong. However, because we are using that many features (which is an often occurring situation in linguistics) this is not a trivial thing to do. Neither does it save us from the largest classes "sucking up" all of the documents that belong to other classes.

4.3.3 From recognition to verification

4.3.3.1 The verification scenario

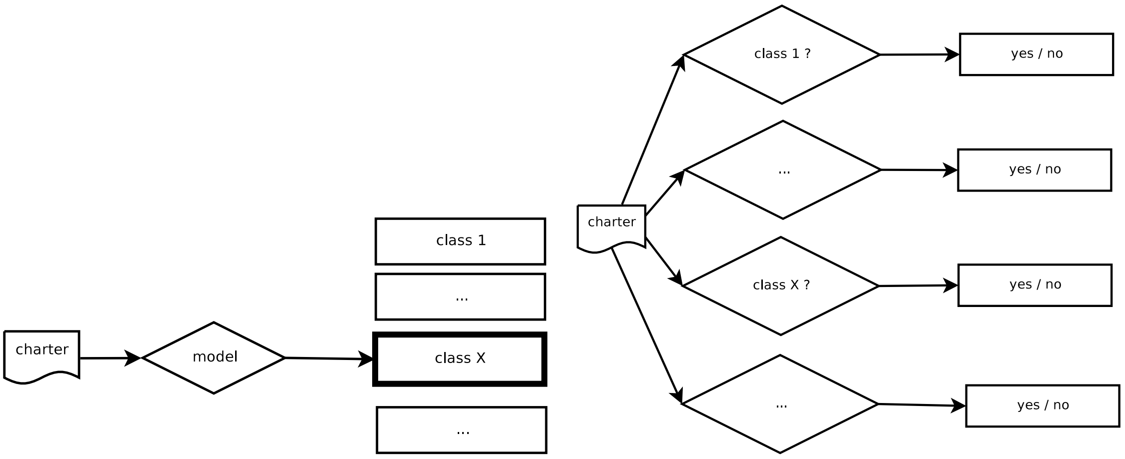
In order to accommodate to the need for a deeper analysis we decided to switch to another classification paradigm. Instead of developing one monolithic model that should recognize every place (or date), we switched to a verification scenario. This means creating a binary model for each class separately that can tell us whether a charter was written in a given location or date. Figure 4.1 illustrates both of these approaches.

An extra advantage that comes with the verification mechanism is the fact that we can spot which models are triggered by a document. This could possibly deliver some more information on the sources of confusion between multiple classes. If a charter is only accepted by its own model and one for a city that is located in the neighbourhood of its real origin, then this might be an indication of the fact that we are modelling the geographical spread. On the other hand, if either no or almost all documents are accepted by a certain model then we know that there is something substantially wrong.

4.3.3.2 Measures to use

Using verification brings with it the possibility to extend the classification accuracy with some extra measures. Amongst them are the False Accept Rate (FAR, the proportion of charters that have been undeservedly accepted by the model) and the False Reject Rate (FRR, documents that do belong to a class but that have not been recognised as such). We define FAR and FRR as follows:

$$FAR = \frac{\text{Number of false positives}}{\text{total number of negative instances}}$$



(a) multiclass model

(b) verification model

Figure 4.1: The recognizer model versus the verification model.

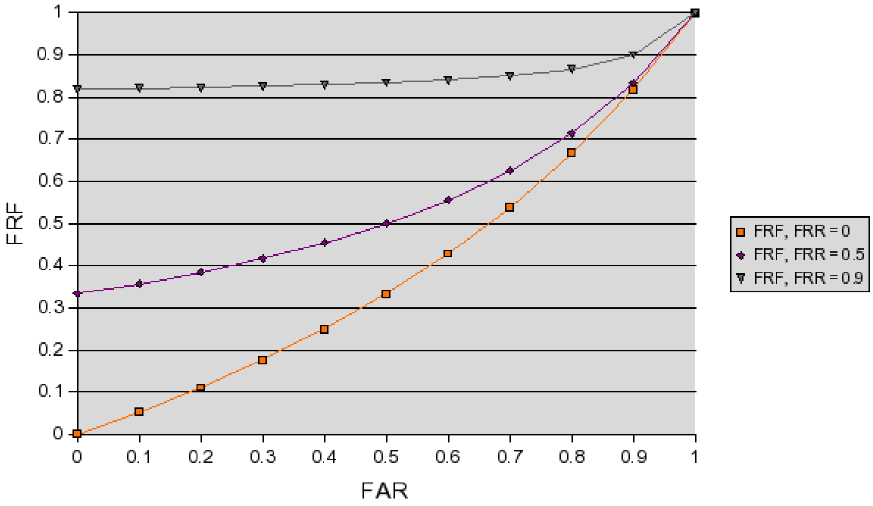


Figure 4.2: FRF-scores in function of the False Accept Rate, for different values of the False Reject Rate.

$$FRR = \frac{\text{Number of false negatives}}{\text{total number of positive instances}}$$

Finally we can combine both FAR and FRR into a self-defined measure called FRF (False Rate F-score), defined as:

$$FRF = 1 - \frac{2(1 - FAR)(1 - FRR)}{(1 - FAR) + (1 - FRR)}$$

FRF gives an indication about the total quality of the model as a whole, between 0 (perfect) and 1 (no predicting power, either all elements are accepted or rejected). As soon as the FAR or FRR equals 1, the FRF score becomes 1 as well. Figure 4.2 illustrates the effect of FAR and FRR on the FRF-score. The same effects can be observed when mutually substituting FAR and FRR in the chart.

4.3.3.3 Linguistic Profiling as an extra classification method

After deciding to use a verification approach, we chose to add specific author verification software to the earlier mentioned list of classifiers to be tested (see section 4.3.1). This made it possible to look if we could get better results using a tool that was specifically designed with verification tasks in mind. Even if this would not be the case, at least it is worthwhile to compare its results with those of the standard classifiers we chose before.

We selected the Linguistic Profiling (henceforth abbreviated as LProf) algorithm, as described in Van Halteren (2007), which is intended to track down plagiarism. This method derives a large number of linguistic features for each document and afterwards compares them to a profile reference corpus. Although a broad range of feature sources can be used (syntactical, pragmatic, etc.) we restricted ourselves for this study to the lexical features. This means de facto that we can reuse the feature vectors that we already created.

As this verification method largely depends on the selection of good parameters, we further on will create a separate parameter tune set while using n-fold cross-validation. That way we can select beforehand a reasonable set of parameters for every fold without the risk of overfitting.

4.4 Class selection

There are two main classes we will consider: time and location. For each of these we need to define exact class boundaries and levels of granularity.

4.4.1 Time classes

4.4.1.1 Decade

We mentioned before the strategy to split up the time dimension (i.e. 1201 - 1401) in intervals of 10 years. However, there is an important issue with this choice as a dating class. As we could see in figure 3.2 the distribution over the decades is seriously skewed, both in the 13th and 14th century corpus. The large majority of the charters in each corpus can be found in the "younger" parts of the corpora. Therefore we would most probably end up in a situation where the only possible conclusion is that we can more accurately predict the decade of writing for younger charters. This suspicion is supported by the preliminary tests we did when making a classifier selection.

As a matter of fact we are in no way obliged to use the completely arbitrary time intervals of ten years. We decided to discard the decade time intervals and looked for some alternatives that were not influenced by the distributional bias.

4.4.1.2 Sliding windows

One of the problems that occurs while using decade-based classification is the arbitrary character of the time slices. In reality diachronic language changes occur continuously. Apart from that the split into decades results in classes whose size differs significantly. To address these issues we also used another categorisation, based on sliding time windows.

The first window starts at the begin of the century. We add step by step the succeeding years to the window, until it contains 100 charters or more. Then we move the begin of the window to the right on the time axis, to the next year for which a charter has been found in this corpus. Again we add the next years, until the limit of 100 documents is reached. This procedure is repeated until the right boundary of the window reaches the end of the century.

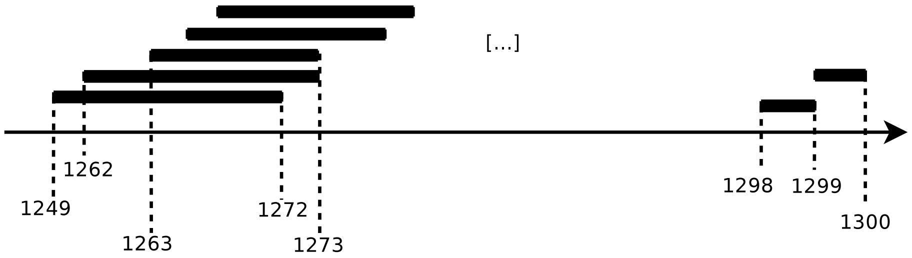


Figure 4.3: Sliding windows as dating classes.

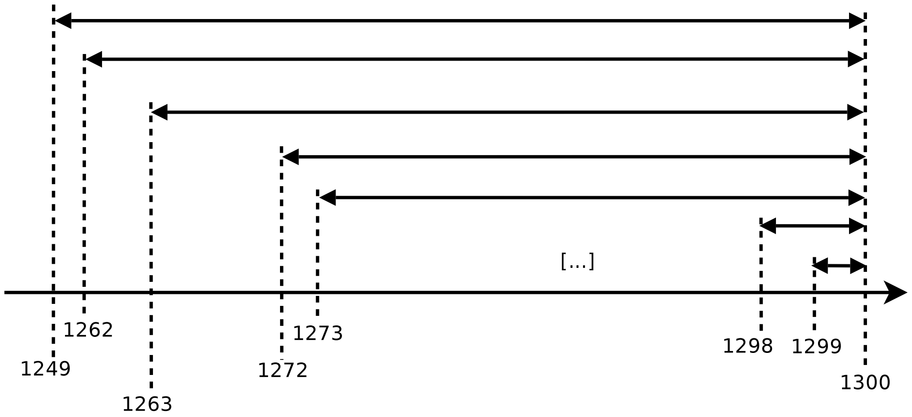


Figure 4.4: Century halves as as dating classes.

This results e.g. in the following time slices for the 13th century: 1236-1272, 1249-1272, 1252-1273, 1253-1273, 1254-1273, 1260-1273, [...], 1298-1299, 1299-1300. An illustration of these time intervals can be found in figure 4.3.

4.4.1.3 Two halves

Yet another approach is splitting a century in two halves and predicting to which part a document belongs. We vary the split point from the begin to the end of the century. In this way we try to find out which years act as a natural border line — i.e. delivers a good prediction result — between the two halves. If such a sharp dividing line is found, it could be the indication of sudden linguistic change, the introduction of a new word, etc. A schematic overview of this approach is given in figure 4.4.

4.4.2 Location classes

Most charters contain the name of the city or village where they have been written. In some cases (especially for the 14th century) this location information could not be derived unambiguously – instead the "region around town X" is used then. This poses a problem towards the creation of homogeneous classes. Therefore we decided to exclude the charters with only a regional localisation. We did the same with all documents from places that occur less than 10 times in one of both corpora.

Although this means we loose some scarce resources it is the only way to keep some consistency with regards to the material used. In the end we depend on the quality of the training data to achieve good classification results.

After this selection we retained the following places:

- For the 13th century: Aalst, Aardenburg, Brugge, Damme, Dordrecht, Eeklo, Evergem, Gent, Geraardsbergen, Grimbergen, Hulst, Maldegem, Mechelen, Ter Doest, Utrecht. These places are indicated on figure 4.5.

- For the 14th century: Amersfoort, Amsterdam, Breda, Brugge, Brussel, Delft, Deventer, Dordrecht, Eersel, Egmond-Binnen, Gemert, Gent, Gouda, Groningen, Haarlem, Halen, Hasselt, Helmond, Heusden, Hoorn, Kampen, Leiden, Lummen, Maaseik, Maastricht, Middelburg, Schelle, Sint-Truiden, Utrecht, Venlo, Vught, Walem, Wijk bij Duurstede, Zutphen, Zwolle, 's-Gravenhage, 's-Gravenzande, 's-Hertogenbosch. These places are indicated on figure 4.6.

4.5 Test methodology

Given the fact the we rely on scarce historical sources we would like to use these in an optimal way. That is why we decided to use 10-fold cross-validation for testing the classifiers. This combines the 100% use of all gathered informaton with a trustworthy estimation of the classification accuracy. Concretely we proceeded as follows for each fold:

- 80% of the instances was used as training data
- 10% of the instances was used as test data
- When using LProf, the final 10% of the instances was used to determine the optimal parameter settings.

For the standard classification algorithms, we trained a verification model with the training data and benchmarked the test data against it.

With LProf, we first tested the training data with all possible parameter settings. The parameters that gave the best results for the instances of the tuning set were selected for this fold. Finally we measured how well the model (using the chosen optimal parameters) worked for the test data.

One could remark that with this setup the LProf verification has a slight advantage. That is partially true, however given its strong reliance on the choice of the parameters (Van Halteren, 2007) this was the only feasible way of comparing it to the other classifiers. Using the last 10% of the instances as extra training material for KNN would give the latter an even larger advantage. Therefore the KNN classifier will not use the tuning instances at all.

4.6 Inducing the location

We now have made choices with regards to the classes, the features, the algorithms and the way of organising the classification experiment. So the time has come to run the experiment and to interpret the outcomes. As we already managed to make an almost flawless distinction between the 13th and 14th century, the decision was made to create separate classification models for the charters from both centuries. This excludes potential unwanted influences of the mutual differences between the corpora used on the results.

First of all we focus on the recognition of the place where a document was written. As separate models are created for the 13th and 14th century we will discuss the associated models separately.

Figure 4.5: Map with the 13th century locations.

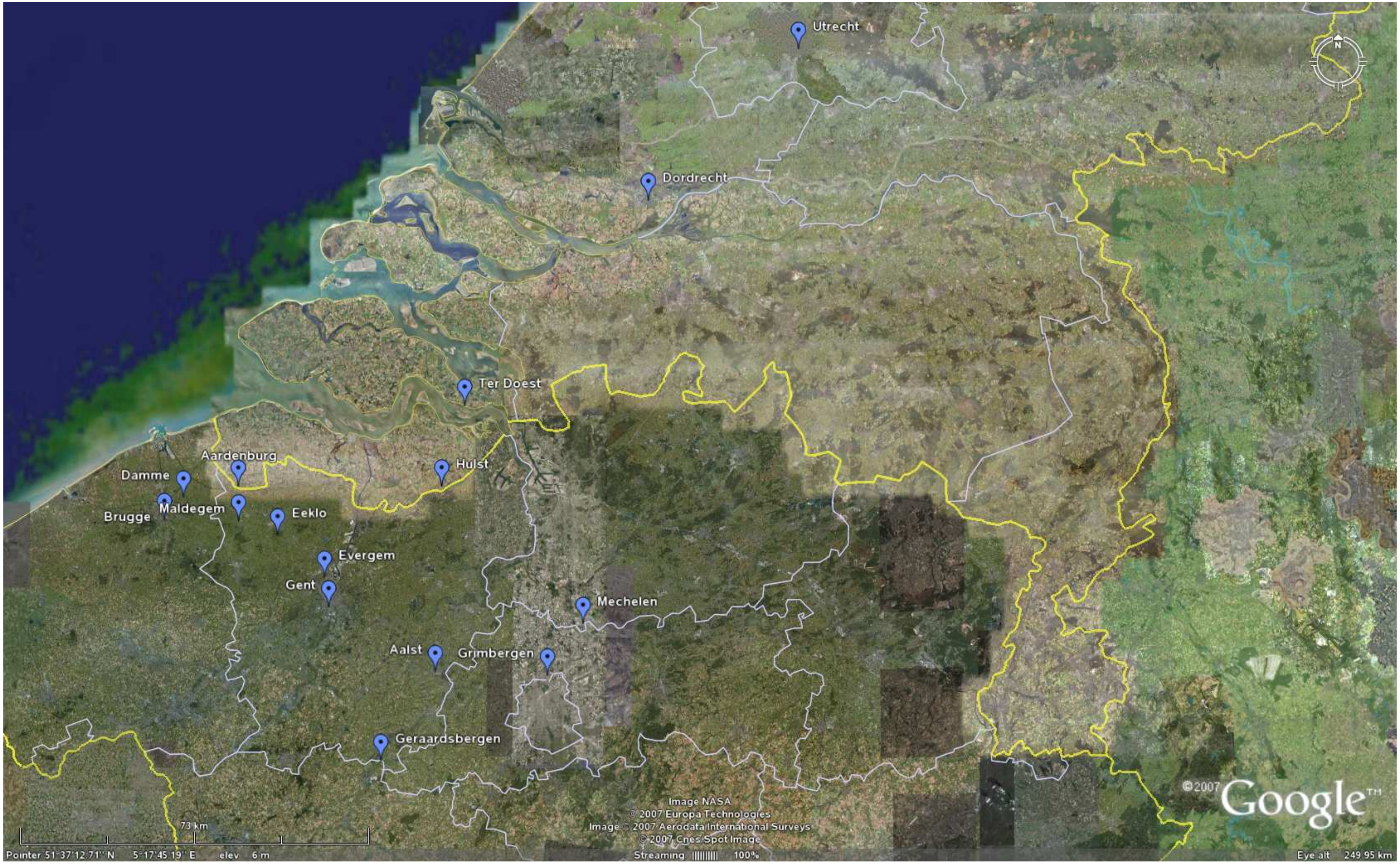


Figure 4.6: Map with the 14th century locations.

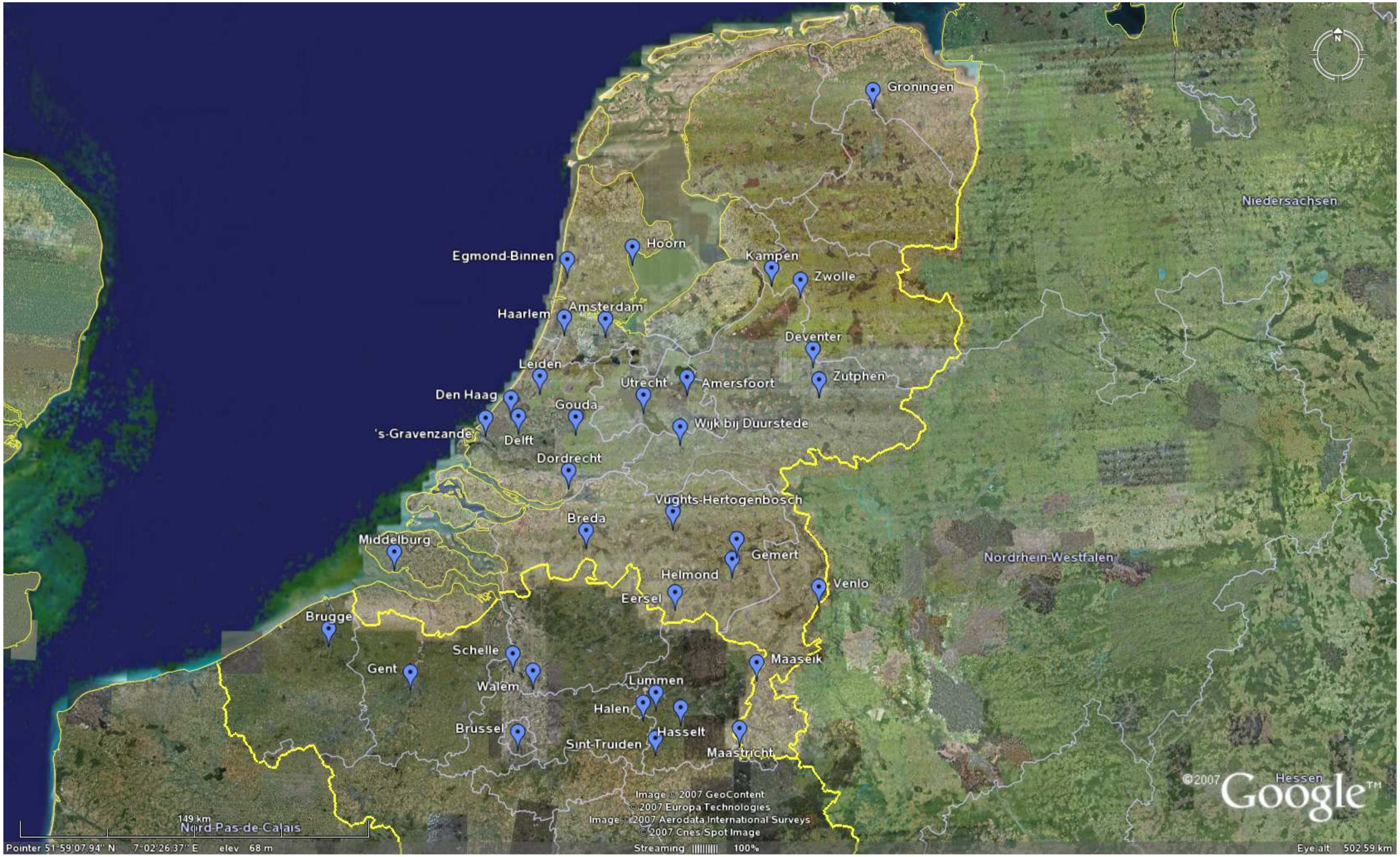


Table 4.4: Location recognition results for the 13th century corpus

		LProf				KNN			
Location	charters	Error rate	FAR	FRR	FRF	Error rate	FAR	FRR	FRF
Aalst	11	0.0038	0.0008	0.4000	**0.2502**	0.0053	0.0000	0.7000	0.5385
Aardenburg	24	0.0045	0.0000	0.3000	**0.1765**	0.0098	0.0046	0.3500	0.2136
Brugge	997	0.1000	0.2364	0.0545	**0.1551**	0.1030	0.3606	0.0172	0.2252
Damme	10	0.0258	0.0214	0.6000	**0.4321**	0.0076	0.0023	0.7000	0.5387
Dordrecht	57	0.0083	0.0016	0.1800	0.0995	0.0121	0.0087	0.1000	**0.0565**
Eeklo	12	0.0061	0.0023	0.5000	**0.3338**	0.0053	0.0008	0.6000	0.4287
Evergem	10	0.0038	0.0000	0.5000	**0.3333**	0.0053	0.0000	0.7000	0.5385
Gent	78	0.1576	0.1608	0.1000	**0.1315**	0.0409	0.0080	0.6286	0.4595
Geraardsbergen	16	0.0038	0.0000	0.5000	0.3333	0.0038	0.0000	0.5000	0.3333
Grimbergen	20	0.0061	0.0000	0.4000	0.2500	0.0053	0.0031	0.1500	**0.0824**
Hulst	11	0.0045	0.0000	0.6000	0.4286	0.0045	0.0000	0.6000	0.4286
Maldegem	12	0.0053	0.0023	0.4000	**0.2506**	0.0038	0.0000	0.5000	0.3333
Mechelen	75	0.0371	0.0312	0.1429	**0.0904**	0.0159	0.0016	0.2714	0.1576
Ter Doest	20	0.0311	0.0246	0.4500	**0.2966**	0.0106	0.0023	0.5500	0.3798
Utrecht	18	0.0083	0.0038	0.6000	**0.4292**	0.0076	0.0000	1.0000	1.0000

4.6.1 Location for the 13th century

4.6.1.1 Quantitative analysis

For all of the 14 locations that were mentioned at least 10 times in the 13th century corpus, we created a recognition model. In other words, we measured for each document the likelihood that it was written in one of these 14 places. This means that one charter can possibly assigned to multiple places. If this is the case, we can attempt to explain this confusion. Ideally this confusion can be attributed to the geographical proximity as this shows that we are effectively modelling the spatial spread.

