NAMES 1.1 corpus

Clariah Core project CC 17-012
(April 2017 – June 2019)

Manual v1

Project members:
Gerrit Bloothooft & David Onland, UiL-OTS Utrecht
Martin Reynaert, TiCC / Tilburg University
Katrien Depuydt, Tanneke Schoonheim & Matthieu Fannee, INT Leiden
Jauco Noordzij, Huygens Institute for Dutch History, Amsterdam

I Introduction

The NAMES corpus is a collection of 189,707 given names (61,9 million tokens) and 562,676 surnames (54,6 million tokens) as present in 19th century certificates for birth, marriage and decease (wiewaswie.nl as of 2011). This manual describes data properties, the approaches for assigning standards to the names , and the data structure of the resulting tables. The assignment of standards to given names and surnames is a complicated issue, some considerations and choices made for the NAMES corpus are discussed in section II. It is important to note that for the NAMES project standard should be interpreted as a tag of a limited search space, which not necessarily coincides with an etymological basis. Section III describes the provenance of the onomastic data, while their preparation for the NAMES project is given in section IV. The approaches in the assessment of standards by experts and machine is treated in section V. Details of the data structure of the resulting tables in the NAMES CORPUS release 1.1 are given in section VI. The INT LexiconService for a part of the NAMES Corpus is described in section VII and Appendix II, the RDF conversion and access through ANANSI in section VIII.

II Standards

The assignment of standards to given names and surnames is a difficult issue for what defines a correct assignment? The ideal supporting information is when name variants have been used for the same individual (which is part of our underlying material, see section V), but even this may result in ambiguities, pointing to more than one possible standard for some names. This is resolved in part by allowing for relations between standards, see section V.

The role of the expert in this project was to review the (automatic) assignment of names to standards, to assign new variants to existing standards, or to decide that an additional standard is needed. No detailed motivation is asked for these decisions, which implicitly assumes that the expert has some internal rule system (knowledge) upon which decisions are based. If such a system would exist it may also be modelled computationally. However, it is not at all sure that expert decisions always have a systematic basis. The best compromise therefore is an interactive approach in which both expert decisions and (statistical) computational modelling are involved. This actually was the paradigm in this project, but it came only limited to implementation. There is room for further research on statistical learning on variants of personal names.

Ambiguities and uncertainties are an inherent property of the assignment of standards to person names. Whereas for many names there is little doubt that names are variants, such as *Willemina* and *Wilhelmina*, for other names various options may exist, such as *Margje,* which could be a variant of the standards *Maria* or *Margaretha*. The probabilities of these options could be estimated from the number of proven variants (for instance, through record linkage), or from the statistical likelihood of character/string operations. In any case, the assignment of a standard can vary between highly likely to some level of likelihood. The notion of *gold standard* (without doubt) can therefore only be applicable for a limited part of an onomastic corpus. If an expert is asked to make a final decision this is actually an impossible task when more interpretations are possible. Nevertheless, in the development phase the expert view is the only feedback/test option when information on proven variant pairs is missing.

We have distinguished four levels of quality in expert decisions, of which only level 1 is based on a thorough analysis of underling information of variant pairs, and used for given names only. Level 2, the highest level for surnames, concerned an evaluation of the standards which were automatically

derived by clustering of variant pairs, without considering in detail the underlying information. Due to the larger number of standards this simply would be too time consuming. Levels 3 and 4 concerned standards which were automatically derived by a few simple rules based on a Levenshtein edit distance of 1 with standardized names, and only marginally reviewed for the names with the highest frequency. But even this, the review of several ten thousands of surnames took many weeks of expert analysis. In all, there is room for considerable improvement, for which intelligent computational assistance is an absolute necessity.

Results are presented as a percentage coverage of names (type) and their occurences (tokens) at the four levels of quality. It should be emphasized that whereas the token coverage is as high as 99% for both given names and surnames, this does not imply that 99% of the tokens has a correct standard. It implies that for 99% of the tokens it has been possible to propose a standard (with various levels of quality).

III Data provenance

Personal names were obtained from the Catch LINKS project (2009-2015, https://socialhistory.org/nl/node/5451) and concern all given names and surnames which appear in the certificates of the 19th century Civil Registration (started in 1811; version Genlias 2011, which currently is a subset of the Wiewaswie corpus (accessible online at http://www.wiewaswie.nl). The data used in the NAMES project were not corrected for spelling errors, but efforts were made to identify and remove names which were originally digitized in the wrong fields, i.e. family names as given names and reversely, patronymics, titles and prefixes in both given names and family names.

The full NAMES corpus consist of information derived from 53.839.590 million person attestations, which results in 189,707 different singular given names (with 61.9 million tokens - due to multiple given names) and 562,676 different singular surnames (with 54,625,914 tokens - due to a limited number of multiple surnames).

As a reference, the count of both given names (by gender) and surnames in the current Civil Registration has been used. These data originate from the Basisregistratie Personen (BRP), formerly the Gemeentelijke Basisregistratie (GBA), 509,561 given names (case sensitive, selection 2014: 17,804,960 tokens) and 326,946 surnames (case sensitive, selection 2007: 16,534,465 million tokens). The count of names (or its elements) (>4 because of privacy reasons) can be provided without violating privacy issues, as they are also shown online in the Voornamenbank (www.meertens.knaw.nlk/nvb) and the Familienamenbank (http://www.cbgfamilienamen.nl/nfb/).

IV Preparation of the original data

IV.1 Given names

Given names may consist of multiple names (*Wilhelmus Franciscus Maria*). These are parsed into individual name elements (*Wilhelmus, Franciscus* and *Maria*), each of which have the gender of the bearer. Only singular given names and their gender are provided in a table.

For each name three counts are provided: occurrences in the first (or only) position, occurrences in a later (unspecified) position, and occurrences in the last position. The latter two figures apply only

to multiple given names. These numbers are given both for the genlias2011 data and the BRP2014 data. Whereas a name in first position is usually indicative for gender, this is not necessarily the case for later positions. Furthermore the name in last position can be a patronym. This may originate from digitization when the name was not recognized as a patronym or when there was no separate field for patronyms.

IV.2 Surnames

Surnames can consist of prefixes, such as *de, van der, ten, uiten* and the surname itself. These were parsed as *prefix(es)* and *surname*. Most surnames are singular (a single word) but sometimes surnames have a complex structure and can consist of a prefix followed by a complex name (*van Heekeren van der Schoot*) which are (1) parsed as a prefix (*van*) and a complex surname (*Heeckeren van der Schoot)*, while the latter is (2) parsed into the singular surnames (*Heeckeren* and *Schoot*) without an infix (*van den*). Three tables are provided: (1) the full surname with fields for prefix and surname, (2) singular surname, (3) prefix. There are 39,334 cases with a space in the complex surname (5,8%), indicating a multiple surname (not always, such as in rare cases like *ook wel de Graaff*).

For each singular surname the counts in both the genlias2011 and BRP2007 corpora are provided for occurrence in the first position (i.e. in the example above, *Schoot* is not counted as it shows in second position).

V Standards

The basis for the derivation of standards is a corpus of name variant pairs which were derived in the LINKS project on the basis of near exact record linkage. Two certificates in the civil registration have a high likelihood of concerning the same individuals if at least four out of five names of child, father and mother match exactly (given name child, surname child, given name father, given name mother, surname mother). When the fifth name is not the same, this creates a (supposed) name variant pair (*Ruijter/Ruyter*), see Bloothooft en Schraagen (2015). This name variant corpus has been extensively cleaned in the LINKS project, and is part of the NAMES corpus. It consists of 328,411 surname variant pairs and 122,526 first name variant pairs. Pairs are stored bi-directionally in tables, i.e. both as name1-name2 and name2-name1, for easier lookup.

Bloothooft, G. and Schraagen, M. (2015),' Learning name variants from inexact high-confidence matches', in Bloothooft et al. (Eds.), *Population Reconstruction*, Springer, Switzerland.

The name variant pairs are the basis for clustering of names which have been seen as mutual variants. Each cluster is then characterized by a standard name, usually the name with the highest frequency. This has also been done by Bloothooft en Schraagen (2015). As not all names are found in variant pairs, only 42,033 given names (22.2%) could be clustered into 926 gender-independent LINKS standards, while 119,906 surnames (21.3%) formed 15,112 LINKS standards.

