# GF WordNet

Krasimir Angelov
GF Summer School 2025

# Overview

# GF WordNet

- A parallel lexicon with 264 languages
- WordNet style semantic relations
- Integrated with the RGL whenever possible
- ~ 110 000 core lexical entries
- ~ 600 000 people names
- 3.7 million location names
- > 10 000 examples in abstract syntax

https://cloud.grammaticalframework.org/wordnet/

# 40 Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 9 Germanic Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 7 Romance Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 5 Slavic Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 3 Finno-Ugric Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 2 Turkic Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 2 Bantu Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

*Note: Noun classes are all wrong*

# 2 Indo-Aryan Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 2 Semitic Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

*Note: Inflection probably wrong*

# 7 More Languages

- integrated with the RGL

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Afrikaans | Chinese | Finnish | Icelandic | Kazakh | Nynorsk | Russian | Swedish |
| Albanian | Danish | French | Interlingua | Korean | Bokmål | Slovenian | Thai |
| Arabic | Dutch | German | Italian | Macedonian | Polish | Somali | Turkish |
| Bulgarian | English | Hindi | Japanese | Maltese | Portuguese | Spanish | Urdu |
| Catalan | Estonian | Hungarian | Latvian | Mongolian | Romanian | Swahili | Zulu |

# 21 Languages

- not integrated with the RGL yet

| | | | | |
|---|---|---|---|---|
| Amharic | Egekusii | **Latin** | Punjabi | Telugu |
| Ancient Greek | **Greek** | Lithuanian | Rukiga | |
| **Basque** | Greenlandic | Malay | Sindhi | |
| Croatian | Hebrew | Nepali | Slovak | |
| Czech | Kikamba | **Persian** | Tamil | |

# 264 languages in total

- The full list of languages with statistics:

https://github.com/unipv-larl/GWC2025/releases/download/papers/GWC2025_paper_2.pdf

- Languages are also searchable from the web interface

- Only lemmas available for most languages

# Project 1: Linguistic Typology

- For each pair of languages compute the average Levenshtein distance between two words.
- Construct a 264 x 264 matrix with all distances between languages
- Use t-SNE to embed all languages in 2 or 3 dimensional space
- Do you detect the language families?
- Can you detect incorrect translation as pairs of words which are far apart?

# Synsets

## Synonyms

| Abstract | Bulgarian | English | Finnish | Portuguese | Swedish |
|---|---|---|---|---|---|
| family_1_N | семейство | family | suku | casa | familj |
| home_8_N | дом | home | perhe | casa | hem |
| household_N | домакинство | household | kotitalous | casa | hushåll |

# Morphology

| Abstract | Bulgarian | English | Finnish | Portuguese | Swedish | *f* |
|---|---|---|---|---|---|---|
| 1. horny plate covering and protecting part of the dorsal surface of the digits | | | | | | |
| + nail_1_N | нокът | nail | kynsi | unha | nagel | |
| 2. a thin pointed piece of metal that is hammered into materials as a fastener | | | | | | |
| - nail_2_N | гвоздей | nail | naula | prego | spik | |

## Substantiv (utr)

| | | obest | best |
|---|---|---|---|
| **nom** | **sg** | spik | spiken |
| | **pl** | spikar | spikarna |
| **gen** | **sg** | spiks | spikens |
| | **pl** | spikars | spikarnas |

# Examples

| | |
|---|---|
| **Bulgarian** | Вода бликна през улиците. |
| **Catalan** | Aigua adollà mitjançant els carrers. |
| **Danish** | Vand vældede på grund af gaderne. |
| **Dutch** | Water opwelde door de straten. |
| **English** | Water gushed through the streets. |
| **French** | L'eau jaillissait par les rues. |
| **German** | Wasser strömte durch die Straßen. |
| **Italian** | L'acqua sgorgò per le vie. |
| **Norwegian Nynorsk** | Vatn strøymde på gatane. |
| **Norwegian Bokmål** | Vann strømma gjennom gatene. |
| **Portuguese** | A água jorrou pelas ruas. |
| **Romanian** | Apă a țâșnit prin stradele. |
| **Russian** | Вода хлынула через улицы. |
| **Spanish** | La agua brotó por las vías. |
| **Swedish** | Vatten forsade genom gatorna. |

- Literal translations via the RGL abstract syntax
- Manually checked for Swedish and Bulgarian
- Major factor when choosing the correct translations