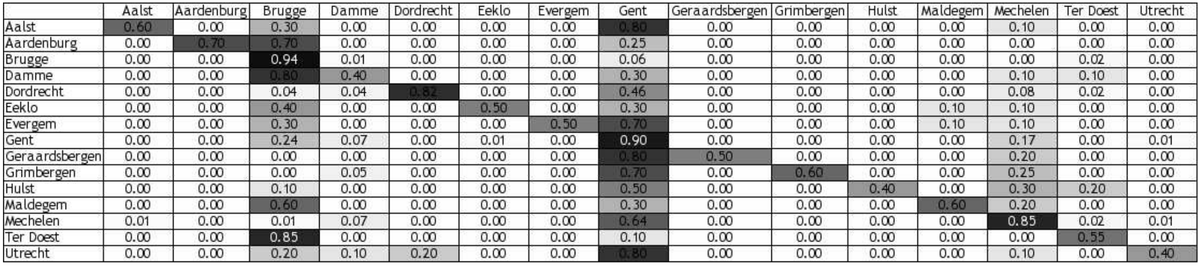
We created these location recognizers with the Linguistic Profiling (LProf) and K Nearest Neighbours (KNN) algorithms. Both use the relative character trigrams described above as feature vectors. The outcome for each location model is given in table 4.4. Apart from the Error rate and FAR, FRR and FRF scores we also added the exact number of the charters in each location.

However, we also want to give a more intuitive presentation of the classification results. Therefore we present table 4.5 and 4.6, where we visually mark the acceptance of a set of documents (the rows) by a model (the columns) with a degree of darkness. The darker a cell, the more of these charters were recognised by the model mentioned in the column heading. Ideally, the cells on the diagonal are all black while the others are white as this would mean that all models perfectly recognize "their" documents.

Looking at the classification results, we can draw the following cautious conclusions:

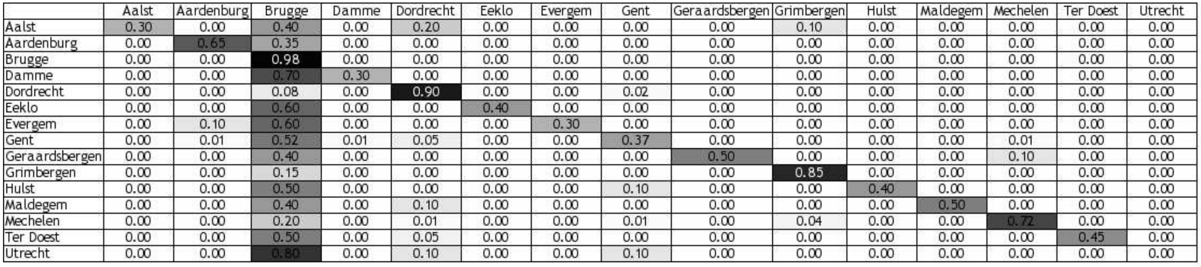
- Both classifiers manage to provide a reasonable to good recognition. In terms of the mean FRF-score Linguistic Profiling (0.27) outperforms K Nearest Neighbours (0.40).
- The more charters available, the better the place recognition.
- LProf tends to have a higher false positive rate than KNN, especially for those locations with many documents (Brugge, Gent). However the latter more frequently fails to recognize a document's real class.

Table 4.5: Confusion matrix for the location recognition in the 13th century material (LProf). Each column represents a recognition model.



	Aalst	Aardenburg	Brugge	Damme	Dordrecht	Eeklo	Evergem	Gent	Geraardsbergen	Grimbergen	Hulst	Maldegem	Mechelen	Ter Doest	Utrecht
Aalst	0.60	0.00	0.30	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.10	0.00	0.00
Aardenburg	0.00	0.70	0.70	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brugge	0.00	0.00	0.94	0.01	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.02	0.00
Damme	0.00	0.00	0.80	0.40	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.10	0.10	0.00
Dordrecht	0.00	0.00	0.04	0.04	0.82	0.00	0.00	0.46	0.00	0.00	0.00	0.00	0.08	0.02	0.00
Eeklo	0.00	0.00	0.40	0.00	0.00	0.50	0.00	0.30	0.00	0.00	0.00	0.10	0.10	0.00	0.00
Evergem	0.00	0.00	0.30	0.00	0.00	0.00	0.50	0.70	0.00	0.00	0.00	0.10	0.10	0.00	0.00
Gent	0.00	0.00	0.24	0.07	0.00	0.01	0.00	0.90	0.00	0.00	0.00	0.00	0.17	0.00	0.01
Geraardsbergen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.50	0.00	0.00	0.00	0.20	0.00	0.00
Grimbergen	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.70	0.00	0.60	0.00	0.00	0.25	0.00	0.00
Hulst	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.40	0.00	0.30	0.20	0.00
Maldegem	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.60	0.20	0.00	0.00
Mechelen	0.01	0.00	0.01	0.07	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	0.85	0.02	0.01
Ter Doest	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.55	0.00
Utrecht	0.00	0.00	0.20	0.10	0.20	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.10	0.00	0.40

Table 4.6: Confusion matrix for the location recognition in the 13th century material (KNN).



	Aalst	Aardenburg	Brugge	Damme	Dordrecht	Eeklo	Evergem	Gent	Geraardsbergen	Grimbergen	Hulst	Maldegem	Mechelen	Ter Doest	Utrecht
Aalst	0.30	0.00	0.40	0.00	0.20	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
Aardenburg	0.00	0.65	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brugge	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Damme	0.00	0.00	0.70	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dordrecht	0.00	0.00	0.08	0.00	0.90	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eeklo	0.00	0.00	0.60	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Evergem	0.00	0.10	0.60	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gent	0.00	0.01	0.52	0.01	0.05	0.00	0.00	0.37	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Geraardsbergen	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.10	0.00	0.00
Grimbergen	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.00
Hulst	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.40	0.00	0.00	0.00	0.00
Maldegem	0.00	0.00	0.40	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00
Mechelen	0.00	0.00	0.20	0.00	0.01	0.00	0.00	0.01	0.00	0.04	0.00	0.00	0.72	0.00	0.00
Ter Doest	0.00	0.00	0.50	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.00
Utrecht	0.00	0.00	0.80	0.00	0.10	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00

4.6.1.2 Qualitative analysis

Now that we have an impression on the classification accuracy, the question remains how the models that we use actually manage to recognize the originating place. The algorithms that we are using do not allow a quick insight in their internal behaviour, certainly not given the fact that we are using thousands of features. In order to cope with this, we try to derive the most distinctive and informative attributes using the RELIEF score (Kononenko, 1994). This heuristic supervised attribute selection algorithm has proved to be very efficient in a large set of machine learning problems.

We searched for the "top 5 features" (those with the highest RELIEF score) in all the feature vectors for the 13th century model for the location Mechelen, as this class contains a reasonable amount of charters (75) and reaches a good FRF score (0.0904 with LProf). For each of these top features we looked up the most frequent words that contain the character trigram (i.e. the feature). This gives an impression of each trigram's most natural context. The five most informative features for the Mechelen verification model are:

1. *mac* [relief-score 0.221]: mach (315), macht (168), mach. (34), machline (26), machenaers (25)

2. *ech* [0.200]: wech (643), sculdech (335), recht (275), rechte (233), rechten (220)

3. *wi* [0.198]: wie (2937), wi (1905), willem (1201), willen (723), wijf (648)

4. *nen* [0.197]: sinen (1099), scepenen (1065), ghenen (583), enen (563), binnen (436)

5. *rth* [0.196]: berthout (20), northalf (14), marthe. (10), maerthe (10), berthoude (10)

Most of theses character trigrams are directly related to the name of the location (including variations): Machline, Macheline, Mechelen, etc. The *rth* trigram can be linked to the name Jan Berthout, who was an important knight living in Mechelen, as the first sentence of charter 1269 shows: *Wi Jan berthout riddre [...].*[3]

For *nen* and *wi* ("we"), we don't have a direct explanation. According to Mooijaart (1992), the variant *wi* does almost not occur in West Flanders, so that might give the classifier a location hint. However, the main conclusion of this sample study is (not really unexpected) that proper nouns form the most important clue for the location classifier. The trigrams directly derived from the city name are the most indicative.

4.6.2 Location for the 14th century

4.6.2.1 Quantitative analysis

Analogously to the procedure for the 13th century documents we also assessed the performance of a set of localisation models for the 14th century corpus. Again we tested both LProf and KNN, of which the respective scores can be found in table 4.8 and 4.9. An comparison of both methods is shown in table 4.7.

The results for the localisation of the 14th century charters mainly come down to:

- Very accurate recognition of the originating place, on average better than that for the 13th century.
- Contrary to the mean FRF-scores for the 13th century, KNN scores better (0.11) than Linguistic Profiling (0.21).
- Again most confusions occur between locations that are situated near to each other, especially for the KNN models.
- In general, a recognizer scores worse when there is less training data. This clearly is the case for Egmond-Binnen, Hasselt and Sint-Truiden – both for LProf and KNN. However for some other places (like the KNN model for Zutphen), there is no such a relation.
- The Schelle KNN-based model seems to suffer from quite a few false accepts.[4]

4.6.2.2 Qualitative analysis

Like we did for the 13th century locations, we also took a closer look at one location that reached a good verification result, Gouda in this case (112 charters, FRF score 0.0185 for KNN).

1. rfr [relief-score: 0.594]: erfrenten (84), erfrecht (22), eyrfrogghen (1), erfrogghen (1), erfrogghe (1)
2. fre [0.331]: erfrenten (84), erfrecht (22), frederic (20), frederikes (5), lijfrente (4)
3. yed [0.327): lyede (88), lyeden (14), belyede (14), verlyeden (10), meyedach (8)
4. lye [0.317]: lyede (88), lyen (18), lyeden (14), lye (14), belyede (14)
5. ork [0.250]: orkonden (157), orkonde (95), orkunde (80), oorkonden (18), sporkele (15)
6. gou [0.246]: goude (246), vergouden (35), gouden (35), gouuerneerres (9), gouts (8)

[3] We found extra evidence for the function of Jan Berthout in the publication *Brabantsche yeesten* on the cd-rom on Middle Dutch (Instituut voor Nederlandse Lexicologie, 1998).

[4] In first instance we thought this might be due to the similarity of the place name and an often occurring money unity, called *schelling*. However, an alternative verification model that was trained without the proper names (to which Schelle obviously belongs) still showed a high false accept rate. This falsifies the assumption that the occurrence of *schelling* triggers the acceptation of Schelle as location.

Table 4.7: Location recognition results for the 14th century corpus

		LProf				KNN			
Location	charters	Errors	FAR	FRR	FRF	Errors	FAR	FRR	FRF
Amersfoort	23	0.0053	0.0015	0.2500	0.1434	0.0045	0.0038	0.0500	0.0275
Amsterdam	98	0.0500	0.0415	0.1667	0.1084	0.0076	0.0041	0.0556	0.0305
Breda	88	0.0174	0.0102	0.2000	0.1152	0.0152	0.0047	0.2800	0.1645
Brugge	115	0.0076	0.0000	0.1667	0.0909	0.0083	0.0008	0.1667	0.0912
Brussel	67	0.0220	0.0126	0.2600	0.1540	0.0144	0.0016	0.3400	0.2053
Delft	76	0.0144	0.0072	0.1429	0.0800	0.0106	0.0016	0.1714	0.0944
Deventer	69	0.0083	0.0000	0.1833	0.1009	0.0114	0.0095	0.0500	0.0302
Dordrecht	85	0.0212	0.0113	0.1750	0.1005	0.0174	0.0145	0.0625	0.0391
Eersel	61	0.0053	0.0000	0.7000	0.5385	0.0000	0.0000	0.0000	0.0000
Egmond-Binnen	13	0.0038	0.0000	0.5000	0.3333	0.0038	0.0000	0.5000	0.3333
Gemert	16	0.0045	0.0008	0.5000	0.3335	0.0030	0.0008	0.3000	0.1767
Gent	42	0.0114	0.0038	0.5000	0.3342	0.0114	0.0069	0.3000	0.1788
Gouda	112	0.0386	0.0314	0.1182	0.0768	0.0030	0.0000	0.0364	0.0185
Groningen	73	0.0053	0.0008	0.1200	0.0642	0.0015	0.0000	0.0400	0.0204
Haarlem	58	0.0076	0.0000	0.2500	0.1429	0.0068	0.0000	0.2250	0.1268
Halen	17	0.0045	0.0008	0.5000	0.3335	0.0015	0.0000	0.2000	0.1111
Hasselt	14	0.0098	0.0061	0.5000	0.3347	0.0045	0.0000	0.6000	0.4286
Helmond	41	0.0068	0.0016	0.2333	0.1327	0.0023	0.0008	0.0667	0.0348
Heusden	42	0.0038	0.0000	0.2500	0.1429	0.0000	0.0000	0.0000	0.0000
Hoorn	16	0.0053	0.0015	0.5000	0.3337	0.0015	0.0000	0.2000	0.1111
Kampen	39	0.0038	0.0000	0.2500	0.1429	0.0023	0.0000	0.1500	0.0811
Leiden	117	0.0462	0.0388	0.1273	0.0852	0.0189	0.0149	0.0636	0.0399
Lummen	20	0.0053	0.0000	0.3500	0.2121	0.0023	0.0015	0.0500	0.0264
Maaseik	23	0.0061	0.0015	0.6000	0.4288	0.0008	0.0000	0.1000	0.0526
Maastricht	45	0.0098	0.0008	0.4000	0.2502	0.0008	0.0000	0.0333	0.0169
Middelburg	41	0.0144	0.0107	0.5000	0.3357	0.0038	0.0000	0.5000	0.3333
Schelle	14	0.0038	0.0000	0.5000	0.3333	0.0545	0.0550	0.0000	0.0283
Sint-Truiden	15	0.0083	0.0038	0.6000	0.4292	0.0053	0.0000	0.7000	0.5385
Utrecht	80	0.0061	0.0016	0.1000	0.0533	0.0121	0.0119	0.0167	0.0143
Venlo	14	0.0045	0.0000	0.6000	0.4286	0.0000	0.0000	0.0000	0.0000
Vught	11	0.0030	0.0000	0.4000	0.2500	0.0000	0.0000	0.0000	0.0000
Walem	16	0.0038	0.0000	0.5000	0.3333	0.0015	0.0000	0.2000	0.1111
Wijk bij Duurstede	13	0.0023	0.0000	0.3000	0.1765	0.0015	0.0008	0.1000	0.0530
Zutphen	93	0.0159	0.0078	0.2750	0.1622	0.0227	0.0000	0.7500	0.6000
Zwolle	56	0.0144	0.0094	0.1750	0.0997	0.0083	0.0039	0.1500	0.0827
′s-Gravenhage	53	0.0114	0.0055	0.1600	0.0893	0.0045	0.0008	0.1000	0.0530
′s-Gravenzande	46	0.0045	0.0000	0.1500	0.0811	0.0045	0.0031	0.0500	0.0271
′s-Hertogenbosch	29	0.0053	0.0000	0.3500	0.2121	0.0045	0.0023	0.1500	0.0821

Table 4.8: Confusion matrix for the location recognition within the 14th century material, using LProf

	Amersfoort	Amsterdam	Breda	Brugge	Brussel	Delft	Deventer	Dordrecht	Eersel	Egmond-Binnen	Gemert	Gent	Gouda	Groningen	Haarlem	Halen	Hasselt	Helmond	Heusden
Amersfoort	0.75	0.05	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amsterdam	0.00	0.83	0.00	0.00	0.00	0.04	0.00	0.03	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00
Breda	0.00	0.02	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Brugge	0.00	0.02	0.00	0.83	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Brussel	0.00	0.00	0.02	0.00	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Delft	0.00	0.06	0.01	0.00	0.00	0.86	0.00	0.03	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00
Deventer	0.00	0.02	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00
Dordrecht	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00
Eersel	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Egmond-Binnen	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemert	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gent	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.05	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gouda	0.00	0.05	0.04	0.00	0.00	0.03	0.00	0.02	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00
Groningen	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00
Haarlem	0.00	0.15	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.75	0.00	0.00	0.00	0.00
Halen	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00
Hasselt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.50	0.00	0.00
Helmond	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.00
Heusden	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75
Hoorn	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00
Kampen	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leiden	0.00	0.06	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.01	0.00	0.00
Lummen	0.00	0.05	0.05	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
Maaseik	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maastricht	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
Middelburg	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Schelle	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sint-Truiden	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Utrecht	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Venlo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vught	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
Walem	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wijk bij Duurstede	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zutphen	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.00	0.00
Zwolle	0.00	0.08	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
's-Gravenhage	0.00	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00
's-Gravenzande	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
's-Hertogenbosch	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00

	Hoorn	Kampen	Leiden	Lummen	Maaseik	Maastricht	Middelburg	Schelle	Sint-Truiden	Utrecht	Venlo	Vught	Walem	Wijk bij Duurstede	Zutphen	Zwolle	's-Gravenhage	's-Gravenzande	's-Hertogenbosch
Amersfoort	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amsterdam	0.02	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Breda	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brugge	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brussel	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Delft	0.00	0.00	0.17	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deventer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.00	0.00	0.00
Dordrecht	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eersel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Egmond-Binnen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemert	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gent	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gouda	0.00	0.00	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
Groningen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00
Haarlem	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Halen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hasselt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.20	0.00	0.00	0.00
Helmond	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Heusden	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hoorn	0.50	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kampen	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.00	0.00
Leiden	0.00	0.00	0.87	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lummen	0.00	0.00	0.00	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maaseik	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maastricht	0.00	0.00	0.00	0.00	0.07	0.60	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
Middelburg	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Schelle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sint-Truiden	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Utrecht	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Venlo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00
Vught	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Walem	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
Wijk bij Duurstede	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00
Zutphen	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.73	0.05	0.00	0.00	0.00
Zwolle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.83	0.00	0.00	0.00
's-Gravenhage	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.00
's-Gravenzande	0.00	0.00	0.05	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.85	0.00
's-Hertogenbosch	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.65

Table 4.9: Confusion matrix for the location recognition within the 14th century material, using KNN

	Amersfoort	Amsterdam	Breda	Brugge	Brussel	Delft	Deventer	Dordrecht	Eersel	Egmond-Binnen	Gemert	Gent	Gouda	Groningen	Haarlem	Halen	Hasselt	Helmond	Heusden
Amersfoort	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amsterdam	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Breda	0.00	0.00	0.72	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.02	0.00
Brugge	0.00	0.00	0.01	0.83	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brussel	0.00	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Delft	0.01	0.00	0.00	0.00	0.00	0.82	0.00	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deventer	0.00	0.00	0.01	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dordrecht	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eersel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Egmond-Binnen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemert	0.10	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gouda	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00
Groningen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00
Haarlem	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00
Halen	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00
Hasselt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.40	0.00	0.00
Helmond	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00
Heusden	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Hoorn	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kampen	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leiden	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lummen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maaseik	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maastricht	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Middelburg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Schelle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sint-Truiden	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Utrecht	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Venlo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vught	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Walem	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wijk bij Duurstede	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zutphen	0.05	0.00	0.02	0.00	0.00	0.00	0.22	0.07	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zwolle	0.02	0.00	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
's-Gravenhage	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
's-Gravenzande	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
's-Hertogenbosch	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	Hoorn	Kampen	Leiden	Lummen	Maaseik	Maastricht	Middelburg	Schelle	Sint-Truiden	Utrecht	Venlo	Vught	Walem	Wijk bij Duurstede	Zutphen	Zwolle	's-Gravenhage	's-Gravenzande	's-Hertogenbosch
Amersfoort	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amsterdam	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Breda	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.16	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brugge	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brussel	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Delft	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
Deventer	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dordrecht	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eersel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Egmond-Binnen	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.20	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemert	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
Gent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gouda	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Groningen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
Haarlem	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.02	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Halen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hasselt	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Helmond	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
Heusden	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hoorn	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kampen	0.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00
Leiden	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lummen	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maaseik	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maastricht	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
Middelburg	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.20	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
Schelle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sint-Truiden	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Utrecht	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Venlo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vught	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Walem	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00
Wijk bij Duurstede	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00
Zutphen	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.25	0.05	0.02	0.00	0.05
Zwolle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.00	0.00
's-Gravenhage	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.00
's-Gravenzande	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00
's-Hertogenbosch	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85

The most informative trigrams are derived from *erfrenten* ("hereditary rent-charge") and *lyede* ("to execute"). Then *orkonden* ("charters" or "to publish a charter") follows. We added the sixth ranked feature here on purpose, because it is derived from the name of the location (*Goude*).

Erfrenten appears to give a semantic indication about the location: it seems most charters about this subject (that still exist today) were written in Gouda.

Lyede seems to be a very rarely used form outside Gouda. It only occurs three times in other locations in the whole 14th century corpus: in E724r (region IJsselstein), F585r (region Hasselt) and L237p (Helmond). As this word does not have a particular unique meaning and as its orthographic variants (*lijde*, *lide*) are more widely spread, we are inclined to look at it as a good orthographic indicator of the location.

The same more or less goes for *orkonde(n),* which has a wide variety of spelling variants and is obviously very frequently used.

Summing up, we found the following categories of localisation clues in the character trigrams:

- Proper nouns (specifically city and village names).
- Semantic clues: unique subjects in the corpus.
- Orthographic clues: deviations from a frequently used spelling.

Especially the last category shows it might be possible to derive a charter's location from the orthography.

4.6.3 Optimising KNN

One of the remarkable results we found is that for the 13th century charters the KNN classifier performs worse than LProf while it is the other way around for the 14th century. To find out whether the performance of KNN can be improved by tuning some parameters, we carried out an extra experiment. Instead of choosing the class of an instance's nearest neighbours (k = 1), we checked the 10 nearest neighbours and their classification. This was done for the 13th century charters from Gent, as these got a low FRF-score using KNN (0.34) in comparison with LProf (0.06).

When we have the 10 nearest neighbours at our disposal, we can try to find an optimal threshold for the verification of the document (e.g., as soon as 2 neighbours out of the 10 belong to this class we accept the document under investigation). However, a thorough comparison of the manually selected optimal threshold with the results for k = 1 revealed that the latter is already the optimal parameter setting for this test case. In fact, for almost all the charters from Gent there only was one positive neighbour (out of the 10). Only charter 210 had 2 positive neighbours. Thus, we can conclude that it is not worth the effort to select some threshold before applying KNN classification to our data.

4.6.4 The influence of the size and number of the charters on the verification quality

A potential reason for differences in the verification quality of the models we developed is the difference in size of the charters. A longer document could reveal more information. The number of the charters in each class might influence the classification as well: one could expect that more training material delivers better results. To see whether these suppositions might be true, we performed a correlation analysis[5] between both factors mentioned (mean document size and class size) and the outcome of the location verification (using the FRF score).

[5] We calculated the Pearson product-moment correlation coefficient using R (R Development Core Team, 2006).

