Multilingual Natural Language Generation on a Large Scale

Krasimir Angelov, Andrea Carrión del Fresno, Ekaterina Voloshina,
Evan Xingye Geng, Aarne Ranta

International Conference on Mathematical and Computational Linguistics for Proofs

GF WordNet

GF WordNet

- A parallel lexicon with 264 languages
  - like Wiktionary but it is not a wiki – it is a database
- WordNet style semantic relations
  - like Princeton WordNet but with translations
- Integrated with syntax whenever possible
  - includes syntactic combinators
- Server side scripting language for applications
- Free, open-source, editing on the web

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

45 Languages

- integrated with syntax

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

11 Germanic Languages

- integrated with syntax

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

Note: Scots has just been started
Note: Faroese mostly morphology, little syntax

8 Romance Languages

- integrated with syntax

Afrikaans

Albanian

Arabic

Belarusian

Bulgarian

Catalan

Chinese

Danish

Dutch

English

Estonian

Faroese

Finnish

French

German

Hindi

Hungarian

Icelandic

Interlingua

Italian

Japanese

Latin

Latvian

Kazakh

Korean

Macedonian

Maltese

Mongolian

Nynorsk

Bokmål

Polish

Portuguese

Romanian

Russian

Slovenian

Scots

Somali

Spanish

Swahili

Swedish

Thai

Turkish

Ukrainian

Urdu

Zulu

7 Slavic Languages

- integrated with syntax

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

Note: Ukrainian and Belarusian are not online yet
Note: Macedonian mostly morphology, little syntax

3 Finno-Ugric Languages

- integrated with syntax

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

2 Turkic Languages

- integrated with syntax

Afrikaans

Albanian

Arabic

Belarusian

Bulgarian

Catalan

Chinese

Danish

Dutch

English

Estonian

Faroese

Finnish

French

German

Hindi

Hungarian

Icelandic

Interlingua

Italian

Japanese

Latin

Latvian

Kazakh

Korean

Macedonian

Maltese

Mongolian

Nynorsk

Bokmål

Polish

Portuguese

Romanian

Russian

Slovenian

Scots

Somali

Spanish

Swahili

Swedish

Thai

Turkish

Ukrainian

Urdu

Zulu

Note: Kazakh mostly morphology, little syntax

2 Bantu Languages

- integrated with syntax

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

Note: Noun classes are all wrong

2 Indo-Aryan Languages

- integrated with syntax

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

2 Semitic Languages

- integrated with syntax

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

Note: Arabic inflection probably wrong

7 More Languages

- integrated with syntax

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

264 languages in total

- The full list of languages with statistics:

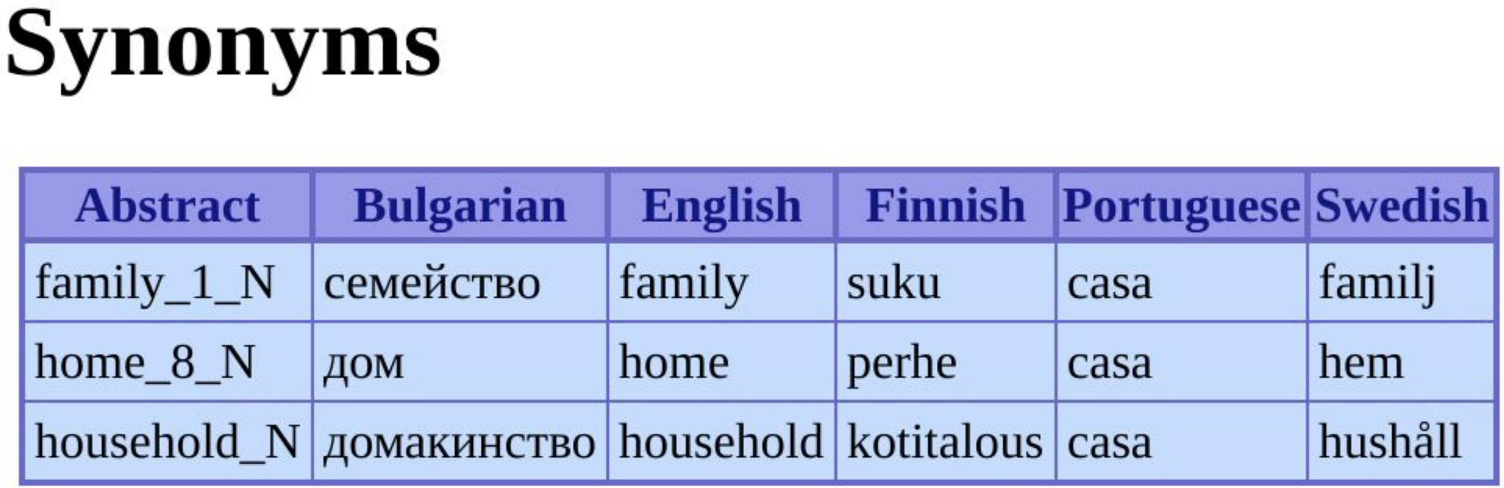
https://github.com/unipv-larl/GWC2025/releases/download/papers/GWC2025_paper_2.pdf

