Learning name variants from true person resolution

Gerrit Bloothooft
UiL-OTS
Utrecht University
The Netherlands
g.bloothooft@uu.nl

Marijn Schraagen
LIACS
Leiden University
The Netherlands
m.p.schraagen@liacs.leidenuniv.nl

Abstract

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

1 Introduction

In entity resolution, 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, 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 real cases.

There are circumstances where sources are rich enough to allow for true person resolution while not using all available information. Names that are not needed in the resolution process, may contain true variants (but errors as well). This paper investigates the procedures needed to harvest a corpus of true name variant in a largely automated way, applied to 14.8 million records from the 19th century vital registration of the Netherlands.

The paper is structured as follows: Section 2 describes the source corpus, section 3 describes the method to harvest name variants, while section 4 discusses the options to differentiate true variants from errors, and section 5 concludes with results and discussion.

2 Material

The data used in the investigation is extracted from the Dutch *Wie-was-wie* (who-was-who) database (www.wiewaswie.nl, release November 2011 as Genlias). *Wie-was-wie* contains civil certificates from the Netherlands, for the events of birth, marriage and death, which registration started in 1811. Most documents originate from the 19th and early 20th century. A record consists of the type of event, a serial number, a date and a place, and information about the participants. The parents are also 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. No links are provided between documents or individuals. The digitization 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.

3 Method and variant pair harvesting

To meet the requirement of true person resolution, we had to estimate the minimum matching information between two records to make this decision. The record combination was taken from birth, marriage and death certificates in the *Wie-was-wie* database. We required exact matching of the first name of a 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 woman always keep their maiden name in the administration. The fourth name of the mother or the father was not part of the resolution 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 true person resolution by selecting matches between birth and death certificates for which *all* five names and year of birth were available and matched. We assumed that this would generate only true matches in the Netherlands. 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 names of parents were a sufficient condition for true person resolution, as violation would generate only a few errors.

An example of a rare match where the condition did not hold is: 07-09-1850 birth of *Jannigje Kool* in Schoonrewoerd, from the parents *Arie Kool* and *Cornelia van Gent*, and her decease in 31-12-1918 in Lexmond, at 68 years of age with mention of the same parents. Competing is the birth of *Jannigje Oosthoek* on 25-04-1850 in Charlois, from the parents *Arie Oosthoek* and *Cornelia van Gent*. 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 be a composition of more than one name, such as the first name *Johan Willem Frederik*. Although we required identity of four out of five (full) names, 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.

Also the differing name could be a composition of a number of names. In this case it was identified which of these names was different in both records. Since the order of the names could be different in both cases, all possible options were analyzed on the basis of Levenshtein distance. For instance, in the pair *Anna Christina Elizabeth / Christiena Elizabeth*, in the second name *Anna* is missing, while the variant pair is *Christina / Christiena*. Sometimes the name order was different as in *Virgin Thomasa Franken / Thomasa Virginia* with *Virgin / Virginia* as variants, while in other cases there was no variant pair at all, but a missing name only: *Adriana Agnita Cornelia / Adriana Cornelia.*

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 different order were taken together. This resulted in 48,684 pairs of single female first names with a total of 246,519 occurrences (or tokens), 31,885 pairs of single male first names with a total of 183,050 tokens, and 177,258 pairs of single family names with a total of 374,901 tokens.

The most frequent name variant pairs, which have only minor spelling differences that mostly do not influence pronunciation, are the female first names *Elisabeth / Elizabeth, Willemina / Wilhelmina, Geertrui / Geertruij*, the male first names *Johannes / Johannis, Jacob / Jakob, Arij / Arie,* and the surnames *Jansen / Janssen, Bruin / Bruijn, Ruijter / Ruyter*.

4 Variants and errors

In this research we wish to harvest a clean corpus of name variant pairs, but name errors complicate the process, also under the condition of true person resolution. Name errors can originate from the writing of the original certificates, but also from misreading or typing errors in the recent digitization 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 person resolution under less favorable conditions, name errors should be recognized 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. But his death certificate mentions *Grietje Spruit* as mother, which leads to the erroneous first name variant pair *Aafje / Grietje*. Additional evidence that the records concern the same person comes from the name of his wife *Aaltje Kort*, mentioned both in the marriage and death certificate (although this information is not used in the resolution 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 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.

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 *Alexander Adolph Edward Johan Reinoud*, son of *Dirk / Derk Willem Gerard Johan Hendrik Brantsen van de Zijp* and *Jacoba Charlotte Juliana van Heeckeren van Kell* while in another registration of these persons, the mother´s family name was found as *Well.* 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 generalized 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. Most frequent erroneous variants for male first names are for instance combinations of popular names, such as *Jacob / Jan, Jacobus / Johannes, Willem / Jan, Gerrit / Hendrik, Jan / Hendrik, Gerrit / Jan, Klaas / Jan, Gerrit / Cornelis, Willem / Hendrik, Dirk / Jan*. The use of rules and manually inspection (active learning) is unavoidable to make a distinction between variants and errors, but we have tried to keep it to a minimum.