Table 4.10: correlations between the class size, document length and FRF-score for LProf

	number of documents in a class	mean document length
FRF 13th century	-0.5231, p=0.0454	-0.3132987, p=0.2555
FRF 14th century	-0.6460083, p=0.00001184	0.4232611, p=0.008103

Table 4.11: correlations between the class size, document length and FRF-score for KNN

	number of documents in a class	mean document length
FRF 13th century	0.1773261, p=0.5272	-0.02463928, p=0.9305
FRF 14th century	-0.1013910, p=0.5447	0.2150307, p=0.1948

It should be stressed that if such a correlation is found, it does not imply by any means a causal effect. In other words: if there is a connection between the classification quality and one of these parameters, the better results are not necessarily caused by the mere size of the documents and/or classes. However, if such a correlation is not found, we can falsify the assumptions on such a causal effect.

The results of our analyses can be found in table 4.10 (for LProf) and 4.11 (for KNN). For the KNN verification, no significant correlation can be found. The LProf verification table shows a negative correlation[6] between the number of documents within a class and the FRF score. So the more documents we have at our disposal, the better the verification. This fits with the intuitive assumption that more data leads to a better model. However there also appears to be a small positive correlation between the mean document length and the FRF score for the 14th century corpus. Strangely enough, this implies that shorter documents lead to better verification results than longer ones. It is difficult to find a satisfying explanation for this result. It might be the case that the trigrams derived from a location name become more prominent if charter contains fewer other (distracting) words. Nevertheless, there are some clear conclusions we can draw from the correlation results as a whole:

- Larger charters do not lead to better verification results.
- For Linguistic Profiling, a good verification score seems to be related to a higher number of charters in a class.

4.7 Inducing the date

Now that we have proved that it is possible to predict the location using character trigrams with satisfactory results, we intend to do the same for the date of the charters. As explained before we will consider verification models for succeeding overlapping time intervals that contain at least 100 charters and for all the charters that were written before/after a specific year.

4.7.1 Interpretation of the confusion matrices

Because the year intervals result in a large number of verification models, the resulting confusion matrices won't show the recognition rate anymore as a number, but only a corresponding colour code. If a model correctly recognizes a document the corresponding cell becomes green. A false acceptance is marked by a red-coloured cell. Like the grayscale tables, the darkness of a colour is proportional to the degree of charters accepted by a certain model.

[6] Assuming a significance level of $p < 0.05$.

This approach allows us to analyse the outcome of all verification models as one entity. The darkness of the green cells then gives an indication about the recall, while the redness of the table displays the precision of the system as a whole (the less red, the higher the precision).

It should also be noticed that at the left side we do not present the year intervals but each year that is observed in the corpus separately, in increasing order. Ideally such a confusion matrix has a dark green diagonal with a width that is proportional to the number of years in each interval.

4.7.2 Overlapping intervals

4.7.2.1 Intervals for the 13th century

In table 4.12 an overview is given of the verification task for the overlapping intervals of the 13th century.

First, Linguistic Profiling delivers quite reasonable results[7]. We can clearly recognize the green diagonal, indicating that the verification works for the largest part. The upper left corner contains far more red than the rest of the table. This can be explained by the fact that there are not that many documents for the earliest years of the 13th century in our corpora, so the intervals are broader and thus the classification becomes more diffuse. Another observation is the fact that most interval models bring along false positives, as shown by the number of red columns. The extreme results for the top lines (completely red when all documents are accepted or white if they all are rejected) are due to the fact that the first years are only represented by a few charters.

Second, KNN completely misses the ball[8]. Its interval models do not recognize anything. Between the intervals [1270, 1277] and [1277,1280] all charters are — incorrectly — accepted. The same thing, although on a smaller scale, happens for [1295, 1296] and [1296, 1297].

Because KNN gave disappointing results, we ran an extra experiment to find out if the same would happen for a more consistent data set. As a very large subset of the 13th century corpus originates from Brugge, we set up interval verification for only this city. So instead of looking at diachronic variation over a set of documents from all locations, we kept the location factor constant and focused at variation over time. This resulted in a set of models that gives acceptable results[9], as the recognisable green diagonal proves. Still, a lot of charters are not accepted by the models – the diagonal is only light green and in fact hardly any recognition rate surpasses 50%.

Ideally we would apply the same principle as we did for the dating of the Brugge charters to the localisation task: measuring the localisation performance for a fixed point in the time. However, our data is unfortunately too scarce to achieve this: there is no single year for which there is an equally large set of locations available.

4.7.2.2 Intervals for the 14th century

Table 4.13 shows the results for the sliding year intervals for the 14th century. The picture is more or less the same for the LProf outcome[10], the charters are for the largest part correctly accepted by their model (93% for the first intervals, about 70% for the last ones), but too much charters are falsely accepted by other models as well.

Unfortunately time and computing resource limits forced us to restrict the assessment for the KNN dating models to a sample[11]. The KNN confusion matrix differs significantly from

[7] average FRF-score: 0.29

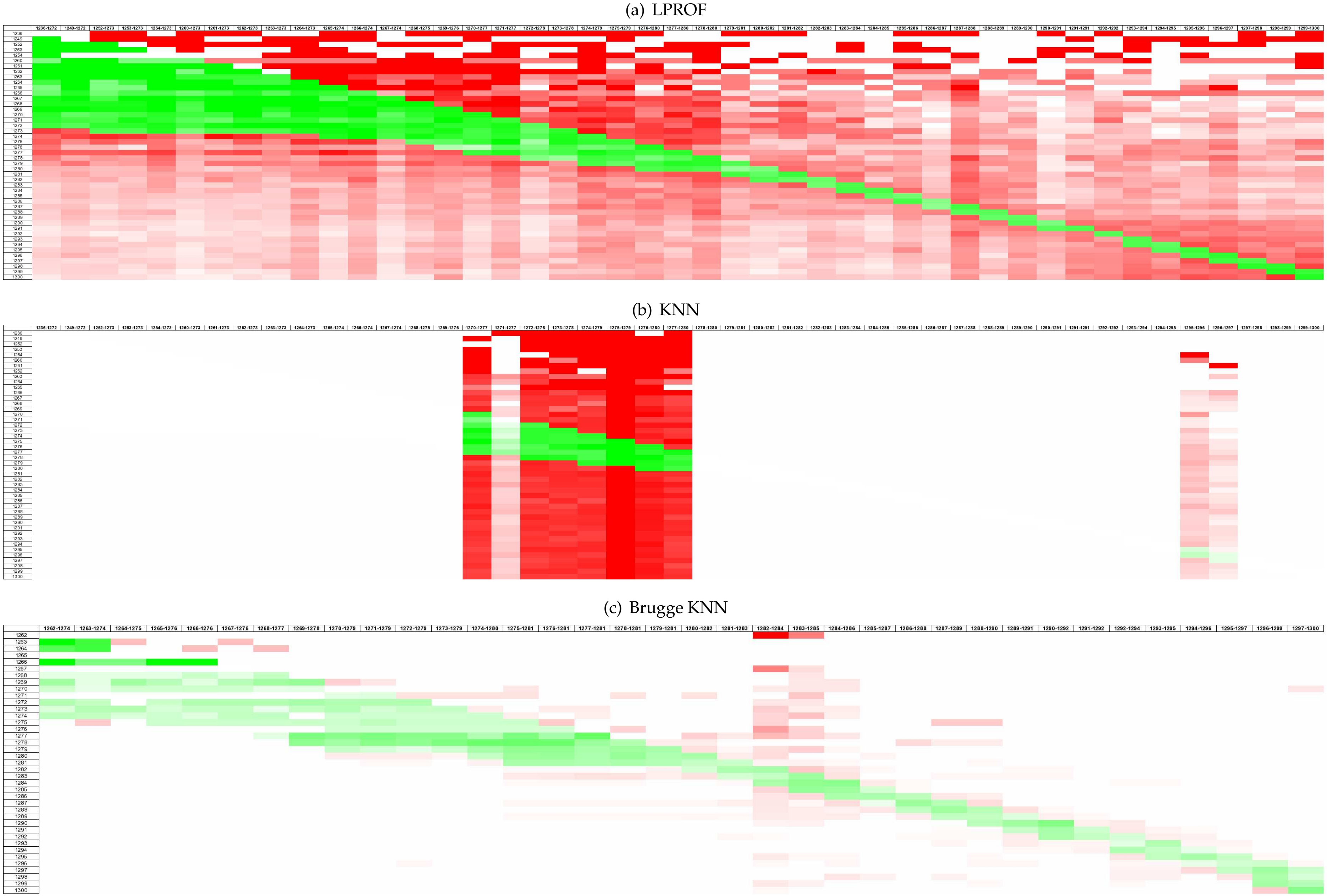
[8] average FRF-score: 0.94

[9] average FRF-score: 0.60

[10] average FRF-score: 0.32

[11] This is mainly due to the combination of large floating point feature vectors and the amount of models (about 1000 for the 14th century dating) we created. Testing some of these took several weeks on a computing cluster with

Table 4.12: Confusion matrix for the year interval recognition in the 13th century material. Each column stands for a recognition model, each row for a set of documents that were written during a single year.



(a) LPROF

(b) KNN

(c) Brugge KNN

that for the previous century, as we do get sensible verification models[12]. The amount of false positive accepts is lower than the LProf models. This comes however with a disadvantage: not all charters are recognized by "their" time model. For the broad intervals (14 years around the begin of the century) the recognition rate peaks at about 50%, where it drops to 25% for smaller intervals (2 years at the end of the century).

4.7.3 Dividing the century

In this section we attempt to create models for splitting up a century in two parts ("was a charter written after year X?") , progressively making the second part smaller. The main question is whether we can draw such a sharp line between the two resulting parts. To give an impression about the outcomes of these models we plot the FRF-measure in function of the interval. A sudden drop of the FRF thus would indicate that there is a sharp division between the two resulting halves. This in turn could be caused by e.g. a drastic language change or the introduction of a new word.

For technical reasons[13] these classifications were not performed with Orange but with Timbl (Daelemans et al., 2007). As both implement the KNN learner scheme, this should not have a significant influence on the results.

4.7.3.1 Division boundaries for the 13th century

The results for the century division modelling are given in figure 4.7.

Using LProf, most intervals get a FRF score of between 0.30. At the begin the few oldest charters result in a 100% false accept rate; this explains the high FRF score at the left hand side of the graph.

The similarity between the outcome for the KNN model for the whole century and that for Brugge only is remarkable. As more than half of the charters from the 13th century come from Brugge, this should not cause any surprise. We therefore take a closer look at the Brugge model for an explanation of the 13th century KNN model.

The KNN model for Brugge shows an interesting FRF decrease around 1285. As from 1290, the FRF raises again. This can possibly be explained by the fact that the 13th century corpus contains a lot of similar documents from Brugge, the so-called *lakenkeuren* (documents on the quality etc. of cloth) that all were written after 1284. It seems that in this case we built a recognizer for this specific document type instead of a dating model.

4.7.3.2 Division boundaries for the 14th century

To conclude, we consider the division models for the 14th century, as shown in figure 4.8. For KNN, we only took a sample to prove this as the creation and testing of all these models demands a lot of time and computing power. The outcome, both for KNN and LProf, is not too spectacular. A reasonable recognition rate is attained (a slightly better one using KNN). The improving results towards the end of the century can be possibly explained by the fact that more charters originate from this time period. Given this larger amount of training data, the classification results can be of a higher quality as well.

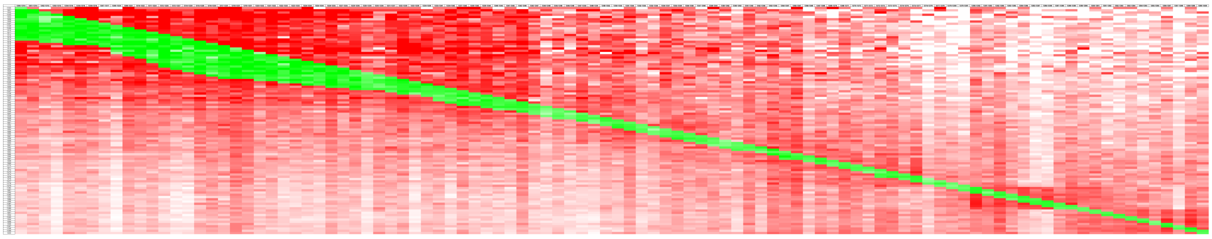
tens of nodes.

[12] average FRF-score: 0.49

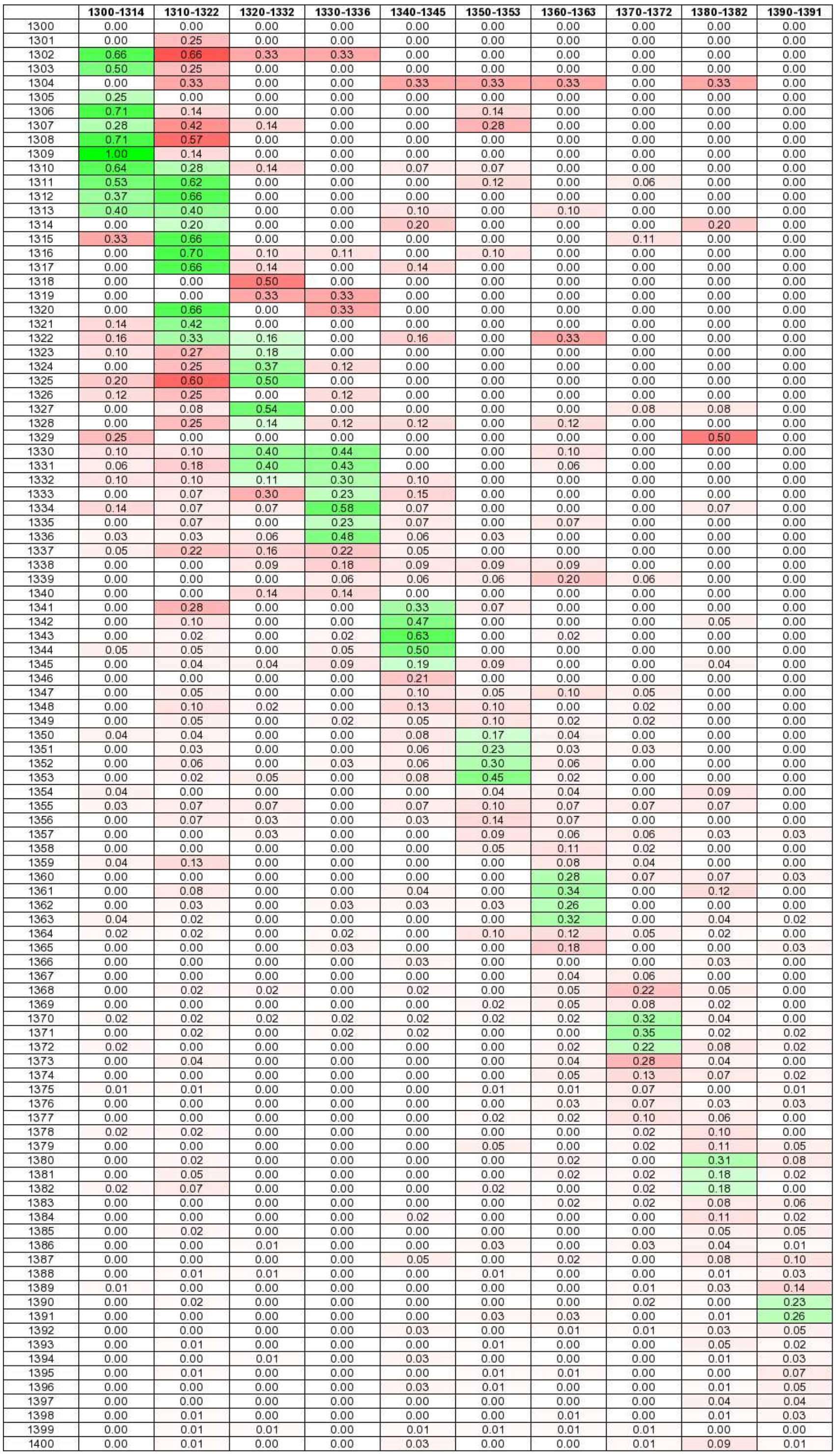
[13] A bug in Orange for linux, which prevented us from using a linux cluster to perform the large number of classifications.

Table 4.13: Confusion matrix for the year interval verification in the 14th century material.