These LINKS standards have been extensively reviewed and extended in the NAMES project. This has been done at four levels of quality, which level is added to the data. It should be emphasized that a ‘standard’ in the NAMES project should be understood as a level of clustering useful for search and nominal linkage. The aim was to minimize the likelihood of missed candidates (false negatives). This means that names which are etymologically different may have got the same standard. A two-step procedure has been followed: expert review (quality level 1 – based on proven variant pairs, and level 2 – based on review of LINKS standards only), and automatic assignment of standards followed

by marginal expert review (quality levels 3 and 4).

V.1 Expert review

V.1.1 Given names

A given name was analysed firstly on the basis of the ratio between the number of variant pairs for which the variant name has the same LINKS standard, relative to the total number of pairs found for that name. When this ratio was below 0.5 it was investigated whether a combination of some standards improved this fraction (for the combination). This result was the starting point for an extensive onomastic expert review of all underlying variant pairs of a standard. This resulted in revisions of the LINKS standards and re-ordering of names under INT standards, with quality level 1. As the gender of a name bearer is not always given or deducible from an attestation, it was decided to use gender-independent standards.

The names of the INT standards were chosen in such a way that they referred to their etymological origin (variants of *Hendrik* got the standard HEIMRIJK; this was preferred over the name of the most-frequent variant in genlias2011/wiewaswie, although the latter variant is also added as gn_standard_hf to the the table NAMES_GN _STANDARD for easier understanding), and by using the lemma's in J. van der Schaar (2010, 20th edition; 1st edition 1964), 'Dictionary of given names'. When more than one etymological origin could apply a generic form was chosen, such as ADELB- which summarizes origins such as ADELBERT and ADELBRAND.

The assignment of 784 standards can be found in the table NAMES_GN_PAIRS in the field gnp_standard_INT. In various cases a single standard did not apply to both names of a pair which was solved by using 708 double standards such as ADELWOLF/AVE, and 15 triple assignments such as DAMASUS/DAMIANUS/DANIËL.

The double and triple standards would not be a problem if there is a one-to-one relation between a name and a standard. But ambiquity shows in different standards assigned to a single name originating from name pairs of which the name is assumed to be derived from the other name and inherits its standard. A maximum of 7 different standards was found related to the female name *Matje* (MATTHIAS, MARIA, MARTINUS, MARTHA, MARGARETA, METTE). The number of names with 1 or more INT standard is as follows:

number of INT standards	number of given names
1	38,946
2	1,535
3	166
4	33
5	6
6	3
7	1

For the NAMES project it was decided to generalize this ambiguity at the level of standards and not at the level of individual names (although it can be derived from the NAMES_GN_PAIRS table). This required to enforce a single standard on a name and to introduce a table with relations between standards. The optimal decision of a single standard is not self-evident. Especially reduced or abbreviated given names pose a problem. Such short names may originate from full names with a different origin, such as *Jo* from *Jozef* or *Johannes* and *Mina* from *Wilhelmina, Hermina* of *Jacomina*. This is (partly) solved by defining additional NAMES standards for short names. It is helpful that given

names of Germanic origin (and others as well) are often composed of a limited set of two (or more) syllables. If they are abbreviated this often results in variants related to a single syllable. For many of the two-syllabe Germanic-based standards there exists both abbreviated standards associated to the first and second syllable . For the first syllable for instance ADEL next to ADELBERT, ADELBRAND, ADELFONS, ADELFRIED, ADELHARD, ADELHEID, ADELLAND, ADELMAR, ADELRIJK, ADELTRUIDA, ADELWIJN, ADELWOLF, and ANS to ANS*, DIED to DIED*, GOS to GOS* and others, where for instance ANNE* summarizes longer standards starting with with ANNE (ANNEBELLA, ANNEBET, ANNEGRIET, ANNELEN, ANNELIES, ANNEMARIE). In several other cases abbreviated names are still uniquely related to the full name, and no separate standard is needed, such as *Bram* from *Abraham*. To show this relation better, several INT standards were given a slightly adapted NAMES standard.

The introduction of (1) additional standards for abbreviated names, (2) the assignment of a single standard to a given name, (3) the assignment of standards to name pairs which did not get an INT standard, and (4) the choice of a slightly adapted standard name, resulted in 21.869 name pairs (18%) that got a different NAMES standard than the INT standard.

The expert review process and the following adaptations resulted in 813 gender-independent standards for given names related to 42,705 given names (coverage 22.2%, quality level 1, i.e. expert review at the level of variant pairs) with 61,030,898 tokens (coverage 98.60%). From the 120,104 (bi-directional) variant pairs, 108,496 got the same NAMES standard for both names (90.3%).

Table 1. A few examples of variant pairs with unequal standards, showing various problems which can be partly solved by establishing relations between standards.

names	**standards**
AARTJE – ANTJE	arnoud – anne
FROUKJE – FRANKJE	vrouw – franciscus
ALFINA - AFINA	adelwolf – ave
BAARTJE – BAATJE	bert – beatrix
MEINKE – MINKE	mein – mina
KLAASJE – KLAARTJE	nicolaas - clara

For easier interpretation of a standard, the given name with the highest frequency (in genlias2011/wiewaswie) under a standard is also added as field in table NAMES_GN_STANDARD.

Table 2. Examples of differences between the NAMES standard and the name with the highest frequency under that standard.

standard	**given name with highest frequency in standard**
ab	ABE
adel	AALTJE
adelb-	EBELTJE
adelbert	ALBERT
adelbrand	ALBRAND
adelfons	ALPHONSUS
adelfried	ALFRED
adelgonda	ALLEGONDA
adelhard	ALDERT
adelheid	ALIDA
adelland	ALMA
adelmar	JELMER
adelrijk	AALDRIK
adeltruida	AALTRUID

V.1.2 Surnames

The onomastic expert review of surnames was limited to the standards themselves without going into details of the variant pairs underlying them. Nevertheless, the CBG corpus of Dutch family names (http://www.cbgfamilienamen.nl/nfb/) and F. Debrabandere (2003), 'Dictionary of surnames in Belgium and Northern France' have been used in the review process as well, to support the choice of the standard (lemma). [If copy rights allow, the digitized version of Debrabandere, as realized in the NAMES project, will be made available to Clariah]. This resulted in a reduction from 15,112 to 10,162 standards for 119,906 surnames (coverage 21.31%, quality level 2, i.e. expert review at standard level) with 26,114,275 tokens (coverage 47.81%).

V.2 Automatic assignment of standards

V.2.1 Semi-phonetic transcription

The first step in the automatic assignment of standards was the use of the semi-phonetic transcription of names [Bloothooft, G. (1995 (http://www.gerritbloothooft.nl/Publications/VREGELS.pdf , internal report)]. This transformation is phonetically inspired, but is deliberately limited because spelling of context in historical text can be highly variable, which hampers the usage of context-dependent rules.

Table 3. Example of the given names that all have the semi-phonetic transcription ELYSABET (which form generates most variants), plus their frequency in genlias11.

ELISABETH	858284	ELISAABETH	14	ELIZAABETH	2	ELIZABETH'	1
ELIZABETH	240116	ELISABEETH	12	ELIZABETHJ	2	ELIZABETHTH	1
ELIZABET	11984	ELIISABETH	11	ELLISABET	2	ELIZSABET	1
ELISABET	9777	ELIZABEHT	11	ELLIZABET	2	ELIZZABETH	1
ELIJSABETH	958	ELISABET	11	EEEEELISABETH	1	ELJZABET	1
ELIJZABETH	890	ELYZABET	10	EELEISABETH	1	ELÏESABETH	1
ELISABETH	591	ELISABBETH	9	EELIJZABETH	1	ELÏSABETH	1
ELIEZABETH	553	EL;IZABETH	8	EELISAABET	1		
ELIESABETH	436	ELISABETT	8	EELIZABET	1		
ELIESABET	261	ELIZABET	8	EELYSABETH	1		
ELIJZABET	254	ELIIZABETH	7	EL;IZABET	1		
ELIJSABET	251	ELIZABEDT	7	ELEISABET	1		
ELIEZABET	187	ELIZABEETH	7	ELIESABEDT	1		
ELIZABETH	150	ELEIZABETH	6	ELIESABEHT	1		
ELEIJSABETH	124	ELIZABETT	6	ELIJSABEHT	1		
ELYSABETH	119	ELIZABETTH	6	ELIJSABETT	1		
ELISABEHT	74	EELISABET	5	ELIJSZABET	1		
ELLISABETH	73	ELISABEDT	5	ELISAABET	1		
ELYZABETH	65	EELIESABET	4	ELISABED	1		
ELEIJSABET	43	EELIESABETH	4	ELISABEDTH	1		
ELISSABETH	43	EELIJSABETH	4	ELISABEET	1		
ELISZABETH	41	EELIZABETH	3	ELISABET4H	1		
ELISABETTH	19	ELIESAABET	3	ELISABET6H	1		
ELYSABET	19	ELISABETHJ	3	ELISABETH6	1		
ELEISABETH	17	ELIZABETJ	3	ELISABETH?	1		
EELISABETH	16	ÉLISABETH	3	ELISSABEHT	1		
ELISABETJ	16	(ELIZABETH)	2	ELISZABET	1		
ELLIZABETH	16	ELIESABED	2	ELIZABBETH	1		
ELIZSABETH	15	ELISABEHTH	2	ELIZABED	1		
EL;ISABETH	14	ELISSABET	2	ELIZABEET	1		

