Chapter 4
Learning Name Variants from Inexact High-Confidence Matches

Gerrit Bloothooft and Marijn Schraagen

Abstract Name variants which differ more than a few characters can seriously hamper record linkage. A method is described by which variants of first names and surnames can be learned automatically from records that contain more information than needed for a true link decision. Post-processing and limited manual intervention (active learning) is unavoidable, however, to differentiate errors in the original and the digitised data from variants. The method is demonstrated on the basis of an analysis of 14.8 million records from the Dutch vital registration.

4.1 Introduction

In record linkage, the decision to make a link between two instances of information can be complicated by spelling variation, variants (translation, suffix variation, changes in order of name elements, etc.) and errors. A usual approach to cope with this kind of variation is to define a distance or similarity metric (at written or phonemic level) to describe the spelling difference between two names. If the difference between the names is less than some threshold, they are considered to be variants and may indicate the same person (Christen 2012). This approach has the limitation that (1) the same threshold is used for all names, while a name-dependent threshold may be more effective (although distance measures may incorporate this to some extent), and (2) the threshold is chosen arbitrarily or is at best decided upon by its overall effect: the linkage process should not produce too much overlinking,

G. Bloothooft (✉)
Utrecht Institute of Linguistics-OTS, Utrecht University, Trans 10, 3512 JK, Utrecht
The Netherlands
e-mail: g.bloothooft@uu.nl

M. Schraagen
Leiden Institute of Advanced Computer Science, Leiden University, Leiden
The Netherlands
e-mail: m.p.schraagen@liacs.leidenuniv.nl

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_4

i.e. too many false links. Although small variation in names can be identified in this way, larger variation, such as between *Jan* and *Johannes*, is usually beyond a threshold. On the other hand, small differences in names, like the surnames *Bos* and *Vos*, do not always imply a genuine variant. These observations indicate the need for a corpus which explicitly describes name variants that could have been used for the same person. Experts could help in the laborious task to construct such a corpus, but it would be efficient if these variants could be learned at least in part automatically from data.

There are circumstances where sources are rich enough to allow for record linkage while not using all available information. Names that are not needed in the matching process, may contain true variants (but errors as well). This chapter investigates the procedures needed to construct a corpus of true name variants in a largely automated way, applied to 14.8 million records from the nineteenth century vital registration of the Netherlands.

To derive name variant pairs, record links based on several elements or fields (e.g., the names of various people mentioned in a record) are examined. In case one of the fields differs between the records while the other elements are exactly equal, the differing field values are assumed to contain a name variant. After variant construction, post-processing using rules and heuristics takes place to remove erroneous variant pairs.

The chapter is structured as follows: Sect. 4.2 describes related work, Sect. 4.3 describes the source data, Sect. 4.4 describes the method to collect name variants, while Sect. 4.5 discusses the options to differentiate true variants from errors. In Sect. 4.6 results are presented and evaluated. The possibility to use name variants for clustering and name standardisation is explored in Sect. 4.7, including an extra iteration of the main name pair construction method. In Sect. 4.8 a comparison is made with a name variant corpus of FamilySearch, and Sect. 4.9 concludes.

4.2 Related Work

The basic name variant derivation procedure can be compared to corpus-based stemming of regular text using co-occurrence of terms in a document. An example is discussed in Xu and Croft (1998), where the basic assumption is that word variants that should be conflated will co-occur in a (typically small) text window. The approach of Xu and Croft is intended to address issues in rule-based and dictionary-based stemming. A text window in a document can be compared to a pair of linked records, in the sense that in both cases sufficient information is present to conflate variants. However, the construction of record links is non-trivial, which complicates the current approach. On the other hand, the structure of a record is given by the division into fields, in contrast to the structure of a natural language sentence. This reduces the need for complex co-occurrence statistics in the current approach.

The current method of extracting name variants from record matches is essentially a network approach, in which ambiguous nodes can be combined if the sets of connected nodes (in this case other names in a record) are similar (Malin 2005; Getoor 2007). However, the current dataset is represented as a simple network in which records are small, equally sized unconnected subgraphs, therefore reducing the need for elaborate graph traversal algorithms.

Name variant construction from data has been discussed by Driscoll (2013), who uses text patterns to search for name-nickname variants on web pages, combined with various morphological rules and matching conditions, partly automatically induced from the initial variant pairs. The results are promising, especially when all methods are combined using an automatically derived weight for each method. A key component of the current approach is however not used by Driscoll, which is the selection of a large amount of candidate pairs from record links. The overall quality of the candidates in this selection allows the rules applied in the current method to be less strict, which improves coverage without a large increase in error rate.

4.3 Material

The data used in the investigation is extracted from the Dutch *WieWasWie* (who was who) database (www.wiewaswie.nl, release November 2011 as Genlias). *WieWasWie* contains civil certificates from the Netherlands, for the events of birth, marriage and death, of which the registration started in 1811. Most documents originate from the nineteenth and early twentieth centuries. A record consists of the type of event, a serial number, a date and a place and information about the participants. The parents are listed for the main subject(s) of the document, i.e. the newborn child at birth, the bride and groom at marriage, and the deceased person at death, respectively. The documents do not contain identifiers for individuals and no links are provided between documents. The digitisation of the certificates is an ongoing process that is performed by volunteers. For the 2011 release it is estimated that key information from 30 % (4.1 million) of the birth certificates, 90 % (3.1 million) of the marriage certificates, and 65 % (7.6 million) of the death certificates has been made available. This concerns about 55 million references to individuals. These references include 101,830 different male first names, 128,800 different female first names, and 565,647 different surnames (all singular elements, if necessary derived from composite names).

4.4 Method and Variant Pair Construction

The three different types of certificates (birth, marriage and death) all contain information about individuals and their parents. This information can be used for matching, as illustrated in Fig. 4.1. The method described in this chapter uses the

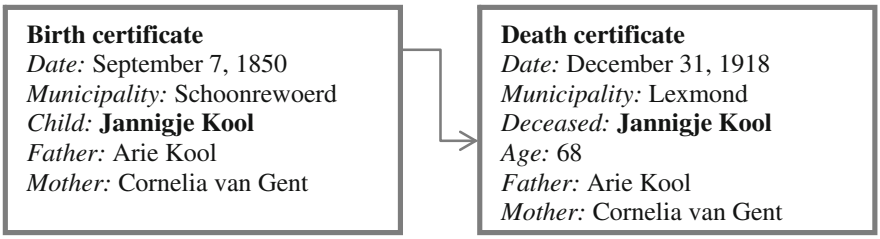

Fig. 4.1 Example match between birth and death certificates with *Jannigje Kool* as main person

assumption that true record matches can be found using a subset of the available information, which enables construction of variant pairs based on the remaining information.

In the application of the method we required exact matching of the first name of the main person and equal year of birth (derived from age in marriage and death certificates, plus or minus one year). Moreover, three out of the four names of the mother and father should match exactly as well. Note that in the Netherlands women always keep their maiden name in the administration. The fourth name of the mother or the father was not part of the linkage decision and open to variation and, if differing between the two records, generated a name variant pair. This could concern a male (father) or female (mother) first name, or a surname (father or mother).