# VerbNet Frames

| | |
|---|---|
| **Bulgarian** | Вода се изля на растенията. |
| **Catalan** | Aigua corregué a les plantes. |
| **Danish** | Vand strømmede til planterne. |
| **Dutch** | Water stroomde op de vegetaties. |
| **English** | Water poured onto the plants. |
| **French** | L'eau coulait aux plantes. |
| **German** | Wasser strömte in die Pflanzen. |
| **Italian** | L'acqua scorse a le piante. |
| **Norwegian Nynorsk** | Vatn rennadde på plantane. |
| **Norwegian Bokmål** | Vann strømma på plantene. |
| **Portuguese** | A água correu a as plantas. |
| **Romanian** | Apă a curs în plantele. |
| **Russian** | Вода [pour_4_V]лась на растения. |
| **Spanish** | La agua fluyó a las plantas. |
| **Swedish** | Vatten hällde på växterna. |
| pour_4_V: flow in a spurt<br><br>**roles:** Theme, DestPrep, Destination | |

- 25% of the VerbNet frames are also integrated in the GF WordNet
- 750 frame examples
- Generally verb frames need more work

# Linking with Wikidata

| Abstract | Afrikaans | Bulgarian | Catalan | Danish | Dutch | English | French | German | Hungarian | Icelandic | Italian | Macedonian | Norwegian Nynorsk | Norwegian Bokmål | Polish | Portuguese | Romanian | Russian | Slovenian | Spanish | Swedish | Turkish | f |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. any of various burrowing animals of the family Leporidae having long ears and short tails; some domesticated and raised for pets or food | | | | | | | | | | | | | | | | | | | | | | | |
| rabbit_1_N | konyn | заек | conill | kanin | konijn | rabbit | lapin | Kaninchen | nyúl | kanína | coniglio | зајак | kanin | kanin | królik | lebre | iepure | кролик | kunec | conejo | kanin | tavşan | |

# European rabbit

92 languages

Article Talk Read Edit source View history

From Wikipedia, the free encyclopedia

*This article is primarily concerned with the wild animal. For detailed information on domesticated varieties, see Domestic rabbit. For general information on all rabbit species, see Rabbit.*

The **European rabbit** (***Oryctolagus cuniculus***) or **coney**[5] is a species of rabbit native to the Iberian Peninsula (Spain, Portugal and Andorra) and southwestern France.[3] It is the only living species in *Oryctolagus*, a genus of lagomorphs. The average adult European rabbit is smaller than the European hare, though size and weight vary with habitat and diet. Due to the European rabbit's history of domestication, selective breeding, and introduction to non-native habitats, wild and domesticated European rabbits across the world can vary widely in size, shape, and color.

European rabbits prefer grassland habitats and are herbivorous, mainly feeding on grasses and leaves, though they may supplement their diet with berries, tree bark, and field crops such as maize. They are prey to a variety of predators, including birds of prey, mustelids, cats, and canids. The European rabbit's main defense against predators is to run and hide, using vegetation and its own burrows for cover. It is well known for digging networks of burrows, called warrens, where it spends most of its time when not feeding. The European rabbit lives in social groups centered around territorial females. European rabbits in an established social group will rarely stray far from their warren, with female rabbits leaving the warren mainly to establish nests where they will raise their young. Unlike hares, rabbits are born blind and helpless, requiring maternal care until they leave the nest.

The European rabbit has had major agricultural and biological impacts as an invasive species, and has been hunted and raised as a food source since medieval times. It is the only domesticated species of rabbit, and all known breeds of rabbit are its descendants. It has often been introduced to exotic locations

**European rabbit**

**Temporal range:** Chibanian–Recent[2]
~0.6–0 Ma

# Linking with Wikidata - Motivation

Supports the development of the lexicon:

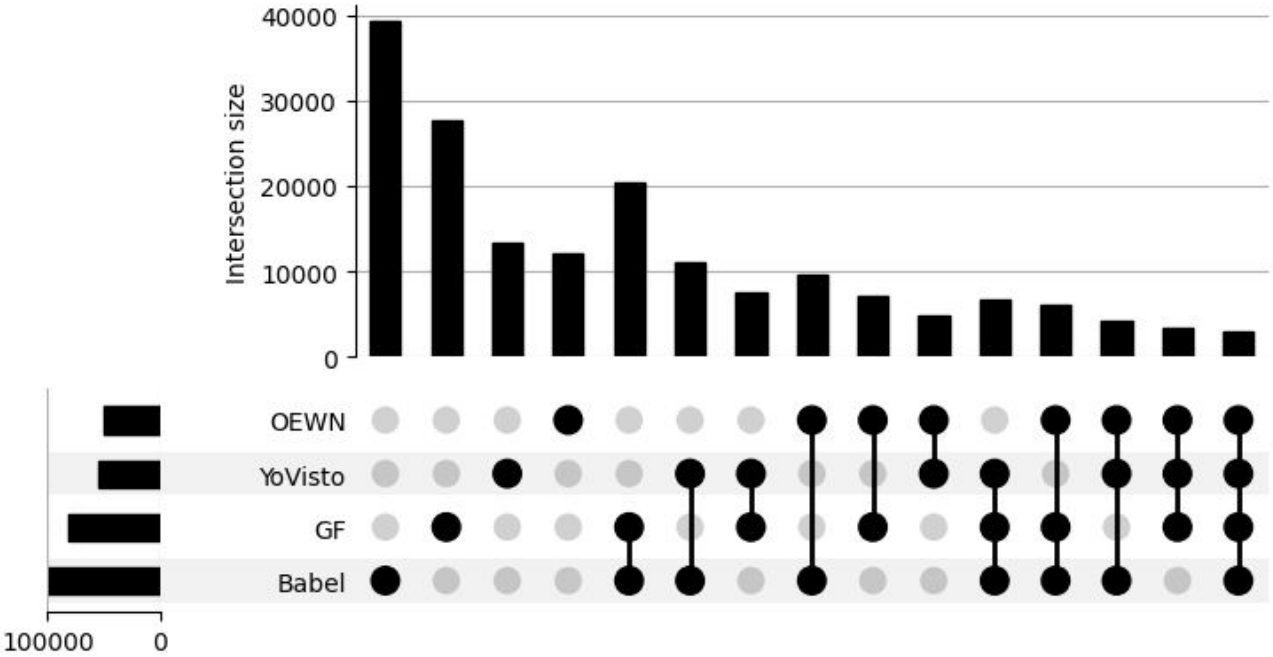
- A picture tells a thousand words
- Nice to be able to read the article
- Source of automatic translations

Supports NLG with Wikidata

- The NLG API can generate abstract trees from a QID
- More precise alignment is sometimes needed

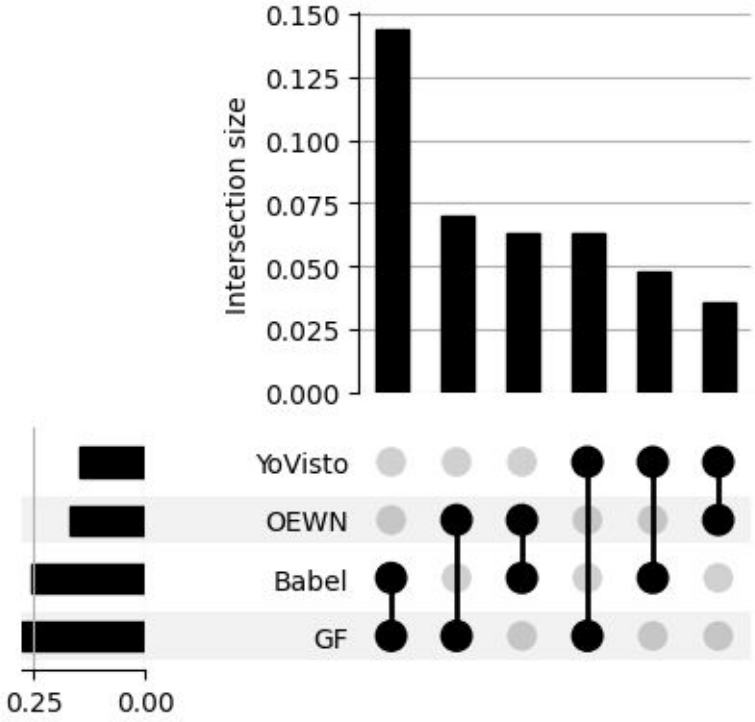
# Linking with Wikidata (Comparative size of the resources)

