Genusidator
(genus + elucidator)

A Rule-Based System to Explain Grammatical Gender Assignment in German Nouns

Simon Zuberek
2023 IALLT Conference
June 15th, 2023
The Gender System in German

- Three grammatical genders (noun classes): **Masculine**, **Feminine**, **Neuter**
- Definite articles: *der, die, das*
- Indefinite articles: *ein, eine, ein*
- Declined by cases:

<table>
<thead>
<tr>
<th>Case</th>
<th>Masculine</th>
<th>Feminine</th>
<th>Neuter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative</td>
<td>der / ein</td>
<td>die / eine</td>
<td>das / ein</td>
</tr>
<tr>
<td>Genitive</td>
<td>des / eines</td>
<td>der / einer</td>
<td>des / eines</td>
</tr>
<tr>
<td>Dative</td>
<td>dem / einem</td>
<td>der / einer</td>
<td>dem / einem</td>
</tr>
<tr>
<td>Accusative</td>
<td>den / einen</td>
<td>die / eine</td>
<td>das / ein</td>
</tr>
</tbody>
</table>
Nouns constitute over 70% of the words in the German language.\(^1\)

Collectively, nouns and the corresponding articles are the most frequently-used words in the German language.\(^2\)
Acquisition of German Grammatical Gender

- **By the age of 2** children distinguish between grammatical gender, but prefer to use the indefinite (ein/eine) over the definite article (der/die/das).³
- **By the age of 5** the definite articles are left out in situations where the grammatical gender is not clear.⁴
- **By the age of 7** in tests using nonce nouns, children tend to assign the same gender to those nonce nouns that adults.⁵
- **By the age of 10** the acquisition of the noun gender is complete.⁶
The grammatical gender in German isn’t explicitly taught. Students are told to learn it by heart.

Native speakers of German and/or the majority of German language instructors were never taught the principles that determine gender.

German language instructors tend to believe that the grammatical gender assignment is arbitrary.
“Every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart. There is no other way.”

Twain, Mark. 1880. “The Awful German Language”, Appendix D in A Tramp Abroad, Chatto and Windus
I THINK WE CAN DO BETTER THAN THAT
The Rules behind the Gender Assignment

Ruleset 1: Semantic Categories
Nouns of similar categories of things or concepts tend to have the same gender.

Ruleset 2: Morphophonemic Categories
Nouns that have the same affixes tend to have the same gender.
Masculine

Ruleset 1: Semantic
- animals
- times of the day
- days of the week
- months
- seasons
- points on the compass
- precipitation and wind
- celestial bodies
- types of soil, minerals, and rock
- dirt and waste
- etc.

Ruleset 2: Morphophonemic
- **Suffixes:**
  - -aal
  - -ag
  - -al
  - -am
  - -an
  - etc.
- **Prefixes:**
  - Kn-
  - Schwa-
Feminine

Ruleset 1: Semantic
- numbers and mathematics
- time
- authority, power, governance
- rules, permissions, limits
- knowledge and wisdom
- communication
- musical instruments
- hollow shapes
- food
- gestures and motions
- etc.

Ruleset 2: Morphophonemic
- Suffixes:
  - -a
  - -acht
  - -ade
  - -age
  - -anz
  - etc.
Ruleset 1: Semantic
- higher-level categories
- letters of the alphabet
- languages
- grammatical terms and POS
- colors
- human and animal babies
- pieces and particles
- types of metals
- materials
- units of measurement
- etc.

Ruleset 2: Morphophonemic
- Suffixes:
  - -en
  - -ien
  - -land
  - -reich
  - -stan
  - etc.
- Prefixes:
  - Ge-
Noun + Article → Rules
The Pipeline

Preprocessing

- **User inputs the noun** (argparse)
- **Output the gender and lemmatize** (spaCy German transformer pipeline)
- **Parse compound words** (German Compound Splitter)
- **Translate to English** (deepL API)
- **Generate a taxonomy** of hypernym synsets going all the way to the root node of the semantic ontology graph (nltk and wordnet).
WordNet 3.0

- A hierarchically organized lexical database (a knowledge graph)
- A thesaurus + some aspects of a dictionary
Senses of ‘bass’ in WordNet 3.0

**Noun**

- **S**: (n) *bass* (the lowest part of the musical range)
- **S**: (n) *bass*, *bass part* (the lowest part in polyphonic music)
- **S**: (n) *bass*, *basso* (an adult male singer with the lowest voice)
- **S**: (n) *sea bass*, *bass* (the lean flesh of a saltwater fish of the family Serranidae)
- **S**: (n) *freshwater bass*, *bass* (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S**: (n) *bass*, *bass voice*, *basso* (the lowest adult male singing voice)
- **S**: (n) *bass* (the member with the lowest range of a family of musical instruments)
- **S**: (n) *bass* (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

