

WHITE PAPER: NAME GENDERIZING

Name genderizing is the process of identifying the gender based on a person's name.

Last update: 2017-05-30

A. Arn, G. Achim, E. Dubencu

Optimaize GmbH | Im Oberdorf 16 | CH-8602 Wangen Zürich | www.nameapi.org

TABLE OF CONTENTS

1.	NAME GENDERIZING	3
2.	THE NAMEAPI GENDERIZER SOFTWARE.....	4
2.1	UNDERSTANDING THE DATA	4
2.2	LARGE INTERNATIONAL DATABASES, CONTINUED MAINTENANCE.....	5
2.3	GENDER INFORMATION IN NAMES BY CULTURE	6
2.4	DIFFICULTIES	8
2.5	THE RETURN VALUE.....	8
3.	USE CASES.....	9
4.	FINAL CONCLUSIONS.....	9
5.	TRY IT.....	9
6.	NAMEAPI FROM OPTIMAIZE GMBH (LTD).....	10
6.1	NAME DATA.....	10
6.2	SOFTWARE	10

1. NAME GENDERIZING

Names contain various pieces of information. In many cultures, a person's name tells his or her gender.

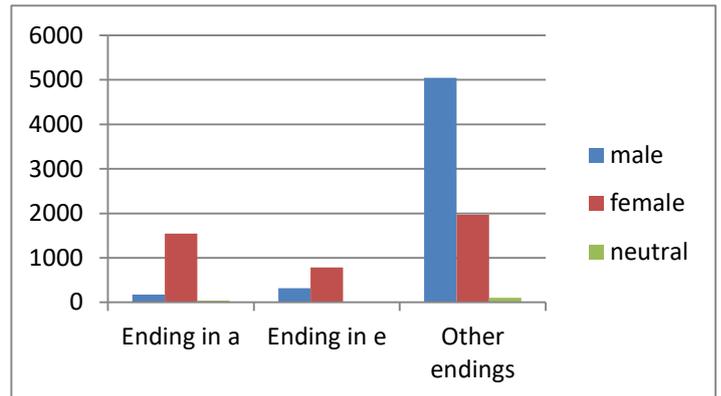
Simple! I just look at the last character!

A simple approach: names ending in -a and -e are female, the rest are male.

Example: Daniel => male, Daniela => female

From a sample set of 10k given names:

- Ending in a:
174 male, 1544 female, 36 neutral
- Ending in e:
321 male, 788 female, 20 neutral
- Ending in another character:
5046 male, 1969 female, 102 neutral



Conclusion: For some use cases it's "better than nothing". While having many correct guesses, it also causes many wrong results.

This approach is similar to the zero-cost weather forecast of predicting that tomorrow's weather is the same as today's - at a 75% success rate. The required effort to achieve better results grows exponentially from there.

Ok, then I can make a list of known names and look them up!

This approach works quite well for the names that are on the list.

It fails when:

- the name is not on the list
- the name on the list is marked as neutral
- the list contains errors

In a western country, and with a reasonably large list, this quickly achieves a success rate of 85% with little false results.

	A	B
1	NAME	GENDER
2	MARY	f
3	PATRICIA	f
4	LINDA	f
5	BARBARA	f
6	ELIZABETH	f
7	JENNIFER	f
8	MARIA	f
9	SUSAN	f
10	MARGARET	f
11	DOROTHY	f
12	LISA	f
13	NANCY	f
14	KAREN	f

What else can be done?

Increasing the success rate requires a better understanding of the name.

2. THE NAMEAPI GENDERIZER SOFTWARE

This is the approach taken by the genderizer module.

2.1 UNDERSTANDING THE DATA

To understand a sentence, one must first recognize the language it is written in. For names, this means indentifying the culture it is from, so that each part can be understood correctly.

Prof Dr Mary-Louise D. Miller Jr => *English*
Abd al-Masih Al-Azawi => *Arabic*
Dong Hua Lee => *Chinese*

The more information available, the better:

- The whole name, not just the given name
- Context: where the application is used (country)
- Nationality, correspondence language of the person
- Age, birth year

Parsing the input is advised even if only the given name is available (titles, multiple first names, ...).