V.2.2 Given names

As a first automatic assignment the standard of a given name was extended to names with the same semi-phonetic transcription. These names also got quality level 1. The coverage of standardized given names increased from 42,705 to 72,871 (from 22.2% to 38.4%) which shows the power of the semi-phonetic transcription [the 72,871 names only have 163,973 different tokens because the initial set already included names with the same semi-phonetic form]. The number of covered tokens increased from 98,60% to 98,87% (see table 1).

In addition, two additional automatic approaches were followed to identify more variants. In the second automatic assignment all non-standardized names were compared at the semi-phonetic level to names with a standard (brute force, courtesy Marijn Schraagen). For a non-standardized name, related standardized names were selected on the basis of a Levenshtein distance of 1, a length of over 5 characters, and equal two initial characters. The standard shared by most of the selected names was chosen (without taking the frequency of the names into account), and these additional names were given a quality level of 3 (automatically assigned without further expert review). For given names this resulted in an increase to 120,294 standardized names (63.4%) with a token coverage of 99,43%.

The third automatic analysis was based on the property that less spelling variation is found in the first part of a name. For non-standardized names, the set of standardized names was determined which had the *first five semi-phonetic symbols* in common. Five symbols usually cover the first syllable and the beginning of the second, which is distinctive for many names with equal standard. From this set the most occurring standard (again without taking frequency of the names into account) was chosen for the non-standardized name. The same procedure was subsequently followed for remaining non-standardized names for sets which have the *first four semi-phonetic symbols* in common. The resulting total of 48,219 additional assignments were manually reviewed for serious flaws which resulted in 8,383 updates. All assignments got quality level 4. Quality levels 1 up to 4 then have a total coverage of name types of 88.8% while the token coverage is 99.82% (see table 4).

Table 4. Summary of counts for quality levels of assigned given name standards (813 standards). The NAMES corpus consists of 189,707 different given names with a token frequency of 61,894,242 (over all positions in a given name).

quality level	given names	cum %	tokens	cum %
1 (expert)	42,705	22.2	61,030,898	98.60
1 (semi-phon)	30,166	38.4	163,973	98.87
3 (edit dist = 1)	47,423	63.4	348,846	99.43
4 (equal start)	48,219	88.8	241,378	99.82
none	21,193	100.0	108,763	100.0

Most of the remaining 21,193 names have a low frequency in the 19th century corpus, but several are highly frequent in the current vital registration (basisregistratie personen, brp). It then concerns fashion names and immigrant names, with as top: *Kim* (8 in genlias, 25,489 in brp2014)*, Mohamed, Tom, Iris, Astrid, Roy, Pim, Sven, Sonja, Naomi, Sem, Daphne, Maud, Bob, Twan, Ilona, Noor, Mohammed, Fatima, Zoë*, … As the focus of the current project is on 19th century names, a further analysis of these 20th century names (with additional standards) was not pursued, and left for a subsequent project.

Table 5. Example of all given names with standard *Abraham*, with gender, frequency in genlias11 (all positions in name), and quality level; ordered by quality level and frequency.

ABRAHAM	M	135137	1
ABRAM	M	11322	1
ABRAHAM	F	859	1
ABRAHAMINA	F	560	1
ABRAMINA	F	523	1
BRAM	M	159	1
ABRAAM	M	64	1
ABRHAM	M	60	1
ABRAHAN	M	49	1
ABRAHAMINE	F	48	1
ABRAHAMMINA	F	41	1
ABRAMMINA	F	37	1
ABRAHEM	M	35	1
ABRAM	F	30	1
ABRAHAMUS	M	29	1
ABARAHAM	M	20	1
ABRAHAMA	F	19	1
ABAHAM	M	15	1
ABRAHM	M	14	1
ABARHAM	M	11	1
ABERAM	M	10	1
AABRAHAM	M	9	1
ABRAHAMMA	F	8	1
ABRAMMA	F	8	1
ABRAMIENA	F	6	1
ARAHAM	M	6	1
ARBRAHAM	M	6	1
AABRAM	M	5	1
ABRAHMA	M	5	1
ANBRAHAM	M	5	1
ARBAHAM	M	5	1
ABRAHAMN	M	3	1
ABRAMA	F	3	1
ABRAHAMA	M	3	1
AABRAAM	M	2	1
ABBRAHAM	M	2	1
ABRAAHAM	M	2	1
ABRABAM	M	2	1
ABRAHAMIENA	F	2	1
ABRAHAMMIENA	F	2	1
ABRAMAN	M	2	1
ADBRAHAM	M	2	1
ASBRAHAM	M	2	1
BRAM	F	2	1
ABANDRINA	F	1	1
ABRA?M	M	1	1
ABRAHAAM	M	1	1
ABRAHAM'	M	1	1
ABRAHAM-	M	1	1
ABRAHAMINNA	F	1	1
ABRAHAMM	M	1	1
ABRAHAMMINE	F	1	1
ABRAHANIUS	M	1	1
ABRAHMAM	M	1	1
ABRAHMAN	M	1	1
ABRAHÁM	M	1	1
ABRANDINA	F	1	1
ABRÄM	M	1	1
ABTRAM	M	1	1
AHAHAM	M	1	1
AMBRAHAM	M	1	1
ARAHAMMINA	F	1	1
AVRAHAM	M	1	1
BRAAN	M	1	1
BRAN	M	1	1
ABRA	F	721	3
ABRINA	F	80	3
ABRAHA	F	67	3
ABRAHANNA	F	55	3
ABRAMINE	F	47	3
ABRADINA	F	30	3
ARAMINA	F	20	3
ABRIJNA	F	7	3
ABRA	M	7	3
ABRAHAMDINA	F	6	3
ABRAMSITO	M	6	3
ABERHAM	M	5	3
ABRAHANM	M	5	3
ABRAMSIE	F	5	3
ABREHAM	M	5	3
ABRAMMINE	F	4	3
ABARAM	M	3	3
ABRAHAMIA	F	3	3
ABRAJETTA	F	3	3
ABRABRAM	M	2	3
ABRAHAIN	M	2	3
ABRAHAMI	F	2	3
ABRAHAMSITO	M	2	3
ABRAMIENE	F	2	3
ABRAMLINA	F	2	3
ABRAMWIENE	F	2	3
ABRAMWINE	F	2	3
ABRAN	M	2	3
ABTRAHAM	M	2	3
ALBRAHAM	M	2	3
ANRAHAM	M	2	3
ABRAHA	M	2	3
ABERAHAM	M	1	3
ABRABRAHAMINA	F	1	3
ABRADINNE	F	1	3
ABRAGAM	M	1	3
ABRAHAMAN	M	1	3
ABRAHAMBOTHALL	M	1	3
ABRAHAMENA	F	1	3
ABRAHAMIENTJE	F	1	3
ABRAHAMIJ	M	1	3
ABRAHAMIN	F	1	3
ABRAHAMISRA?L	M	1	3
ABRAHAMJACOB	M	1	3
ABRAHAMJACOBUS	M	1	3
ABRAHAMMIA	F	1	3
ABRAHAMOVITCH	M	1	3
ABRAHAMX	M	1	3
ABRAHMMIANA	F	1	3
ABRAHRAM	M	1	3
ABRAHUM	M	1	3
ABRAJAM	M	1	3
ABRAK	M	1	3
ABRAKAM	M	1	3
ABRAMANA	F	1	3
ABRAMCINA	F	1	3
ABRAMDINA	F	1	3
ABRAMDINE	F	1	3
ABRAMIENTJE	F	1	3
ABRAMJACOB	M	1	3
ABRAMMARINE	M	1	3
ABRAMOW	M	1	3
ABRAMT	M	1	3
ABRAMËUS	M	1	3
ABRANAM	M	1	3
ABRANS	M	1	3
ABRANSI	F	1	3
ABRASIUS	M	1	3
ABRATA	F	1	3
ABREAHAM	M	1	3
ABREAM	M	1	3
ABREMINA	F	1	3
ABRHAMA	M	1	3
ABRHAMINA	F	1	3
ABRIMA	F	1	3
ABROMINA	F	1	3
ADRAHAM	M	1	3
AGRAHAM	M	1	3
ABANDONE	F	3	4
ABREE	F	2	4
ARBACES	M	2	4
ABAN	M	1	4
ABANA	F	1	4
ABANDANA	M	1	4
ABANDIA	F	1	4