- Languages are also searchable from the web interface

- Only lemmas available for most languages

Synsets

Like in Princeton WordNet, but here synsets are two dimensional, and we preserve translation relations



Synonyms

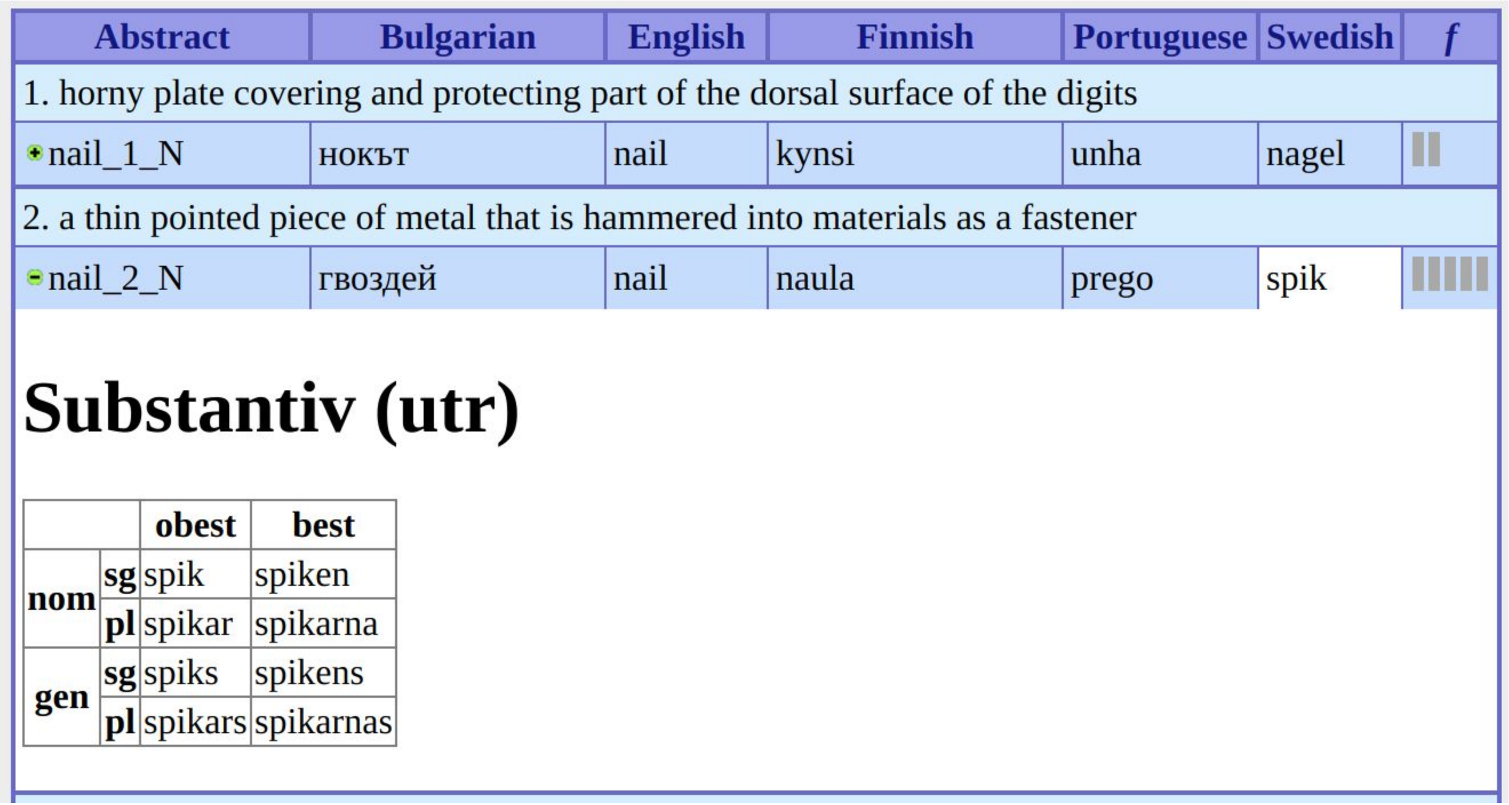
Abstract	Bulgarian	English	Finnish	Portuguese	Swedish
family_1_N	семейство	family	suku	casa	familj
home_8_N	дом	home	perhe	casa	hem
household_N	домакинство	household	kotitalous	casa	hushåll