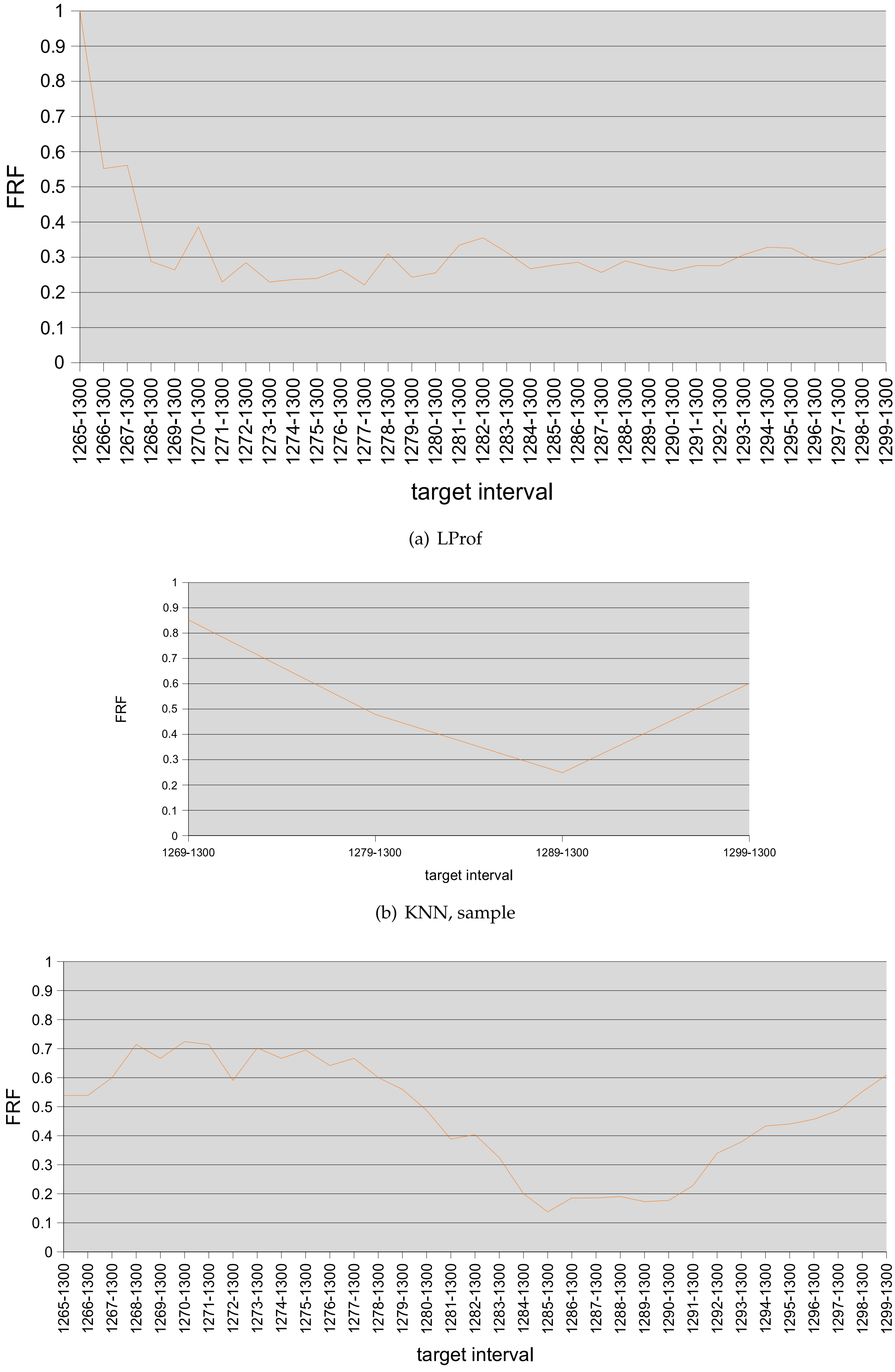
(a) LPROF



(b) KNN, sample



	1300-1314	1310-1322	1320-1332	1330-1336	1340-1345	1350-1353	1360-1363	1370-1372	1380-1382	1390-1391
1300	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1301	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1302	0.66	0.66	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00
1303	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1304	0.00	0.33	0.00	0.00	0.33	0.33	0.33	0.00	0.33	0.00
1305	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1306	0.71	0.14	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00
1307	0.28	0.42	0.14	0.00	0.00	0.28	0.00	0.00	0.00	0.00
1308	0.71	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1309	1.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1310	0.64	0.28	0.14	0.00	0.07	0.07	0.00	0.00	0.00	0.00
1311	0.53	0.62	0.00	0.00	0.00	0.12	0.00	0.06	0.00	0.00
1312	0.37	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1313	0.40	0.40	0.00	0.00	0.10	0.00	0.10	0.00	0.00	0.00
1314	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.20	0.00
1315	0.33	0.66	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00
1316	0.00	0.70	0.10	0.11	0.00	0.10	0.00	0.00	0.00	0.00
1317	0.00	0.66	0.14	0.00	0.14	0.00	0.00	0.00	0.00	0.00
1318	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1319	0.00	0.00	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00
1320	0.00	0.66	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00
1321	0.14	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1322	0.16	0.33	0.16	0.00	0.16	0.00	0.33	0.00	0.00	0.00
1323	0.10	0.27	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1324	0.00	0.25	0.37	0.12	0.00	0.00	0.00	0.00	0.00	0.00
1325	0.20	0.60	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1326	0.12	0.25	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00
1327	0.00	0.08	0.54	0.00	0.00	0.00	0.00	0.08	0.08	0.00
1328	0.00	0.25	0.14	0.12	0.12	0.00	0.12	0.00	0.00	0.00
1329	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
1330	0.10	0.10	0.40	0.44	0.00	0.00	0.10	0.00	0.00	0.00
1331	0.06	0.18	0.40	0.43	0.00	0.00	0.06	0.00	0.00	0.00
1332	0.10	0.10	0.11	0.30	0.10	0.00	0.00	0.00	0.00	0.00
1333	0.00	0.07	0.30	0.23	0.15	0.00	0.00	0.00	0.00	0.00
1334	0.14	0.07	0.07	0.58	0.07	0.00	0.00	0.00	0.07	0.00
1335	0.00	0.07	0.00	0.23	0.07	0.00	0.07	0.00	0.00	0.00
1336	0.03	0.03	0.06	0.48	0.06	0.03	0.00	0.00	0.00	0.00
1337	0.05	0.22	0.16	0.22	0.05	0.00	0.00	0.00	0.00	0.00
1338	0.00	0.00	0.09	0.18	0.09	0.09	0.09	0.00	0.00	0.00
1339	0.00	0.00	0.00	0.06	0.06	0.06	0.20	0.06	0.00	0.00
1340	0.00	0.00	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.00
1341	0.00	0.28	0.00	0.00	0.33	0.07	0.00	0.00	0.00	0.00
1342	0.00	0.10	0.00	0.00	0.47	0.00	0.00	0.00	0.05	0.00
1343	0.00	0.02	0.00	0.02	0.63	0.00	0.02	0.00	0.00	0.00
1344	0.05	0.05	0.00	0.05	0.50	0.00	0.00	0.00	0.00	0.00
1345	0.00	0.04	0.04	0.09	0.19	0.09	0.00	0.00	0.04	0.00
1346	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.00
1347	0.00	0.05	0.00	0.00	0.10	0.05	0.10	0.05	0.00	0.00
1348	0.00	0.10	0.02	0.00	0.13	0.10	0.00	0.02	0.00	0.00
1349	0.00	0.05	0.00	0.02	0.05	0.10	0.02	0.02	0.00	0.00
1350	0.04	0.04	0.00	0.00	0.08	0.17	0.04	0.00	0.00	0.00
1351	0.00	0.03	0.00	0.00	0.06	0.23	0.03	0.03	0.00	0.00
1352	0.00	0.06	0.00	0.03	0.06	0.30	0.06	0.00	0.00	0.00
1353	0.00	0.02	0.05	0.00	0.08	0.45	0.02	0.00	0.00	0.00
1354	0.04	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.09	0.00
1355	0.03	0.07	0.07	0.00	0.07	0.10	0.07	0.07	0.07	0.00
1356	0.00	0.07	0.03	0.00	0.03	0.14	0.07	0.00	0.00	0.00
1357	0.00	0.00	0.03	0.00	0.00	0.09	0.06	0.06	0.03	0.03
1358	0.00	0.00	0.00	0.00	0.00	0.05	0.11	0.02	0.00	0.00
1359	0.04	0.13	0.00	0.00	0.00	0.00	0.08	0.04	0.00	0.00
1360	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.07	0.07	0.03
1361	0.00	0.08	0.00	0.00	0.04	0.00	0.34	0.00	0.12	0.00
1362	0.00	0.03	0.00	0.03	0.03	0.03	0.26	0.00	0.00	0.00
1363	0.04	0.02	0.00	0.00	0.00	0.00	0.32	0.00	0.04	0.02
1364	0.02	0.02	0.00	0.02	0.00	0.10	0.12	0.05	0.02	0.00
1365	0.00	0.00	0.00	0.03	0.00	0.00	0.18	0.00	0.00	0.03
1366	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.00
1367	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.06	0.00	0.00
1368	0.00	0.02	0.02	0.00	0.02	0.00	0.05	0.22	0.05	0.00
1369	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.08	0.02	0.00
1370	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.32	0.04	0.00
1371	0.00	0.02	0.00	0.02	0.02	0.00	0.00	0.35	0.02	0.02
1372	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.22	0.08	0.02
1373	0.00	0.04	0.00	0.00	0.00	0.00	0.04	0.28	0.04	0.00
1374	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.13	0.07	0.02
1375	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.07	0.00	0.01
1376	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.07	0.03	0.03
1377	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.10	0.06	0.00
1378	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.10	0.00
1379	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.02	0.11	0.05
1380	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.31	0.08
1381	0.00	0.05	0.00	0.00	0.00	0.00	0.02	0.02	0.18	0.02
1382	0.02	0.07	0.00	0.00	0.00	0.02	0.00	0.02	0.18	0.00
1383	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.08	0.06
1384	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.11	0.02
1385	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05
1386	0.00	0.00	0.01	0.00	0.00	0.03	0.00	0.03	0.04	0.01
1387	0.00	0.00	0.00	0.00	0.05	0.00	0.02	0.00	0.08	0.10
1388	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.03
1389	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.14
1390	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.23
1391	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.01	0.26
1392	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.01	0.03	0.05
1393	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.02
1394	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00	0.01	0.03
1395	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.07
1396	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.01	0.05
1397	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04
1398	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03
1399	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00
1400	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.01	0.09	0.01

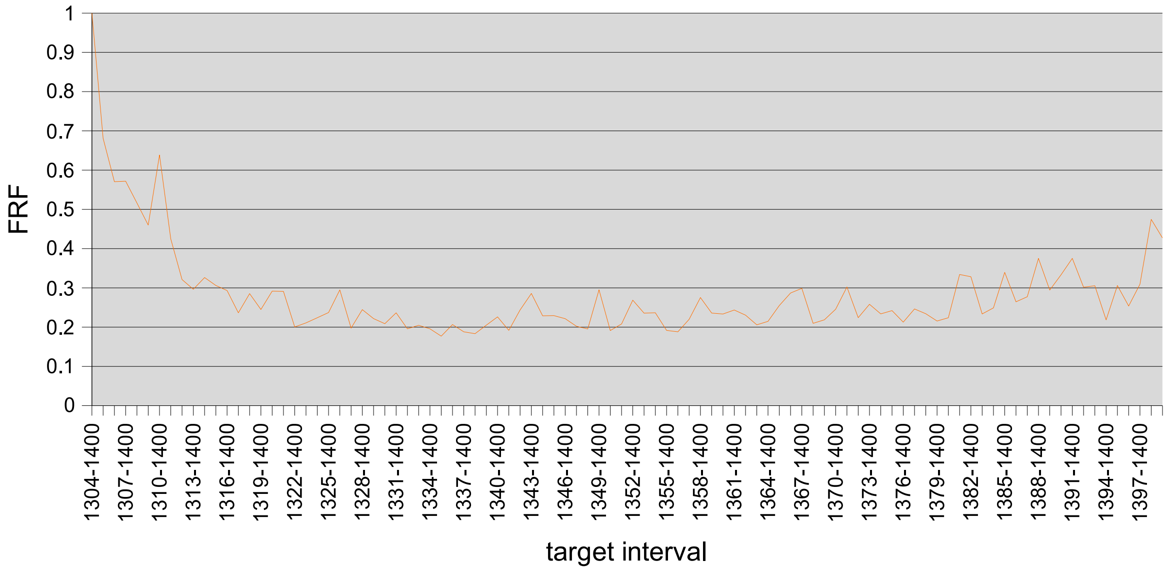


(a) LProf

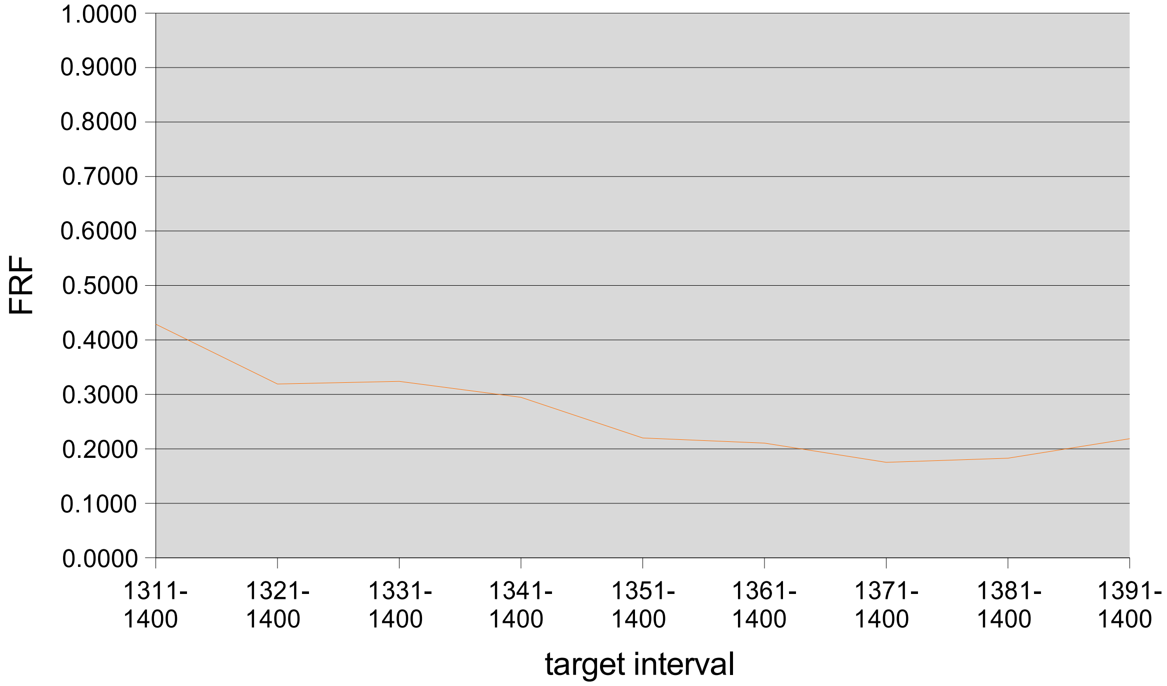
(b) KNN, sample

(c) Brugge KNN

Figure 4.7: FRF scores for the century halves recognition in the 13th century material.



(a) LProf



(b) KNN, sample

Figure 4.8: FRF scores for the century halves recognition in the 14th century material.

4.8 Summary of the results

At the end of this chapter, we ask ourselves what can be learnt from all the experiments that we have performed.

First of all, the charters from the 13th century differ significantly from those from the 14th century. The century (or the originating corpus) can be well distinguished and recognized automatically. Therefore we will work further on with separate models for both centuries.

Secondly, the location where a charter was written appears to be a property that can be predicted quite accurately using character trigrams. As for the optimal classifier algorithm the results are mixed. For the 13th century, LProf takes the first place. However KNN outperforms in turn LProf for the 14th century. In general the location verification works better for the 14th century charters.

Thirdly, the time interval verification gives worse results. Again the 13th century is more problematic, especially when using KNN. This can be partially explained by the fact that the time slices span a longer period for the first part of this century, because of the low number of charters available. That could lead to more diffuse classes, which are harder to model. The 14th century combined with LProf results in a high false accept rate, while in contrast KNN tends towards an elevation of false rejections.

An attempt to model the date information in another way — trying to split each century in two parts and predicting the matching part for each document — did result in alternative view on the dating task. For the 13th century 1285 seems to be a clear time boundary. This is probably related to the similar content of a large number of charters written after that date in Brugge. The 14th century corpus does not show such a sharp dividing line, though its second part can be recognized with a higher accuracy.

The use of character trigrams as the basis for the features proves to be a reasonable choice. It captures largely the information that can be given by orthographic variation, although it also contaminates the machine learning data with proper nouns[14]. As such, the good results we achieved can be partially explained by the presence of place and person names. To counter these side effects, we will focus on a method to extract purely orthographic variation in the next chapter.

[14] One way of eliminating this "contamination" would be the exclusion of all proper nouns from the data. We still plan to do this for all locations in the future. In the meantime an experiment based on some 14th century sample locations showed that although the FRF-score slightly decreases by excluding the proper nouns, the overall results are still in line with the outcome we described in this chapter.

Chapter 5

Towards a spelling-based classification

We have previously set up a methodology for constructing and testing a set of verification models (both for location and date). Now we will apply this approach to the spelling-related information we extract from the charters, instead of the plain character n-gram frequencies. The characteristic orthography will be represented in two ways: the presence of specific variants and a set of automatically derived rewrite rules. The latter indicate in which way a variant differs from a normal form that is to be defined in this chapter. Both approaches will be employed to date and localise the charters.

5.1 Determining orthographic variants

5.1.1 Exact versus heuristics

In the first place we ask ourselves what is meant by orthographic variation. For this study we will define it as follows: "The orthographic variants of a word is the set of all the ways in which that inflected word is written within a reference corpus."

Now that we have defined orthographic variation, it has become clear that extracting such variants from a corpus is not a trivial thing to do. The most demanding part of filtering out the spelling variants is grouping together all tokens that correspond to one word form. This requires that we know for every token the lemma and the part of speech tag. The latter is especially useful to make a clear distinction between among cases and conjugated forms, as we do not consider inflection as a source of orthographic variation.

These prerequisites lead us to an important question: do we want to base our variant detection on exact criteria or is it more desirable to use heuristics? Ideally, one should use the most exact information available. However, in practice this is not always available (Kempken, 2005). In our case, the situation is mixed. For the 13th century we do not have any annotations, as where the 14th century corpus contains for every word a POS tag and a lemma (a modern Dutch lemma, but that can be used equally well for extracting the variants).

As we want to investigate the information value of orthography in an optimal way, our choice goes to the 14th century. This guarantees that we are really dealing with spelling variants and not with the partially wrong outcomes of heuristics used to determine the variants. As a consequence we will only use the charters from the corpus Van Reenen-Mulder for further experiments.

For this study we used an interim version of this corpus as far as the POS tags concerns. This resulted in e.g. the tagging of the affixes of separable verbs as verbs. This has some consequences, which we will deal with in section 5.2.2.2. In the course of the research corrections were made to the tags. Nevertheless, we decided to go on with the version we started with, to maintain the continuity and the comparability with the results from the previous chapter.

5.1.2 Proper nouns

Recall that we asked ourselves before in what way the spelling of proper nouns differs from other words.

At this stage we want to explore this issue. Therefore we take a closer look at two measures that give an impression of the way proper nouns are written through the 14th century charters.

First, the mean number of orthographic variants is calculated. Proper nouns have on average 1.93 variants, compared to 2.88 for other words. Still this clear difference does not necessarily mean that proper nouns are less prone to variation. Suppose that all of these proper noun variants — although smaller in number — are used equally often, and that for the other words only one variant is frequently used, then the average number of variants still does not give a good impression of the spelling variation. Therefore we also measured how often the most frequent form appears: 85.68% for the proper nouns, while for the other words in 80.50% of the cases the most common form was used.

The measurements we carried out thus prove the assumption that proper nouns are not as often the subject of spelling variation as other words. The spelling variation of proper nouns thus cannot deliver as much information on the provenance of a charter as other words can. Therefore we will not use them during the orthographic feature extraction further on.

5.1.3 Extracting the variants from the corpus

In a first attempt to capture the information that is included in the orthographic variation, we create a list of all tokens that have at least one alternative way of writing within the corpus.

5.1.3.1 Procedure

We proceeded as follows to extract the orthographic variants from the corpus:

- We selected the relevant charters (depending on place/date verification[1]).
- For all tokens in all charters, we looked up the modern lemma and the part of speech tag.
- All proper nouns were removed from the corpus word list.
- For all words with the same lemma and POS tag (the variant list), we calculated the frequency count.
- All elements from each variant list that were found only once were removed.
- If any of the variant lists only contained one element, it was removed as well.

In the end we have a list of all orthographic variants, together with their frequency. For the year induction, this gives a list with 15,229 variants (for 3327 words), the location induction can be performed using 10,846 variants (that correspond to 2552 words).

Note that abbreviated forms of a word are also included in these lists. Proper nouns were excluded, as we explicitly want to neutralise their effect on the classification. Using the spelling of a city name as a pointer for the origin of the charter would be too obvious and diverts the attention from the real influence of orthography.

Because, as we pointed out before, the use of capitals is very inconsistent in Middle Dutch and because we removed the proper nouns anyway, all words were converted into lower case. This is in line as well with the earlier approach for the extraction of the character trigrams.

As an illustration we here list all forms of the word altaar ("altar", nominative singular) in the charters for which the year of writing is known: *altaer* (30 occurrences), *altoer* (3), *elt* (2), *elter* (6), *outaer* (18), *outhaer* (3).

[1] Recall that we discarded all charters that were not exactly dated or localised for respectively date or location verification.

5.1.3.2 Spelling variants as features

All orthographic variants now have been extracted from the 14th century corpus. We still need to convert this information into feature vectors that can be used for machine learning purposes.

The occurrence of each of these variants can be considered as a binary feature: either it is used in a charter or it is not. We don't use the relative frequency count anymore because we are dealing with complete words instead of trigrams, and the former have lower frequency counts. The lower the frequency, the less relevant the exact number of occurrences becomes: the main clue is whether something is encountered at all. Besides, the choice for Boolean features enhances the computation efficiency for further classification operations.

5.1.4 Deriving spelling variation rules from the corpus

The extraction of orthographic variants as a whole still brings along the influence of the words themselves. Ideally we would like to abstract away from the words (and their meaning) and keep the focus on the spelling itself. That is why we now take a further step by deriving spelling rules from the variants, in an attempt to capture more general tendencies instead of particular word use.

Obviously, the task of deriving such rules for historical documents is not something completely new. In Kamps et al. (2007) some interesting approaches are given, although with another objective in mind: searching in a corpus of historical Dutch using a query in contemporary language. In order to improve the recall for these modern queries, the search operation is expanded with extra (historical) word variants. These are produced using rewrite rules that were automatically derived beforehand from a corpus. The rewrite rules apply to consonant or vowel clusters. This strategy relies on the observation that most orthographic variation occurs within such clusters. A manual inspection of a sample of the variants for the 14th century corpus showed this observation is valid for our material and task as well.

5.1.4.1 Algorithm

For the extraction of the spelling variation rules, we rely upon the following algorithm:

- Select the relevant charters (depending on place/date verification)
- For all tokens in all charters, look up the modern lemma and the part of speech tag
- Remove all proper nouns from the corpus word list
- Only keep those words for which at least 2 orthographic forms exist
- For all word forms with the same lemma and POS tag (the variants), a normal form is created:
  - find the most frequent variant, analyse its consonant/vowel-structure
  - for each consonant/vowel-cluster:
    - select the cluster in this position that is observed the most among the variants for this word in the whole corpus

Now, to derive rules from a variant we encountered, we do the following:

- We check if the cluster structure matches that of the normal form for this variant. One final deletion or insertion is allowed. If there is no match, don't create any rule for this variant.

- If the structure matches, the variant's clusters are aligned with those from the normal form.
- Create rewrite rules for each cluster, with the cluster from the normal form at the left hand side
- Remove rewrite rules of the form X → X, if X never changes for any of the variants.

Some characters can be used both as either consonant or vowel, especially *v* and *j* (as alternative for *u* and *i* respectively). We have therefore made the (arbitrary) choice to consider *v* and *j* as vowels. Other ambiguous characters, as *w* instead of *uu*, are only used very rarely and therefore regarded as consonants.

5.1.4.2 Illustration of the algorithm

To illustrate the effect of the algorithm we described above, we here analyse how rules are derived for the the verb *laten* ("let", "leave"), with the modern Dutch lemma *laten* and POS tag 250 (infinitive). Within the whole corpus we can find the following forms: *loeten* (3 occurences), *late* (13), *laten* (67) and *laeten* (9).

- First, the most frequent form is selected, in this case *laten*
- Then we determine the clusters of vowels and consonants: C-V-C-V-C (l-a-t-e-n)
- Now all forms with the same cluster structure are chosen. Note that one final cluster insertion or deletion is allowed. So we retain the following forms:
  - *late* (with one final deletion of the C-cluster "n")
  - *laten*
  - *loeten*
  - *laeten*
- Among these variants, the most common form for every cluster is looked up:
  - C: *l* occurs in all variants, no alternative
  - V: *a* (frequency: 67+13 = 80), *oe* (frequency: 3), *ae* (frequency: 9)
  - C: *t* occurs in all variants, no alternative
  - V: *e* occurs in all variants, no alternative
  - C: *n* occurs in all variants (the final deletion is ignored), no alternative
- Now the normal form is built with the most frequent clusters: l-a-t-e-n. Here it corresponds to one of the existing variants, but that is not necessarily the case.
- From here on, we can derive rules from any of the forms that occurs, using the transformation from the normal form to the encountered variant:
  - *laten*: l-a-t-e-n → l-a-t-e-n, yields: a → a
  - *laeten*: l-a-t-e-n → l-ae-t-e-n, yields: a → ae
  - *loeten*: l-a-t-e-n → l-oe-t-e-n, yields: a → oe
  - *late*: l-a-t-e-n → l-a-t-e, yields: deletion of final n

Note that this example does not include a variant with a non-matching cluster structure. If e.g. the form *gelaten*[2] would have been found, it would not have been used to create the normal form, nor would it have been used to derive variation rules. This due to its C-V-C-V-C-V-C structure, which does not match that of the normal form, C-V-C-V-C, even when allowing one deletion or insertion at the end.

This rule derivation procedure results in 3581 rules for the charters from a known location and 4595 for those for which the year of writing is available. These numbers are considerably smaller than those for the variants as a whole, because one single rule can apply to multiple cases of orthographic variation.

5.1.4.3 Spelling variation rules as features

In analogy with the attributes for the orthographic variants as a whole, we also create a binary feature for every variation rule, indicating whether a specific rule occurs in a single charter. One side-note should be made here: we first stored for each type all of the rules that can be derived of it (e.g. laeten results in 1 rule, a → ae), . Then, whenever the type ("laeten") was found in a document, we added this set of rules as positive features (the feature "a → ae" gets the value "true").

In most cases this strategy has the same effect as deriving all of the rules for each token. However, in the case of homographs (with possibly another part of speech and modern lemma) this results in the over-generation of rules. An extensive example of such a case will be given in section 5.2.2.2. Numerically speaking, 2110 out of 11796 words with at least one orthographic variant are homographs; most of them (1424) are based on one form with 2 possible tags[3].

The main question that comes with this decision is how it influences the verification accuracy. Therefore we performed an extra experiment, storing the rules for every single combination of type, part of speech and modern lemma (e.g. "laeten" + 250 + "laten"). This obviously resulted in a lower number of rules that was generated for each charter. Interestingly enough, it also caused a significant decrease of the verification accuracy. Therefore we can conclude that storing rules for each type, while neglecting POS tag and modern lemma, seems to have a beneficial influence on the overall classification performance. In this case, overgeneralising clearly brings an advantage. Apart from that, this approach also makes it possible to classify a charter that has not been tagged.

5.2 Induction of the location

As we did before using the character trigrams, we now turn to induction of the location where the charters have been written. We run all tests using first the spelling variants as the learning data, followed by the variation rules. As classification method we choose for KNN, because (as described in the previous chapter) it proved to give the best results (certainly for the localisation) and can deal with binary features.

5.2.1 Quantitative analysis

In table 5.1, 5.2 and 5.3 the outcome of the location verification models can be found. We used the KNN algorithm for the creation of these models, as this proved to give the best results for our earlier experiments on the 14th century material.