Joint work with John McCrae and Johann Bergh



# Linking with Wikidata (The percentage of disagreements)

Joint work with John McCrae and Johann Bergh



# Location and People Names from Wikidata

For NLG purposes the grammar is extended with names

| | | | |
|---|---|---|---|
| WordNet | adjectives, nouns, verbs, etc. | 100 thousand | |
| Wikidata | Given names | 64 thousand | Describing 7.3 million people |
| | Family names | 531 thousand | |
| | Place names | 3.7 million | |
| | total | 4.3 million | |

# Constructions

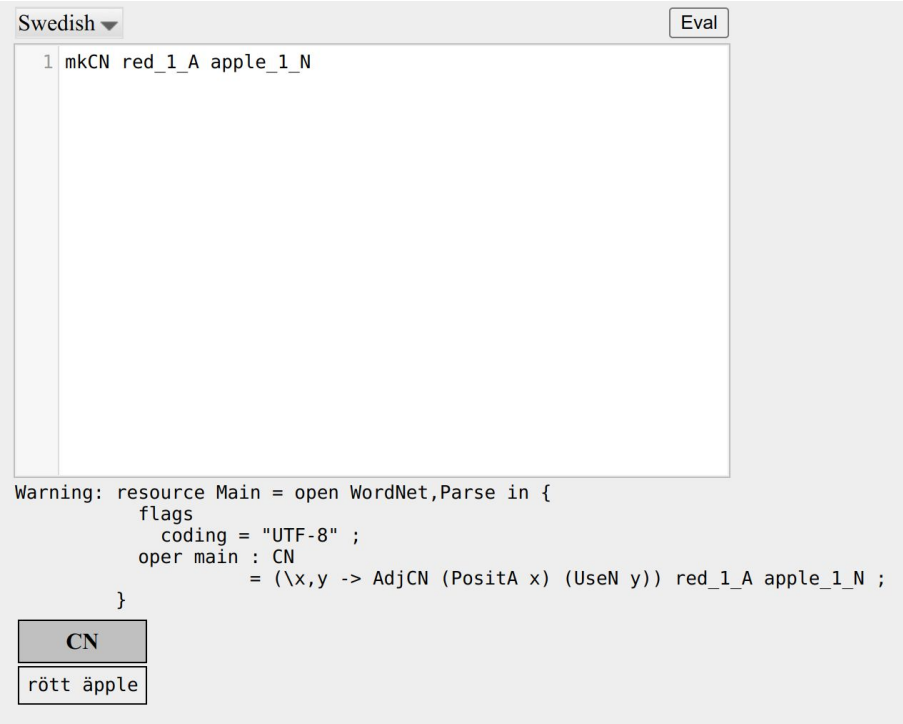
- A collection of multiword expressions attached to a synset or QID

abs: UseN (CompoundN square_1_N kilometre_1_N)
fre: kilomètre carré
spa: kilómetro de cuadrado
swe: kvadratkilometer
fin: neliökilometri
key: Q712226

abs: AdjCN (PositA square_1_A) (UseN kilometre_1_N)
key: Q712226

# Functions Service

https://cloud.grammaticalframework.org/wordnet/gf-functions.html



# Grammar Size

The WordNet grammar:

- 264 languages
- 40 syntaxes
- 4-5 million abstract lexemes
- 78 Gb in total

# Python NLTK style

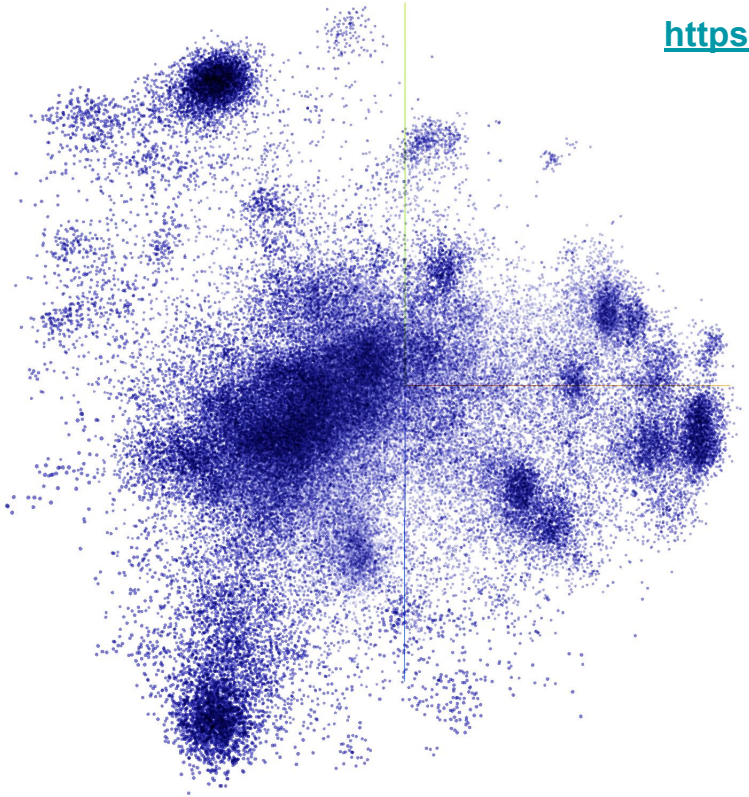
```
$ pip3 install gf-wordnet
$ python3

>>> import wordnet
Either use wordnet.download(['ISO 639-2 code1', ...]) to download the grammar,
or use wordnet.symlink('path to a folder') to link the library to an existing grammar.
If download() is called without an argument it will download all languages.

>>> wordnet.download(['eng'])
Download and boot the grammar 355MB (Expanded to 2637MB)
Download the semantics database 2733MB done
Reload wordnet
```

More information: https://pypi.org/project/gf-wordnet/

# Abstract Sense Embedding



https://cloud.grammaticalframework.org/wordnet/embedding.html

64 dimensional Graph2Vec embedding

Graph2Vec is a variant of Word2Vec which learns a vector for each node.

# Predicting the category

Given a vector embedding is it possible to predict the category of the word?

- Single Layer Perceptron predicts the right type in 75% of the cases
- Two Layer Perceptron predicts the right type in 80% of the cases

# No Correlation with the Graph-based similarity

# Sense Disambiguation

Sense disambiguation by using the Word2Vec model.

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^{\top} v_{w_I}\right)}$$

- v' is the vector learned by Graph2Vec
- v=f(v') transformed by a 4 layered ReLU network

10% split for evaluation

| | |
|---|---|
| Context-Free | 64% |
| Word2Vec | 76% |

evaluation on training data

| | |
|---|---|
| Context-Free | 77% |
| Word2Vec | 78% |

# Project 2: use LLM to estimate a better probabilistic model

- The GF WordNet corpus has only ~10 000 examples
- LLMs can be probed to estimate the probabilities of combinations of words that we don’t have in the corpus.

# Bootstrapping

# Open Multilingual WordNet

Preference is given to translations witnessed in corresponding synset in the Open Multilingual WordNet

Pro:

- We know that the translation has the right sense

Cons:

- For many languages the data is too small. Gives unfair advantage to some words

# PanLex

An aggregation of thousands of manually created dictionaries for hundreds of languages.

When you already have a number of languages in GF WordNet, you can lookup translations from each language to the new target language. The translation that gets the most hits wins.

Pro:

- Available for many languages

Cons:

- Not always sure that the translation is for the right sense
- Sometimes it confuses parts of speech
- Some dictionaries contain explanations as well as translations

# Wikidata

For senses that are linked with Wikidata, pick the translation from there

Pro:

- The linking is sense aligned
- Available for many languages

Cons:

- Wikidata labels are not always translations
- Sometimes there are more than one labels

# Wiktionary

68 844 lexemes from GF WordNet are aligned with their Wiktionary entry based on the SBERT similarity of the glosses:

| | |
|---|---|
| GF WordNet | fruit with red or yellow or green skin and sweet to tart crisp whitish flesh |
| Wiktionary | A common, firm, round fruit produced by a tree of the genus Malus. |

Pro: sense aligned, good translations
Cons: some mistakes still possible

# Project 3: Large Language Models and Transformer MT

Use models to fill in gaps and check entries

Example:

Water poured from the bowl into the cup.
Вода [pour_4_V]лась из миски в чашку.

Google Translate:

Вода из миски перелилась в чашку.

# Verification Status

Uncertain entries are labeled with:

- red - possible translation but might be for a different sense
- yellow - has the right sense, may not be the best translation

# Learning Morphology

What do we do with all the 200+ languages for which there is no grammar?

Learn automatically?

So far so good!