WordNet Style Semantics

Semantic relations:

- Hypernym/Hyponym
- Holonym/Meronym
- Antonym
- Attribute
- DomainOfSynset
- MemberOfDomain
- Entailment
- Cause
- AlsoSee
- VerbGroup
- SimilarTo
- Derived
- Male
- Female

Morphology

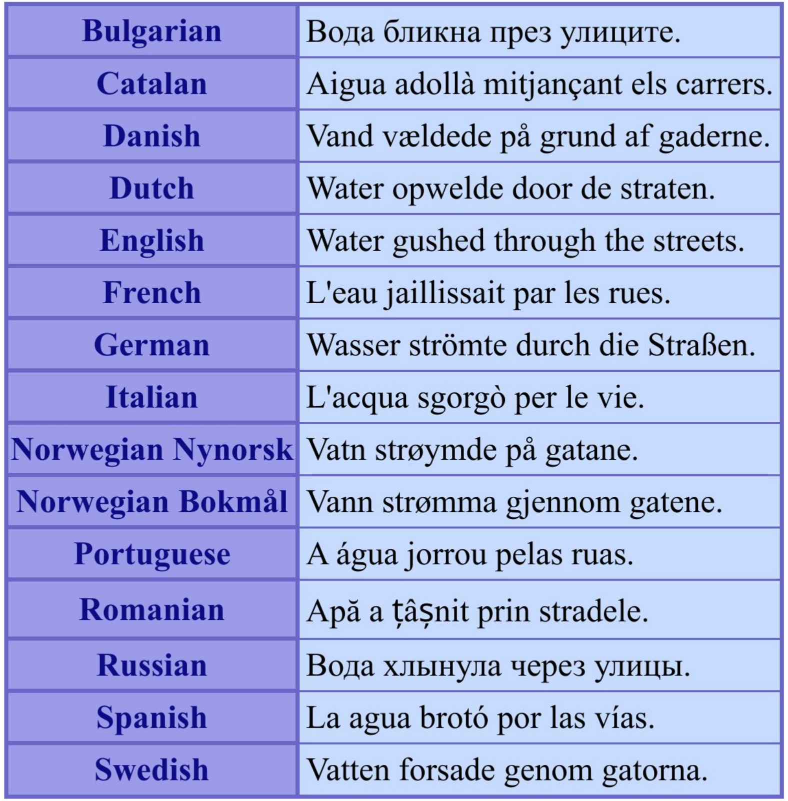


Abstract	Bulgarian	English	Finnish	Portuguese	Swedish	*f*
1. horny plate covering and protecting part of the dorsal surface of the digits						
⊕ nail_1_N	нокът	nail	kynsi	unha	nagel	▮▮
2. a thin pointed piece of metal that is hammered into materials as a fastener						
⊖ nail_2_N	гвоздей	nail	naula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

Substantiv (utr)