We tested whether the requirement of matching four out of five names plus year of birth was sufficient for accurate record linkage by selecting matches between birth and death certificates for which *all* five names and year of birth were available and matched, under the assumption that an exact match on all available names would generate only true matches in our dataset. Subsequently, one of the four names of the parents was ignored and it was counted in how many cases more than one link was generated. This was the case for only 85 out of 1,107,162 matches. We considered this as sufficient support for our assumption that three out of four equal names of parents were a sufficient condition for accurate linkage, as violation would generate only a few errors.

An example of a rare match where the condition did not hold is: 7 September 1850 birth of *Jannigje Kool* in Schoonrewoerd, from the parents *Arie Kool* and *Cornelia van Gent*, and her decease in 31 December 1918 in Lexmond, at 68 years of age with mention of the same parents (see Fig. 4.1). Competing is the birth of *Jannigje Oosthoek* on 25 April 1850 in Charlois, from the parents *Arie Oosthoek* and *Cornelia van Gent* (see Fig. 4.2). Although there are matches of the names *Jannigje*, *Arie*, *Cornelia*, and *van Gent*, and the year of birth 1850, this leads to the erroneous surname variant pair *Kool/Oosthoek*. The—disentangling—place of birth was not used in the matching decision, as this information is error prone, especially when mentioned in the death certificates.

A name can consist of several elements, such as the first name *Johan Willem Frederik*. Although we required identity of four out of five (full) names of a person

Birth certificate
Date: April 25, 1850
Municipality: Charlois
Child: **Jannigje Oosthoek**
Father: Arie Oosthoek
Mother: Cornelia van Gent

Fig. 4.2 Competing birth certificate for the birth certificate in Fig. 4.1

Table 4.1 Examples of variant pairs constructed from composite names

Name 1	Name 2	Differences	Name pair
Anna Christina Elizabeth	*Christiena Elizabeth*	Missing name, name variant	*Christina/Christiena*
Virgin Thomasa Franken	*Thomasa Virginia*	Reversed order, missing name, name variant	*Virgin/Virginia*
Adriana Agnita Cornelia	*Adriana Cornelia*	Missing name	None

and parents, in case of composite names, more single names were involved and thus provide stronger support for a true link. A considerable 50 % of the name pairs were accompanied by five or more identical single names in the comparison, instead of the four identical names minimally required.

The selection on differences in the fourth name of a parent over all combinations of birth, marriage and death records resulted in 897,426 name pairs. For composite names the difference could be caused by different name order, missing names and/or actual name variation. For these cases variants are identified using alignment of name elements based on minimal edit distance, see Table 4.1 for examples.

After this compositional analysis, which was also performed for surnames, pairs of single names remained. Since the order of names in a pair is unimportant, name pairs with opposite order were taken together. The results of this step are shown in Table 4.2. The most frequent name variant pairs, which have only minor spelling differences that mostly do not influence pronunciation, are also shown in Table 4.2.

Table 4.2 Variant construction results and examples

Female first name		Male first name		Surname	
Pairs	Tokens	Pairs	Tokens	Pairs	Tokens
48,684	246,519	31,885	183,050	177,258	374,901
Most frequent variant pairs					
Elisabeth/Elizabeth		*Johannes/Johannis*		*Jansen/Janssen*	
Willemina/Wilhelmina		*Jacob/Jakob*		*Bruin/Bruijn*	
Geertrui/Geertruij		*Arij/Arie*		*Ruijter/Ruyter*	

4.5 Variants and Errors

In this research we wish to construct a clean corpus of name variant pairs, but name errors complicate the process, also when the record links themselves are correct. Name errors can not only originate from the writing of the original certificates, but also from misreading or typing errors in the recent digitisation process, or result from violation of the assumption that four out of five equal names and equal year of birth describe a person uniquely (rare, but shown before). Where true name variants can replace each other in any condition and thus help record linkage under less favourable conditions, name errors should be recognised as such and not be propagated.

As an example of a registration error we consider *Pieter*, born in 1808 as son of *Jacob Houtlosser* and *Aafje Spruit*, as mentioned in the marriage certificate (see Fig. 4.3). But his death certificate mentions *Grietje Spruit* as mother, resulting in the erroneous first name variant pair *Aafje/Grietje*. Additional evidence that the records concern the same person comes from the partner name *Aaltje Kort*, mentioned in both certificates, and the correspondence in municipality (although place and partner information is not used in the matching process).

A distinction between a true variant (*Dirk/Derk*), and an error (*Dirk/Klaas*) is not at all easy to make. We chose for a definition of true variants as names that belong to the same lemma, while errors do not. A lemma [see, e.g., Bratley and Lusignan (1976)] is a usually etymologically based name from which by processes of pronunciation, suffixation, abbreviation, etc., derivate forms can be generated. These processes are very difficult to model or to predict and therefore it is hard, if not impossible to differentiate automatically between a true variant and an error. In many cases onomastic or linguistic expertise is required.

