Exploring Gender Variation of German Rivers using TüNDRA

Author: Daniel Gallagher

Date: February 24, 2025

1 Introduction

TüNDRA (Tübingen aNnotated Data Retrieval Application) is a lightweight query language designed for linguists wishing to search through any of the 1004 available treebanks at the time of writing. These treebanks come in one of two potential forms: *dependency* or *constituency*. Dependency trees display direct relations between words on a single plane and therefore tend to quicker to create and more adaptable to other languages due to a lack of a fixed hierarchical phrase structure. On the other hand, constituency trees represent the hierarchy that is present in sentence structure, allowing phrase structure and more complex syntactic structures to be clearly represented e.g coordination. Dependency grammar was first developed in detail by French Linguist Lucien Tesnière in the early 20th Century, while constituency grammar was a more recent development by Noam Chomsky.

2 Querying in TüNDRA

The syntax of TüNDRA is quite similar to *grew* and allows us to create queries with complex syntactic relations as well as the definition of hierarchical information for constituency trees. Let's walk through a number of examples. A phrase category for instance can be specified with the *cat* parameter.

```
[cat="VX"] -- Verb Phrase
[cat="NX"] -- Noun Phrase
[cat="PX"] -- Prepositional Phrase
```

Boolean operators can also be specified in a similar manner to other query languages.

```
-- word = Oder OR Donau:
[word="Oder"] | [word="Donau"]
-- (word = Rhein) AND (cat = NP):
[word="Rhein"] & [cat="NX"]
```

```
-- part-of-speech = article (ART) followed directly by the lemma "Donau":
[pos="ART"] . [word="Donau"]
```

Hierarchical information can be intuitively encoded in TüNDRA like in the following examples. Non-terminal nodes are represented by *NT* and terminal nodes by *T*.

```
-- A non-terminal node directly dominates
-- a node with part-of-speech preposition
[NT] > [pos="APPR"]

-- A noun phrase node dominates the noun Berlin
[cat="NX"] >* [word="Berlin"]
```

Variables can also be assigned a name which can be used as column headers if your query results are exported to a format suitable for further data analysis such as *csv*.

```
-- Assign variable name 'prep' to the preposition
[NT] > #prep:[pos="APPR"]
```

3 Analysis of River Gender in German

Nouns in German can have one of three genders: *masculine*, *feminine*, or *neuter*. The German word for river is "der Fluss", which indicates that this is a masculine noun. Despite this however, named rivers can vary in terms of which gender they are assigned and therefore it cannot be taken for granted that a river should be treated as maculine. In order to demonstrate this, a query will be created which is applied to the TüBa-D/Z Release 11.0 constituency treebank, which consists of over 100,000 German sentences compiled by the University of Tübingen. The query is as follows:

```
#adpos_phrase:[cat="PX"] > #adpos:[T & pos="APPR"]  &
#adpos_phrase > #noun_phrase:[cat="NX=LOC"] &
#noun_phrase > #definite_article:[T & pos="ART"] $
#river:[T & word=("Donau"|"Elbe"|"Oder"|"Rhein"|"Main"|"Neckar")]
```

Six large rivers that run at least partly through Germany can be seen listed in German. In English they correspond to the Danube, Elbe, Oder, Rhine, Main, and Neckar, respectively. The steps are broken down line-by-line like so:

1. A PP *adpos_phrase* must directly dominate an adposition *adpos* which is a terminal node.
2. That *adpos_phrase* must directly dominate a geographical NP *noun_phrase*.
3. That *noun_phrase* must directly dominate an article *definite_article* which is a terminal node and is sister to (designated by the $ operator) a named river which is also a terminal node.

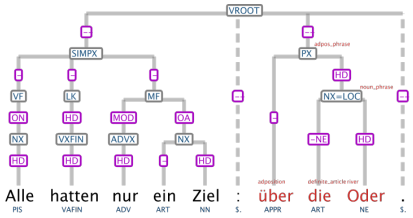

Figure 1: Example From Query Results

One might expect that a query which defines a linear relation between the three words to be easiest, such as one which finds a preposition, followed by an article, followed by a river. However, this would exclude the possibility of including adjectives when speaking about a river (e.g on the beautiful Danube) and does not take into account the possibility of postpositional phrases in German ("der Donau entlang" - along the Danube). The hierarchical nature of constituency treebanks therefore show their advantages here where we are able to easily define domination and sisterhood relations.

3.1 Results

The full results of this query can be viewed in Table 1. Thankfully, we see that there is a consistent gender for each of the rivers, with three being feminine - Elbe, Oder, and Danube - and two being masculine - Rhine and Main. Note that there was the inclusion of the Neckar river in the query but this does not appear in the results. This is due to our query returning no sentences that include this river from this dataset.

Definite Article	**River**	**Occurrences**	**Percentage**
die (feminine)	Elbe	11	39.286
die (feminine)	Oder	8	28.571
die (feminine)	Donau	6	21.429
der (masculine)	Main	2	7.143
der (masculine)	Rhein	1	3.571

Table 1: Article usage and gender for various German rivers.

4 Conclusion

Treebanks are an incredibly useful tool for answering research questions which we may have in linguistics. It has been demonstrated that the TüNDRA query language is a useful tool for linguists in analysing treebanks, especially when we want to define queries which pertain to hierarchical structure as in constituency treebanks. This work has given a brief introduction to this query language and

shown how it can be used to view the varying genders of river names in German. Further work could observe other types of proper nouns in German and model the complex landscape of noun genders in the German language.