ABANIA	F	1	4	ABRESON	M	1	4	AHAHIEROS	M	1	4
ABRENCIENA	F	1	4	ABROM	M	1	4				

V.2.2.1 Relations between standards of given names

Often there is no one-to-one relation between a given name and a standard. This is especially true for abbreviated names, reduced to the first or last syllable of a full given name. Such as *Wil* and *Mina* from *Wilhelmina*. In this example *Wil* could also be a reduction from *Wilfrida* as well, while *Mina* could originate from *Hermina* or *Jacomina*. To accommodate this, short names like *Wil* and *Mina* are associated with an own standard, while multiple relationships between standards are accepted.

Relations between standards are derived from the name variant pairs. As the variants are associated to a standard, this table can be converted to a table of name standard pairs, which describes the relations between standards. In total there are 3,734 relations between the 813 given name standards (bi-directional, table: names_gn_standard_relations).

From the token frequency of the underlying name pairs, the ratio of this frequency for some combination of standard1 and standard2 relative to all name pairs associated with standard1 can be computed (field: ratio). For most standards this ratio is the highest for the identity combination (ideally 100% when no relation to other standards exist). For evaluation purpose we included three tests: a) standard pair identity (*ok-id*), b) fraction > 5% (in one of the two directions of the pair) (*ok-ratio*), and c) whether one standard name is part of the other, to accommodate abbreviations (*ok-suf*), for instance JO and JOHANNES and JOZEF (note that the name of standards is essential to allow for this rule). When applying the test criteria, requiring non-identity, and additional acceptance when the token frequency for a standard combination was higher than 4, only 606 (bi-directional) relations would remain.

Table 6. Example of relations between the standard MINA and other standards. Types indicates the number of different variant pairs with a standard combination, tokens is the total number of variant pairs concerned.

standard1	standard2	types	tokens
MINA	MEIN	50	245
MINA	MINA	48	224
MINA	WILHELM	5	87
MINA	HEERMAN	9	67
MINA	MARIA	4	14
MINA	MEINHARD	4	5

Table 7. Details of the name pairs with as name 1 a name with standard MINA, and paired name 2 with the same or other standard; plus the number of tokens of the variant pair.

name 1	name 2	standard2	tokens
MINA	HERMINA	HEERMAN	31
MIENTJE	HERMINA	HEERMAN	16
MINA	HARMINA	HEERMAN	10
MIENA	HERMINA	HEERMAN	3
MIENTJE	HARMINA	HEERMAN	2
MINA	HARMANNA	HEERMAN	2
MINA	HERMINE	HEERMAN	1
MIENA	HARMINA	HEERMAN	1
MIENA	HERMINE	HEERMAN	1
MIENTJE	MIETJE	MARIA	9
MIENTJEN	MIETJEN	MARIA	3
MIENKE	MIEKE	MARIA	1
MINA	MIA	MARIA	1
MINKE	MENKE	MEIN	95
MIENTJE	MIJNTJE	MEIN	17
MIENTJE	MEINTJE	MEIN	17
MYNTJE	MIJNTJE	MEIN	15
MINTJE	MEINTJE	MEIN	8

MINTJE	MENTJE	MEIN	6
MYNTJE	MIJNTJE	MEIN	6
MIENTJE	MEENTJE	MEIN	5
MINKE	MEINKE	MEIN	5
MIJNE	MEINE	MEIN	5
MINA	MEINA	MEIN	4
MINTJE	MINKJE	MEIN	4
MINTJE	MENSJE	MEIN	4
MINNA	MENNA	MEIN	4
MIENTJEN	MEINTJEN	MEIN	3
MINTJE	MIJNTJE	MEIN	3
MINKE	MINKJE	MEIN	3
MINA	MIJNTJE	MEIN	3
MINKE	MINK	MEIN	3
MIENTJE	MIJNTJEN	MEIN	2
MINKE	MINKIE	MEIN	2
MINSKE	MINKSKE	MEIN	2
MINTJE	MENTJE	MEIN	2
MINTJE	MEENTJE	MEIN	1
MYNA	MIJNA	MEIN	1
MIENTJE	MAINTJE	MEIN	1
MINTJE	MENSCHJE	MEIN	1
MIENTJE	MENTJE	MEIN	1
MINKE	MINNKE	MEIN	1
MINJE	MENTJE	MEIN	1
MINA	MINK	MEIN	1
MINKE	MINK	MEIN	1
MINA	MIJNTJEN	MEIN	1
MINKE	MINKIEN	MEIN	1
MIENTJEN	MEENTJEN	MEIN	1
MINTJEN	MEINTJE	MEIN	1
MIENTJEN	MIJNTJE	MEIN	1
MIENA	MEINE	MEIN	1
MJINTJE	MIJNTJE	MEIN	1
MIJNE	MIJNA	MEIN	1
MINKE	MIJNTJE	MEIN	1
MIENA	MEINA	MEIN	1
MINNA	MINNE	MEIN	1
MINKE	MENKE	MEIN	1
MINA	MINNE	MEIN	1
MIJNE	MEIJNE	MEIN	1
MIENTJE	MEINTJE	MEIN	1
MYNTJE	MIJNTIE	MEIN	1
MINE	MINNE	MEIN	1
MINNA	MINNE	MEIN	1
MYNTJE	MEINT	MEINHARD	2
MINTJE	MEINDERTJE	MEINHARD	1
MINTJE	MINTO	MEINHARD	1
MIENTJE	MIENT	MEINHARD	1
MIENA	MINA	MINA	27
MINA	MIENA	MINA	27
MIENTJE	MIENTJEN	MINA	23
MIENTJEN	MIENTJE	MINA	23
MINE	MINA	MINA	15
MINA	MINE	MINA	15
MINE	MIENE	MINA	8
MIENE	MINE	MINA	8
MIENA	MIENE	MINA	5
MINKE	MINTJE	MINA	5
MIENE	MIENA	MINA	5
MINTJE	MINKE	MINA	5
MINKA	MINKE	MINA	4
MINKE	MINKA	MINA	4
MINA	MIENE	MINA	3
MIENTJE	MIENTIEN	MINA	3
MIENE	MINA	MINA	3
MIENTIEN	MIENTJE	MINA	3
MINE	MIENA	MINA	2
MIENTJE	MIENSJE	MINA	2
MIENTJE	MINTJE	MINA	2
MINKE	MINJKE	MINA	2
MIENA	MINE	MINA	2
MIENSJE	MIENTJE	MINA	2
MINTJE	MIENTJE	MINA	2
MINJKE	MINKE	MINA	2
MINNA	MIENA	MINA	1
MINA	MINNA	MINA	1
MIENET	MIENE	MINA	1
MIENSKE	MIENSJE	MINA	1
MINTJE	MINA	MINA	1
MINSKE	MINKE	MINA	1
MINTJE	MINJE	MINA	1
MINKO	MINKE	MINA	1
MIJNKE	MYNKE	MINA	1
MINI	MINE	MINA	1
MINKEN	MINKA	MINA	1
MIENA	MINNA	MINA	1
MINNA	MINA	MINA	1
MIENE	MIENET	MINA	1
MIENSJE	MIENSKE	MINA	1
MINA	MINTJE	MINA	1
MINKE	MINSKE	MINA	1
MINJE	MINTJE	MINA	1
MINKE	MINKO	MINA	1
MYNKE	MIJNKE	MINA	1
MINE	MINI	MINA	1
MINKA	MINKEN	MINA	1
MINA	WILHELMINA	WILHELM	48
MINA	WILLEMINA	WILHELM	36
MINA	WILLEMPJE	WILHELM	1
MINA	WILLEMINE	WILHELM	1
MINA	WILLEMIJNTJE	WILHELM	1

V.2.3 Surnames

The token coverage of 119,904 surnames with a standard (after expert review) was 48.5% for surnames [quality level 2]. The much lower percentage for surnames than for given names is due to relatively less surname variant pairs in the LINKS corpus leading to missed clusters (from the onomastic point of view).