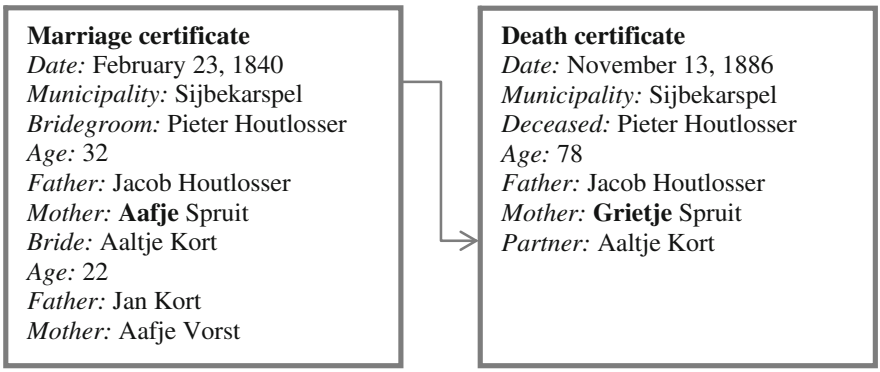

Fig. 4.3 Record match resulting in the erroneous variant pair *Aafje/Grietje*

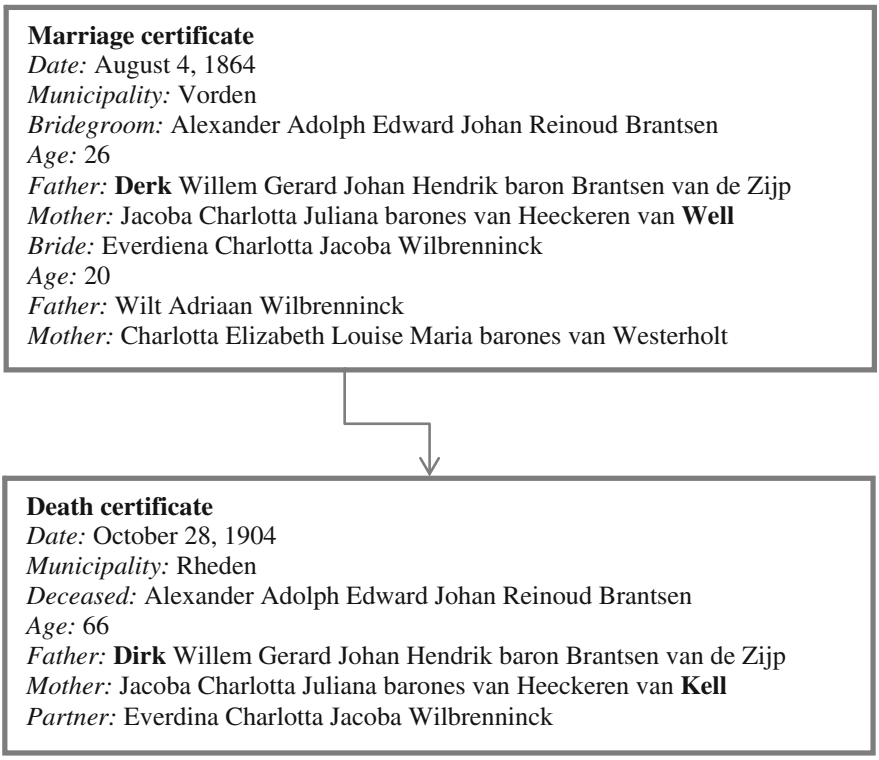

Fig. 4.4 Record match resulting in the correct variant pair *Derk/Dirk* and the erroneous variant pair *Well/Kell*. Note that the variant pair *Everdiena/Everdina* is not considered in the procedure

There can be substantial evidence that the same persons are concerned (for instance in case of long composite names), but this does not exclude errors. An example is shown in Fig. 4.4 where our onomastic knowledge tells us that *Dirk/Derk* is a genuine first name variant, while *Kell/Well* relates to miswriting, misreading or mistyping and should not be generalised beyond this single occurrence. Unfortunately, this differentiation between a true and erroneous name variant pair cannot be made automatically.

Also the frequency of a variant pair (or its probability) is of limited help. Both errors and variants can be rare or frequent. Frequent erroneous variants for male first names are for instance combinations of popular names, see Table 4.3. The use of rules and manual inspection (active learning) is unavoidable to make a distinction between variants and errors.

Table 4.3 Erroneous name pairs consisting of popular names

Name 1	Frequency (in millions)	Name 2	Frequency (in millions)	Interchange frequency
Jacob	0.50	*Jan*	2.39	331
Hendrik	1.11	*Jan*	2.39	212
Jacobus	0.39	*Johannes*	1.37	149
Willem	0.88	*Jan*	2.39	138
Gerrit	0.73	*Hendrik*	1.11	104
Gerrit	0.73	*Jan*	2.39	99
Willem	0.88	*Hendrik*	1.11	82
Gerrit	0.73	*Cornelis*	0.86	63
Klaas	0.33	*Jan*	2.39	60
Dirk	0.35	*Jan*	2.39	59

The frequency of the individual names in the WieWasWie 2011 corpus, as well as the frequency that name 1 is erroneously replaced by name 2 (or reversely) is given

4.5.1 Name Pair Cleaning

The name pairs resulting from the automatic analysis are post-processed in order to remove erroneous pairs. Three different methods have been applied, of which the first method uses an external manually compiled name lexicon, the second method developes and uses a corpus of non-variants, and the third method is based on manually designed variant classification rules (see diagram in Fig. 4.5). The methods are described in detail in the remainder of this section. In case of acceptance by the first method or rejection by the second method application of the third method was not needed. Additional manual review of a limited selection of variant pairs was applied to correct post-processing errors. The selection of pairs for review has been performed in an Active Learning setting (see Olsson (2009) for an overview, Sarawagi and Bhamidipaty (2002) for an application in nominal record linkage), considering pairs based on the frequencies of both the name pair and the individual names in the pair. The frequency values act as a confidence score which allows the algorithm to automatically single out pairs for which manual review is useful without the need to manually evaluate every single pair.

4.5.1.1 Using Name Dictionaries

Variants share a lemma and errors do not. The decision that names share a lemma can be based on expert onomastic knowledge, as laid down in name dictionaries. If available, the content of the dictionaries is usually much more limited than the name variation found in current resources. For the Netherlands, a dictionary of first names (van der Schaar 1964, first edition) is available which associates about 20,000 first names to 3737 gender-independent lemmas. This could be helpful as a starting

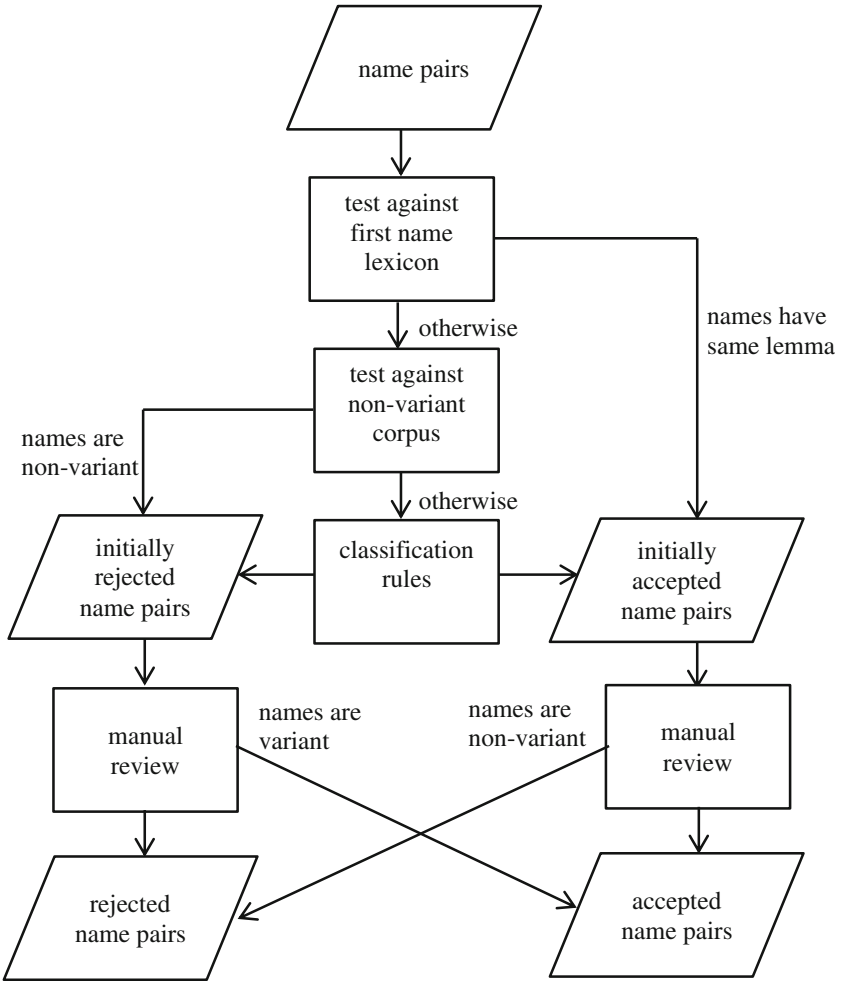

Fig. 4.5 Flow chart of post-processing of name pair variants

point for the identification of many more name variants, but there are a number of limitations. Many (abbreviated) names in the dictionary are associated to more than one lemma, especially short names. For instance, *Aai* with lemmas *Aai, Aalt, Adriaan*. Furthermore, lemmas can be too refined (given our observations of variation in practice), such as *Adagonda, Adelgonde* and *Aldegonde*. Sometimes association has subtle differentiation, such as *Nelie* with the lemma *Cornelis*, *Nelly* with the lemma *Cornelis* or *Petronius* (as *Petronella*), and *Nella* with lemma *Petronius*, which does not seem to conform to the use of names in practice either.