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

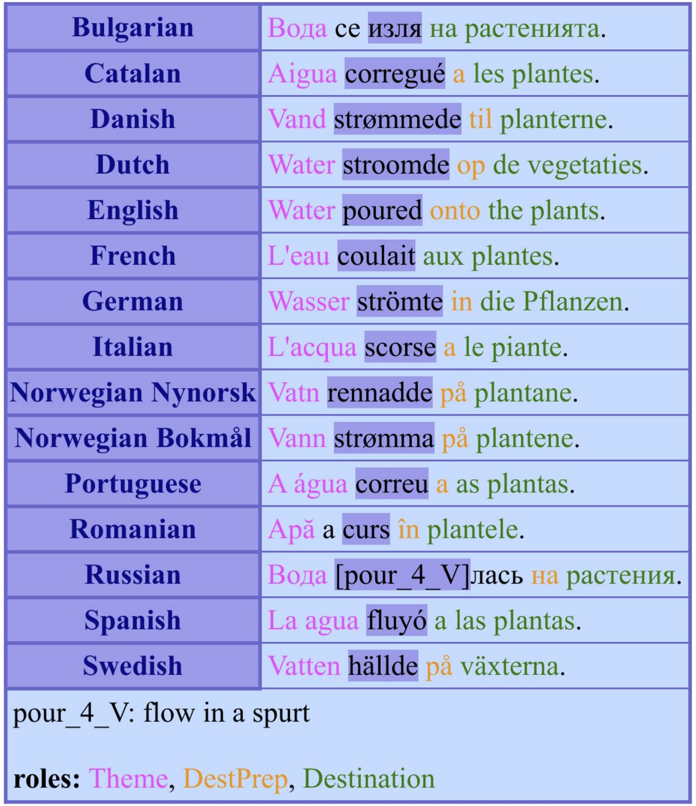
Examples



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 a common abstract syntax
- Manually checked for Swedish and Bulgarian
- Major factor when choosing the correct translations

VerbNet Frames



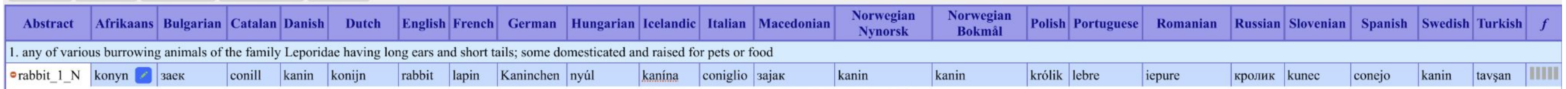
Bulgarian	Вода се изля на растенията.
Catalan	Aigua corregué a les plantes.
Danish	Vand strømmede til planterne.
Dutch	Water stroomde op de vegetaties.
English	Water poured onto the plants.
French	L'eau coulait aux plantes.
German	Wasser strömte in die Pflanzen.
Italian	L'acqua scorse a le piante.
Norwegian 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