When standards were extended to surnames with the same semi-phonetic transcription the number of surnames increased to 194.607 (coverage 34.6%) while the token coverage increased to 87.1%. As with the given names, the semi-phonetic transcription proved to be very effective. This standard assignment is given level 2 as well.

Subsequently, the same brute force method as used for given names was applied for surnames, i.e. on the basis of selection of all non-standardized names that have a Levenshtein distance of 1 to surnames with a standard. A more refined filtering was applied in which the required agreement in the initial (semi-phonetic) characters became less with increasing length of the non-standardized surname:

length = 3 three equal initial characters
length >=4 four equal initial characters
length > 5 two equal initial characters
length > 9 same initial character

When more than one standardized surname fulfilled the conditions, the standard that showed most in this set was assigned to the non-standardized surname. This automatic assignment of a standard got quality level 3 and considerably increased the coverage of surnames from 34.6% to 63.3%, while the token coverage increased from 87.1% to 92.4%.

When the remaining non-standardized surnames were considered, it showed that many of them were not seen as part of a variant pair (some others were not standardized due to errors in the analysis flow), and as such not identified as possible standards. Out of these, 23,332 surnames (with 9,771 different semi-phonetic transcriptions) were selected which have a frequency over 10 both in the NAMES corpus and in the current Civil Registration (BRP) (an arbitrary threshold). These surnames were manually reviewed and 8,844 additional standards were identified. Because of this expert review, quality level 2 was assigned. By this extension of the number of standards, the token coverage increased from 92.4% to 98.7%.

With these larger group of standards, the brute force method was applied again for the remaining non-standardized names. This added a standard to another 62,338 surnames (quality level 3), which raised the coverage of surnames to 79,0% and the token coverage to 99.1%. Table 8 summarizes the results of the subsequent steps and their assigned quality levels.

Table 8. Summary of counts for quality levels of surname standards. The NAMES corpus consists of 562,676 different singular surnames with a token frequency of 53.839.590 (over all positions in a surname).

quality level	surnames	cum %	tokens	cum %	standards
2 (expert)	119,904	21.3	26,114,275	48.5	10,162
2 (semi-phon)	74,703	34.6	20,757,824	87.1	10,162
3 (edit dist = 1)	161,567	63.3	2,903,765	92.4	10,162
2 (extra standards)	25,930	67,9	3,354,706	98.7	19,016
3 (edit dist = 1 phase 2)	62,338	79.0	220,615	99,1	19,016
none	118,334	100.0	488,505	100,0	

V.2.3.1 Relations between standards of surnames

In analogy with given names there are also relations between standards of surnames. These can be derived from the surname variant pairs and result in 47,111 relations between 19,016 standards, which is a relatively very low number (2 to 3 relations per standard).

For evaluation purpose we again included three tests: a) standard pair identity (*ok-id*), b) fraction > 5% (in one of the two directions of the pair) (*ok-ratio*), and c) whether one standard name is part of the other, to accommodate abbreviations (*ok-suf*), for instance AKKER and AKKERHUIS (note that the name of standards is essential to allow for this rule). When applying the test criteria, requiring non-identity, and acceptance only when the token frequency for a standard combination was higher than 4, only 3,074 (bi-directional) relations remain. Such a set could be used to further merge standards. Especially the required token frequency is sensitive: the higher this frequency the more likely a merger, the lower this frequency the more combinations are found which should not be merged. This is shown in table 9, which gives related standards and the number of surname variant pair tokens seen with this combination. The higher the number of these tokens, the more likely there is a true relationship between standards. But several low frequency combinations seem likely as well (for instance LINNEKAMP and LINNENKAMP, LINDENHOLS and LINDENHOUT). It is hard to define rules or thresholds which result in a valid distinction .

A way out (to be investigated) is to derive statistical information from proven variant surname pairs and to use it to underpin merging of standards, rather than to proceed with extensive expert decisions, which have a great risk not to be consistent.

Table 9. Examples of related standards with LIN* derived from surname variant pairs, and the token frequency of the latter (>1)

SN1_STANDARD	SN2_STANDARD	variant pair tokens
LINDEN	LINGEN	36
LINGEN	LANG	19
LINDEN	LONDEN	10
LINDEN	LANDE	8
LINSSEN	LEUSEN	7
LINCKENS	LINTJENS	5
LINTJENS	LENT	5
LIN	LEUNISSEN	4
LINDEN	LIJNEN	4
LINDEN	LINK	4
LINSSEN	LENS	4
LINTERMAN	LAKEMAN	4
LIN	LINDEN	3
LINSKENS	LINSSEN	3
LINTJENS	LITJENS	3
LIN	LIENOS	2
LINDEBOOM	LINDEN	2
LINDEN	LINGTON	2
LINDENHOLS	LINDENHOUT	2
LINGBEEK	LIMBEEK	2
LINGSMA	LEIJDSMAN	2
LINKENHOF	LINKERHOF	2
LINKER	LINKVIS	2
LINNEKAMP	LINNENKAMP	2
LINSSEN	LESSCHEN	2

To further demonstrate the variation encountered in surnames, table 10 gives all surnames that formed a proven pair with the surname LINDEN (i.e. both names were used for the same individual), and the frequency of the pair (from the table names_sns_pairs). It shows deletion of the final n (LINDE), misreading of vowels (LONDEN), likely typos (LINDENF, LINDERN, LINDNE, LOINDEN), or additional second word (LINDENBOOM).

Table 10. Surnames which are found (in the LINKS project) as true variant of the surname LINDEN, and their frequency of occurrence.

LINDE	552	LIJNDEN	3	LINT	2	LINDNE	1
LINGEN	35	LINEN	3	LEENDEN	1	LINDNEN	1
LIENDEN	11	LEIDEN	2	LEINDEN	1	LINDOM	1
LENDEN	10	LENDE	2	LIDEN	1	LINDSEN	1
LINDON	10	LIN	2	LIENDE	1	LINSEN	1
LONDEN	7	LINDENF	2	LINDAM	1	LLINDEN	1
LIDNEN	4	LINDER	2	LINDEBOOM	1	LNDEN	1
LINDERS	4	LINDERN	2	LINDEN.	1	LOINDEN	1
LANDEN	3	LINDT	2	LINDENM	1		

Equally interesting are all surnames which have got the standard LINDEN (table 11). Surnames with attached prefixes are also included under this standard (such as VERLINDEN, UITTERLINDEN) – which was based on an expert decision. Note that also surname (variants) of LISDONK show, which arguably could have got an own standard with a relation between the standards LINDEN and LISDONK.

Table 11. Surnames with the standard LINDEN, their token frequency in genlias11 and the quality level of the standard assignment. .