In our case, the dictionary has been used to accept first name variants that share a lemma, while making no decision on names that are associated to different lemmas.

A total of 3615 female and 2878 male first name pairs were accepted, which is about 5 % of all pairs found. The main gain of this approach is that we can accept name pair variants that differ so strongly that they would not make it through the rules we apply later. This avoids manual intervention for them. For surnames, a comparable dictionary is not electronically available for the Netherlands.

4.5.1.2 Data-Driven Identification of Erroneous First Name Pairs

A data-driven option to identify first name variant *errors* is based on the assumption that first name variants do not show up together in a composite name (Oosten 2008). This would imply that names that *do* show up in a composite name are not variants. From the first name *Anne Maria Helena* we may then conclude that *Anne*, *Maria* and *Helena* are no variants from each other.

This method was tested using all (possible composite) first names from the Dutch WieWasWie 2011 release. These names have 55 million tokens. From the composite first names in this set, all combinations of two names were determined, keeping the order of appearance from left to right in the composite name. This resulted in a non-variant-corpus of 907,660 pairs of first names, with 18 million tokens.

However, the dataset contains errors introduced in the digitisation phase. Patronymics [referring to the first name of the father, such as *Jansz*, short for *Jan's zoon*, son of *Jan,* cf. Anderson (2007)] and parts of the surname, or the whole surname, were sometimes included in the first name field. For example *Aagtje van Eck*, with *van Eck* as surname, is present as a first name, which results in the incorrect non-variant pairs (*Aagtje/van*), (*Aagtje/Eck*), (*van/Eck*). To exclude these errors, we required that name pairs should be seen in both orders, under the assumption that it is unlikely to find a patronymic or a surname before the first name.

Another problem was first name fields with descriptive content, such as *zoon van Geertruida* (*son of Geertruida,* 1 time) and *Aleida Geertruida van (Aleida Geertruida from,* 55 times), which resulted in the erroneous first name pair (*Geertruida/van*), seen in both orders. These name pairs were excluded by requiring a capital initial and a name length of at least three characters (which also excluded single initials). After this, a non-variant-corpus of 118,532 first name pairs resulted (only 13 % of the originally collected pairs), with 15 million tokens (83 %).

In conflict with the assumption of the approach, however, also true variant pairs show up jointly in a composite first name. Frequent examples were *Jan Johannes, Neeltje Cornelia, Arie Adrianus, Jannetje Johanna*, indicating that parents did not mind or even did not realise the common basis of both names. After removal of the pairs that have the same lemma in the dictionary of first names and a few manual corrections, the no-variant-corpus was held against the name variant set and resulted in the removal of 2458 female first name pairs and 2343 male first name pairs. The advantage of this approach is that we can exclude erroneous name pair variants that would pass the rules we apply in the subsequent step.

4.5.1.3 Rules That Accept Name Variant Pairs

If there were no errors in the source material, our method would not require additional cleaning methods. But since the source material is not error free, additional methods are needed, and the application of rules is one of them. Our rules can be much more relaxed, however, than rules that apply on any pair of names as they are used on a pre-selected corpus of name pairs.

Two sets of rules are applied: a first set of rules converts a name into a semi-phonetic form, while a second set of rules compares the differences between two names on the basis of Levenshtein distance and additional requirements that resemble the Jaro-Winkler distance measure (Winkler 1990). Both sets will be explained in this section.

In the past, the lack of spelling rules has promoted variation in spelling. Attempts can be made to apply rules on written forms that result in a version that has close correspondence to the original pronunciation. Since it is impossible to catch all spelling variation (especially under the presence of all kinds of errors), a limited but robust rule set was developed that converts names from Dutch sources into a semi-phonetic form (Bloothooft 1995). Semi-phonetic implies that although the coding is inspired by the conversion of written characters into speech sound symbols, no attempt has been made to arrive at a correct phonetic transcription, which is impossible under the presence of unpredictable writing or digitisation errors. This rule set resembles other approaches to phonetic encodings [Soundex (Russell 1918), Double Metaphone (Philips 2000), cf. also Dolby (1970)], but is tailored towards properties of the relation between spelling and pronunciation for Dutch.

Major rules are (1) symbol simplification by ignoring diacritics, (2) reducing all character replications to a single symbol, (3) reducing all vowel combinations to single symbols, (4) rules for resolving the ph, gu, ch and ck combinations, and (5) rules for the letters c, d, h, j, q, v, x, z. Examples are *Jannigje* > JANYGJE, *Cornelia* > KORNELYA, *Jozeph* > JOSEF. In further processing, this semi-phonetic form of a name was used.

A second set of heuristic rules was adopted that limits the acceptable differences between two names. A variant pair that complies with a rule was accepted. Major ingredients were the Levenshtein distance between the names, the name lengths, and number of identical (semi-phonetic) initials (at least one). These rules have some relationship to the Jaro-Winkler distance measure, but are more relaxed.

There is a considerable group of name pairs that result from (understandable) misreading of the initial. Frequent misreadings are found between the initials *T* and *F, P, J, S* or *K; F* and *P* or *J; I* and *J;* and *M* and *H, W,* or *Al.* The difficulty of misreading (at the digitisation phase) is that there is often a bias towards an (erroneous) existing name on the basis of the knowledge of the person who digitises (for instance, the first name pairs *Pietje/Tietje, Jannetje/Tannetje, Wessel/Hessel,* and the surname pairs *Tol/Pol, Meijden/Heijden, Noort/Voort*). If this misreading

Table 4.4 Heuristic rules containing thresholds for variant pair acceptance

Levenshtein distance	Length	Minimum length of shared prefix	Example
1	Shortest >4	1	*Joanna, Johanna*
2	Shortest >4	2	*Gerrit, Geurt*
3	Longest >5	3	*Annegien, Annigje*
4	Longest >7	4	*Laurentius, Laurijs*
5	Longest >8	4	*Franciscus, Frans*
Total length of pair minus Levenshtein distance >16		1	*Lingmandus, Luigmondus*

In addition to the six rules specified in this table, a more complex seventh rule on suffixes is explained in the text. Rules are applied both to the original and semi-phonetic form of a name

happens systematically, the resulting name confusion needs not even be rare. Automatic detection of them is difficult because the Levenshtein distance is small (only 1 because of the initial). Therefore we required by rule the same initial in the name pair, and more equal initial characters for more relaxed conditions of the Levenshtein distance between the names (at the semi-phonetic level, which already takes care of the major genuine spelling variation of the initials).