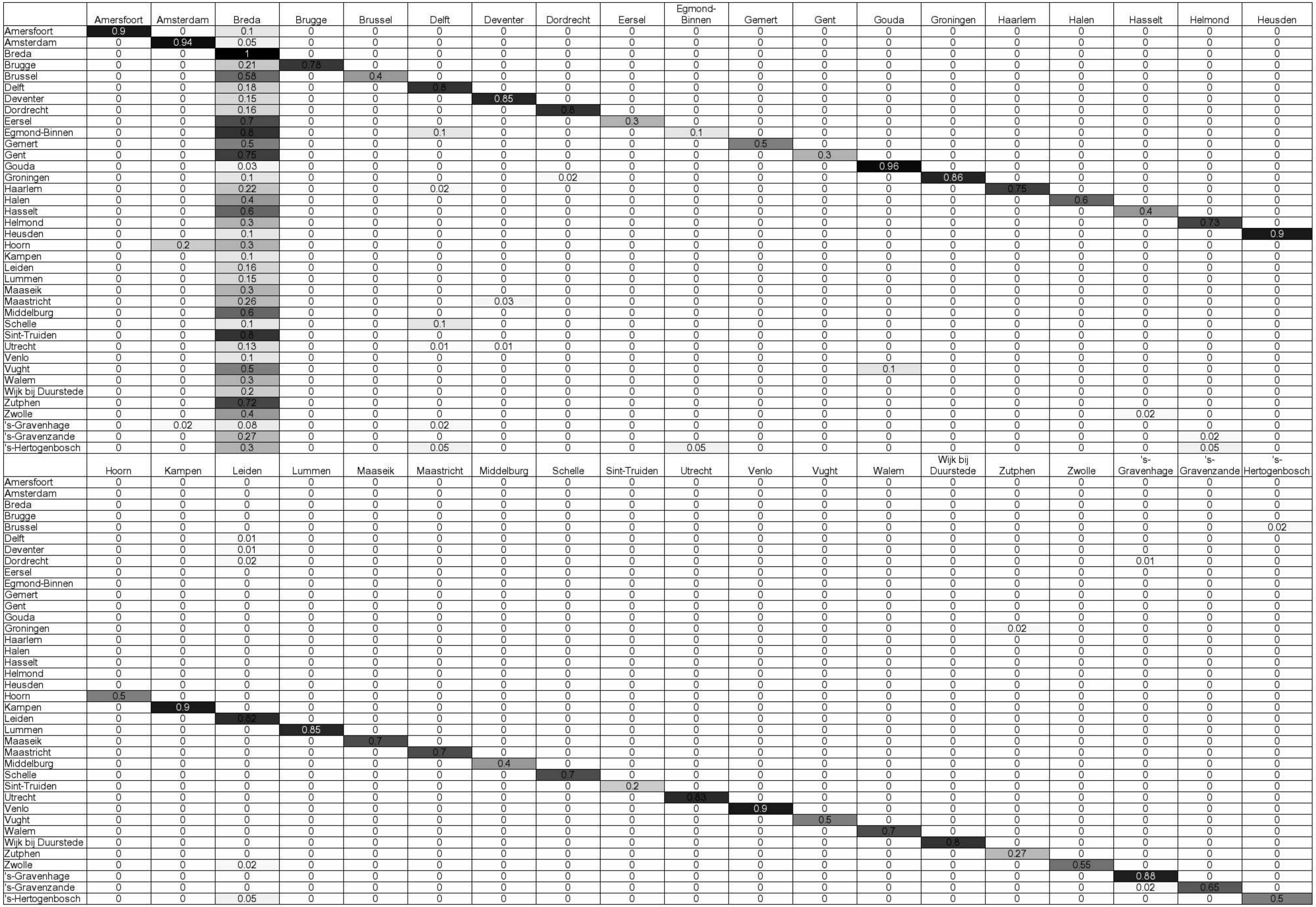
[2] This only is an example to illustrate the algorithm used, this variant was not found in our corpus.

[3] The homographs with most variants are:
af, dat, des, dies, groete, oere, of, soe, stede, voers, voirs, vp (9)
die, hare, hem, hore (10)
op, wt (11)
en (12)

Table 5.1: Location verification results for the 14th century corpus, using orthographic variants and variant-derived rules as features.

		Variants				Rules			
Location	charters	Errors	FAR	FRR	FRF	Errors	FAR	FRR	FRF
Amersfoort	23	0.0015	0.0000	0.1000	0.0526	0.0061	0.0054	0.0500	0.0282
Amsterdam	98	0.0068	0.0033	0.0556	0.0301	0.0159	0.0138	0.0444	0.0294
Breda	88	0.2235	0.2323	0.0000	0.1314	0.0152	0.0031	0.3200	0.1915
Brugge	115	0.0098	0.0000	0.2167	0.1215	0.0053	0.0032	0.0500	0.0272
Brussel	67	0.0227	0.0000	0.6000	0.4286	0.0083	0.0000	0.2200	0.1236
Delft	76	0.0152	0.0048	0.2000	0.1130	0.0106	0.0080	0.0571	0.0332
Deventer	69	0.0083	0.0016	0.1500	0.0818	0.0076	0.0040	0.0833	0.0453
Dordrecht	85	0.0129	0.0008	0.2000	0.1114	0.0265	0.0169	0.1750	0.1029
Eersel	61	0.0053	0.0000	0.7000	0.5385	0.0008	0.0000	0.1000	0.0526
Egmond-Binnen	13	0.0076	0.0008	0.9000	0.8182	0.0061	0.0000	0.8000	0.6667
Gemert	16	0.0038	0.0000	0.5000	0.3333	0.0061	0.0023	0.5000	0.3338
Gent	42	0.0106	0.0000	0.7000	0.5385	0.0091	0.0008	0.5500	0.3795
Gouda	112	0.0045	0.0017	0.0364	0.0193	0.0068	0.0058	0.0182	0.0120
Groningen	73	0.0053	0.0000	0.1400	0.0753	0.0061	0.0039	0.0600	0.0328
Haarlem	58	0.0076	0.0000	0.2500	0.1429	0.0038	0.0000	0.1250	0.0667
Halen	17	0.0030	0.0000	0.4000	0.2500	0.0015	0.0008	0.1000	0.0530
Hasselt	14	0.0053	0.0008	0.6000	0.4287	0.0038	0.0015	0.3000	0.1770
Helmond	41	0.0076	0.0016	0.2667	0.1544	0.0098	0.0085	0.0667	0.0385
Heusden	42	0.0015	0.0000	0.1000	0.0526	0.0045	0.0046	0.0000	0.0023
Hoorn	16	0.0038	0.0000	0.5000	0.3333	0.0038	0.0023	0.2000	0.1120
Kampen	39	0.0015	0.0000	0.1000	0.0526	0.0000	0.0000	0.0000	0.0000
Leiden	117	0.0189	0.0050	0.1727	0.0966	0.0220	0.0149	0.1000	0.0594
Lummen	20	0.0023	0.0000	0.1500	0.0811	0.0015	0.0000	0.1000	0.0526
Maaseik	23	0.0023	0.0000	0.3000	0.1765	0.0000	0.0000	0.0000	0.0000
Maastricht	45	0.0068	0.0000	0.3000	0.1765	0.0045	0.0008	0.1667	0.0912
Middelburg	41	0.0045	0.0000	0.6000	0.4286	0.0038	0.0000	0.5000	0.3333
Schelle	14	0.0023	0.0000	0.3000	0.1765	0.0008	0.0000	0.1000	0.0526
Sint-Truiden	15	0.0061	0.0000	0.8000	0.6667	0.0038	0.0000	0.5000	0.3333
Utrecht	80	0.0076	0.0000	0.1667	0.0909	0.0106	0.0056	0.1167	0.0644
Venlo	14	0.0008	0.0000	0.1000	0.0526	0.0008	0.0008	0.0000	0.0004
Vught	11	0.0038	0.0000	0.5000	0.3333	0.0023	0.0008	0.2000	0.1114
Walem	16	0.0023	0.0000	0.3000	0.1765	0.0023	0.0008	0.2000	0.1114
Wijk bij Duurstede	13	0.0015	0.0000	0.2000	0.1111	0.0023	0.0008	0.2000	0.1114
Zutphen	93	0.0227	0.0008	0.7250	0.5687	0.0121	0.0055	0.2250	0.1289
Zwolle	56	0.0136	0.0000	0.4500	0.2903	0.0083	0.0023	0.2000	0.1120
′s-Gravenhage	53	0.0061	0.0016	0.1200	0.0645	0.0227	0.0205	0.0800	0.0512
′s-Gravenzande	46	0.0106	0.0000	0.3500	0.2121	0.0091	0.0031	0.2000	0.1123
′s-Hertogenbosch	29	0.0083	0.0008	0.5000	0.3335	0.0083	0.0023	0.4000	0.2506

Table 5.2: Confusion matrix for the location verification within the 14th century material, using orthographic variants.



	Amersfoort	Amsterdam	Breda	Brugge	Brussel	Delft	Deventer	Dordrecht	Eersel	Egmond- Binnen	Gemert	Gent	Gouda	Groningen	Haarlem	Halen	Hasselt	Helmond	Heusden
Amersfoort	0.9	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Amsterdam	0	0.94	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Breda	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brugge	0	0	0.21	0.78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brussel	0	0	0.58	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Delft	0	0	0.18	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0
Deventer	0	0	0.15	0	0	0	0.85	0	0	0	0	0	0	0	0	0	0	0	0
Dordrecht	0	0	0.16	0	0	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0
Eersel	0	0	0.7	0	0	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0
Egmond-Binnen	0	0	0.8	0	0	0.1	0	0	0	0.1	0	0	0	0	0	0	0	0	0
Gemert	0	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0
Gent	0	0	0.75	0	0	0	0	0	0	0	0	0.3	0	0	0	0	0	0	0
Gouda	0	0	0.03	0	0	0	0	0	0	0	0	0	0.96	0	0	0	0	0	0
Groningen	0	0	0.1	0	0	0	0	0.02	0	0	0	0	0	0.86	0	0	0	0	0
Haarlem	0	0	0.22	0	0	0.02	0	0	0	0	0	0	0	0	0.75	0	0	0	0
Halen	0	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0.6	0	0	0
Hasselt	0	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0
Helmond	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.73	0
Heusden	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9
Hoorn	0	0.2	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kampen	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Leiden	0	0	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lummen	0	0	0.15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maaseik	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maastricht	0	0	0.26	0	0	0	0.03	0	0	0	0	0	0	0	0	0	0	0	0
Middelburg	0	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Schelle	0	0	0.1	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
Sint-Truiden	0	0	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Utrecht	0	0	0.13	0	0	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0
Venlo	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Vught	0	0	0.5	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0
Walem	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wijk bij Duurstede	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zutphen	0	0	0.72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zwolle	0	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0
's-Gravenhage	0	0.02	0.08	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0
's-Gravenzande	0	0	0.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0
's-Hertogenbosch	0	0	0.3	0	0	0.05	0	0	0	0.05	0	0	0	0	0	0	0	0.05	0

	Hoorn	Kampen	Leiden	Lummen	Maaseik	Maastricht	Middelburg	Schelle	Sint-Truiden	Utrecht	Venlo	Vught	Walem	Wijk bij Duurstede	Zutphen	Zwolle	's- Gravenhage	's- Gravenzande	's- Hertogenbosch
Amersfoort	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Amsterdam	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Breda	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brugge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brussel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02
Delft	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Deventer	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dordrecht	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0
Eersel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Egmond-Binnen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gemert	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gouda	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Groningen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0
Haarlem	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Halen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hasselt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Helmond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Heusden	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hoorn	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kampen	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Leiden	0	0	0.82	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lummen	0	0	0	0.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maaseik	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maastricht	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0
Middelburg	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0
Schelle	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0
Sint-Truiden	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0
Utrecht	0	0	0	0	0	0	0	0	0	0.83	0	0	0	0	0	0	0	0	0
Venlo	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0
Vught	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0
Walem	0	0	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0
Wijk bij Duurstede	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0	0	0
Zutphen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.27	0	0	0	0
Zwolle	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0.55	0	0	0
's-Gravenhage	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.88	0	0
's-Gravenzande	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0.65	0
's-Hertogenbosch	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5

Table 5.3: Confusion matrix for the location verification within the 14th century material, using orthographic rules.

	Amersfoort	Amsterdam	Breda	Brugge	Brussel	Delft	Deventer	Dordrecht	Eersel	Egmond-Binnen	Gemert	Gent	Gouda	Groningen	Haarlem	Halen	Hasselt	Helmond	Heusden
Amersfoort	0.95	0	0	0	0	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0
Amsterdam	0	0.95	0	0	0	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0	0
Breda	0	0.04	0.68	0	0	0	0	0.04	0	0	0	0	0.02	0	0	0	0	0	0
Brugge	0	0	0.01	0.95	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0
Brussel	0.02	0	0.02	0	0.78	0.02	0	0.04	0	0	0	0.02	0	0	0	0	0	0.04	0
Delft	0	0	0.01	0	0	0.94	0	0.01	0	0	0	0	0	0	0	0	0	0	0.01
Deventer	0	0	0	0	0	0	0.91	0	0	0	0	0	0.01	0.01	0	0	0	0	0
Dordrecht	0.02	0.02	0	0	0	0	0	0.82	0	0	0	0	0	0	0	0	0	0.01	0.01
Eersel	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0.1	0
Egmond-Binnen	0	0	0	0	0	0.1	0	0.1	0	0.2	0	0	0	0.1	0	0	0	0.1	0.1
Gemert	0	0	0	0	0	0.2	0.1	0	0	0	0.5	0	0	0	0	0	0	0	0
Gent	0	0	0	0.2	0	0.05	0	0.1	0	0	0	0.45	0	0	0	0	0	0	0.05
Gouda	0	0	0	0	0	0	0	0	0	0	0	0	0.98	0	0	0	0	0	0
Groningen	0	0	0	0	0	0	0.02	0.02	0	0	0	0	0	0.94	0	0	0	0	0
Haarlem	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0.87	0	0	0	0
Halen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0
Hasselt	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0.7	0	0
Helmond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	0
Heusden	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Hoorn	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kampen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Leiden	0	0.02	0	0	0	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0	0
Lummen	0	0	0	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0
Maaseik	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maastricht	0	0	0	0	0	0	0.03	0	0	0	0	0	0.03	0	0	0	0	0.03	0.03
Middelburg	0	0	0	0	0	0	0	0.2	0	0	0	0	0.1	0	0	0	0	0.1	0
Schelle	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sint-Truiden	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0
Utrecht	0.05	0.01	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0.01
Venlo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Vught	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
Walem	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wijk bij Duurstede	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Zutphen	0	0.05	0	0	0	0	0	0.05	0	0	0	0	0	0.05	0	0	0	0	0
Zwolle	0.02	0	0	0	0	0	0	0	0	0	0.02	0	0.02	0.02	0	0	0.05	0	0
's-Gravenhage	0	0.02	0	0	0	0.02	0	0.02	0	0	0	0	0	0	0	0	0	0	0
's-Gravenzande	0	0.05	0	0	0	0	0	0	0	0	0.02	0	0.05	0	0	0	0	0.02	0
's-Hertogenbosch	0	0	0	0	0	0	0.05	0	0	0	0.05	0	0	0	0	0	0	0.05	0

	Hoorn	Kampen	Leiden	Lummen	Maaseik	Maastricht	Middelburg	Schelle	Sint-Truiden	Utrecht	Venlo	Vught	Walem	Wijk bij Duurstede	Zutphen	Zwolle	's-Gravenhage	's-Gravenzande	's-Hertogenbosch
Amersfoort	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Amsterdam	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Breda	0	0	0.1	0	0	0	0	0	0	0	0	0.02	0	0	0.02	0	0.08	0	0
Brugge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0
Brussel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0.02	0	0
Delft	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0
Deventer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0	0	0	0
Dordrecht	0	0	0.01	0	0	0	0	0	0	0.01	0	0	0.01	0	0	0	0.06	0	0
Eersel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Egmond-Binnen	0	0	0.1	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0.1
Gemert	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0.1
Gent	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
Gouda	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Groningen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0
Haarlem	0.02	0	0.02	0	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0	0
Halen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
Hasselt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0
Helmond	0	0	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
Heusden	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hoorn	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
Kampen	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Leiden	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0
Lummen	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0	0.05	0	0	0
Maaseik	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maastricht	0	0	0	0	0	0.83	0	0	0	0	0	0	0	0	0	0.03	0	0	0
Middelburg	0	0	0.1	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0
Schelle	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0
Sint-Truiden	0	0	0	0	0	0.1	0	0	0.5	0	0.1	0	0	0	0	0	0.1	0	0
Utrecht	0	0	0	0	0	0	0	0	0	0.88	0	0	0	0	0	0	0.01	0	0
Venlo	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Vught	0	0	0.1	0	0	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0
Walem	0	0	0.1	0	0	0	0	0	0	0	0	0	0.8	0	0	0	0.1	0	0
Wijk bij Duurstede	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0	0	0
Zutphen	0	0	0.02	0	0	0	0	0	0	0.02	0	0	0	0	0.77	0	0.02	0	0
Zwolle	0	0	0.02	0	0	0	0	0	0	0.02	0	0	0	0	0	0.8	0	0	0
's-Gravenhage	0	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0.92	0	0
's-Gravenzande	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.8	0
's-Hertogenbosch	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0.05	0.05	0.05	0.1	0	0.6

5.2.1.1 Variants

Generally, the results are slightly worse than for the trigram-based models. Still the outcome is pretty good, more than half of the location models recognises 70% or more of its charters, with hardly any false positives. One noteworthy exception is Breda, whose model accepts documents from almost all other locations as well.

Overall the outcomes for the models trained and tested on spelling variation seem to be in line with those for the character trigrams. For some locations (Zutphen, Hasselt, Sint-Truiden, Egmond-Binnen) the models perform below average. The mean FRF-score for all locations is 0.2327, which means that it performs worse than the corresponding character trigram models we tested before (FRF-score: 0.1148).

5.2.1.2 Rules

The main question when analysing the results for the spelling variation rules is if they can capture orthographic tendencies more closely than the full variants. The answer clearly is yes. The generalisation effect of these rules results in a mean FRF-score of 0.1180, which is close to that of the best performing trigram verification model, although the rules version is not aware of proper nouns. This encouraging outcome indicates that the rules that we derived indeed can be used to model diatopic language use. From this perspective it would be interesting to investigate which rules have the highest discriminating power, and if they fit with the manually selected classification criteria used in earlier studies. This will be done in the next section.

5.2.2 Qualitative analysis

Now that we pointed out it is possible to build spelling-based models for location verification, we try to find which orthographic properties are the best counsellors.

5.2.2.1 Variants

As before, the Gouda model performs very well (96% of its charters are accepted, almost no false accepts). When looking at the feature relevance we get the following top five features:

1. *lyede* [relief-score: 0.8303]: "to execute"
2. *orkonden* [0.8293]: "to publish with a charter"
3. *kennen* [0.7949]: "to know"
4. *m* [0.6730]: abbreviation of *met* ("with"), *men* ("(some)one") or *maar* ("but")[4]
5. *panden* [0.5686]: "to confiscate"[5]

These attributes match for the largest part with the trigrams found before for Gouda (especially *lyede* and *orkonden*). To broaden our view on the differences between locations we also calculated the most relevant features for Gent, a location that performed worse in terms of verification:

1. *de* [0.5035]: "the"
2. *wiue* [0.4706]: "wife"

[4] Note that the roman number M is also encountered in this corpus, but no orthographic variants were found for this particular use.

[5] Source: De Vries et al. (1882)

3. *lieden* [0.4649]: "men"
4. *sente* [0.4443]: "saint" or "admission"[6]
5. *vp* [0.4180]: "on"

The lower RELIEF scores already imply the features for Gent are less indicative than those for Gouda. For both locations, we clearly can see that function words and typical "charter vocabulary" (man, wife, charter, etc.) makes up the list of most relevant location indicators. This fits with the notion that words that are frequently used also tend to have more orthographic variants.

There is a large agreement between the variants that we found for the two examples and the words that were manually selected by Rem (2003) as locators, i.e. words whose variants are known to be a good indicator of a location. Among them are *orkonden*, *lieden*, *sente* and *vp*. The first one is a positive indicator for Gouda, the last three are positively associated with Gent. This clearly shows that our approach, the automatic selection of spelling-based features, resembles the outcome of a careful manual selection process.

5.2.2.2 Rules

For the rule-based approach there are no surprises either for our two example locations: Gouda performs well (98% is correctly accepted), Gent below average (only 45%). Again we list the 5 most important features, accompanied by their RELIEF-score. For the rules that do not result in another cluster, we added a list with alternative rewrite options within variants of the same word. These are labelled *(not: [...])*, indicating that they don't apply to this rule.

The top 5 rules for Gouda are:

1. d → nd [0.7342]
2. rc → rk [0.7096]
3. lls → lls (not: llns/llnts/llts/ll) [0.5191]
4. ll → lls [0.5191]
5. ye → ye (not: ei/ie/ii/ij/ue/ey/ee/je/y/e/j/i/ije) [0.4951]

We looked up in the documents from Gouda for which words these rules apply. An example where the *rk* cluster is used as alternative to the more common *rc* is *orkonden*. The rule for *ye* fits with the spelling variant *lyede*, which was found before to be very distinctive. There is at least a partial overlap between the extracted rules and the full variants. The combination *nd* instead of *d* was found a lot in the name of days in the Gouda charters: *sonnendaghes* (instead of *sonnedaghes*), *manendaghes* (instead of *manedaghes*). The name of days are extensively covered in Rem (2003), although the focus there goes mainly to the number of syllables, e.g. *maendaghes* versus *manendaghes*. Our results suggest that the final n at the end of the bisyllabic variant[7] is also an important piece of information.

Finally *lls* stems from *Hollants* (the adjective for the region Holland), a frequent form that has a broad number of abbreviations (*Holls*, *Hollnts*, *Hollts*, etc.). Apparently *Holls* is a characteristic abbreviation for the charters from Gouda. It is found in fixed expressions for amounts of money, e.g. *viif ende twintich scellinghe holls buerseghelts*[8].

For Gent the following features have the most influence:

[6] Source: De Vries et al. (1882)

[7] For Gouda, we could not find any of the monosyllabic variants.

[8] 25 "shilling" as rated on the Holland stock exchange

1. u → ie [0.5319]
2. rs → r [0.5038]
3. uu → uu (not: uuy/ov/ou/ui/i/y/u/uy/ue/ij/vv/oe/uij) [0.4906]
4. scr → scr (not: sc/cr/c/schr) [0.3799]
5. v → o [0.36208]

Again we see that at least one rule is closely related to the top five variants we found earlier for Gent: *ie* from *lieden* (instead of *luden*). As stated before we can see a clear analogy with the features selected on purely linguistic grounds in the literature. The use of *uu* in forms like *huus* is indeed indicated as very frequent for the area of Gent: about 82% of the observations features *uu*, according to Mooijaart (1992). The cluster *scr* (appearing in among others *screef, "wrote")* seems to be a good pointer for Gent as well, although we could not find any reference to this in the literature.

One rule cannot be explained directly by looking at the literature or the previously found important variants: rs → r . A closer examination learnt that this can be explained by the existence of a non-standard abbreviation form of *vorseiden* ('aforementioned'). It is most frequently abbreviated with *vors*, but in one place, Brussel, *vor* is used. This single occurrence results in the rule rs → r, which then is associated with the orthographic form *vor*. However, *vor* also is a high frequent preposition ('before'). Due to the homography between the abbreviation and the preposition, everywhere the latter occurs, the rule rs → r is added.

The fifth most important rule, v → o, could be tracked down to a tagging peculiarity in the 14th century corpus. In the (intermediary) version we used the affix of separable verbs are tagged as a verb. A separable verb with the affix *op* (actually the preposition "on") was found in a charter from Gent[9], *op [te] rechtene* ("to be established"). Now in general *op* has a higher frequency then *vp* in the whole corpus[10], so one would expect *op* to be the normal form. Therefore rewrite rules — both for *vp* and *op* — should start with *o* and not *v*. However within the (small) class of op/vp-occurrences that are tagged as a verb, *vp* is more frequent. So the normal form for the verb-tagged *vp* or *op* becomes *vp*, and therefore when *op* is encountered with a verb tag, the rule v → o is generated. Then this rule is associated to the form *op*, and as soon as *op* is encountered (without regard to its POS tag), v → o is added as a feature.

5.3 Cluster analysis for the locations

In the aforementioned verification task we looked at all charters separately. It was demonstrated that the specific spelling can be used to build a verification model for a location. However, this still does not necessarily mean that we were in fact modelling the location only. It might very well be possible that we based our models partially on the specific behaviour of certain authors.