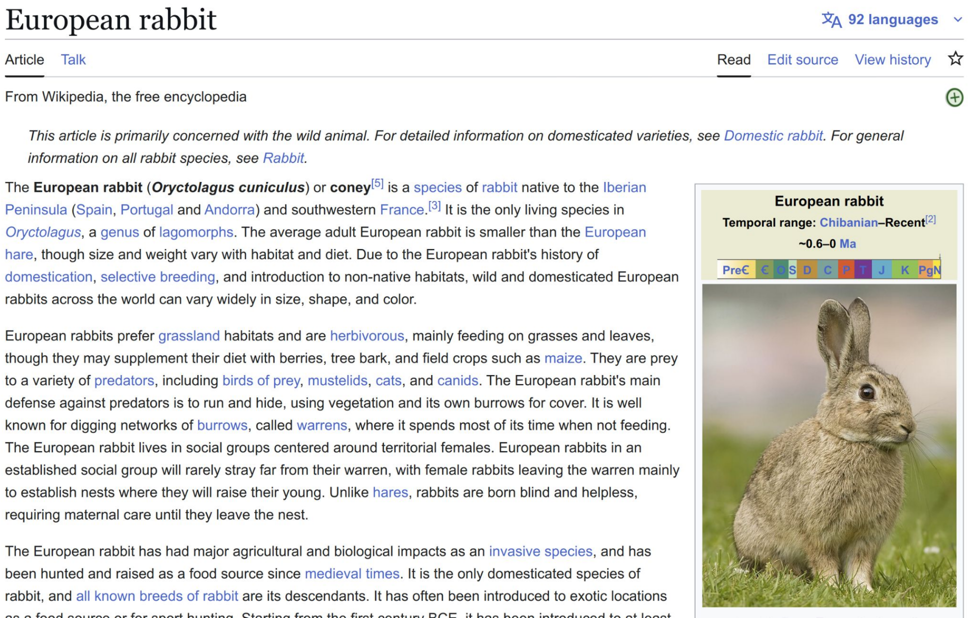
roles: Theme, DestPrep, Destination

- 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

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 as a food source or for sport hunting. Starting from the first century BCE, it has been introduced to at least

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

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	

Grammatical Framework

- Statically Typed Functional Programming Language
- Specialized for the description of Natural Languages
- Reversible – parsing and generation by the same grammar

Abstract syntax	API
AdjCN (PositA open_11_A) (UseN set_2_N)	mkCN open_11_A set_2_N

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