Rules are summarised in Table 4.4. These rules were applied to both the original and the semi-phonetic name form. If a variant pair passed a rule in either the original or semi-phonetic form, the pair was accepted. There was a final rule—applied to the semi-phonetic name form only—which required two identical initial characters, while the name ends in (any part of) the semi-phonetic suffixes TSJEN, TJEN, TYN, KJEN, KEN, KYN, YA, PJEN, PY or was empty. For instance: *Eva*/*Eefje* > EFA/EFJE > EF + A/EF + JE is accepted as variant pair.

From Table 4.4 it can be seen that variant pairs with a Levenshtein distance well over 2 can be accepted by a rule, which also holds for the additional suffix rule discussed above. A general threshold of 2 or 3 is common, the gain of the current method is in the conditional acceptance of a wider range of edit-distances.

It is impossible to fully automate the decision on the status of name variant pairs by rules. For instance, the genuine name pair *Willem/Guillaume* differs as a Dutch–French translation too much in spelling. Manual decisions, on the basis of expert knowledge, are unavoidable but should be kept to a minimum. An additional manual review was critical, and concentrated on true variants of low frequency and rejected variant pairs with a high frequency. If there was any doubt on the status of a variant pair, the name pair was not accepted. A manual decision could imply a rejection of a name pair that was accepted by rule, of acceptance of a name pair that did not pass the rules (for instance because the initials were not equal).

4.6 Results and Evaluation

A summary of the results of all phases in the cleaning process is presented in Table 4.5 for first names and in Table 4.6 for family names. For the accepted name variant pairs, the percentage with a certain Levenshtein distance is given in the tables as well, both for the original and the semi-phonetic form of the names. A Levenshtein distance equal or larger than 3 (usually too large to be accepted in straightforward record linkage as this generates abundant overlinking), is found—in original form—for 15.7 % of the female first names, 11.4 % for the male first names, and 7.0 % for the surnames (10.9, 8.3, 3.9 % for the semi-phonetic form, respectively). In terms of tokens the percentages are somewhat lower. This may be considered the gain of the method. As expected, the Levenshtein distance in the semi-phonetic form is lower for than in the orthographic form, but mainly for distances up to 2. Larger name pair differences originate in suffix variation or translation rather than in spelling differences for the same pronunciation.

Table 4.5 Overview of cleaning results for first name variant pairs

	Female first names		Male first names	
	Name pairs	Tokens	Name pairs	Tokens
Initial name pairs	48,684	246,519	31,886	183,050
Accepted by dictionary	3610	94,551	2877	90,761
Rejected as non-variant	2412	12,041	2289	6538
Rejected by rules	11,336	18,716	7077	10,079
Rejected manually	118	414	42	126
Accepted manually	1001	3917	563	2458
Total accepted	34,818	215,438	22,478	166,307
Total rejected	13,866	31,081	9408	16,743
Levenshtein distance (*original*)				
1	58 %	69 %	65 %	70 %
2	26 %	20 %	24 %	18 %
3	9 %	5 %	7 %	5 %
>3	7 %	6 %	4 %	7 %
Levenshtein distance (*semi-phonetic*)				
0	19 %	29 %	22 %	29 %
1	52 %	45 %	53 %	46 %
2	18 %	15 %	17 %	13 %
3	7 %	7 %	5 %	7 %
>3	4 %	4 %	3 %	5 %

The exclusion and acceptance mechanisms as described in Sect. 4.5.1 are detailed. For all accepted name variant pairs the percentage with a certain Levenshtein distance is given, both for original and semi-phonetic name forms

Table 4.6 Overview of the results of the various steps in cleaning the initial corpus of name pair variants for family names. Details as in Table 4.5

	Family names	
	Name pairs	Tokens
Initial name pairs	177,258	374,901
Accepted by dictionary		
Rejected as non-variant	103	199
Rejected by rules	56,694	79,079
Rejected manually	346	507
Accepted manually	783	2410
Total accepted	120,115	295,116
Total rejected	57,143	79,785
Levenshtein distance (*original*)		
1	69 %	77 %
2	24 %	19 %
3	5 %	3 %
>3	2 %	1 %
Levenshtein distance (*semi-phonetic*)		
0	29 %	44 %
1	53 %	45 %
2	14 %	8 %
3	4 %	2 %
>3	0.2 %	0.2 %

As mentioned in the previous section the heuristics used in the classification process resemble the well-known Jaro-Winkler similarity, as both methods compute similarity based on shared prefixes and number of edit operations relative to the length of the string. To compare both methods, the Jaro-Winkler similarity (which is expressed as a similarity value between 0 and 1) is computed for all candidate variant pairs of first names and surnames together that have been selected by the basic method outlined in Sect. 4.4. In Fig. 4.6 the amount of pairs is presented for different similarity values, using separate curves for pairs accepted or rejected by the joint post-processing methods. Both the similarity in the original names and the similarity in the semi-phonetic forms are shown in the graph.

Figure 4.6 shows that the two methods are indeed correlated: accepted pairs generally receive a higher Jaro-Winkler score than rejected pairs. The score at the intersection of the curves of accepted and rejected name pairs is around 0.85 and could be taken as a threshold. This value is consistent with those used in the literature (see e.g. de Vries et al. 2009). The area under the curve for rejected pairs >0.85 (20 % false acceptances) and <0.85 under the curve for accepted pairs (13 % false rejects) is the gain of the current post-processing methods over the application of the Jaro-Winkler similarity. The curves in Fig. 4.6 do not differ much for names in original and semi-phonetic form. This implies that the Jaro-Winkler similarity does not improve by application on the semi-phonetic name form.

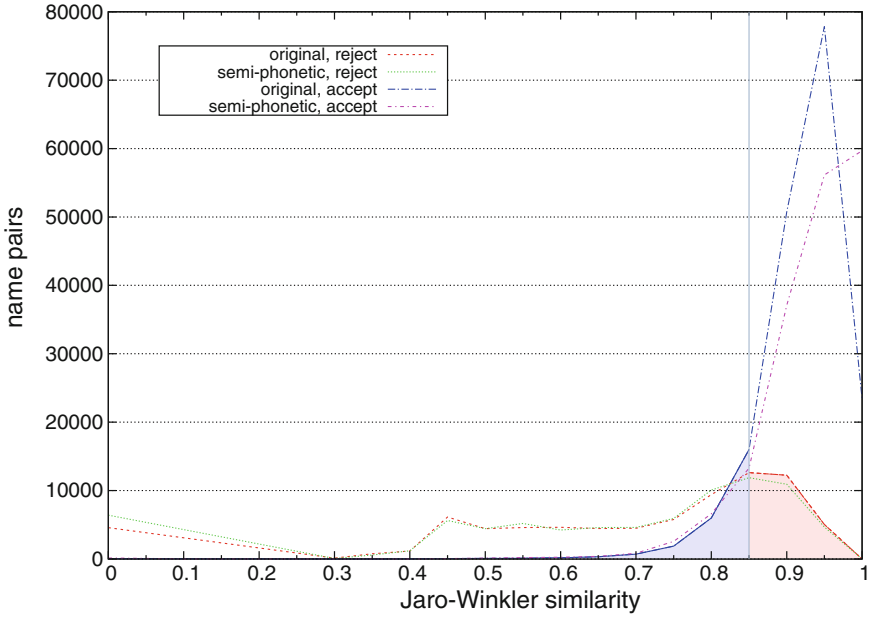

Fig. 4.6 Jaro-Winkler similarity for candidate variant pairs. Values are binned with interval 0.05 for readability. The vertical line at 0.85 is the optimal decision threshold for acceptance as similar names