To counter this argument, we need to show that there is a high degree of coherence between the models for neighbouring places. From this perspective we will create an orthographic profile for each location and perform a cluster analysis on these fingerprints. If the outcome matches with the geographical reality, the claim that only individual authors are being modelled can be falsified.

[9] I241p-399-01

[10] Overall there are 1283 occurrences of *vp* versus 4658 of *op.*

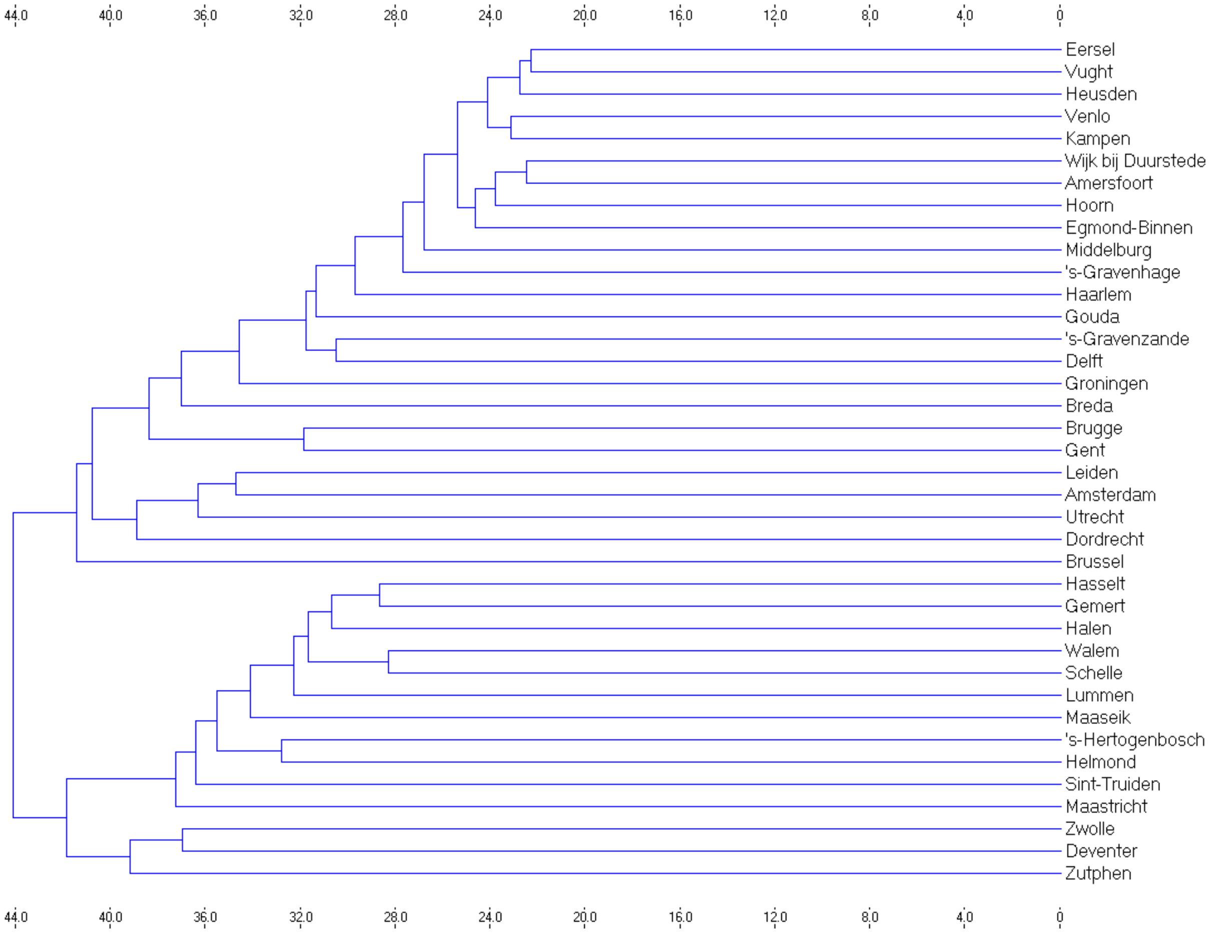


Figure 5.1: Cluster analysis using the orthographic variants for each location in the 14th century corpus.

5.3.1 Orthographic profiles and analysis

When creating an orthographic blueprint for a location we want all peculiarities of that place to be included. This means we need to ensure that the observations done for a single charter also have an influence on the profile for the place where that charter was written. Therefore we decided to create a feature vector containing all orthographic observations for any document from that place. In other words, we combined all of the binary vectors for a location's charters with an OR filter. Both for the features based on the variants and the rules we created such location-wide vectors and performed a cluster analysis.

These analyses are all made with Orange (Demsar & Zupan, 2004). First we calculated all distances between the feature vectors representing a location, using the standard Euclidean distance metric. Afterwards, an agglomerative hierarchical cluster analysis was performed on the resulting distance matrix. As the inter-cluster distance measure we chose for complete linkage[11], this proved to give the clearest results.

5.3.2 Cluster analysis using spelling variants

Using these "orthographic profiles" we calculated the distances between all places and performed a cluster analysis, of which the outcome is shown in figure 5.1.

The outcomes of this cluster analysis fit amazingly well with the geographical reality. Some clear tendencies that can be observed are:

[11] Also known as the farthest neighbour method. For details, see Johnson (1967).

- The largest clusters are roughly formed by the eastern (Hasselt, ..., Zutphen) and western (Eersel, ..., Brussel) locations.
- Places in or close to Limburg form one large group in the cluster: Hasselt, Gemert, Halen, Lummen, Maaseik, Helmond, Sint-Truiden, Maastricht. Two locations that are grouped together, Walem and Schelle, form a part of this Limburg cluster although they are located outside this region. They constitute however the only representations of the region between Antwerpen and Mechelen, which can explain the outcome.
- Large cities are brought together, e.g. Amsterdam, Leiden and Utrecht.
- Some smaller groups also catch the eye, for all of which the grouping can be explained by the minor geographical distance:
  - Zwolle, Deventer, Zutphen
  - Brugge, Gent
  - ’s-Gravenzande, Delft
  - Eersel, Vught, Heusden
  - Wijk bij Duurstede, Amersfoort

These results indicate we can state that the spelling profiles we created for each location largely match with the geographical reality.

5.3.3 Cluster analysis using spelling rules

In figure 5.2 the outcome is given for the cluster analysis based on the spelling rules. While largely in line with the clusters based on the full orthographic variants, we can see an even higher degree of detail here:

- The eastern triangle Zwolle - Deventer - Zupthen, together with southeastern places in Limburg:
  - Maaseik, Sint-Truiden, Maastricht
  - Hasselt, Halen, Lummen
- Brussel acts again like an island. The same goes for the combination of Brugge and Gent.
- The coast area of Holland (’s-Gravenzande, ..., Delft) is clearly recognizable.
- The triangle Amsterdam-Utrecht-Leiden appears here as well.
- The cluster Eersel, ..., ’s-Hertogenbosch largely spans the area of North Brabant and contains mainly subclusters that make sense:
  - Eersel, Vught
  - Gemert, Helmond
  - Heusden, Wijk bij Duurstede

The only cluster that is rather strange is Venlo and Kampen, as the distance between these places is larger then 130 km from each other. On the map in figure 5.3 we indicated some of the most prominent clusters we mentioned here.

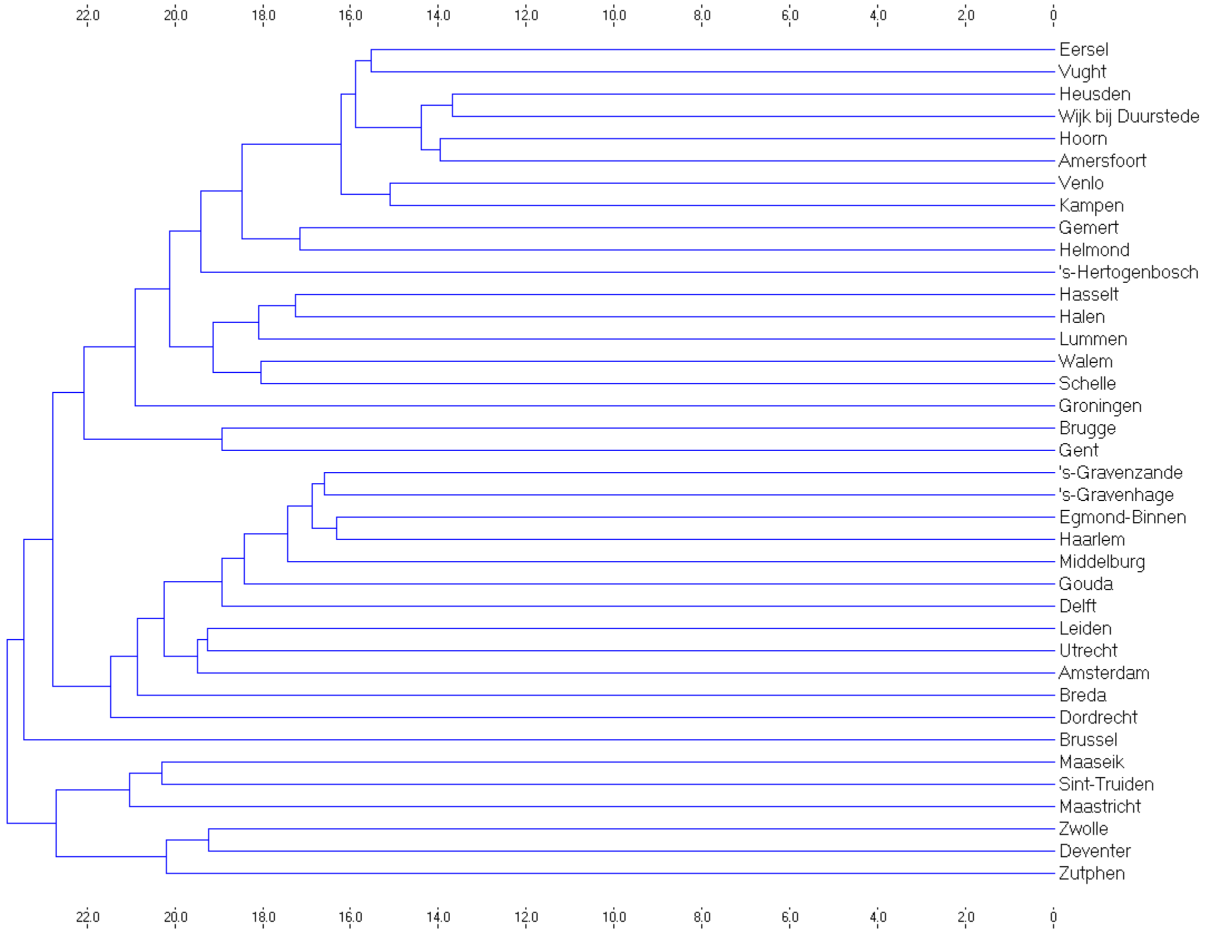


Figure 5.2: Cluster analysis using the orthographic rules for each location in the 14th century corpus.

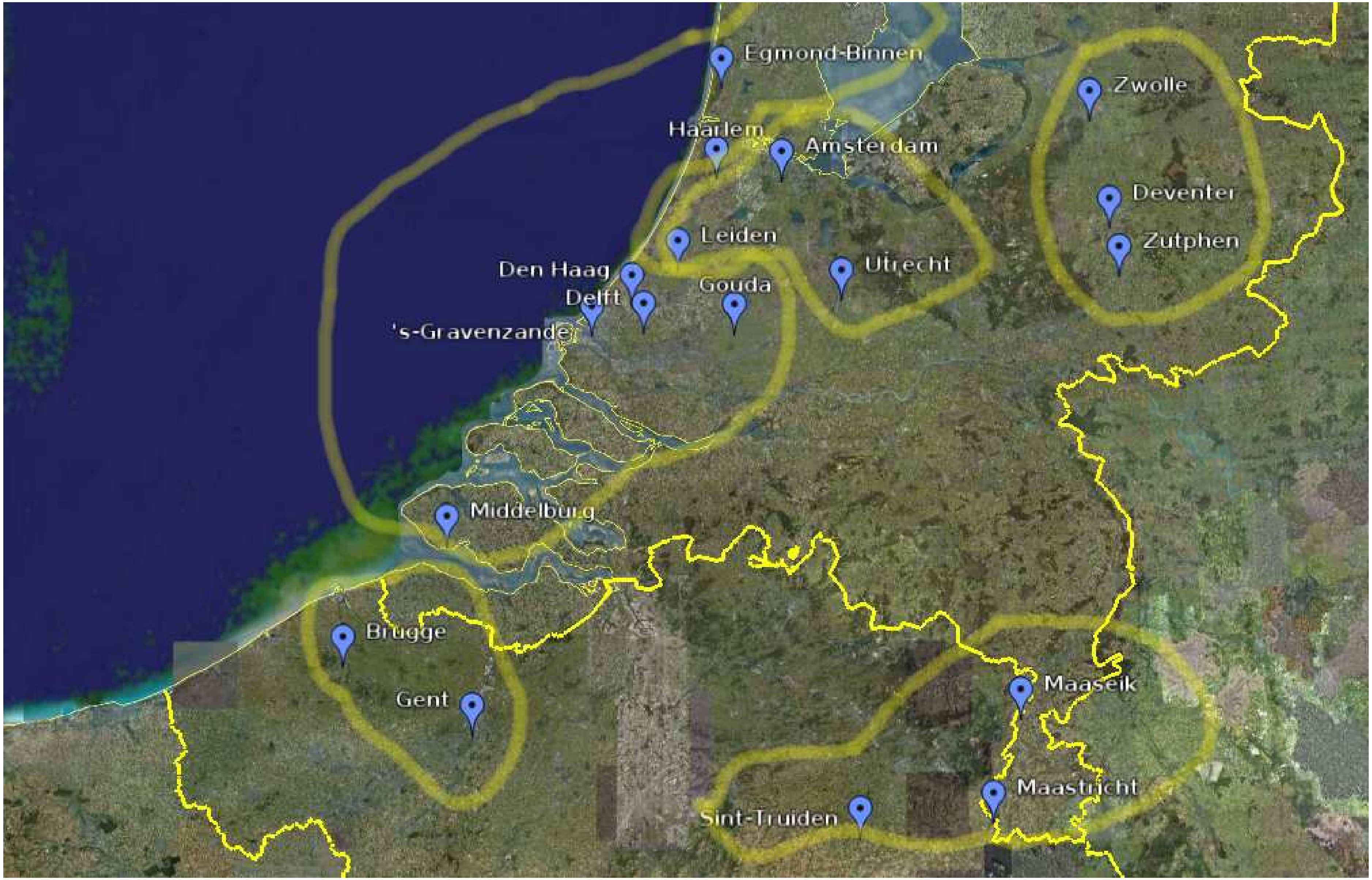


Figure 5.3: Some clusters that were found using the spelling rules features for the 14th century locations.

5.4 Induction of the date

To complete the assessment of the spelling profiles as the basis for classification we turn again to the dating task.

5.4.1 Variants

Figure 5.4 shows the verification results for the 14th century year intervals, using the full orthographic variants as attributes. Overall, the charters that are accepted by a year interval model do belong to that or a closely related interval. However a majority (at least 60%) of the documents belonging to a class is not correctly recognized as such, which is indicated as well by the relatively high average FRF-score of 0.61. It seems that the orthographic variant features are not optimally suited to build a dating model. Therefore we also decided to skip the century halves classification with this data.

5.4.2 Rules

5.4.2.1 Intervals

In comparison to figure 5.4, the rule-based induction of the time slices (figure 5.4) performs better. Both the darker diagonal and the lower mean FRF-score of 0.49 prove this.

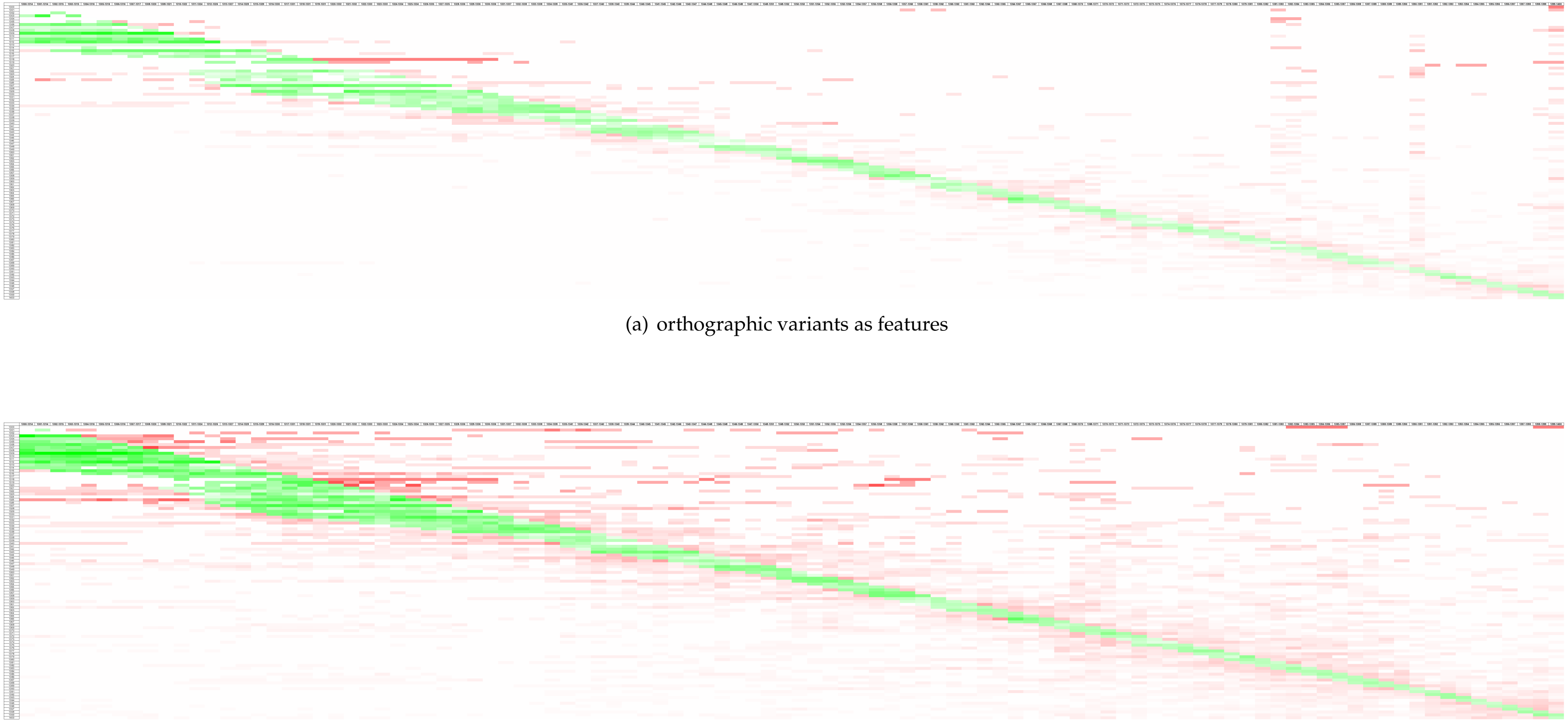
5.4.2.2 Century halves

The result of the century halves verification using the spelling rules, as shown in figure 5.5, is very similar to that based on the character trigrams. The FRF-score decreases (and thus the prediction quality increases) slowly, reaching the lowest level between 1370 and 1380, and then starts to raise again. The peaks at the begin and end of the chart can be explained by the fact that the models there tend to accept or reject all documents, leading to a high FRF-score.

Possibly the good results when 1370 is used as a boundary are due to the fact that the charters that belong to the last part of the century are somehow more consistent (with regards to the language, the subject or the use of fixed expressions). Therefore they might be better easier to recognize.

5.4.3 Summary

The results mentioned above remind us of the confusion table 4.13 for the same task, using the character trigram count as features. There the same tendency — a high false reject rate — can be observed. Summing up it seems that the KNN machine learner, independent of the features used, tends to overlook an important part of the charters for the induction of year intervals. At the other side of the spectrum we saw LProf, which in turn designates too many charters to each class.



(a) orthographic variants as features

(b) orthographic rules as features

Figure 5.4: Confusion matrix for the year interval verification in the 14th century material, using KNN.

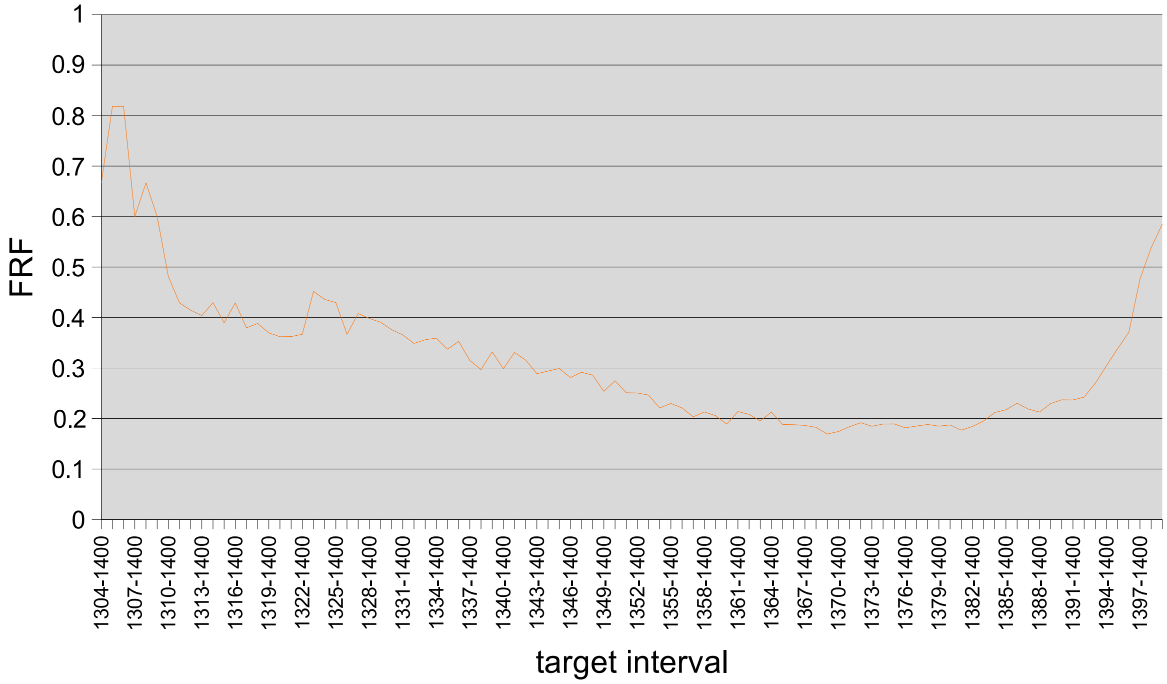


Figure 5.5: FRF scores for the century halves recognition in the 14th century material, using the spelling rule features.

Chapter 6

Results and discussion

In this chapter we provide a summary of the experiments that were carried out. We look at how the classification based on orthography compares to that based on character n-grams. Afterwards, the overall outcomes are discussed and related to the existing literature. At last we make some suggestions for future research on this topic.