**Adjective**

- **S**: (adj) *bass*, *deep* (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"
Hyponymy Hierarchy for ‘bass’

- **S:** (n) bass, *basso* (an adult male singer with the lowest voice)
  - *direct hyponym / inherited hyponym / sister term*
    - **S:** (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
      - **S:** (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
      - **S:** (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
      - **S:** (n) entertainer (a person who tries to please or amuse)
      - **S:** (n) person, individual, someone, somebody, mortal, soul (a human being) "there was too much for one person to do"
    - **S:** (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
      - **S:** (n) living thing, animate thing (a living (or once living) entity)
        - **S:** (n) whole, unit (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?", "the team is a unit"
        - **S:** (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
      - **S:** (n) physical entity (an entity that has physical existence)
        - **S:** (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
The Pipeline (continued)

**Rule 1: Semantic**
1. Start with the taxonomy of hypernyms for the given noun.
2. Generate an intersection of the set representing the noun’s taxonomy and the set entailing the semantic categories associated with the noun’s gender.
3. If no intersection is generated, parse the noun and recursively run the process again for the base noun.

**Rule 2: Morphophonemic**
1. Iterate over the lists of affixes associated with the gender of the input noun.
2. Check if the noun includes said affixes.
3. In case of nested suffixes, output the longest suffix.
The Pipeline (continued)

Preprocessing

Evaluate Masculine
- **Rule 1** (generate a closure over a hypernym taxonomy and search for the masc. categories)
- **Rule 2** (check the affixes)
- **Check if monosyllabic** (EN syllables counter)

Evaluate Neuter
- **Rule 1** (generate a closure over a hypernym taxonomy and search for the neut. categories)
- **Rule 2** (check the affixes)
- **Check if a foreign borrowing** (langdetect module)

Evaluate Feminine
- **Rule 1** (generate a closure over a hypernym taxonomy and search for the fem. categories)
- **Rule 2** (check the suffixes)
Evaluation:

Rules → Article
A list of **102,444** German nouns was scraped from Wiktionary DE.
The list was filtered for duplicates, retaining **100,064** unique nouns.
The nouns were analyzed for the grammatical class yielding **90,623** nouns that were successfully identified as:
- **31,164** masculine (~34%)
- **36,306** feminine (~40%)
- **22,153** neuter (~26%)
Evaluation - Extract the Features

- Four sets of features were extracted to describe each noun:
  - **Semantic** (Which categories does it belong to?)
  - **Morphological** (What prefixes and suffixes does it feature?)
  - **Etymological** (Is it a borrowing?)
  - **Syllabic / Phonological** (Is it monosyllabic?)
Evaluation - Train the Model

- A ML classifier (multinomial regression)
- Train the model on 81,561 (90%) nouns.
- Evaluate the model on 9062 (10%) nouns.
- Baseline accuracy is ~40% based on a dummy model.
Next Steps

- With GermaNet available, redevelop the program to employ a native German ontology, rather than WordNet.
- Find a better alternative to the Free German Dictionary for compound noun parsing.
- Develop a web app.
- Keep debugging.
Challenges and Lessons Learned

- **GermaNet** licensing takes time.
- English **WordNet** was substituted based on the assumption that semantic taxonomy is largely overlapping (i.e. a fork is a hyponym of a “pointy utensil” in either language).
- DeepL is better than Google Translate.
- Due to the lack of synsets certain semantic categories had to be excluded (e.g. proper nouns, various types of shapes, hot and cold things, etc.).
- **spaCy**'s morphological parser is 97% accurate (relevant for gender detection).
- **spaCy**'s lemmatizer is 99% accurate (relevant for lemmatization).
- Composite parsing is based on [Free German Dictionary](https://www.dwds.de/dictionary.php) (and it’s not the best).
- Syllable count approximation was done with syllables, an EN syllable counter requiring the following g2g rewrites: ‘ä’→‘ae’, ‘ö’→‘oe’, ‘ü’→‘ue’, ‘ß’→‘ss’.
References

1. Based on an analysis of around 100,000 nouns listed in the Duden Deutsches Universalwörterbuch, as of mid-2015. Source: Duden - Deutsches Universalwörterbuch.

2. Based on an analysis of around 16 million words included in the Duden German language database, as of mid 2015. Source: Duden - Deutsches Universalwörterbuch.

3. The source for the ages by which German children master aspects of German gender comes from the studies referenced in Mills, A.E. 1986. The Acquisition of Gender: A Study of English and German. Springer-Verlag.

4. Ibid.

   Köpcke, Klaus-Michael. January 2009. Genus, p. 137, references the findings of four separate such experiments.

6. See reference number 3.


8. As per Vayenas, Constantin. 2019. Der, Die, Das - The Secrets of the German Gender. Self-Published.
Code:
github.com/zoobereq/genusidator

Contact:
s.zuberek@columbia.edu

Web:
zuberek.net
Thank you!