In addition, it is of interest to consider the name pairs that are not accepted although they have a Levenshtein distance ≤2. These figures are not presented in Table 4.5, but amount to 39 % of the 13,866 erroneous female first name pairs, 49 % of the 9408 erroneous male first name pairs, and 43 % of the 57,143 erroneous surname pairs in original form, and 48, 42 and 48 % in their semi-phonetic form, respectively. If encountered in record linking, and considered by Levenshtein distance only, these names will be incorrectly accepted. This demonstrates the need for explicit knowledge of name variants.

A comparison of the name pair types and tokens shows that the rules mainly exclude rare name pairs as there are about only 1.5 times more tokens than types. On the other hand, first name pairs that were accepted because the names share the same lemma in a dictionary are frequent with on average about 30 times more tokens per pair. The latter variants are obviously well-known and made it to the dictionary.

Although we collected more surname variant pairs (120,115) than first name variant pairs (57,296 in all), the tokens of first name variant pairs were more frequent. On average, a first name variant pair was observed 6.7 times, while this was 2.4 times for surnames. This shows that there is much more variation in first names than in surnames.

The analysis of name pairs does not show how many different names are involved. This is shown in Table 4.7, together with the figures found in the full release (WieWasWie 2011). The number of singletons (i.e., name types with frequency 1) in both collections is presented as well, as they constitute 50 % of all first

Table 4.7 Number of different names (and singletons among them) in the accepted name pairs, and in the full WieWasWie corpus (release 2011)

	In accepted name pairs		WieWasWie 2011	
	All	Singletons	All	Singletons
Female first names	28,574	5766	128,800	63,132
Male first names	20,234	4048	101,830	51,000
Family names	129,929	22,361	565,647	225,389

names and 40 % of all surnames in the full corpus while they are present in 20 % of the accepted first name pairs, and in 15 % of the accepted surname pairs.

Given the constraints applied to arrive at accepted name pairs, the number of different names in the current analysis is relatively limited. Names might have been missed if they did not meet the required conditions, or if they were consistently written in the same way for any person and do not have a variant (the latter names will not present problems in record linkage). However, the selected names have a high coverage as will be shown in the next section. Coming to grips with the variation in these names can have a highly positive effect on record linkage.

4.7 Name Clusters and Standardisation

A corpus of true name variant pairs can be used to create clusters of names for which variant pairs are only found within a cluster. On this basis, yet unseen name variant pairs can be anticipated. We applied a non-standard clustering technique on the derived variant pairs which involved the following steps. Initially, for every name it was counted how many variants (types) were available. Based on the assumption that names with many variants likely constitute a cluster kernel, names were analysed in the order of the number of variants they had. The cluster procedure was applied on the semi-phonetic form of a name. For female first names, *Elisabeth* (including equal semi-phonetic forms) had most variants (148), for male first names this was *Hendrik* (69 variants), and for surnames *Tijssen* (50 variants).

As a first step, all variants of a name under investigation were added to the cluster. As a second step, for each of the names in the cluster it was analysed whether they had variants themselves that were not already in the cluster. If such a new variant name shared more than 60 % of the own variant *tokens* with names already in the cluster, the new name was added to the cluster. The value of 60 % was arbitrarily chosen on the basis of pilot analyses. A higher and more restrictive percentage would result in too many missed true variants in the cluster, while a lower and more permissive percentage would result in incorrect inclusion of variants. This process was continued until no new names could be added to the cluster. Subsequently, the next potential kernel name (with the then highest number of different variants) was analysed. The procedure continued until no names could be clustered anymore.

In the final phase of the procedure, variant pairs remained of which both names had no relation to other names and could be considered as mini-clusters. For the current analysis we left them out of consideration. For male first names 1221 clusters resulted comprising 16,487 names, for female first names there were 1530 clusters with 23,813 names, and for surnames 11,696 clusters with 93,793 names resulted. The largest clusters per name type were found for *Maria* (669 names), *Hendrik* (392) and the patronymic surname *Klaassen* (168). These figures are higher than the maximum number of variants found for an individual name since not all possible combinations of names in a cluster were present in the corpus. The cluster for the female first name *Elisabeth* is shown in Fig. 4.7 as an example.

In ideal clusters, the two names of a variant pair should both be members of the same cluster. This was the case for 91 % of the male first name pairs, 89 % of the female first name pairs and 85 % of the surname pairs. The remaining pairs could be erroneous and overlooked during the manual inspections, or the names were

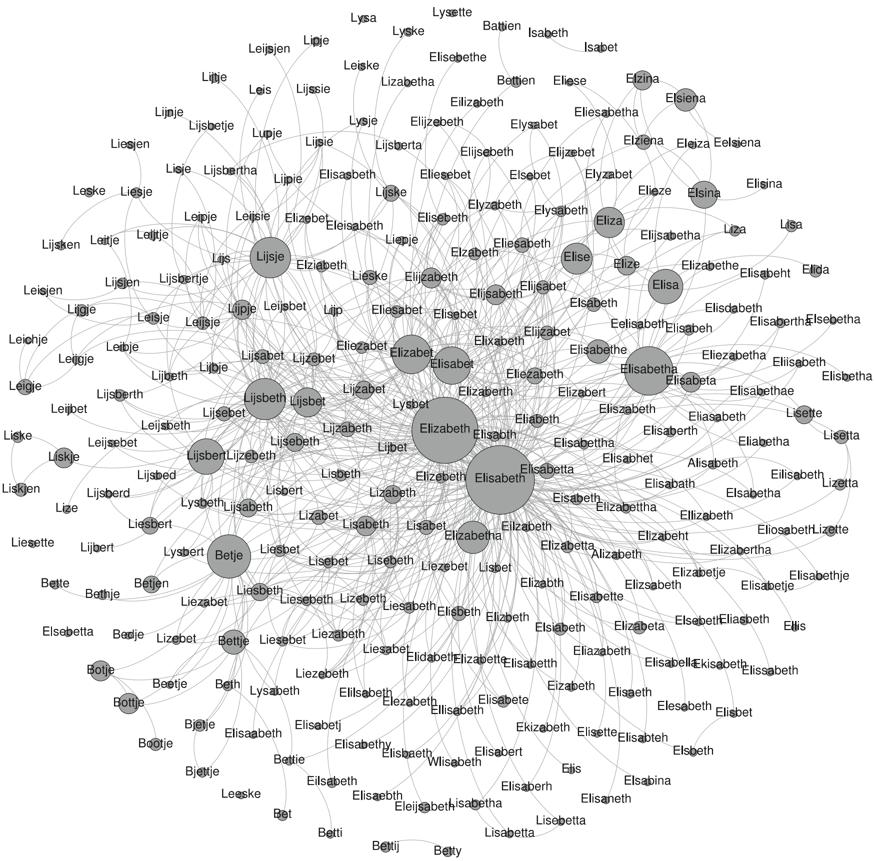

Fig. 4.7 Example cluster for *Elisabeth*. Edges denote variant pairs found in data. Node size is proportional to name frequency. Only names with a frequency ≥10 are shown

incorrectly associated in the cluster procedure. But there also can be genuine reasons that the names in a pair are not associated to the same cluster.