Surname	tokens genlias11	quality level			
LINDEN	78755	2	UIJTERLINDEN	64	2
LINDE	24647	2	LENDEN	63	2
LINT	5280	2	LEENDEN	60	2
VERLINDEN	3795	2	LIEMDE	58	2
LIENDEN	3269	2	LIEMPDE	56	3
LENDE	1189	2	LINOT	52	3
VERLINDE	1042	2	UIJTTERLINDE	50	2
LINDT	957	2	UIJTTERLINDEN	45	2
LIND	715	2	UITERLINDE	43	2
UITTERLINDEN	601	2	UITTERLINDE	42	2
LISDONK	582	2	LINNEKEN	40	3
LIJNDEN	511	2	LYNDEN	40	2
LINDONK	360	2	VERLIENDEN	29	2
UIJTERLINDE	306	2	LINDEIJER	28	3
LINDON	241	2	LEIJSDONK	26	2
LINDAUER	209	3	LINDOLL	26	3
VERLANGEN	206	3	VERWINDEN	24	3
LINDERN	197	2	LIMDE	22	2
LIENDE	189	2	LINDOR	22	3
LIESDONK	172	2	LIJSDONCK	20	2
LINDO	151	3	LEIJSDONCK	17	2
UITERLINDEN	148	2	LIENDO	17	3
TERLINDEN	142	2	LISDONCK	16	2
LINDER	127	2	LEENDE	15	2
LENDI	95	3	LIENDER	13	2
LINDOORN	83	3	LIJNDE	12	2
LINDA	81	3	LEINDEN	11	2
LINDOM	81	2	LINDESAY	11	3
LIJSDONK	68	2	LINKTON	11	3
			LINDORFF	10	3

TERLINDE	10	2	LEUSDONK	6	2
LINTH	9	2	LIGNON	6	3
VERLIJNDEN	9	2	LINDAU	6	3
LEISDONK	8	2	LINTON	6	3
LINKEN	8	2	LIJNDONK	5	2
LINON	8	3	LIJNOHR	5	3
LYANDER	8	3	LINCKEN	5	2
LIENT	7	2	LINDONCK	5	2
LYSDONK	7	2			

VI The NAMES CORPUS release 1.1, data structure

The data are provided in the following tables (available as MySQL 5.5, MySQL 8.0 dump, and tab-separated files (UTF8, enclosed by " ", with CR/LF)):

NAMES_GN_1	singular given names (189,706)
NAMES_GN_PAIRS_1	variant pairs of given names (120,104)
NAMES_GN_STANDARD_1	standards of given names (813)
NAMES_GN_STANDARD_RELATIONS_1	relations between standards of given names (3,655)
NAMES_SNF_1	full surnames (673,784)
NAMES_SNS	singular surnames (562,676)
NAMES_SNS_PAIRS	variant pairs of singular surnames (328,410)
NAMES_SNS_STANDARD	standards of singular surnames (19,016)
NAMES_SNS_STANDARD_RELATIONS	relations between standards of singular surnames (47,111)
NAMES_SNF_PREFIX	full surname prefixes (1,536)

For convenience sake, the two RELATIONS files are also provide in LIMITED form, in which a threshold is set on the records according to the rule: TEST<>'' AND TEST<>'' AND (SNS OR GN)_RELATION_TOKENS>4 AND (SN OR GN)1_STANDARD<>(SN OR GN)2_STANDARD.

NAMES_GN_STANDARD_RELATIONS_LIMITED_1	(606)
NAMES_SNS_STANDARD_RELATIONS_LIMITED_1	(3,074)

To convert the MySQL 8.0 dump into a MySQL 5.5 dump the following changes should be edited, for all occurences:

COLLATE=utf8mb4_0900_ai_ci	in	COLLATE=utf8_unicode_ci
utf8mb4	in	utf8

The order of fields in the tables deviates from the order listed in the sections VI.1 and VI.2, which is no problem for the MySQL tables. The tab-separated files have a header record which describes the subsequent fields.

In most tables more fields than necessary are given for easier interpretation of the information. These extra fields can be derived from other fields in the same or other tables. They are indicated as '*may be omitted*'.

The tables of name variant pairs are given in a bi-directional way, i.e. variant1 – variant2 is given, but variant2 – variant1 as well. They can be easily de-duplicated by filtering on id1<id2. The chosen presentation has the advantage that by ordering on variant1 immediately all versions of the related variant2 are seen.

privacy

Because of reasons of privacy all frequencies between 1 and 5 from data from the BRP are mapped to the value of 4, which is an internationally used threshold.

VI.1 Given names

Table which gives all given names in the NAMES corpus, their gender and standard id, quality level of standard assignment, and various counts.
189,706 records out of which 168,513 with a standard.

NAMES_GN_1 **table of given names**

fields:

gn_id	identifier of the given name
gn	given name (singular, upper case)
gn_gender	{F (female), M (male)}
gn_pho	semi-phonetic transcription of given name
gn_standard	given name standard, *may be omitted (can be derived from names_gn_standard)*
gn_standard_id	identifier of the given name standard
gn_standard_quality	quality level of standard assignment {1,3,4}
gn_tokens_genlias11	sum of gn_tokens_genlias11_1 and gn_tokens_genlias11_m, *may be omitted*
gn_tokens_genlias11_1	number of name bearers in genlias2011 corpus, initial name
gn_tokens_genlias11_m	number of name bearers in genlias2011 corpus, middle name
gn_tokens_genlias11_e	number of name bearers in genlias2011 corpus, end name, if middle name exist
gn_tokens_brp14	sum of gn_tokens_brp14_1 and gn_tokens_brp14_m, *may be omitted*
gn_n_brp14_1	number of name bearers in brp2014 corpus, initial name
gn_n_brp14_m	number of name bearers in brp2014 corpus, middle name
gn_n_brp14_e	number of name bearers in brp2014 corpus, end name, if middle name exist

NAMES_GN_STANDARD_1 **table of standards of given names**

Table of the given name standards, with identifier and counts. 813 records.

fields:

gn_standard_id	identifier of the standard
gn_standard	standard name (lower case)
gn_standard_hf	standard name (uppercase) as name with highest token frequency
gn_standard_syll1	first syllable in standard
gn_standard_syll2	second syllable in standard
may be omitted (can be derived from names_gn):	
gn_standard_types_f	number of female names with standard
gn_standard_types_m	number of male names with standard
gn_standard_tokens_genlias11_f	number of female bearers of standard in genlias2011
gn_standard_tokens_genlias11_m	number of male bearers of standard in genlias2011
gn_standard_tokens_brp14_f	number of female bearers of standard in brp2014 corpus
gn_standard_tokens_brp14_m	number of male bearers of standard in brp2014 corpus

NAMES_GN_PAIRS_1 **table of variant pairs of given names**

Table which gives the identifiers of pairs of given name variants (bi-directional).
120,104 records.

fields:

gnp_id	identifier of the variant pair {vpv-00001}
gnp_tokens	number of variant pair tokens derived from genlias2011
gnp_gender	{F(female),M(male)}
gn1_id	identifier of name 1
gn2_id	identifier of name 2
may be omitted (can be derived from names_gn):	
gn1	given name1 (singular, upper case) , gender = gnp_gender
gn2	given name2 (singular, upper case), gender = gnp_gender
gn1_standard	given name1 standard
gn2_standard	given name2 standard
gn1_standard_id	identifier of the standard of name 1
gn2_standard_id	identifier of the standard of name 2
gnp_standard_INT	standard assigned by INT

NAMES_GN_STANDARD_RELATIONS_1 **table of relations between standards of given names**

Table which summarizes the major relations between standards of given names (bi-directional).
3,655 records, including the identity standards (incomplete because for some standard no variants pairs have been seen at all).

fields:

gn1_standard_id	identifier of standard1
gn2_standard_id	identifier of standard2
may be omitted (can be derived):	
gn1_standard	standard1
gn2_standard	standard2
gn_relation_types	number of name pairs with this combination of standards
gn_relation_tokens	number of name pair tokens with this combination of standards COULD BE USED TO REDUCE THE NUMBER OF RELATIONS, FOR INSTANCE GNS_RELATION_TOKENS>4, IN COMBINATION WITH THE FIELD <TEST>
ratio	percentage of name pair tokens in this combination relative to all pair tokens with gn1_standard
test	evaluation indication: OK-ID IDENTICAL STANDARDS OK-RATIO RATIO>5% OK_SUF ONE STANDARD IS PART OF THE OTHER LIKE *JOHANNES - JO*

NAMES_GN_STANDARD_RELATIONS_LIMITED_1 **table with limited number of relations between standards of given names**

Table (bidirectional), derived from NAMES_GN_STANDARD_RELATIONS_1 , limited by the rule: TEST<>'' AND GN_RELATION_TOKENS>4 AND GN1_STANDARD<>GN2_STANDARD . Records that do not fulfil this condition likely concern a spurious relation.
606 records.

fields:

gn1_standard_id	identifier of standard1
gn2_standard_id	identifier of standard2
may be omitted (can be derived):	
gn1_standard	standard1
gn2_standard	standard2
gn_relation_types	number of name pairs with this combination of standards
gn_relation_tokens	number of name pair tokens with this combination of standards

VI.2 Surnames

NAMES_SNF_1: table of full surnames

A service table of full surnames (including prefix) from which the singular surname and prefix tables are derived. 673,784 records. The prefixes have been cleaned from erroneous content (see the NAMES_SNF_PREFIX table) but the surname itself is original. Problematic content can be found from snf-672,799 onwards. For patronymic surnames the estimated underlying given name is added, but not thoroughly analysed and likely incomplete.