Grammar Size

The WordNet grammar:

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

Python NLTK style

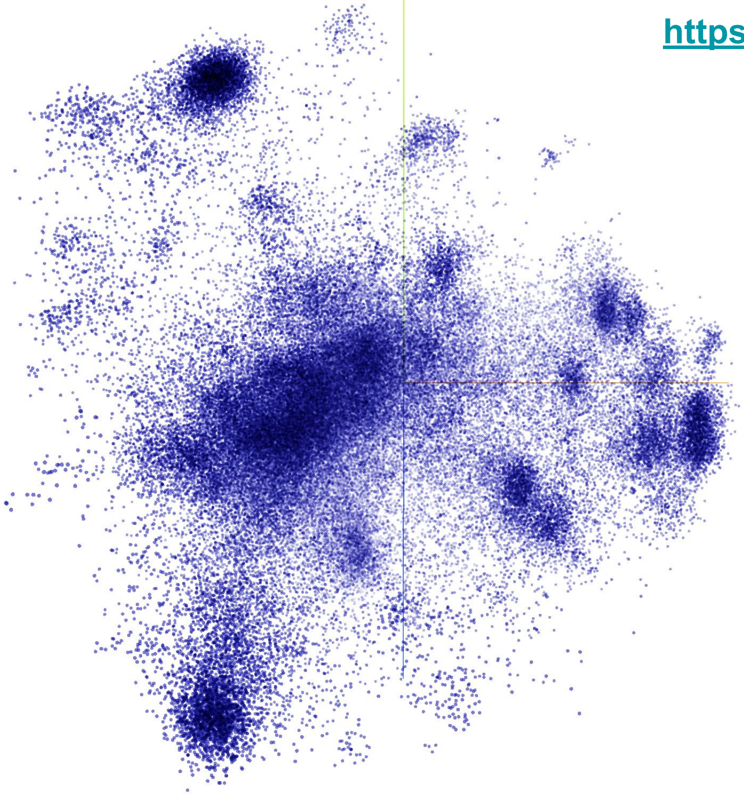
```
$ pip3 install gf-wordnet
$ python3

>>> 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.

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

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 Grammars

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

- Learn automatically from data
- Generalize from an existing language

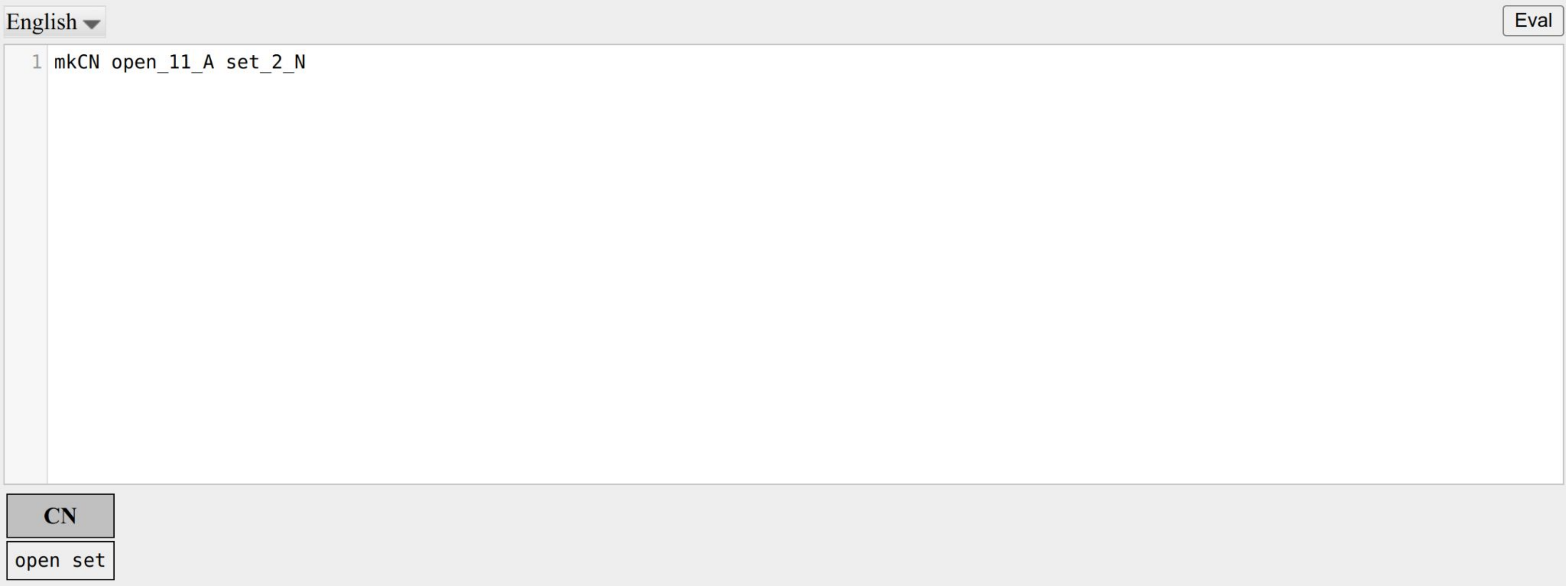
Pilot languages:

- Albanian
- Belarusian
- Faroese
- Kazakh
- Macedonian
- Ukrainian

NLG Scripting

GF Functions Service

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