Names can reside in two closely related clusters although the current analysis did not provide sufficient evidence for such a merger, for instance the clusters *Egidius* and *Gilles*. Also, names can have more than one interpretation in terms of clusters, for instance *Louwis* as *Louis* or *Laurens*. The same holds for abbreviations, where the short form can be derived from several distinctive names. Productive are first names based on a suffix such as *Dina*, with possible derivations from *Alberdina, Berendina, Gerdina*, and so on. Whereas most names are associated to a single cluster, names in the latter categories are better associated to a group of clusters and analysed separately in record linkage procedures.

Once one has arrived at name clusters, it is a small step to name standardisation. The name in the cluster with the highest frequency was arbitrarily chosen as the standard name. This opens the possibility to perform a second round in discovering name variant pairs. Names in the original corpus were replaced by their standard if this could lead to four equal names of a person and parents, and a different fifth name. On this basis a total of 1,433,707 variant pairs (tokens) were collected, an increase of 60 % over the first round. A new analysis was performed on these pairs (while keeping all earlier manual decisions). This resulted in 26 % more accepted variant pairs (types) for both male and female first names, and 37 % more surname variant pairs.

After clustering of these variant pairs, 19,757 male first names were distributed in 1334 clusters, 28,509 female first names in 1694 clusters and 127.194 surnames in 15,114 clusters. The gain relative to the first round was 20 % for first names and 36 % for surnames, while the number of clusters increased less: 10 % for first names and 29 % for surnames. The ratio between the relative gain of newly created clusters and newly associated names is lower for first names than for surnames, indicating that for first names many new names were added as variant to existing clusters, while for surnames they created more new clusters.

If we use the clusters as a basis for the standardisation of names, about 20 % of all 101,830 male and 128,800 female first names get a standard, and 23 % of all 565,647 surnames. Whereas these percentages are rather low, the remaining names are rare and many are even singletons (hapaxes). In terms of tokens, the standardised first names describe 98.2 % of all 63 million element tokens and the standardised surnames 89.1 % of all 56.6 million element tokens. The number of standardised names can be extended by including non-standardised names with the same semi-phonetic form. This increases the number of standardised names by 63 %, but since these names are of low frequency the increase in the number of tokens is 0.2 % for first names and 1.5 % for surnames (to 98.4 and 87.4 % respectively).

4.8 Comparison to the FamilySearch Name Variant Corpus

The quality of the variant pairs derived above (referred to as the *test database*) can be estimated by a benchmark comparison with another, independent variant database. This section describes such a comparison with the name variant database of FamilySearch[1] which is the research department of the Church of Jesus Christ of Latter-day Saints (LDS, more commonly known as the Mormon Church). This database will be referred to as the *LDS Database*.

The LDS database is created as a by-product of genealogical research conducted by the LDS Church. Genealogies have been constructed from a large variety of sources, including census data, church records, court and inheritance records, land ownership records and migration records. The sources and resulting records have been reviewed by church clerks and linguists between the 1940s and mid-1980s in order to record name variation. This review has been mainly a manual process, based on general phonetic, syntactic and etymological guidelines with name variants resulting from genealogy research as a starting point. Source data originated mostly from North America, the British Isles (including Ireland) and continental Europe, but also some Central and South America and a small amount of sources from Asia. An estimated total of one billion name tokens has been used for the name variant database.

In order to be informative, the comparison setup needs to satisfy at least the following conditions:

1. The benchmark database and the test database have been constructed using the same definition of name variation.
2. The set of names contained in the test database is a subset of or equal to the set of names contained in the benchmark database.
3. The authority of the benchmark database is established.

If the first condition is not satisfied, a different classification due to a difference in definition is expected for an unknown number of name pairs. If the second condition is not satisfied then the accuracy of classification of name pairs containing names which are not present in the benchmark database cannot be established. If the third condition is not satisfied, differences in classification may be due to errors in the benchmark database instead of errors in the test database. In all three cases a comparison of name pair classifications is less informative. For the current comparison these three conditions will be discussed.

The main method described in this chapter has the aim to decide on variant pairs on the basis of true record matches without further requirements. However, the subsequent cleaning required a definition of a true variant pair:

[1]A web-based query interface is available on https://familysearch.org/stdfinder/NameStandardLookup.jsp

"A true *name variant pair* is a pair of names which can be traced to the same lemma". A similar notion of onomastic variation has been used in the LDS database, therefore, the first condition seems to be satisfied. However, the association of a name to a lemma may be uncertain or ambiguous and different interpretations may be used. If a name variant is based on a spelling error (reading error or typo), morphological or etymological information that could trace to the lemma may be lost. This may especially occur in short names, e.g., *Aatje*, which could be a variant of *Ada/Adriana* (morphological) but also could be a spelling error of *Aafje* or *Aaltje/Alida* which have different lemmas. Furthermore, the granularity at the lemma level can be different. In the LDS database the names *Gerrit* and *Geurt* (mentioned as variants in Table 4.4) are considered to belong to different name groups. Conversely the name group for *Sophie* in the LDS database contains *Fae, Feetje, Feye* which are remote or even unlikely variants that may be considered as belonging to different lemmas such as *Feie*. These issues indicate that the first condition of the benchmark procedure is not entirely satisfied.

In the LDS database many names from WieWasWie cannot be found (see Table 4.8) and vice versa, therefore the second condition is clearly violated.

Also, the authority of the benchmark is not fully clear. The procedures and guidelines used in the construction of the LDS database have not been documented in detail and manual decisions have influenced the database to a large extent. In the experience of the authors of the current chapter the overall quality of the LDS database is high, but a significant amount of classifications seems debatable or plain incorrect. The violation of all three conditions should be kept in mind in assessing the value of the benchmark validation. Obviously it would be preferable to use a benchmark database which does satisfy the conditions. Alternatives include *NameX* (commercial, namevariants.co.uk), *NamepediA* (community based, namepedia.org) or *JRC-Names* [named entities, see Steinberger et al. (2011)]. However, considering coverage, availability and technical accessibility, the LDS database is the most suitable for the current benchmark comparison.

