Genusidator
(genus + elucidator)

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

Simon Zuberek
May 16th, 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:

	Masculine	**Feminine**	**Neuter**
Nominative	der / ein	die / eine	das / ein
Genitive	des / eines	der / einer	des / eines
Dative	dem / einem	der / einer	dem / einem
Accusative	den / einen	die / eine	das / ein

70%

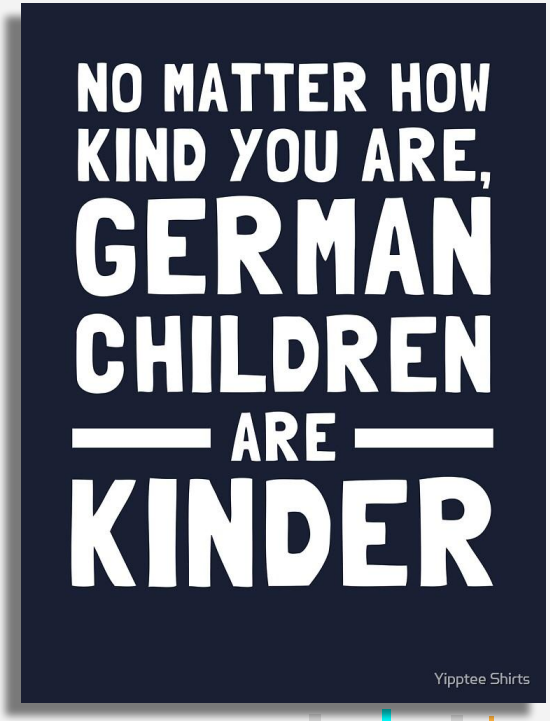
Nouns constitute over 70% of the words in the German language.[1]

German nouns occur in one of the three grammatical genders: feminine, masculine, neuter

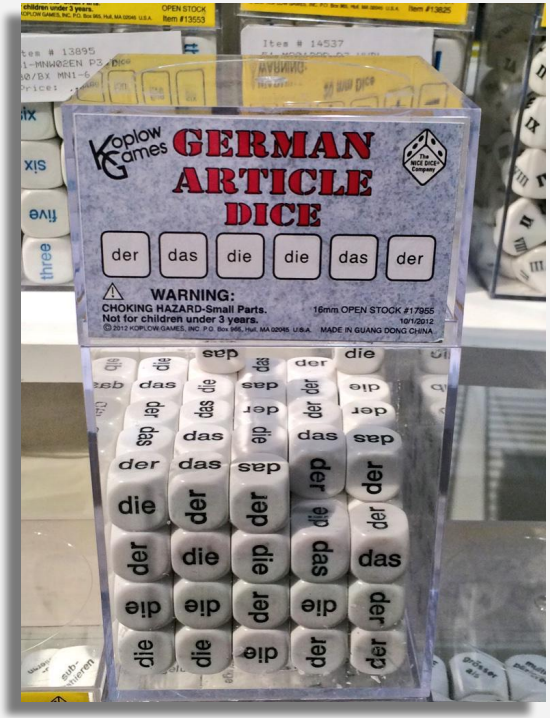
Collectively, nouns and the corresponding articles are the most frequently-used words in the German language.[2]

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]