```
mkCN open_11_A set_2_N
```

CN

```
open set
```

Lookup for Abstract Expressions

Web-based shell at:

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

- An entity in Wikidata is identified by QIDs, e.g. Q34 is Sweden, Q1 is the universe
- A simple API to the abstract expression (lexical item) for a QID:
  - expr "Q142"
  - mkNP theSg_Det (expr "Q1")

Querying for Entities

```
let e = entity "Q142"
in mkCl (mkNP (expr "Q142"))
         (mkNP aSg_Det (mkCN (mkCN (expr e.P31.id))
               (mkAdv with_Prep
                         [select: -1 | <mkNP (mkDecimal e.P1082.amount) inhabitantMasc_1_N,
                                        e.P1082.P585.time>
                         ]
               )))
```

Control Structures

Make it possible to manipulate and aggregate variants:

[<keyword>: <opt. argument> | <expr. with variants>]

Examples:

- [select: -1 | <e.P1082.amount, e.P1082.P585.time>]
- [list: and_Conj | mkCN (expr e.P37.id)]

Content Planning

```
let e = entity "Q142"
   cn0 = mkCN (expr e.P31.id)
   cn1 = mkCN cn0
              (mkAdv with_Prep
                 [select: -1 | <mkNP (mkDecimal e.P1082.amount) inhabitantMasc_1_N,
                                 e.P1082.P585.time>
                 ])
in [one | (cn1 | cn0)]
```

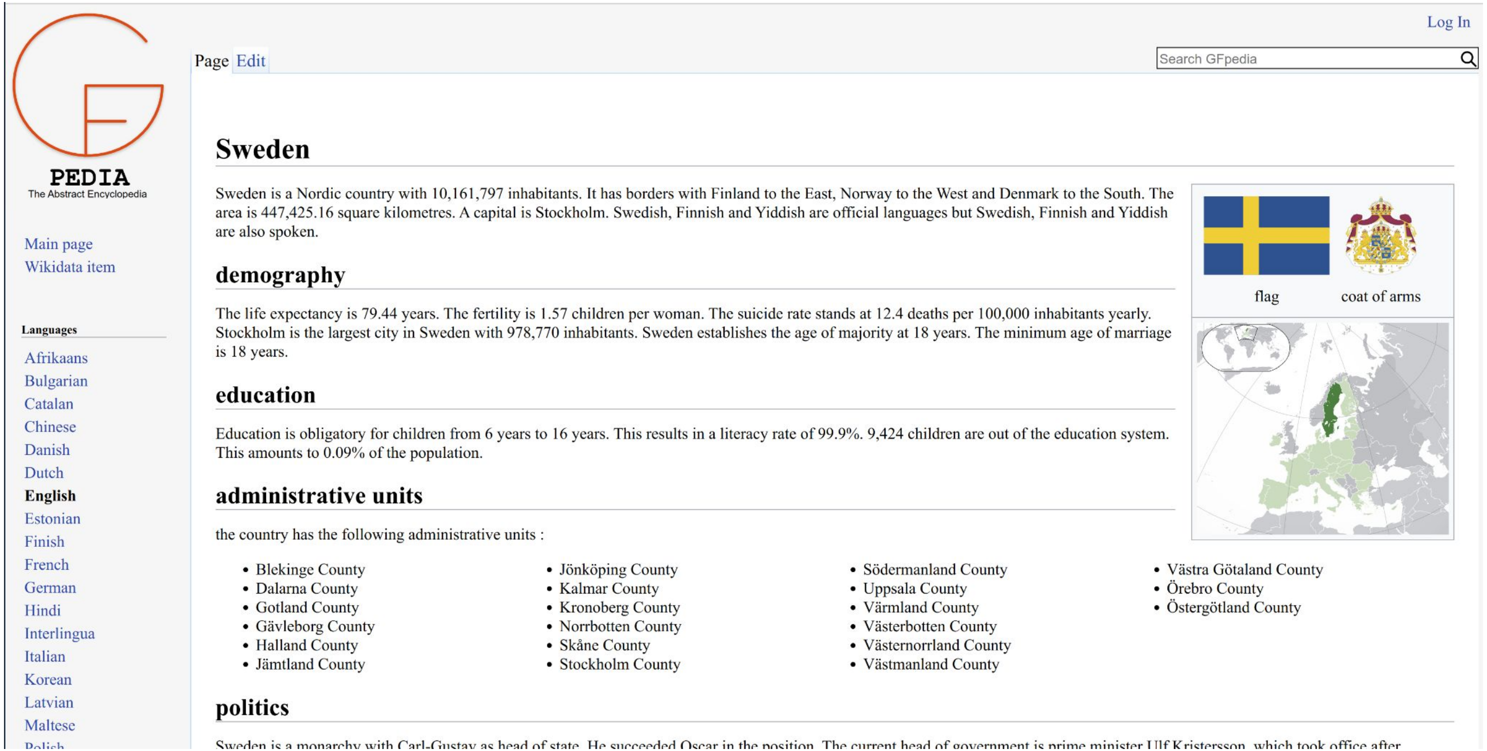
Markup

First Class Support for Markup