Table 4.8 Results of a benchmark comparison with the name variant database of FamilySearch (LDS)

	Post-processing result	First names	%	Surnames	%
Total name pairs		80,570	100.0	177,258	100.0
Not in LDS		37,624	46.7	124,902	70.5
Present in LDS		42,946	53.3	52,356	29.5
Variant in LDS	Rejected	936	2.2	609	1.2
	Accepted	*18,675*	*43.5*	*12,414*	*23.7*
Non-variant in LDS	Rejected	*13,882*	*32.3*	*22,622*	*43.2*
	Accepted	9453	22.0	16,711	31.9

4.8.1 Comparison Results

The basic method described in Sect. 4.4 resulted in 257,828 name pairs, of which 80,417 pairs have been rejected by post-processing (see Tables 4.5 and 4.6). All pairs have been compared to the LDS database, of which the results are summarised in Table 4.8.

In Table 4.8 a name pair is considered not in LDS if either one or both names are missing from the LDS database. This category applies to around 47 % of the first name pairs and 71 % of the surname pairs. For these names the benchmark is unable to provide an indication of the accuracy of the basic algorithm or the post-processing procedure. This is partly caused by the presence of spelling errors in the test database and the coding of diacritic marks. However, also many valid names consisting of only basic characters are not present in the LDS database, mostly low-frequent names such as *Elijzebet* (frequency: 22), *Edcko* (frequency: 1 as part of a composite name) or *Ruighaven* (frequency: 12). More common names are missing from the LDS database as well, for example the surname *Paardekooper* (English: *horse merchant*, frequency: 1888) is not included while the variants *Paardekoper, Paardenkooper, Paerdekooper, Paerdekoper, Parrdekooper, Peerdekooper, Peerdekoper* are present. The omissions include several high-ranked names, such as *Pieters* as a surname (frequency: 38,005).

In Table 4.8 the relatively high values for variant-accepted and non-variant-rejected, as well as the low values for variants-rejected show a reasonable agreement between the benchmark and the test database (indicated by italic numbers in the table). However, the high values for non-variant-accepted show disagreement: according to the LDS database many of the names in these pairs belong to different name groups while the post-processing algorithms consider these names as valid variant pairs. This result could be interpreted as an indication that the current algorithms are too permissive. However, as noted above, the databases are subject to granularity differences which influence the classification. For the combination non-variant-accepted a high value indicates a more fine-grained clustering in the LDS database, which is consistent with limited manual browsing of the database.

In general it can be concluded that the amount of agreement is higher than the amount of disagreement, which means that both the LDS database and the approach of this chapter are capable of capturing a significant amount of person name variation.

4.9 Discussion

This research was based on a set of record matches with a very high confidence level. We focused on extraction of true name variants from these record matches to apply them later under less favourable conditions. Although the method automatically produced name variants, even with very large edit-distances, errors in the

source data implied the extraction of erroneous variant pairs as well. The need for detection of these erroneous name pairs compromises the method, especially because manual inspection proved unavoidable. Nevertheless, the automatic selection, although not error free, assists enormously to identify true name variants from real data.

From Tables 4.5 and 4.6 it can be seen that the level of name errors that are present in this corpus concerns about 30 % of both the first name and surname variant pairs (and 9.2 % of the female, 13.0 % of the male first name pair tokens, and even 21.3 % of the surname pair tokens). These error levels may be worrying, but the reassuring observation is that they were detected by post-processing and evaluation procedures. In sources that are less rich in information on individuals, these errors cannot be traced that easily. In such cases it may only be hoped that more complex decision strategies (other than pair-wise comparison of records) can be developed, to perform accurate matching and error detection.

Part of the errors we identified are likely reading/transcription errors of the type *Pietje/Tietje*, in which *P* and *T* are confused, or typing errors like *Bos/Vos* with *B* and *V* as neighbouring keys. The additional problem with these errors is that they can result in existing names. Because we focused on name variants that have an onomastic basis, these pairs were labelled as errors. However, if we could estimate the likelihood of these errors, this could be incorporated in a linkage decision model (rather than requiring excess of information to be able to circumvent these reading errors).

Name variants can be summarised by clustering and name standardisation. Such a standardisation could realise a large efficiency gain in nominal record linkage, but can also be very helpful in search procedures. Whereas a considerable percentage of name variants indeed has names which belong to the same cluster, there are also variant pairs of which the names are associated to different clusters. This may indicate an erroneous name pair, but often it is an indication of ambiguity: names can be associated to more than one lemma. This is particularly true for first names which are derived from suffixes (for instance *Fien* from *Afien, Adolfien, Josefien, Rudolfien,* etc.) and for abbreviated names. Linkage and search procedures should then test all possible options for interpretation.

The method of using reliable record links to discover name variants is promising, but the process is complicated by the cleaning of errors in the data. This can only be partially performed by automatic procedures. Manual inspection and expert judgement, implemented in an active learning setting, is unavoidable. An explorative comparison of the results with the name variant corpus of FamilySearch revealed that many variants are not shared by both corpora, indicating the enormous scale of name variation (usually of low frequency). In name standardisation, choices for standards are not at all obvious, but influence the results of a comparison.

The final proof of the gain of the presented method is in application of the results in record linkage (through acceptance of names within a cluster of variants and the acceptance of specific name pair variants of which the names reside in different clusters). Such an evaluation requires a golden standard of linked records on which various linkage methods can be applied.

Acknowledgments This work is part of the research programme LINKS (LINKing System for historical family reconstruction, http://www.iisg.nl/hsn/projects/links.html), which is financed by the Netherlands Organisation for Scientific Research (NWO), grant 640.004.804.

References

Anderson, J. M. (2007). *The grammar of names*. Oxford: Oxford University Press.

Bloothooft, G. (1995). *Rules for semi-phonetic conversion of first names and family names*. Uil-OTS internal report (in Dutch).

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1, (Article 5).

Bratley, P., & Lusignan, S. (1976). Information processing in dictionary making: Some technical guidelines. *Computers and the Humanities*, *10*(3), 133–143.

Christen, P. (2012). *Data matching—Concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-centric systems and applications*. Berlin: Springer.

Dolby, J. L. (1970). An algorithm for variable-length proper-name compression. *Journal of Library Automation, 3*(4), 257–275.

Driscoll, P. (2013). Computational methods for name normalization using hypocoristic personal name variants. In *Multi-source, multilingual information extraction and summarization* (pp. 73–91), Springer.

Malin, B. (2005). Unsupervised name disambiguation via social network similarity. In *Proceedings of the workshop on link analysis, counterterrorism, and security* (pp. 93–102).

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. *SICS Technical Report T, 2009*, 06.

Oosten, M. (2008). *Past names, family relation based on data from Genlias*, MSc thesis, LIACS, Leiden University (in Dutch).

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, *18*(6), 38–43.

Russel, R. (1918). Index. US Patent 1261167.

Sarawagi, S., & Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 269–278). ACM.

Schaar, J. van der (1964). *Woordenboek van voornamen*, Aula (since 1992 edited by D. Gerritzen).

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., & van der Goot, E. (2011). JRC-NAMES: A freely available, highly multilingual named entity resource. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 104–110).

Vries, T. de, Ke, H., Chawla, S., & Christen, P. (2009). Robust record linkage blocking using suffix arrays. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM)* (pp. 305–314).

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)* (pp. 354–359).

Xu, J., & Croft, W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *Transactions on Information Systems (TOIS), 16*(1), 61–81.