Now that the culture has been detected, a specialized genderizer comes into play.

2.2 LARGE INTERNATIONAL DATABASES, CONTINUED MAINTENANCE

Optimaize maintains dictionaries with all parts of names, especially first and last names, with attributes such as gender and language.

Statistical sources are mainly census publications, birth lists, and telephone books.

Given names

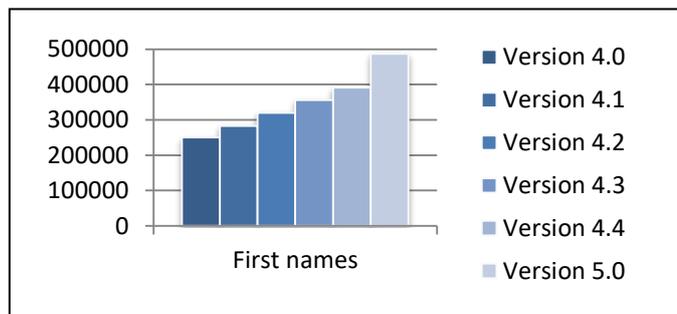


Number of given names by culture, illustrated as tag cloud (June 2015)

English

The dictionaries grow from version to version.

However: more names, more problems.
Larger databases does not automatically mean better results.



2.3 GENDER INFORMATION IN NAMES BY CULTURE

Western cultures

Given names are mostly gender-specific, with only a small percentage of names being gender-neutral. Given names can be combined, abbreviations are common, and it's not rare to find titles and qualifiers aligned with them.

Middle names are mostly just additional given names. If the first given name is gender neutral, the second may provide clarity. An exception is the USA, where the middle name is again chooseable freely, for example the family name from the mother's lineage as in (George *Walker* Bush).

Surnames are gender-agnostic.

Slavic

There are very few unisex given names. In many slavic countries, the surname has a gender-specific ending (*). Examples: Mikhail Gorbachev, Raisa Gorbachova.

In some countries the middle names are patronyms, that means, names formed from the father's first name with a gender specific suffix.

Icelandic

Given names are gender-specific. Surnames are patronyms and contain gender information (*).

Turkish

The majority of the Turkish given names are either clearly assigned to one gender, or have few exceptions only. About 10% of the names are gender neutral with a tendency towards one side. Names with meanings of beauty or natural elements like "Çiçek" (flower) or "İlknur" (first light) are rather female, whereas names with meanings of desired virtues like strength and courage, for example "Erdoğan" (robust man) or "Metin" (strong), are rather male.

The surname is gender agnostic.

Arabic

Almost all given names are gender-specific. The difficulty is that there is a magnitude of transcription variants to Latin. For example, the name Muammar can be rendered in at least 24 ways.

In the traditional Arabic naming system, there are some name elements that show the gender (*). Colonized countries switched to the western naming system.

Abd al-Masih bin Al-Azawi => bin = "son of" => male

Indian

India has as many naming systems as it has cultures. Many given names are unisex, or male and female forms are spelled the same in the Latin script. For Sikh people, the middle name provides information about the gender. Example: Gurpreet Singh Dhaliwal => male, Gurpreet Kaur Dhaliwal => female.

Chinese

For Chinese, there is no gender detection, there is only gender guessing. Surnames are gender-agnostic. Given names have an inclination towards a certain gender, which can be seen in statistics. In Chinese, anything goes for a child's name. Names are usually made up of two syllables (two Chinese characters). It is not feasible to create a definitive list of Chinese given names, because any combination of the 100k characters is possible.

Names related to beauty, gentleness, or plants are usually female, while names related to strength, courage or the country are usually male. Example: 健雄 Chien-Shiung "*courageous hero*".

Vietnamese

Surnames are gender-agnostic. The first and middle names that carry notions of strength and bravery, such as "Mạnh" = "strong", "Quyết" = "willful" are male, whereas female ones have a softer meaning like "Thuỳ" = "gentle", "Thu" = "autumn". Older generations customarily had the middle name "Thị" for women and "Văn" for men. In official documents, the middle name is usually combined with the surname, if there is no dedicated space for it.