```
let e = entity "Q142"
in <div>
    <h1>expr "Q142"</h1>
    <p>mkCl (mkNP (expr "Q142"))
             (mkNP aSg_Det (mkCN (mkCN (expr e.P31.id))
                          (mkAdv with_Prep
                                    [select: -1 | <mkNP (mkDecimal e.P1082.amount)
                                                         inhabitantMasc_1_N,
                                                 e.P1082.P585.time>
                                    ]))</p>
  </div>
```

GF Pedia 2.0

https://cloud.grammaticalframework.org/wordnet/gf-wikidata.wiki



Sweden

Sweden is a Nordic country with 10,161,797 inhabitants. It has borders with Finland to the East, Norway to the West and Denmark to the South. The area is 447,425.16 square kilometres. A capital is Stockholm. Swedish, Finnish and Yiddish are official languages but Swedish, Finnish and Yiddish are also spoken.

demography

The life expectancy is 79.44 years. The fertility is 1.57 children per woman. The suicide rate stands at 12.4 deaths per 100,000 inhabitants yearly. Stockholm is the largest city in Sweden with 978,770 inhabitants. Sweden establishes the age of majority at 18 years. The minimum age of marriage is 18 years.

education

Education is obligatory for children from 6 years to 16 years. This results in a literacy rate of 99.9%. 9,424 children are out of the education system. This amounts to 0.09% of the population.

administrative units

the country has the following administrative units :

- Blekinge County
- Dalarna County
- Gotland County
- Gävleborg County
- Halland County
- Jämtland County
- Jönköping County
- Kalmar County
- Kronoberg County
- Norrbotten County
- Skåne County
- Stockholm County
- Södermanland County
- Uppsala County
- Värmland County
- Västerbotten County
- Västernorrland County
- Västmanland County
- Västra Götaland County
- Örebro County
- Östergötland County

politics

Sweden is a monarchy with Carl-Gustav as head of state. He succeeded Oscar in the position. The current head of government is prime minister Ulf Kristersson, which took office after

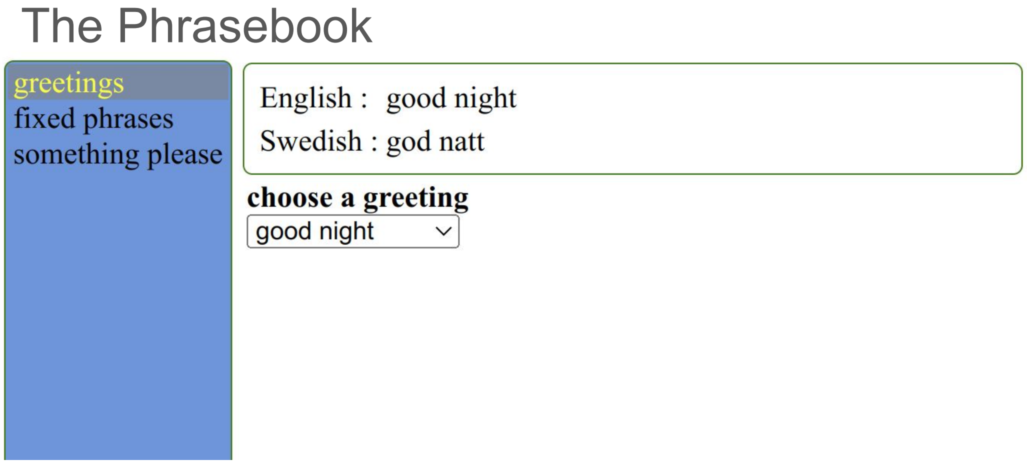
User Adaptations

Options

```
let qual = option quality_1_N of {
                italian_A ;
                swedish_A ;
                delicious_1_A
              } ;

    item = option food_1_N of {
                apple_1_N ;
                wine_1_N ;
                pizza_N
              }

in mkNP (mkCN qual item)
```



The Phrasebook

English : good night
Swedish : god natt

choose a greeting

good night

Linguistic Variations

Language adaptations for the best results

```
case lang of {
  "fin" => mkPhrMark (mkCl (mkNP (mkQuant it_Pron) NumPl abutterMasc_N) neighbours);
  "rus" => mkPhrMark (mkCl (mkNP it_Pron) (mkVP (mkVP border_5_V) (mkAdv with_Prep neighbours)));
  "fre" => mkPhrMark (mkCl (mkNP theSg_Det country_2_N) (mkVP have_1_V2
                                 (mkNP num (mkCN (mkCN border_1_N) (mkAdv with_Prep neighbours)))));
  "bul" => mkPhrMark (mkCl (mkNP (ProDrop she_Pron)) (mkVP have_1_V2
                                 (mkNP num (mkCN (mkCN border_1_N) (mkAdv with_Prep neighbours)))));
 "spa" => mkPhrMark (mkCl (mkNP (ProDrop it_Pron)) (mkVP have_1_V2
                                 (mkNP num (mkCN (mkCN border_1_N) (mkAdv with_Prep neighbours)))));
 _    => mkPhrMark (mkCl (mkNP it_Pron) (mkVP have_1_V2
                                 (mkNP num (mkCN (mkCN border_1_N) (mkAdv with_Prep neighbours)))))
}
```

So far so good!

A never ending journey is going on …