As expected, there are only minor differences in the top-10 surnames in genlias11 and brp07:

rank	genlias11	brp07
1	de Jong	de Jong
2	de Vries	Jansen
3	Jansen	de Vries
4	Janssen	van den Berg
5	van Dijk	van Dijk
6	Bakker	Bakker
7	van den Berg	Janssen
8	Visser	Visser
9	Smit	Smit
10	de Boer	Meijer

fields:

snf_id	identifier of the full surname {snf-000001}
snf	full surname (both cases)
pf_id	identifier of the surname prefix {pf-0000001}
pf	prefix of the surname, *may be omitted*
snf_pf_tokens_genlias11	number of bearers of the full surname in genlias2011/wiewaswie corpus
snf_pf_tokens_brp07	number of bearers of the full surname in brp2007 corpus
snf_patronym	in case of patronymic surname, the possible given name

NAMES_SNS: table of singular surnames

A table with singular surnames (no prefix), and their identifier. 562,676 records.

sn_id	identifier of the singular surname {sn-0000001}
sn	singular surname (upper case)
sn_pho	semi-phonetic transcription of singular surname
sn_standard_id	identifier of the surname standard {sns-000001}
sn_standard	name of the surname standard, *may be omitted*
sn_tokens_genlias11	number of name bearers in genlias2011/wiewaswie corpus
sn_tokens_brp07	number of name bearers in brp2007 corpus
sn_standard_quality	quality level of standard assignment {2,3,4}

NAMES_SNS_STANDARDS: **table of standards of singular surnames**

Table that gives the singular surname standards, and their identifier. 19,016 records.

fields:

sn_standard_id	identifier of the singular surname standard {sns-000001}
sn_standard	standard name (lower case)
may be omitted:	
sn_standard_types	number of singular surnames with standard
sn_standard_tokens_genlias11	number of singular surname element bearers with standard in genlias2011/wiewaswie corpus
sn_standard_tokens_brp07	number of singular surname standard bearers in brp2007 corpus

NAMES_SNS_PAIRS: **table of variant pairs of singular surnames**

A table that gives the surname identifiers of variant pairs (bi-directional). 328,410 records (164.205 uni-directional pairs). There are 3,924 records of which one or both surnames do not show in the NAMES_SNS table. They contain a special character (like ‘ ? ,) and are dropped in an earlier phase of the analysis. It was decided not to drop them in this table because they yield information (on the use of the special characters).

fields:

snp_id	identifier of the variant pair {snp-000001}
snp_tokens	number of variant pair tokens derived from genlias2011/wiewaswie
sn1_id	identifier of the surname 1 {sn-0000001}
sn2_id	identifier of the surname 2 {sn-0000001}
sn1_tokens_pairs	number of times surname1 showed in any variant pair
sn2_tokens_pairs	number of times surname2 showed in any variant pair
n_names_linkage	highest number of additional names in record linkage to decide for a variant pair (minimum = 4); the higher this number, the higher the probability of a true link
may be omitted:	
sn1	surname1 (singular, upper case)
sn2	surname2 (singular, upper case)
sn1_standard	surname1 standard
sn2_standard	surname2 standard
sn1_standard_id	identifier of the standard of surname 1 {sns-000001}
sn2_standard_id	identifier of the standard of surname 2 {sns-000001}
sn1_tokens_genlias	number of tokens of surname1 in genlias11/wiewaswie
sn2_tokens_genlias	number of tokens of surname2 in genlias11/wiewaswie

NAMES_SNS_STANDARD_RELATIONS table of relations between standards of given names

Table which summarizes the major relations between standards of given names (bi-directional). 47,106 records, including the identity standards (incomplete because for some standard no variants pairs have been seen at all).

fields:

sn1_standard_id	identifier of standard1 {sns-000001}
sn2_standard_id	identifier of standard2 {sns-000001}
may be omitted (can be derived):	
sn1_standard	standard1
sn2_standard	standard2
sns_relation_types	number of name pairs with this combination of standards
sns_relation_tokens	number of name pair tokens with this combination of standards COULD BE USED TO REDUCE THE NUMBER OF RELATIONS, FOR INSTANCE SNS_RELATION_TOKENS>4, IN COMBINATION WITH THE FIELD <TEST>
ratio	percentage of name pair tokens in this combination relative to all pair tokens with sn1_standard
test	evaluation indication: OK-ID IDENTICAL STANDARDS OK-RATIO RATIO>5% OK_SUF ONE STANDARD IS PART OF THE OTHER LIKE *AKKERHUIS - AKKER*

NAMES_SNS_STANDARD_RELATIONS_LIMITED_1 table with limited number of relations between standards of given names

Table (bidirectional), derived from NAMES_GN_STANDARD_RELATIONS_1 , limited by the rule: TEST<>'' AND SNS_RELATION_TOKENS>4 AND SN1_STANDARD<>SN2_STANDARD . Records that do not fulfil this condition likely concern a spurious relation. 3,074 records.

fields:

sns1_standard_id	identifier of standard1
sns2_standard_id	identifier of standard2
may be omitted (can be derived):	
sn1_standard	standard1
sn2_standard	standard2
sns_relation_types	number of name pairs with this combination of standards
sns_relation_tokens	number of name pair tokens with this combination of standards

NAMES_ SNF_PREFIX: **table of surname prefixes**

A service table which contains all surname prefixes in original version (not cleaned), with standard, counts in genlias11 and brp07, and codes for different classes. 1,536 records. Tokens counts are computed on the basis of *upper case*. In genlias11 1.056 different prefixes were found, in brp07 only 169. 454 prefixes were cleaned in an earlier phase of the analysis of genlias11, and have counts 0; they are included here as examples of erroneous field entries.

The interpretation and count (different types) per code is as follows:

pf_code	**standards**	**count**	**meaning**	**examples**
pf	255	1252	prefix	*van*
nt	17	88	noble title	*Baron Gravin*
ot	8	17	ordinary title	*Vrouwe Meester*
i	25	58	initial	*A.*
al	66	66	alias	*Genaamd*
r	5	20	relation	*kind veuve*
e		41	error	*… ; \ < NN achternaam*

fields:

pf_id	identifier of the prefix {pf-0000001}
pf	prefix of surname (both cases)
pf_uc	prefix of surname (upper case), *may be omitted*
pf_standard	standard of prefix
pf_code	code of class of the prefix
pf_tokens_genlias11	number of (upper) prefix bearers in genlias2011/wiewaswie corpus
pf_tokens_brp07	number of (upper) prefix bearers in brp2007 corpus

Versions and updates

version 1.0	January 5, 2019	
version 1.1	March 18, 2019	• names_gn_1: deduplication of gn_id for different gn_gender, 2360 records • names_gn_1: change of gn_standard_id for 169 records, with consequences in all _gn tables with respect to counts • names_gn_pairs_1: removal of 2426 records with unequal gender for the names in the pair • names_gn_pairs_1: removal of gn1_gender and gn2_gender and addition of field gnp_gender • names_gn_pairs_1: addition of field gnp_standard_INT with standard assignment by INT from 2018-07-02 • all gn tables got the version extention _1
	May 13,2019	• names_snf_1 created without fields snf_tokens_genlias11 en snf_tokens_brp07 which proved to be inaccurate, while correct versions can be found in names_sns.
	May 27, 2019	• names_gn_standard_relations_limited: created as more realistic selection of relations betweens given names. Does not include fields RATIO and TEST. • names_sns_standard_relations_limited: created as more realistic selection of relations betweens family names. Does not include fields RATIO and TEST.

The NAMES corpus can be obtained (in .tsv format or as MySQL database) by a request to Gerrit Bloothooft (g.bloothooft@uu.nl).

VII The Names lexicon in the INT lexicon service

A selection of the NAMES corpus has been made available in the INT lexicon service. The INT LexiconService is a webservice that gives any piece of software quick online access to a lexicon by means of http requests. The LexiconService was originally designed to access a computational lexicon with an Impact Lexicon Database structure.[1] The service is now deployed at INT and gives access to INT's GiGaNT lexicon of 13th to 20th century Dutch. The lexical information provided by the webservice can be given in both XML or JSON format.

