NAMEAPI
INTELLIGENCE IN NAMES

White Paper: Name Genderizing

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

Last update: 2015-03-26
A. Arn, G. Achim, E. Dubencu

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

Table of contents

1. NameAPI from Optimaize GmbH (ltd)

Optimaize GmbH (Optimaize), with its headquarters in Wangen-Zürich (Switzerland) and branch office in Cluj-Napoca (Romania), manages the global leading name database and develops a software around peoples' names. *Optimaize* operates a public portal on www.namepedia.org, and commercial offerings on www.nameapi.org. NameAPI serves customers, including large enterprises, CRM and other software companies, direct marketers, and media companies.

1.1 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.

1.2 Software

On the basis of the name data, *Optimaize* develops the software "NameAPI" 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

2. Name Genderizing

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

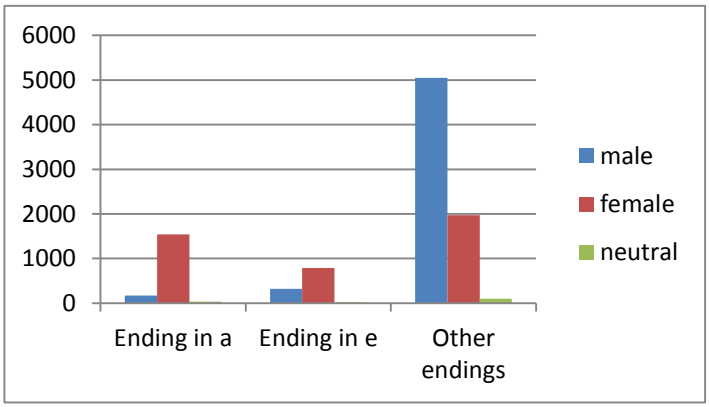
Simple! I just look at the last character!

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, in an expected writing form
- the name on the list is marked as neutral

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 succes rate requires a better understanding f the name.

3. The NameAPI Genderizer Software

This is the approach taken by the genderizer module.

3.1 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
- Nationality, correspondence language
- Age, birth year

Parsing the input is advised even if only the given name is available.

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

3.2 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 alined with them.

Middle names are mostly just additional given names. An exception is the USA, where the middle name may as well be a surname (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 (*). Example: Mikhail Gorbachev, Raisa Gorbachova

Icelandic

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

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. Given names are gender-neutral, as in Chinese. It's the middle name that reveals the gender. In official documents, the middle name is usually combined with the given name. Example: "Tan Dung".

(*) 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.

3.3 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) are used to calculate the chance for each gender.

In the second case, identifying the culture is the key:

Andrea Bocelli => Italian => likely male
Andrea Berg => German => likely female

3.4 The return value

The genderizer 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 percent from 0-100.
Result Confidence	0-100 where 100 is the best.

4. 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.

5. 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 any possibly made up words 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.

Besides new names being created, existing names can also change in how they are used. In western cultures, names can transform from male to unisex, or unisex to female. Examples: Sidney, Morgan, Kelly.

6. Try it

Browse to the live demo and see the name genderizer in action:
http://www.nameapi.org/en/live-demos/name-genderizer/