4.1 Name pair cleaning

The name pairs resulting from the automatic harvesting step 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 developed and uses a corpus of non-variants, and the third method is based on manually designed variant classification rules. The methods are described in detail in the remainder of this section. Acceptance by the first method or rejection by the second method overruled the outcome of the third method. Additional manual review of a limited selection of variant pairs was applied to correct post-processing errors.

4.1.1 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 Dutch dictionary of first names (van der Schaar, 1964, first edition) is available which associates about 20,000 first names to 3,737 gender-independent lemma's.

This could be helpful as a starting 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 lemma's *Aai, Aalt, Adriaan*. Furthermore, lemma's 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 lemma *Cornelis*, *Nelly* with lemma *Cornelis* or *Petronius*, 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 lemma's. 3,615 female and 2,878 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.

4.1.2 Data-driven harvesting 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 no 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 first names from the Dutch Genlias 2011 release. These names consist of 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 no-variant-corpus of 907,660 pairs of first names, with 18 million tokens.

Unfortunately, as any corpus, the Genlias 2011 release contains errors in the records which arose in the digitization phase. Patronyms and parts of the family name, or the whole family name, were sometimes included in the first name field. For example *Aagtje van Eck*, with *van Eck* as family name, is present as a first name, which results in the incorrect first name 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 patronym or a family name before the first name.

Another problem were 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 letter at the beginning of a name, and a name length of at least three characters (which also excluded initials). After this, a no-variant-corpus of 118,532 first name pairs resulted (only 13% of the originally harvested pairs), with 15 million tokens (83%).

In conflict, however, with the assumption of the approach, also true first name 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 realize 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 2,458 female first name pairs and 2,343 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.

4.1.3 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 or rules compares the differences between two names on the basis of Levenshtein distance and additional requirements that resem-

ble the Jaro-Winkler distance measure (Winkler, 1990).

In the past, the lack of spelling rules has promoted the application of ad hoc rules on the same pronunciation, with spelling variation as a result. Reversely, it could now be tried to apply rules on written forms that result in a close correspondence to the original pronunciation. Since it is impossible to catch all spelling variation (certainly not under the presence of all kinds of errors), a limited but robust rule set – much less crude than Soundex - was developed that converts names, from Dutch sources, into a semi-phonetic form (Bloothooft, 1995).

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. Many misreadings concern the combinations *T – F, T – P, T – J, T – S, T – K, F – P, F – J, I – J, M – H, M – W, M – Al*. The difficulty of misreading (at the digitization phase) is that there is often a bias towards an (erroneous) existing name on the basis of the knowledge of the person who digitizes (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 happens systematically, the resulting name confusion needs not even to 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 summarized in Table 1. These rules were applied to both the original and the semi-phonetic name form. If applicable on any of these two forms, the variant 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 1 it can be seen that variant pairs with a Levenshtein distances well over 2 can be accepted by rule, which also holds for the additional suffix rule discussed above. In many cases a Levenshtein distance of 2 is acceptable as a general threshold. The gain of the current method is in the acceptance of higher edit distances.

Levenshtein distance	length	minimum number of identical initial characters	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*

Table 1. Six heuristic rules that determine whether a name variant pair is accepted. An additional, 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.

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 re-

	Female first names		Male first names		Family names	
	name pairs	tokens	name pairs	tokens	name pairs	tokens
initial name pairs	48,684	246,519	31,886	183,050	177,258	374,901
excluded as no variants	*2,412*	*12,041*	*2,289*	*6,538*	*103*	*199*
excluded by rules	*11,336*	*18,716*	*7,077*	*10,079*	*56,694*	*79,079*
excluded manually	*118*	*414*	*42*	*126*	*346*	*507*
accepted by dictionary	*3,610*	*94,551*	*2,877*	*90,761*		
accepted manually	*1,001*	*3,917*	*563*	*2,458*	*783*	*2,410*
total excluded	13,866	31,081	9,408	16,743	57,143	79,785
total accepted	**34,818**	**215,438**	**22,478**	**166,307**	**120,115**	**295,116**
Levenshtein distance in original form:						
1	58%	69%	65%	70%	69%	77%
2	26%	20%	24%	18%	24%	19%
3	9%	5%	7%	5%	5%	3%
> 3	7%	6%	4%	7%	2%	1%
Levenshtein distance in semi-phonetic form:						
0	19%	29%	22%	29%	29%	44%
1	52%	45%	53%	46%	53%	45%
2	18%	15%	17%	13%	14%	8%
3	7%	7%	5%	7%	4%	2%
> 3	4%	4%	3%	5%	0.2%	0.2%

Table 2: Overview of the results of the various steps in cleaning the initial corpus of name pair variants. Three exclusion and two acceptance mechanisms 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.

view was critical, and concentrated on true variants of low frequency and rejected variants 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, or acceptance of a name pair that did not pass the rules (for instance because the initials were not the same).