To enable users to use the Names Lexicon for query expansion, we have updated the lexicon service with the given names and surnames of this project. It is documented in appendix 2 of this document.

In the project *CLARIAH chaining search*, a Names sandbox has been created, https://github.com/INL/chaining-search/blob/master/namesSandbox.ipynb.

Please inform Katrien.depuydt@ivdnt.org when you want to use the lexicon service.

VIII RDF conversion and access through ANASI

Still to be documented.

[1] The current Dutch lexicon consists of a lemmata table with grammatical categories (part of speech), a wordforms table, an attestations table, a documents table with metadata (source information and date).

Appendix 1

Building a search facility with the NAMES corpus 1.1

The NAMES corpus can be used for fuzzy search of name variants. The steps to be considered are:

1 It should be asked explicitly whether the search concerns a given name, a surname, or undecided.
2 The user enters a name (if desired a quality level ; for a given name a gender may be entered as well).
3 Depending on the answers under 1) and 2) the name should be looked up in names_gn_1, names_sns, or both.
4 If the name is not found, the search ends. If found, the corresponding gn_standard_id and/or sn_standard is used.
5 In names_gn_standard_1 (and/or names_sns_standard) all names with the same standard_id are selected (and their gender, and their quality level, and possibly the token frequencies in genlias11 and/or brp14).
6 In names_gn_standard_relations_1 (and/or names_sns_standard_relations) all related standards are selected (*use rules which require <test> is not empty, and a threshold for <gn_tokens_genlias11> (or <sns_tokens_genlias11>), for instance >4*).
7 All names with a related standard are selected in names_gn_1 and/or names_sns (and their quality level, their gender, and possibly the token frequency in genlias11 and/or brp14). In the output it should be indicated that this set has a more indirect relationship to the original name.

For the understanding of the status of a variant it is helpful when the name of the standard is provided with each selected name as well. This could be the NAMES standard, <gn_standard> or <sn_standard>, in all cases.

For given names, in addition <gn_standard_hf> (the name with the highest frequency in genlias11 with that standard, from table names_gn_standard_1)or <gnp_standard_INT> (the standard associated to a name variant pair by INT, if available, from table names_gn_pairs_1) could be provided.

To facilitate the implementation of the rule mentioned under 6), the result of the rule has been made available in the tables: NAMES_GN_STANDARD_RELATIONS_LIMITED_1, NAMES_SNS_STANDARD_RELATIONS_LIMITED_1

Appendix 2

Introduction to the Clariah Names lexicon implementation in the INT LexiconService™

The **Clariah Names lexicon** is loaded into the **INT LexiconService**. In that way, the Names lexicon can be accessed by any computer. Access is provided through so-called http requests carrying queries to the lexicon.[2]

The structure of the Names lexicon can be shortly described in the following terms. The lexicon consists of two distinct sets: given names and surnames. Each set contains a large number of names variants. All variants that represent the same standard name are clustered around that standard name (and share the same standard name ID). The quality – or reliability – of this link between variant and standard name is coded in a so-called quality level (level 1-4 from high to low). Finally, the Names lexicon tells if different standard names are connected (so-called standard relations).

The information here above is available in the INT LexiconService implementation too. Of course, the INT LexiconService has a generic architecture allowing support for different kinds of lexica. As a consequence, the INT LexiconService does not refer to 'standard names', but to 'lemmata'. The same way, a set of 'names variants', which forms the paradigm of a (name) lemma, is called a set of 'wordforms' or a paradigm. For the same straightforward reasons, the INT LexiconService stores the names' genders in the more generic field 'parts-of-speech' as a part-of-speech is normally used to describe the distinctive features of some word. Finally, the INT LexiconService describes the Names lexicon quality levels in the – once again – more generic 'provenance' field. Since the quality levels correspond to different procedures for establishing the connection between a variant and a standard name (like automatic link versus expert review, etc.), 'provenance' seems the legitimate choice.

This is summed up in the following table:

Names lexicon labels	Generic LexiconService labels
Standard name	Lemma
Name variant	Wordform
Name gender	Part-of-speech
Standard name relations	Lemmata relations
Surname vs given name	Dataset
Level of quality	Provenance

The way the Names lexicon data can be accessed in the INT LexiconService is extensively described in the INT LexiconService Manual (separatedly available), but some queries are provided here as a 'Quick start tutorial'.

[2] That is: a URL (like the address of any website) followed by a few parameters, telling what you know (eg. a word form) and what you want (e.g. expand to connected word forms).

First of all, the INT LexiconService must be accessed at the following URL:

http://sk.taalbanknederlands.inl.nl/LexiconService/?database=nameslex

As a first test, enter a name in the field First name/Last name (ignore other fields) which will generate **all the variants** of the test name (in JSON or XML output). In terms of the INT LexiconService, this amounts to asking for all word forms of the lemma of the test name. Most lemma's are uniquely related to given names or surnames, but in case insensitive format 119 lemma's are identical for given names and surnames.

This query can also directly implemented by the URL below, which for example for the lemma 'jansen' generates a list of 658 surname variants:

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**get_wordforms_from_lemma**?database=nameslex&**lemma=jansen**&case_sensitive=false

Three things must be noted here:

- The query specifies the lexicon name (nameslex), which is needed to tell the LexiconService which lexicon to search.
- The query has a 'case_sensitive' parameter set to 'false'. This is needed since the LexiconService is case sensitive by default, while the data isn't in this particular case.
- For the NAMES corpus, the LexiconService does not provide the lemma, but only uses its ID.

If you don't know what the standard name (lemma) is, as you only have a name variant at your disposal, the latter can be used as input too. Just ask the LexiconService to give you the lemma_id corresponding to a specific word form (in this case 'jansen' is now used as such, not as a lemma) :

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**get_lemma_from_wordform**?database=nameslex&**wordform=jansen**&case_sensitive=false

This yields the standard name ID of the corresponding lemmata (in this case ID=8317 for surnames from the table names_sns, and ID=443 for given names from the table names_gn) in return.

These IDs can be used to get their variants, for example for ID=443:

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**get_wordforms_from_lemma_id**?database=nameslex&**lemma_id=443**

which will give all given name variants of the wordform 'jansen'.

Next, we'll see how to get **related standard names (lemmata)**. For this, we send the lemma ID and ask for related lemmata (we use ID=8317 for the lemma of wordform 'jansen'):

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**get_related_lemmata**?database=nameslex&**lemma_id=8317**

The response will contain the 3 standard name IDs of related lemmata (8277, 8318, 8363) (from the table with a limited set of relations). These IDs can be used to get their variants too, for example for ID=8277:

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**get_wordforms_from_lemma_id**?database=nameslex&**lemma_id=8277**

Next, we'll see how to get **related standard names**. This has to be done in two steps. We first need to get the standard name ID of the standard name. Once again, in the LexiconService, this amounts to asking for the lemma_id of a lemma:

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**get_lemma_id_from_lemma**?database=nameslex&**lemma=jansen**&case_sensitive=false

Next, you might be interested in **expanding a name variant to all known variants of the same standard**. This amounts to asking the LexiconService to give the full paradigm (all word forms of both surnames and given names):

http://sk.taalbanknederlands.inl.nl/LexiconService/lexicon/**expand**?database=nameslex&**wordform=jansen**&case_sensitive=false

If you want to **limit your search** to a given dataset, a given gender or a given quality level, you'll just need to add the corresponding parameters (with the required value) to your query.

For example, to search the surnames only (and exclude the given names from the search), just add the 'dataset' parameter to your query:

…&**dataset=names_sns**…

Limiting the search to given names can be done with another value:

…&**dataset=names_gn**…

If you wish to limit your search to some gender, add the 'pos' (part-of-speech) parameter:

…&**pos=F**… or …&**pos=M**…

Finally, quality levels can be addressed by setting the 'paradigm_provenance' parameter (describing the quality of the link between a particular element of the paradigm [a variant] and its standard).

For example, adding the following will limit the search to variants with a 'level 3'-quality link to their standard name:

…&**paradigm_provenance=level_3**…

Further details regarding writing LexiconService queries are fully described in the INT LexiconService manual (separate document).