(*) The gender information from inherited name elements (surnames) is handled with great care. It can only reliably be used in their respective countries. Expatriats sooner or later freeze the grammatical surname forms, passing it on to future generations. For example, the Swedish surname Svenson has lost its original meaning, and a person in the USA with that surname doesn't have to be male.

2.4 DIFFICULTIES

There are three reasons why a writing form of a given name can't be clearly assigned to one gender:

1. True unisex names, for example "Casey". This includes many short forms.
2. Names that exist in multiple cultures, for example "Andrea".
3. Detail is lost through transcription, rendering distinctive names the same.

In all these cases statistics help: name statistics (census, birth lists, phone books) are used to calculate the chances.

In the second case, identifying the culture and respecting the context is key:

Andrea Bocelli => Italian => likely male

Andrea Berg => German => likely female

2.5 THE RETURN VALUE

The *genderizer* software returns the following values:

Gender Result	MALE FEMALE NEUTRAL UNKNOWN (the system cannot tell)
Male Chance	If the Gender result is NEUTRAL then this is the percentage of MALE results. A decimal from 0-1, where the remaining percentage is for the FEMALE results.
Result Confidence	0-1 where 1 is the best.

3. USE CASES

When do you need genderizing?

1. Addressing a person

You don't know the person's gender yet, but would like to address him or her more personally. It just sounds better with a correct salutation than "Hi there!". However, if in doubt, we recommend to use a neutral salutation.

2. Customer segmentation

The gender can be a key criteria in selecting the right target group.

3. Input validation

You know the gender and the name. Does it match? Warning the back office employee or the customer at data entry can prevent the infiltration of wrong information right from the start.

4. Name matching, deduplication

You compare two records to find out if they could be the same. A gender mismatch is a strong indicator.

4. FINAL CONCLUSIONS

Which approach should I take?

This depends on your use case:

- How much accuracy do you need?
- Can you afford mistakes?

Can you compile a complete, definitive list with all names?

The short answer is no. The long answer also. But read on. Some countries maintain lists and rules of permitted given names to either protect the child, or to keep the local traditions. Others are completely open and allow new creations as given names.

Another reason involves the names in cultures that don't use the Latin script. Because of many different transcription rules, freestyle rules, target languages, and source language dialects, often there is a magnitude of possible writing forms.

Sometimes existing names change in how they are used. If names are passed on to female offspring, in time a feminization of the name takes place. Examples: Sidney, Morgan, Kelly.

5. TRY IT

Browse to the live demo and see the name genderizer in action:

<http://www.nameapi.org/en/live-demos/name-genderizer/>

6. NAMEAPI FROM OPTIMAIZE GMBH (LTD)

Optimaize GmbH (*Optimaize*), founded in the year 2000 with its headquarters in Wangen-Zürich (Switzerland) and branch office in Cluj-Napoca (Romania) since 2006, manages the global leading name database and develops a software around persons' names. *Optimaize* operates a public portal on www.namepedia.org, and commercial offerings on www.nameapi.org. NameAPI serves international customers including software companies, financial institutions, direct marketers, and media companies.

6.1 NAME DATA

All parts of personal names are being collected, including given names and surnames from all languages and cultures. These are linked with additional information, including gender, language and frequency. Original spellings in non-Latin scripts (including transcriptions and transliterations) are also recorded.

The used sources are phone books, national government publications, websites on the subject, and local freelancers. NamePedia spiders online news sources from 55 countries and extracts named entities focusing on people's names.

6.2 SOFTWARE

On the basis of the name data, *Optimaize* develops the "NameAPI" software with the following modules:

Name Genderizer:	Identifies the gender of a person's name
Name Parser:	Identifies and orders the parts of names
Email Name Parser:	Extracts names out of email addresses
Name Matcher:	Compares names and computes similarity (duplicates)
Name Extractor:	Extracts names out of plain text
Name Profiler:	Enhances name records with cultural attributes
Name Validator:	Checks correctness (spell-checking, fake names and random typing, gender and name order)
Name Formatter:	Formats names in correct upper/lower case
Name Variant Generator:	Suggests variants and variant spellings for names