5 Results and discussion

A summary of the results of all phases in the cleaning process is presented in table 2. For the accepted name variant pairs, the percentage with a certain Levenshtein distance is given in that table 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.

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 se-

lected by the basic method outlined in Section 3. In Figure 1 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.

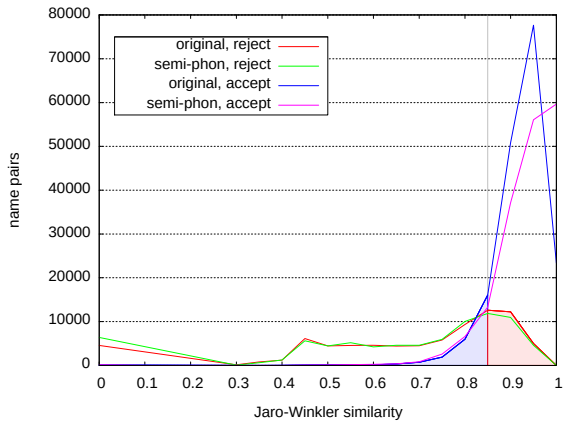

Figure 1. Jaro-Winkler similarity for candidate variant pairs. Values are binned with interval 0.05 for readability.

Figure 1 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, red shaded) and < 0.85 under the curve for accepted pairs (13% false rejects, blue shaded) is the gain of the current post-processing methods over the application of the Jaro-Winkler similarity. The curves in figure 1 do not differ much for names in original and semi-phonetic form. This implies that the Jaro-Winkle similarity does not improve by application on the semi-phonetic name form.

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 2, but amount to 39% of the 13,866 erroneous female first name pairs, 49% of the 9,408 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 matching, 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 learns that the rules mainly exclude rare name pairs as there are about only 1.5 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 tokens per pair. The latter variants are obviously well-known and made it to the dictionary.

Although we harvested 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. 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 3, together with the figures found in the full release (Wie-was-wie 2011). The number of singletons in both collections is presented as well, as they constitute between 40-50% of all names.

	in accepted name pairs		Wie-was-wie 2011	
	all	singletons	all	singletons
female first names	28,574	15,616	61,873	29,912
male first names	20,234	10,984	52,964	26,566
surnames	129,292	72,917	569,063	227,799

Table 3: Number of different names (and singletons among them) in the accepted name pairs, and in the full Wie-was-wie corpus (release 2011).

Given the constraints applied to arrive at accepted name pairs, the number of different names in the current analysis is fairly large. Names can have been 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).

A corpus of true name variant pairs can be used to create clusters of names for which name vari-

ant pairs are only found within a cluster. On this basis, yet unseen name variant pairs can be anticipated. Such a clustering also provides an additional control on the name variant pairs. If variants of the same name are found in more than one cluster, this may be an indication of ambiguity (*Jo* could be a variant of both *Josef* and *Johannes*), or an error.

Once one has arrived at name clusters, it is a little step to name standardization. These name standards could be implemented in the original corpora, and the current analysis can be repeated. The requirement still is that three out of four names of the parents are the same, but now one or more of the names can be a standardized variant. This will generate an additional corpus of name variants – but on a slightly more risky basis (standardization can be incorrect). Research into this direction is in progress.

The foundation of this research was in a corpus of records with a very high confidence level of true person resolution. We focused on the harvesting of true name variants to apply these later under less favorable conditions. But it is also of interest to consider the level of name errors that are present in this corpus. From table 2 it can be seen that this 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 the errors could be identified as such because there was sufficient evidence to decide on true person resolution. Of course, in sources that are less rich in information on individuals, these errors cannot be traced that easily. In that case it may only be hoped that more complex decision strategies than based on pair-wise comparison of records can be developed, to decide on true person resolution.

Part of the errors we identified are likely reading errors of the type *Pietje / Tietje*, in which *P* and *T* are confused, or typing errors like *Bos / Vos* with *B* and *V* as neighboring 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 labeled as errors. However, if we could estimate the likelihood of these errors, this could be incorporated in a resolution decision model (rather than requiring excess of information to be able to circumvent these errors).

The final proof of the gain of the method is in application of the results in record linkage (through pair-wise acceptance of name variants or in the application of name standards). Such an evaluation actually requires a golden standard on which various linkage methods can be applied. We are planning this for a small subset of vital records.

The method of using true person resolution to learn name variants is promising, but the hard work is in the presence of errors in the data and their cleaning. This can only be partially performed by automatic procedures. Active learning, i.e. manual inspection and expert judgement is unavoidable.

Acknowledgement

This research has been supported under the NWO CATCH programme in the LINKS project.

References

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

Christen, P. (2012). *Data matching*, Springer.

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

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

Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". *Proceedings of the Section on Survey Research Methods* (American Statistical Association): 354–359.

De Vries, T., Ke, H., Chawla, S. and Christen, P. (2009) *Robust Record Linkage Blocking using Suffix Arrays*. In *Proceedings of the 18th ACM conference on Information and knowledge management.* (ACM): 305–314,