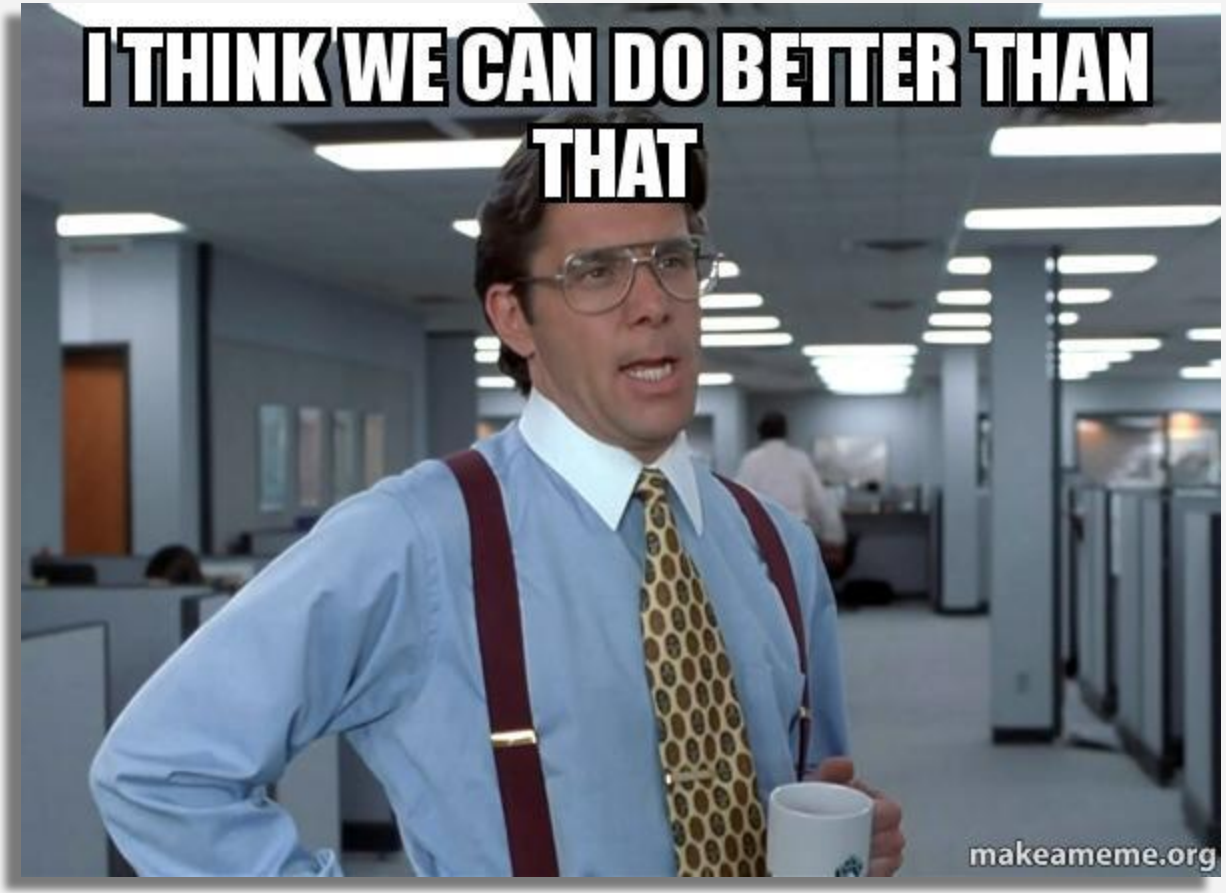
Motivation behind the Project



- The grammatical gender in German isn't explicitly taught. Students are told to learn it by heart.
- German language instructors believe that the grammatical gender assignment is arbitrary.[7]
- Native speakers of German and/or the majority of German language instructors were never taught the principles that determine gender.

“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
makeameme.org

The Rules behind the Gender Assignment [8]

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.**

Neuter

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-*

Article + Noun → Rules

The Pipeline

Preprocessing

- **User-input the noun** (`argparse`)
- **Output the gender and lemmatize** (`spaCy German transformer pipeline`)
- **Parse** (`German Compound Splitter`)
- **Translate to English** (`Google Translate API`)
- **Generate a taxonomy** of hypernym synsets going all the way to the root node of the graph (`nltk` and `wordnet`).

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 Feminine

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

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)

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 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.

Demo

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).
- 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 (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’`.

Next Steps

- Evaluation:
  - Flip the process: **Rules → Grammatical Gender**
  - Train a 3-way ML classifier that predicts the gender of the noun based on the semantic and morphophonemic features.
  - A list of ~100K German nouns available at `german-nouns`
- Once `GermaNet` available, redevelop the program to employ a native German ontology, rather than `WordNet`.
- If `GermaNet` not available, experiment with a `deepL` instead of `Google Translate`.
- Find a better alternative to the Free German Dictionary for compound noun parsing.
- Develop a web app in Flask.
- Keep debugging.

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.
5. Krohn, Dieter and Krohn Karin. 2008. *Der, das, die - oder wie? Studien zum Genuserwerb schwedischer Deutschlerner*. Peter Lang., p. 107.
   Köpcke, Klaus-Michael. January 2009. *Genus*, p. 137, references the findings of four separate such experiments.
6. See reference number 3.
7. Köpcke, Klaus-Michael. 1982. *Untersuchungen zum Genussystem der deutschen Gegenwartssprache*. Max Niemeyer Verlag, page 1. This author cites four language experts of the time, to back up his claim.
8. As per Vayenas, Constantin. 2019. *Der, Die, Das - The Secrets of the German Gender*. Self-Published.

Code available at:

github.com/zoobereq/genusidator

Questions?

Thank you!