6.1 Results

6.1.1 13th versus 14th century

At the time of setting up the experiments, we decided to build separate models for the two corpora we used. This now gives the opportunity to compare the results for both of them, which is necessarily restricted to the verification based on character trigrams. It can be concluded without any doubt that both the dating and the localisation works better for the 14th century material. The exact reason for this is more difficult to indicate; we will come back to this with some hypotheses in the discussion section.

6.1.2 Localisation

Overall, the results of our localisation experiments are encouraging. This means there is evidence for diatopic language variation. As mentioned in Van Dalen-Oskam et al. (2004), this can be caused by dialect influences on the charter language, or by the occurrence of fixed expressions, a kind of "charter language", which are typical for a region. The character n-grams possibly contain traces of such local charter language; the orthographic rules should however not be heavily influenced by it. As these rules also result in a good classification, we dare to say that we are indeed modelling the dialect use in the charters — at least for the 14th century.

6.1.2.1 13th century

For the 13th century charters, Linguistic Profiling outclasses KNN with a mean FRF score of 0.25 versus 0.38. For two cities (Dordrecht and Grimbergen), KNN performs better nevertheless. Recall that we only used the character trigrams for this corpus, as the annotations necessary to extract the spelling variants were not available. A detailed evaluation of both verifiers is available in table 4.4.

6.1.2.2 14th century

Table 6.1 summarises the outcome for all methods we tested on the 14th century corpus. Two winners can be appointed: the trigrams (average FRF: 0.11) and the orthographic rewrite rules (average FRF: 0.12), both in combination with a KNN classifier. However, it should be noted that the rules are not relying on any proper noun, while the trigrams include even the name of the location itself. Moreover, the rule features are Boolean and each feature vector contains only 3581 elements, in comparison to 8407 elements for the trigrams.

Although both methods have a comparable mean FRF score, they sometimes show large differences for the verification of the same locations. Inspired by this observation, we decided to create a combination of both, using a majority voting system. To avoid ties, we also added a third verification model: the one using the full orthographic variants. For every charter the outcome that has been given by a majority of all three verification models is chosen. The FRF-scores for this newly created model are given as well in table 6.1. Only for Breda a better score is reached then those of each individual model; generally it is not worth the effort to explore this method further.

6.1.3 Dating

Dating the charters is task that is of another difficulty order than the localisation: it is much more problematic. Predicting the originating century works almost perfect, using the n-grams — although this might be due to the difference between the two corpora.

As soon as a finer granularity is required, the verification performance decreases dramatically. Time interval models have the tendency to accept or reject a large part from the documents that are tested. They thus seem to be unable to grasp all of the subtleties that can give an indication on the period of writing.

6.1.3.1 13th century

For the whole 13th century corpus, only the LProf-based verification produces a sensible outcome (mean FRF-score: 0.29). However it suffers from a high rate of false accepts. KNN is wide off the mark, as the mean FRF-score of 0.94 clearly shows.

Even when we switched to the much easier task of predicting whether a charter was produced before or after a certain year, the results did not improve dramatically.

Because of the high proportion of charters from Brugge, we also created a specific dating model for the 13th century charters from Brugge. This time KNN gave an acceptable result. Though the FRF-score of 0.60 is worse than the 0.29 of the verification model for all locations, the false accepts are not as prominent for Brugge, resulting in a more trustworthy model.

6.1.3.2 14th century

The outcome for the 14th century models is more or less the same than that for the 13th century, with the noteworthy exception that the KNN model, even when fed with the relative trigram frequencies, delivers results that make sense.

As we also have the spelling-based features available for this century, we ran the dating task with them as well. In table 6.2 a comparison between all of the tested verifiers is given. We can learn from it that the combination LProf and the trigram frequency delivers the best results. Again it should be noted that this should be interpreted with some caution: the general tendency is that LProf accepts too many charters, while KNN does the opposite. These results should hence be compared bearing the purpose of the verification in mind.

The results of the KNN-model trained on the spelling rules are about the same as those for the trigram-based model, although the former does not have access to the proper nouns. In that sense it thus has a higher performance.

Analogously to the 13th century, distinguishing the charters before and after a certain year did not result in important accuracy improvements either, with a mean FRF score of 0.26 (for the best performing combination, a KNN classifier based on n-gram frequency).

Table 6.1: Overview of the FRF-scores for all localisation methods for the 14th century charters.

FRF	trigram, KNN	variants, KNN	rules, KNN	majority vote
Amersfoort	0.03	0.05	0.03	0.03
Amsterdam	0.03	0.03	0.03	0.02
Breda	0.16	0.13	0.19	**0.09**
Brugge	0.09	0.12	**0.03**	0.06
Brussel	0.21	0.43	**0.12**	0.28
Delft	0.09	0.11	**0.03**	0.06
Deventer	0.03	0.08	0.05	0.04
Dordrecht	**0.04**	0.11	0.10	0.07
Eersel	**0.00**	0.54	0.05	0.05
Egmond-Binnen	**0.33**	0.82	0.67	0.82
Gemert	**0.18**	0.33	0.33	0.33
Gent	**0.18**	0.54	0.38	0.33
Gouda	0.02	0.02	0.01	0.01
Groningen	0.02	0.08	0.03	0.03
Haarlem	0.13	0.14	**0.07**	0.1
Halen	0.11	0.25	**0.05**	0.11
Hasselt	0.43	0.43	**0.18**	0.43
Helmond	0.03	0.15	0.04	0.05
Heusden	0.00	0.05	0.00	0
Hoorn	0.11	0.33	0.11	0.18
Kampen	0.08	0.05	**0.00**	0.05
Leiden	0.04	0.10	0.06	0.05
Lummen	0.03	0.08	0.05	0.05
Maaseik	0.05	0.18	**0.00**	0.05
Maastricht	**0.02**	0.18	0.09	0.09
Middelburg	0.33	0.43	0.33	0.33
Schelle	0.03	0.18	0.05	0.05
Sint-Truiden	0.54	0.67	**0.33**	0.43
Utrecht	**0.01**	0.09	0.06	0.05
Venlo	0.00	0.05	0.00	0.00
Vught	**0.00**	0.33	0.11	0.11
Walem	0.11	0.18	0.11	0.18
Wijk bij Duurstede	**0.05**	0.11	0.11	0.11
Zutphen	0.60	0.57	**0.13**	0.45
Zwolle	**0.08**	0.29	0.11	0.13
′s-Gravenhage	0.05	0.06	0.05	0.04
′s-Gravenzande	**0.03**	0.21	0.11	0.11
′s-Hertogenbosch	**0.08**	0.33	0.25	0.29
Average	0.11	0.23	0.12	0.15

Table 6.2: Comparison between interval verification methods for the 14th century (for a sample). The average has been calculated on all test data, except for the trigram & KNN combination.

interval	trigram, KNN	trigram, LProf	variants, KNN	rules, KNN
1300-1314	0.35	0.21	0.59	0.33
1310-1322	0.35	0.23	0.60	0.37
1320-1332	0.59	0.27	0.68	0.40
1330-1336	0.43	0.36	0.56	0.41
1340-1345	0.40	0.31	0.54	0.52
1350-1353	0.53	0.29	0.54	0.47
1360-1363	0.53	0.31	0.64	0.58
1370-1372	0.53	0.39	0.58	0.50
1380-1382	0.63	0.36	0.76	0.61
1390-1391	0.60	0.39	0.68	0.65
Average	0.49	0.32	0.61	0.49

6.2 Discussion

Of course the results mentioned above give some food for thought and comments. We will deal with them in this section.

6.2.1 Rediscovery of known linguistic phenomena

We already mentioned during the qualitative analyses that quite a lot of the most distinguishing features correspond to phenomena that have been documented in the literature on Middle Dutch. Obviously it is impossible to discuss all of the generated top features (of which a list can be found in appendix B). However we found a striking correspondence between a set of rules generated for some locations in Limburg (or by extension the whole east) and the article of Van Reenen & Huijs (1999) on the use of *g* versus *gh* in the 14th century Middle Dutch.

In this article a relation is postulated between the pronunciation and the graphemic representation of the g. The authors state and argument convincingly that the velar fricative g is written as *gh*, as where the palatal g (even nowadays known as the "soft g" which is typical for Limburg) is represented as *g* in writing.

Now in our model analyses[1] we found the following rules for some cities in Limburg (Venlo, Maastricht) and the east (Zutphen, Kampen):

- Maastricht: rg → rg (not: rgh/[...])
- Venlo: rgh → rg, fg → fg (not: fgh), rg → rg (not: rgh/[...])
- Zutphen: ngh → ng
- Kampen: rfg → rfg (not: rfgh/[...])

We thus can conclude that we independently rediscovered the existence of the palatal g in the southeast (or at least its orthographic equal) and that according to the feature importance score this was found to be a linguistic phenomenon with a high distinguishing power.

[1] The top 10 features for each location model, using the spelling rules as features and the relief score to determine the most important features.

6.2.2 The 13th versus the 14th century

In general it was found that the classification tasks were more difficult for the 13th century charters. This could be due to several reasons. First, it is possible that the metadata (the dating and localisation) in general contains more errors in the Corpus Gysseling.

On the other hand, there also might be a tendency in the 14th century documents towards a more unified orthography in each location.

Finally it is possible as well that in the 14th century some place-dependent fixed expressions ("dictaten") found their way to the writing centres. These could make the recognition task a lot easier, while they should not be considered as an indication of language variation.

6.2.3 Corpus purity

As mentioned in the chapter on the corpora we are using, it is not guaranteed that the location and date mentioned for each charter is always correct. That is the price that is paid when working with old historic material. We can however state that the majority of the charters seems to be well localised (and, although with less certainty: dated) in the originating corpora, as otherwise it would not have been possible at all to build any consistent verification model. In that sense, the outcome of our research strengthens the assumption of overall correctness with regards to the provided metadata in both corpora.

6.2.4 Getting grip on the time dimension

The dating of the charters was without any doubt the hardest nut to crack. This is partially related to the circumstances. It was almost impossible to study the time dimension of the charters while keeping the location dimension constant. Only for Brugge in the 13th century corpus this could be done and this resulted in an improved verification model.

For the other dating models it stays difficult to pinpoint the cause of their suboptimal behaviour. It might be related to the fact that instead of a date a set of locations (which accidentally corresponds to a certain time span) is modelled. Most probably it is also inherently more difficult to model diachronic variation in comparison to diatopic variation, as the former has more fuzzy boundaries.

6.2.5 Trade-offs

A final observation can be made on the proportion between the false accepts and the false rejects that come with each model. Quite some difference can be seen for some models, mainly depending on the machine learning algorithm that was used. In general LProf accepts too many charters, while KNN is often too picky. Depending on the exact research aims one of both can be chosen. It would without any doubt be interesting as well to do further research on the selection of a reject/accept threshold for this data.

6.3 Suggestions for future research

Having the results and associated discussions at the back of our mind, there are still some suggestions for further research we would like to mention here.

- On this moment, some highly frequent homographs are causing noise in the list of rules created for each word form. One could think of some manually crafted rules to ensure that very low-frequent homographs do not result in disturbing rewrite rules anymore.

- Somehow related, it would of course be interesting as well to repeat the orthography-based experiments based on the version of the 14th century corpus with the corrected POS tags.

- During the course of this study some new material was added to the corpora we used. Our experiments could be replicated using this broader data.

  - A lot of Flemish charters were added to the 14th century corpus, resulting in a better north-south equilibrium.
  - This addition also features hundreds of charters from Brussels. That makes it possible to create another dating model for a single location.

- The orthography-based feature vectors for a location's charter collection that were used for the clustering could be studied more closely, maybe they reveal more on typical properties for a place.

- The same vectors could maybe be used to create dialect maps with continuous transitions, as was done by Nerbonne et al. (1999).

- For the 13th century corpus, one could try to derive lists of orthographic variants, in spite of the missing tagging information, using advanced orthographic distance measures, as e.g. described by Kempken (2005) and Pilz et al. (2006).

- Some locations (Zutphen, Hasselt, Sint-Truiden, Egmond-Binnen) are apparently more difficult to model. It would be good to know why they cannot be modelled as well as other places, based on an analysis of their feature vectors and the text of the charters in question.

Chapter 7

Conclusion

At the end of this study, we look back at the research questions that were formulated in chapter 2 and answer them.

For those charters whose origin is well known, is it possible to recognize the location in which they were written on the basis of orthographic features, initially character n-grams?

We can recognize almost all locations with a high success rate, relying on character n-grams. The charters from the 14th century corpus give better results with regards to localisation than those of the 13th century.

The same question as above for the dating of the charters.

The dating of the charters shows to be feasible to a certain level, however this does not work as well as the localisation. Each setup we tried either resulted in a non-negligible amount of false positives and/or negatives.

Can we develop a representation for orthographic variance which relates variant spellings to each other or to a canonical base form? If so, how do the classification results compare to the character n-gram approach and how does a combination of both methods perform?

We can define a normal form for each type, and derive rewrite rules for each consonant and vowel cluster, describing the mapping from the normal form towards the encountered type. With this approach, good classification results — comparable to those using the character n-grams — are achieved, even when the proper nouns were excluded from the training and test data. Moreover, these rules also represent more directly the orthography than the lexical information of a word form.

A combination of both methods does not lead to improved classification results.

What is the relation between the granularity of the classes to predict (e.g. years versus decades and cities versus regions) and the accuracy of the classification? Does it pay off to use a multilevel classifier as an alternative to a single classifier?

For the localisation, it turned out that choosing larger classes is undesirable, as the results with the smallest possible classes are already satisfying and selecting larger areas would be highly arbitrary.

As for the dating, larger classes only result in a slightly better class recognition. Even with large classes, e.g. intervals of 50 years, the dating does not work optimally.

On the whole, we opted for a two-level approach by creating separate verification models for the 13th and the 14th century corpora. This choice was inspired by the fact there are near perfect models for century (or corpus) recognition. It was also a necessity, as we only had tagging information available for the 14th century.

It is generally known that proper names and especially toponyms are less prone to diachronic changes than other word classes. From this perspective the question arises whether we should exclude proper names from the material that will be used by the orthographic temporal classifier. Additionally we ask ourselves what their influence on the classification results is – if they are included.

We answer this question for the 14th century corpus, as this contains the necessary part of speech tags to distinguish proper nouns.

Indeed it was found that proper nouns have a lower variation rate than other words. The deviant forms are also less frequently used. Therefore we concluded that — at least for the generation of spelling-based features — proper nouns should not be taken into account.

For the trigram frequency features, the proper names do sometimes play an important role, as we found that in some cases trigrams connected to the location name turned out to belong to the most significant attributes. However, we did not check the influence of proper nouns on the trigram-based classification in depth.

Bibliography

Baayen, H. (2001). *Word Frequency Distributions*. Springer.

Baroni, M. (2007). Corpus linguistics: An international handbook, chap. Distributions in text. Mouton de Gruyter. To appear.

Cavnar, W. B. & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, (pp. 161–175), Las Vegas, US.

Chang, C.-C. & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2007). TiMBL: Tilburg Memory Based Learner, version 6.0, Reference Guide. Tech. rep., ILK Technical Report 07-05.

De Vries, M., te Winkel, L., et al. (1882). Woordenboek der Nederlandsche Taal [Dictionary of the Dutch language]. 's-Gravenhage/Leiden etc. *M. Nijhoff/AW Sijthoff etc. Also available on CD-rom: Het Woordenboek der Nederlandsche Taal op CD-Rom*.

Demsar, J. & Zupan, B. (2004). Orange: From experimental machine learning to interactive data mining, 2004. *White Paper, Faculty of Computer and Information Science, University of Ljubljana*. `http://www.ailab.si/orange`.

Gysseling, M. & Pijnenburg, W. (1987). *Corpus van Middelnederlandse teksten (tot en met het jaar 1300)*. M. Nijhoff.

Instituut voor Nederlandse Lexicologie (1998). *Cd-rom Middelnederlands*. Sdu Uitgevers.

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Kamps, J., Koolen, M., Adriaans, F., & de Rijke, M. (2007). A cross-language approach to historic document retrieval. In L. Burnard, M. Dobreva, N. Fuhr, & A. Lüdeling (eds.), *Digital historical corpora- architecture, annotation, and retrieval*, no. 06491 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Kempken, S. (2005). *Bewertung historischer und regionaler Schreibvarianten mit Hilfe von Abstandsmaßen*. Ph.D. thesis, Diploma Thesis. University of Duisburg-Essen.

Kessler, J. & Oosterman, J. (2007). *Seer scoon ende suyver boeck, verclarende die mogentheyt Gods, ende Christus ghenade, over die sondighe menschen (refereinen 1567), Anna Bijns*. DBNL. `http://www.dbnl.org/tekst/bijn003refe03_01/`.

Kloeke, G.G. (1927). *De Hollandsche expansie in de zestiende en zeventiende eeuw en haar weerspiegeling in de hedendaagsche Nederlandsche dialecten. Proeve eener historisch-dialectgeographische synthese.*. Nijhoff.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *Proceedings of the European Conference on Machine Learning*, (pp. 171–182).

Mooijaart, M. (1992). *Atlas van vroegmiddelnederlandse taalvarianten*. LEd.

Mooijaart, M. & Van der Heijden, P. (1992). Linguïstische en geografische afstand in dertiende-eeuws Middelnederlands. *Taal en tongval*, (2), 188–216.

Nerbonne, J., Heeringa, W., & Kleiweg, P. (1999). Comparison and classification of dialects. In *Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics*, (pp. 281–282).

Pilz, T., Luther, W., Fuhr, N., & Ammon, U. (2006). Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, *21*, 179.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rem, M. (2003). *De taal van de klerken uit de Hollandse grafelijke kanselarij (1300–1340). Naar een lokaliseringsprocedure voor het veertiende-eeuws Middelnederlands.*. Stichting Neerlandistiek VU, Amsterdam.

Van Dalen-Oskam, K., Rem, M., & Wattel, E. (2004). Taal in verandering, chap. De opkomst van bezegelen in de veertiende eeuw: een methodologische vergelijking, (pp. 147–160). Münster: Nodus Publikationen.

Van Halteren, H. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Trans. Speech Lang. Process.*, *4*(1), 1.

Van Reenen, P. & Huijs, N. (1999). De harde en de zachte g. *Taal en tongval*, (12).

Van Reenen, P. & Mulder, M. (1993). Een gegevensbank van 14 de-eeuwse Middelnederlandse dialecten op computer. *Lexikos*, *3*.

Van Uytvanck, D. (2007). Evaluatie, selectie en opwerking van corpora voor onderzoek naar spellingsgebaseerde classificatie bij historisch Nederlands. Unpublished. Available at `http://www.student.ru.nl/dieter.vanuytvanck/thesis/master/verslag.pdf`.

Appendix A

Excluded charters from the CG1

A.1 Chancellery charters

38, 47, 64, 65, 66, 103, 234a, 298, 308, 309, 327, 367, 370, 401, 402, 411, 413, 430, 466, 548, 556, 617, 623, 624a, 626, 627, 628, 629, 630, 715, 734, 740, 757, 758, 759, 767, 778, 780, 787, 807, 824, 832, 844, 846, 860a, 868a, 892, 893, 913, 914, 915, 918, 919, 922, 926, 927, 928, 936a, 937, 969a, 970a, 977, 978, 979, 996, 1001, 1005, 1047, 1090, 1100a, 1100b, 1113a, 1113b, 1114a, 1115a, 1116a, 1116b, 1116, 1118a, 1119a, 1119, 1133, 1136, 1137, 1149, 1163, 1184, 1185, 1221a, 1224a, 1225, 1226, 1240, 1258, 1284a, 1286a, 1293, 1296a, 1318a, 1321, 1338, 1345a, 1351a, 1367, 1368a, 1368, 1377a, 1398, 1402a, 1410, 1416, 1435, 1446, 1451, 1453a, 1460, 1464, 1489, 1490a, 1496b, 1496c, 1496d, 1496e, 1501a, 1507, 1572, 1608, 1613a, 1613b, 1619, 1623, 1625a, 1628, 1630a, 1630, 1631, 1632, 1633, 1634a, 1634b, 1635a, 1637, 1675, 1676, 1686a, 1686b, 1687, 1688, 1690, 1693a, 1695a, 1695b, 1695c, 1696, 1702a, 1704, 1722a, 1729, 1730, 1734, 1735, 1736, 1737, 1738, 1739, 1755, 1760a, 1763, 1764, 1765, 1781a, 1785a, 1788a, 1788b, 1793, 1798a, 1798b, 1798c, 1798d, 1798e, 1798f, 1801, 1802, 1803, 1804, 1805a, 1806a, 1808, 1809a, 1809, 1813, 1826, 1827, 1828, 1829, 1830, 1832, 1833, 1835, 1836, 1837, 1842, 1843, 1844, 1845, 1846, 1847, 1860, 1873a, 1873, 1901, 1912

A.2 Copies

4, 11, 13, 15, 41, 105a, 105b, 140a, 141a, 150a, 184, 241, 256, 339, 347, 388a, 446a, 461, 480, 500a, 504a, 599, 692a, 724a, 724, 727, 749a, 787a, 803, 804, 805a, 807a, 807b, 882a, 884, 987, 1032a, 1052, 1053, 1112a, 1130a, 1176a, 1187, 1217, 1218, 1304, 1522, 1599

Appendix B

Most indicative features per location

Each feature is followed by its RELIEF-score (Kononenko, 1994), between square brackets.

B.1 Spelling variants

's-Gravenhage: houder [0.92637], ghelt [0.73012], staende [0.61525], quam [0.59371], scepen [0.59013], beleghen [0.58544], coept [0.57766], luden [0.5605], briefs [0.52818]
's-Gravenzande: lettren [0.72486], kennessen [0.69324], orconden [0.60465], bezeghelt [0.58391], renten [0.56919], staende [0.47319], zoen [0.43922], verliede [0.40367], dezen [0.39675]
's-Hertogenbosch: aen [0.57848], name [0.54486], wij [0.50439], welker [0.49962], kennissen [0.49538], dekene [0.49338], scepen [0.46544], hanghen [0.45462], waerheyt [0.42888]
Amersfoort: briefs [0.73457], ghecomen [0.62878], scoute [0.61688], hofstede [0.59588], sone [0.5576], ghenen [0.53817], side [0.52483], vor [0.52217], beseghelt [0.51663]
Amsterdam: orconden [0.83574], kennen [0.6832], orconde [0.63367], mit [0.58465], brieue [0.5759], quam [0.54192], tachtich [0.53816], onsen [0.47666], sinte [0.46354]
Breda: en [0.84213], dat [0.80174], van [0.78653], die [0.61283], den [0.55204], te [0.43096], desen [0.40217], tnegentich [0.38706], alle [0.31964]
Brugge: stede [0.64805], vpten [0.59871], tiden [0.569], tachtentich [0.54762], lren [0.546], lieden [0.54557], kennessen [0.50852], quite [0.50529], ghedaen [0.46114]
Brussel: waerheit [0.49248], zeghele [0.47337], sente [0.44839], vors [0.39563], de [0.37227], soe [0.35943], ane [0.35891], letten [0.35243], es [0.34842]
Delft: gherechte [0.63019], driehondt [0.57274], vollen [0.55067], ghecoft [0.51121], dusent [0.47145], ontfaen [0.46966], voir [0.45703], heeft [0.3886], renten [0.37864]
Deventer: dusent [0.61024], steet [0.55761], groete [0.55018], bescreue [0.53361], briefs [0.5078], mit [0.44683], sente [0.4327], sine [0.42627], brieue [0.42626]
Dordrecht: ghegheuen [0.43017], brieue [0.40926], met [0.35715], kennen [0.3284], es [0.29004], an [0.28489], te [0.28136], maken [0.27723], orconde [0.27162]
Eersel: gheheiten [0.73285], ghemeynen [0.55757], vor [0.537], onder [0.52919], tughen [0.51733], zoen [0.4841], vors [0.46862], he [0.44691], onse [0.43919]
Egmond-Binnen: abt [0.8], sullen [0.51339], cont [0.49135], driehondert [0.43368], dusent [0.42973], was [0.41495], vorseyde [0.396], abts [0.396], jc [0.389]
Gemert: openen [0.81547], waerheit [0.70225], tughen [0.63938], comen [0.60648], sijns [0.59505], ordens [0.56771], ghehanghen [0.54696], wij [0.54227], aen [0.54145]
Gent: sente [0.49972], lieden [0.47157], wiue [0.46857], de [0.46332], ouer [0.45795], op [0.43574], huus [0.43191], heeren [0.40605], vp [0.40147]
Gouda: lyede [0.83128], orkonden [0.82969], kennen [0.76837], m [0.72805], panden [0.56344], binnen [0.533], quam [0.51772], beleghen [0.50798], soen [0.45573]
Groningen: stad [0.77393], drehondert [0.69263], opene [0.66839], breue [0.58386], do [0.58044], bekenne [0.56183], de [0.54787], hie [0.49155], wy [0.47091]

Haarlem: zeghelen [0.66321], brieue [0.62748], quam [0.54524], zoe [0.4356], bizeghelt [0.42868], legghende [0.42024], scepene [0.41973], verliede [0.41654], erue [0.40582]
Halen: lren [0.89033], wij [0.83455], halme [0.66033], tughen [0.63433], sien [0.62567], vtiende [0.585], bimpt [0.585], bimpde [0.585], hant [0.58067]
Hasselt: onderpant [0.55733], hof [0.55533], rechte [0.48305], waerheyt [0.46986], hebbe [0.4681], de [0.4573], vors [0.4361], sighel [0.43233], vercocht [0.39933]
Helmond: metter [0.89533], hoem [0.77028], stat [0.74082], deser [0.7259], goet [0.70555], zeghel [0.62297], gheloeft [0.61471], letten [0.60288], vercoft [0.56242]
Heusden: tughen [0.934], dertienhondert [0.84685], seghele [0.82871], onder [0.77], scepen [0.74447], behouden [0.72105], gheloefde [0.67371], vorwaerden [0.67367], onse [0.66419]
Hoorn: bitaelt [0.77333], lesten [0.71714], mitten [0.68098], eersten [0.58602], mit [0.56201], quiit [0.56071], scoudich [0.54905], waert [0.52248], vercoft [0.51283]
Kampen: dies [0.94669], briefs [0.82022], arghelist [0.74404], bekant [0.7242], erfg [0.68855], vor [0.65782], beseghelt [0.63561], oerconde [0.6327], huus [0.60909],
Leiden: seghelen [0.51094], brieue [0.48568], beseghelt [0.4705], voir [0.46746], sone [0.4113], scepene [0.41063], mit [0.38633], onsen [0.34697], cond [0.34175]
Lummen: ocht [0.79033], oude [0.74633], ghehanghen [0.71638], tughen [0.71447], es [0.71314], erve [0.69257], side [0.68957], hant [0.68909], cont [0.65728]
Maaseik: brijf [0.802], eyn [0.77933], aude [0.766], stat [0.74884], end [0.604], eis [0.6], becant [0.6], vorwerden [0.588], mocht [0.588]
Maastricht: sint [0.70895], he [0.61318], getuygen [0.60743], dage [0.59543], joe [0.52076], wir [0.5171], guede [0.49657], hn [0.45557], voir [0.45352]
Middelburg: ghedaen [0.58661], sons [0.582], warent [0.54152], wesene [0.53986], orkonden [0.52867], sone [0.52745], ghecocht [0.47248], scepenen [0.46319], zi [0.43443]
Schelle: late [0.97999], sculdech [0.97942], wale [0.97799], beiden [0.95799], ghesciet [0.952], ochte [0.92833], ptien [0.92633], ane [0.895], wisen [0.815]
Sint-Truiden: onse [0.66125], stad [0.58833], cont [0.57613], meer [0.54657], beden [0.54092], selver [0.48833], oft [0.46703], vrouwen [0.46278], sich [0.44433]
Utrecht: gherechte [0.81757], ghenen [0.7883], wi [0.72236], zellen [0.71352], ghegheuen [0.67551], zien [0.67284], stade [0.66224], bliue [0.63424], quam [0.60643]
Venlo: argelist [0.96966], bekant [0.94466], onder [0.872], gelde [0.80438], ten [0.78757], stat [0.77181], aftedoen [0.74766], suelen [0.72281], to [0.71624]
Vught: sunte [0.63971], ghemeyns [0.59724], zeghel [0.55113], goet [0.50381], aen [0.456], waerheyt [0.45286], schependoms [0.44391], deen [0.44124], scepen [0.43095]
Walem: screef [0.81083], letteren [0.80428], achter [0.75595], waerheit [0.74126], metten [0.70019], scepenen [0.69671], ghelaten [0.62324], saluut [0.61524], liefs [0.60957]
Wijk bij Duurstede: kont [0.93266], briefs [0.89571], alsoe [0.84575], behoudelic [0.78466], eynde [0.76547], scout [0.74924], oerkonde [0.706], gericht [0.69733], voersc [0.69467]
Zutphen: vor [0.73548], vp [0.47014], to [0.45971], heb [0.45238], alse [0.41626], brieue [0.40678], ons [0.39145], dar [0.36962], brief [0.3581]
Zwolle: tijt [0.62671], pond [0.58813], briefs [0.5505], arghelist [0.54276], lude [0.51713], oere [0.49124], oer [0.467], mit [0.44761], erfg [0.44195]

B.2 Spelling rules

's-Gravenhage: lls -> llns [0.57473], ua -> ua (not: uae-va-eua-uee-oe-a-u-ue-uo) [0.38778], ls -> ls (not: lsw-lst-lss-l-s-ss-lz) [0.38471], ve -> voe [0.37955], voi -> voe [0.3768], b -> ft [0.37296], ii -> ii (not: ie-ij-i-y-ije) [0.36762], rt -> r [0.36755], q -> q (not: qw-c) [0.35798]
's-Gravenzande: tt -> ttr [0.49698], ey -> ee [0.48421], gh -> b [0.42274], nts -> nts (not: s-t-nst-nt-ndz-nds-ndts-ntz-ns-ts) [0.41521], vij -> vij (not: viij-vy-vi-ve-vii-vie-y) [0.38098], ve -> voe

[0.35198], u -> e [0.35162], rt -> r [0.34731], vo -> vo (not: voe-voi-vou-oi-oe-v-vue-ve-voo-uo-vv-vu-vi-va) [0.34149]
's-Hertogenbosch: oe -> ij [0.42189], e -> ey [0.39931], rh -> rh (not: rhw-rr-h-r) [0.39206], kn -> kn (not: cn) [0.38867], cn -> kn [0.38067], nc -> ngh [0.37381], ey -> ey (not: oy-ie-ii-ye-ij-ije-j-i-a-e-y-ee-ei-ae-eey-eye-eij) [0.37054], ae -> a [0.36858], ee -> ey [0.36689]
Amersfoort: fs -> fs (not: f-s-ffs) [0.71236], fst -> fst (not: ft-ffst-st) [0.56737], i -> ii [0.54639], rft -> rft (not: rfft-rfdt-rfth-rfc) [0.49825], u -> oe [0.48488], sd -> sd (not: s-d-rd-st-tsd) [0.43426], ll -> l [0.42575], ii -> ie [0.42448], ve -> vo [0.42381]
Amsterdam: rk -> rc [0.61108], u -> o [0.51399], ieue -> ieue (not: eve-ive-yve-ve-yeue-ie-eue-ue-iue-ieve) [0.48207], c -> c (not: nch-ck-tsc-ns-nt-nc-sc-l-k-ct-cs-ch-rst-t-s-x-g-n-ts-tz-tg-tc-gh-tsch-xs) [0.47442], br -> br (not: gh-rbr-g-b) [0.46526], ua -> ua (not: uae-va-eua-uee-oe-a-u-ue-uo) [0.45054], ve -> voe [0.43638], rt -> r [0.42546], final deletion: f [0.40579]
Breda: ue -> ve [0.50705], ee -> e [0.41192], o -> oe [0.39845], k -> gh [0.39536], ls -> ls (not: lsw-lst-lss-l-s-ss-lz) [0.38761], ij -> y [0.38646], ch -> gh [0.36404], i -> j [0.33105], rs -> r [0.3261]
Brugge: o -> v [0.65942], r -> s [0.50221], v -> o [0.45656], u -> ie [0.45474], lr -> lr (not: ll-ld-l-llr) [0.44554], o -> oe [0.41496], ui -> ui (not: u-uij-oue-oui) [0.40525], ft -> ft (not: dt-ftt-fd-cht-ffd-fft-ct-ght-fth-t-l-f-gt-bt-ht-chth) [0.39911], rs -> r [0.39672]
Brussel: u -> e [0.70447], f -> cht [0.49668], ll -> l [0.4433], ei -> ei (not: eij-ii-ij-ayo-y-i-a-e-ee-ey-ie-ye) [0.4189], oue -> oue (not: eue-ave-oi-oe-ou-au-aue-ouue-oeue-ouve-oeve-ouu-ua-ue-ve-oo-o-e-ovi-ove) [0.40567], rs -> r [0.39343], voe -> voe (not: v-voo-voi-voy-voii-vou-vue-vo-ve-uoe-uo-vu-vie-vae) [0.39298], y -> i [0.38585], q -> q (not: qw-c) [0.33785]
Delft: lt -> lt (not: ld-ldt-dt-l-t) [0.46177], lls -> lls (not: llns-llnts-llts-ll) [0.32606], ll -> lls [0.32606], k -> k (not: spr-chl-ch-xs-xsch-lk-ckl-cs-ck-br-rk-sch-l-gh-c-x-cst-csch) [0.31829], ntf -> ntf (not: ndf-nf) [0.31535], q -> q (not: qw-c) [0.30464], z -> s [0.29733], rc -> rc (not: rck-rch-rk-c) [0.27171], oe -> oo [0.26158]
Deventer: ii -> i [0.48118], l -> m [0.46946], u -> a [0.46746], v -> o [0.44585], ae -> ee [0.43564], scr -> scr (not: sc-cr-c-schr) [0.42611], nd -> d [0.39594], fs -> fs (not: f-s-ffs) [0.39241], ee -> ee (not: ye-ai-je-vie-oe-ue-y-uee-veee-ae-eij-vee-eei-eee-eey-j-i-e-a-ei-ey-eue-eve-ii-ij-ie) [0.38746]
Dordrecht: ieue -> ieue (not: eve-ive-yve-ve-yeue-ie-eue-ue-iue-ieve) [0.39783], final deletion: f [0.34515], sc -> sc (not: sk-sch-tsc-ghsc-nsc-s-d) [0.32137], rk -> rc [0.31946], d -> t [0.30777], rd -> rd (not: rdh-rdd-wrd-gh-d-r-rnd-rsd-nd-rt) [0.2939], br -> br (not: gh-rbr-g-b) [0.29051], nt -> t [0.2744], u -> u (not: vi-ee-uie-y-v-i-o-a-e-ae-ye-oy-ou-oe-ij-ie-uue-uuy-uy-uu-ui-ue) [0.26788]
Eersel: e -> ei [0.66365], rfgh -> rfgh (not:) [0.50766], d -> dd [0.4728], lt -> lt (not: ld-ldt-dt-l-t) [0.4652], f -> ft [0.46374], ee -> ey [0.44836], dd -> dd (not: bb-ddr-dt-tt-d) [0.43155], e -> ij [0.39917], ve -> vo [0.39671]
Egmond-Binnen: bt -> bt (not: bdt-bd-bdtz) [0.78524], dsh -> dsh (not: dh-tsh-sh) [0.65914], ii -> ie [0.44754], r -> rs [0.43193], ij -> ie [0.42777], rk -> rc [0.41906], ae -> e [0.41219], u -> e [0.40199], s -> r [0.3751]
Gemert: d -> ds [0.63514], ts -> ds [0.61433], ds -> ds (not: t-ts-dts-tz) [0.58514], vy -> vuy [0.56176], aue -> eue [0.55633], oe -> ij [0.52057], ei -> ei (not: eij-ii-ij-ayo-y-i-a-e-ee-ey-ie-ye) [0.50504], rs -> ns [0.49138], m -> ns [0.49138]
Gent: u -> ie [0.53192], rs -> r [0.50381], uu -> uu (not: uuy-ov-ou-ui-i-y-u-uy-ue-ij-vv-oe-uij) [0.4906], scr -> scr (not: sc-cr-c-schr) [0.37991], v -> o [0.36208], o -> v [0.35408], oue -> oue (not: eue-ave-oi-oe-ou-au-aue-ouue-oeue-ouve-oeve-ouu-ua-ue-ve-oo-o-e-ovi-ove) [0.35404], i -> ij [0.34262], iue -> iue (not: yue-ive-yve-ij-ieue-ieve-ie-ijue-ijve) [0.33354]
Gouda: d -> nd [0.73422], rc -> rk [0.70966], lls -> lls (not: llns-llnts-llts-ll) [0.5191], ll -> lls [0.5191], ye -> ye (not: ei-ie-ii-ij-ue-ey-ee-je-y-e-j-i-ije) [0.49508], c -> c (not: nch-ck-tsc-ns-nt-nc-sc-l-k-ct-cs-ch-rst-t-s-x-g-n-ts-tz-tg-tc-gh-tsch-xs) [0.41169], ue -> ve [0.38068], ve -> voe [0.34787], q -> q (not: qw-c) [0.34134]
Groningen: e -> ie [0.68009], ie -> ie (not: ije-yye-eie-iie-iee-iey-ijey-ii-ij-vie-eue-a-yi-ye-yey-ei-ee-ey-je-eij-ieei-eye-eve-ae-ieue-vi-e-i-u-y-vije) [0.67231], o -> u [0.65536], final deletion: t

[0.56799], ie -> y [0.53827], br -> br (not: gh-rbr-g-b) [0.50831], d -> t [0.50079], dr -> dr (not: tr-d-rdr) [0.45406], ieue -> eue [0.44367]
Haarlem: t -> tw [0.62681], u -> i [0.60748], o -> oi [0.47348], br -> br (not: gh-rbr-g-b) [0.45393], td -> tt [0.44456], ieue -> ieue (not: eve-ive-yve-ve-yeue-ie-eue-ue-iue-ieve) [0.44029], q -> q (not: qw-c) [0.42996], ua -> ua (not: uae-va-eua-uee-oe-a-u-ue-uo) [0.42258], final deletion: f [0.40494]
Halen: lm -> lm (not: rm) [0.66809], ue -> ve [0.62866], oe -> ij [0.621], gn -> gn (not: ngn) [0.60233], lt -> lt (not: ld-ldt-dt-l-t) [0.58322], lr -> lr (not: ll-ld-l-llr) [0.55194], voi -> vo [0.54033], mpt -> mpt (not: mdt-t-mpc-pc-pt-mt-md) [0.523], mpd -> mpd (not: md) [0.52]
Hasselt: e -> ey [0.60742], rfl -> rfl (not: rffl-rf) [0.58088], tn -> n [0.51462], voi -> vo [0.48069], ps -> ps (not: p-pts) [0.47405], aue -> eve [0.43994], ue -> ve [0.43662], oe -> ue [0.40302], i -> ii [0.39402]
Helmond: n -> m [0.69836], gh -> b [0.52284], e -> i [0.44186], rs -> r [0.42062], ue -> oe [0.38619], jae -> jae (not: jo-ioe-ja-ia-jai-iai-iae-joe-joi-yae) [0.37696], ja -> jae [0.35653], vy -> vy (not: ue-uu-u-v-vij-ui-uy-vuy-vi-vj-ve-vu) [0.35452], i -> j [0.35181]
Heusden: nh -> nh (not: h-nsh) [0.76735], fd -> fd (not: d-ft) [0.59946], ct -> ct (not: t-tt-cht-ckt) [0.59114], dd -> dd (not: bb-ddr-dt-tt-d) [0.53317], rs -> r [0.51814], ij -> ii [0.51699], ft -> fd [0.50781], lss -> ls [0.49688], lt -> lt (not: ld-ldt-dt-l-t) [0.45739]
Hoorn: nc -> nc (not: t-p-n-g-c-ns-nx-nt-nck-nch-ng-ngh-ch-ck) [0.58755], uij -> uii [0.58467], ls -> l [0.54815], ll -> ll (not: llll-lls-lll-lr-l) [0.52056], ai -> ae [0.47457], ee -> ae [0.46], rscr -> rsc [0.45381], q -> q (not: qw-c) [0.4503], u -> ou [0.44814]
Kampen: ey -> ie [0.80822], rfg -> rfg (not: rf-rffg-rfgh-rg) [0.79615], fs -> fs (not: f-s-ffs) [0.68555], rtr -> rtr (not: tr-rcr-rt) [0.65759], ve -> vo [0.55282], ct -> ct (not: t-tt-cht-ckt) [0.54379], ls -> l [0.52538], ae -> ae (not: ie-uae-ue-aee-u-o-a-e-ey-ay-aa-ai-ee-ei-vai-vae-oy-oe-ye-oi) [0.49213], rk -> rc [0.46562]
Leiden: u -> uy [0.36558], br -> br (not: gh-rbr-g-b) [0.30895], voe -> voi [0.30443], jae -> jai [0.29664], vo -> voi [0.29643], ua -> ua (not: uae-va-eua-uee-oe-a-u-ue-uo) [0.29065], ieue -> ieue (not: eve-ive-yve-ve-yeue-ie-eue-ue-iue-ieve) [0.2904], q -> q (not: qw-c) [0.28772], rc -> rk [0.28656]
Lummen: ue -> ve [0.82271], f -> cht [0.81476], mts -> ms [0.73138], ie -> i [0.6888], sl -> sl (not: th-t-scl) [0.66971], gn -> gn (not: ngn) [0.64743], rfl -> rfl (not: rffl-rf) [0.6369], r -> rs [0.63305], x -> cs [0.61549]
Maaseik: rp -> rp (not: p) [0.87076], ee -> ey [0.79978], ie -> i [0.58699], u -> oe [0.57642], ie -> ye [0.56798], ue -> ve [0.55814], i -> ey [0.54771], n -> m [0.54685], n -> nt [0.54542]
Maastricht: i -> y [0.69382], oe -> ue [0.53129], ae -> oe [0.46754], ii -> i [0.43761], sc -> sch [0.43726], ie -> ye [0.43269], cht -> chts [0.42577], rg -> rg (not: rgh-rn-r-g-gh) [0.4188], ae -> ae (not: ie-uae-ue-aee-u-o-a-e-ey-ay-aa-ai-ee-ei-vai-vae-oy-oe-ye-oi) [0.41429]
Middelburg: d -> nd [0.64414], oe -> oue [0.59829], ft -> cht [0.5727], rg -> rn [0.55752], rc -> rk [0.49157], fd -> d [0.47043], o -> v [0.46895], nc -> nc (not: t-p-n-g-c-ns-nx-nt-nck-nch-ng-ngh-ch-ck) [0.41793], str -> str (not: st-stt) [0.40749]
Schelle: o -> ou [0.76274], f -> cht [0.72332], ln -> ll [0.708], pp -> ppr [0.70757], sl -> sl (not: th-t-scl) [0.70657], ll -> l [0.66476], h -> t [0.64456], u -> e [0.63755], ey -> ei [0.62874]
Sint-Truiden: ie -> i [0.58565], chs -> chs (not: chts-cts-gs-gh-ghs-ch-cs) [0.545], n -> nt [0.52869], nr -> nr (not: ns-nd-nn-n-r-ndr-rr) [0.5256], ei -> ey [0.50495], sc -> sch [0.48495], x -> x (not: ks-ngs-csch-chs-c-k-ch-cx-cs-cks-xs) [0.44241], m -> ch [0.43167], ie -> ij [0.41651]
Utrecht: u -> e [0.56113], rk -> rc [0.52311], ii -> ie [0.4882], ij -> ie [0.45173], iue -> iue (not: yue-ive-yve-ij-ieue-ieve-ie-ijue-ijve) [0.42874], u -> o [0.42335], ve -> voe [0.42288], md -> d [0.42167], d -> t [0.42148]
Venlo: rgh -> rg [0.95285], fg -> fg (not: fgh) [0.83339], oi -> oe [0.82478], rft -> rft (not: rfft-rfdt-rfth-rfc) [0.78462], rg -> rg (not: rgh-rn-r-g-gh) [0.78374], ngh -> ng [0.76303], oue -> oe [0.73215], ll -> l [0.69232], rtm -> rtm (not: rdm) [0.68763]

Vught: ms -> ms (not: mps-ps-msch) [0.99999], voi -> voe [0.52847], e -> u [0.50248], ln -> ln (not: ll-lr-l) [0.48205], rf -> rf (not: r-rff-rfs) [0.46435], rh -> rh (not: rhw-rr-h-r) [0.45871], ei -> ey [0.40149], i -> ee [0.37825], rt -> r [0.3643]
Walem: rf -> rf (not: r-rff-rfs) [0.70194], nl -> nl (not: k-l-ll-ntl) [0.6897], scr -> scr (not: sc-cr-c-schr) [0.65802], td -> tt [0.62548], ft -> cht [0.57294], u -> e [0.49832], ei -> ei (not: eij-ii-ij-ayo-y-i-a-e-ee-ey-ie-ye) [0.49304], rh -> rh (not: rhw-rr-h-r) [0.4922], iue -> ie [0.48519]
Wijk bij Duurstede: fs -> fs (not: f-s-ffs) [0.71931], final insertion: oe [0.68039], rscr -> rsc [0.64938], rs -> rsc [0.63338], c -> k [0.62938], final insertion: ns [0.60163], nb -> mb [0.58079], n -> nr [0.57764], cht -> chts [0.50596]
Zutphen: ve -> vo [0.68486], o -> v [0.49887], ngh -> ng [0.46187], d -> t [0.45619], v -> o [0.4421], lt -> lt (not: ld-ldt-dt-l-t) [0.42743], voe -> vo [0.39314], voi -> vo [0.39227], eue -> oue [0.36324]
Zwolle: oi -> oe [0.90541], rfl -> rfl (not: rffl-rf) [0.65281], fs -> fs (not: f-s-ffs) [0.57555], vy -> vy (not: ue-uu-u-v-vij-ui-uy-vuy-vi-vj-ve-vu) [0.56062], r -> rs [0.55339], o -> u [0.48098], u -> ie [0.4577], oue -> oe [0.45646], lft -> lft (not: lt-lfd) [0.42323]