Development of Evidence-Based Grammars for Terminology Extraction in OneClick Terms

Marek Blahuš, Michal Cukr, Miloš Jakubíček, Vojtěch Kovář, Vít Suchomel

Lexical Computing

eLex 2023

Working for Sketch Engine

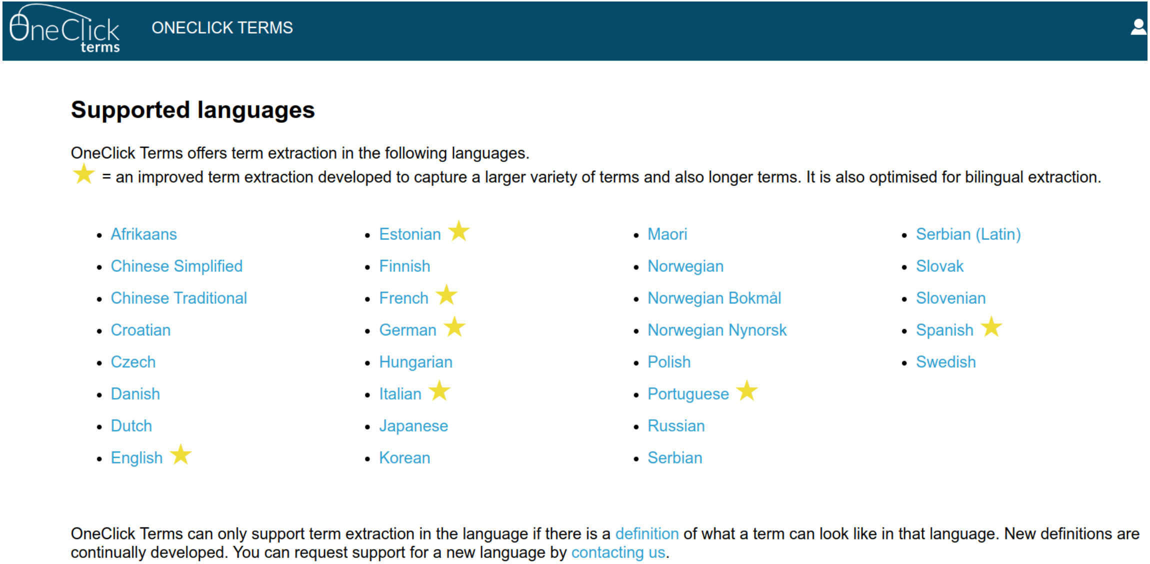
- If you start working for Sketch Engine, you need to un-learn:
  - completeness of algorithms
  - some linguistic theories
- Instead, you learn to:
  - think about accuracy in a „corpus way”
  - prefer corpus evidence

- And you often find yourself working with languages that:
  - you don't speak
  - you may have never heard of before
- Even though I am a polyglot ...
  - I still speak just some 10% of Sketch Engine's languages.
  - I often take aid from native speakers.

Polyglot Gathering 2023
(Yes, I know that `flagsarenotlanguages.com`.)

Terminology Extraction in Sketch Engine

- Keywords & Terms
  - Finding (multi-word) terms in a domain-specific corpus
  - Feature of Sketch Engine since 2013
  - Currently 29 supported languages
- OneClick Terms
  - Single-purpose user-friendly interface to Sketch Engine, built in 2017
  - For translators and terminologists
  - Monolingual or bilingual term extraction
  - `terms.sketchengine.eu`



Supported languages

OneClick Terms offers term extraction in the following languages.
★ = an improved term extraction developed to capture a larger variety of terms and also longer terms. It is also optimised for bilingual extraction.

- Afrikaans
- Chinese Simplified
- Chinese Traditional
- Croatian
- Czech
- Danish
- Dutch
- English ★
- Estonian ★
- Finnish
- French ★
- German ★
- Hungarian
- Italian ★
- Japanese
- Korean
- Maori
- Norwegian
- Norwegian Bokmål
- Norwegian Nynorsk
- Polish
- Portuguese ★
- Russian
- Serbian
- Serbian (Latin)
- Slovak
- Slovenian
- Spanish ★
- Swedish

OneClick Terms can only support term extraction in the language if there is a definition of what a term can look like in that language. New definitions are continually developed. You can request support for a new language by contacting us.

Prerequisites

- Terms are extracted using a corpus-based contrastive technology.
- Key elements for extraction of terminology from a *focus corpus*:
  1. large *reference corpus* in the particular language
  2. generic term extraction algorithm
     („term candidates" are scored by ratio of their normalized frequencies)
  3. language-specific term grammar
     (set of rules defining lexical structures typical of terms)
- Terms are typically noun phrases in canonical form.

Term Grammars

- Not all n-grams containing a noun are noun phrases.
- Each rule in a term grammar consists of:
  1. a labeled query in the CQL language which matches some term candidates,
     e.g. `2:[tag="JJ" | tag="NN.*" | tag="VVG.*"] 1:[tag="NN.*"]`
     matches *black cat, assistance dogs, flying elephant’s*
  2. a preceding directive defining how the term candidates are output,
     e.g. `*COLLOC "%(2.lc) %(1.lemma)"`
     outputs *black cat, assistance dog, flying elephant*
- For easier orientation and maintenance, rules make use of:
  - macros defined in the `m4` language, e.g. `noun` stands for `[tag="NN.*"]`
  - comments which explain a rule or provide an example of term matched by it

Evidence-Based Term Grammars

- Rules inspired by patterns observed in an existing terminology database
  - for EU languages: gold standard = IATE
  - for other languages: maybe Wikipedia titles?
- *This is „the corpus way” of doing it!*
  - descriptive, not prescriptive
  - maximization of coverage for top-ranked lexical structures

Development

- Filtering and cleaning the term base data
  - HTML markup, quotation marks, brackets, ellipses, lists, chemical formulas...
- Single-purpose *term corpus* (i.e. corpus of terms) in Sketch Engine
  - terms as sentences
  - standard PoS tagging, lemmatization, morphological annotation
- Two-level frequency distribution on the full term through Sketch Engine API
  - 1st level: part of speech
  - 2nd level: morphological tag

Frequency distribution in the term corpus

2. adjective + noun (119236 terms, 18.75%)

2.1. JJ NN (109240 terms, 17.18%)

Nuclear housing • active site • aero-medical centre • allelopathic chemical • armed neutrality • back chute • bacterial bed • calcareous grassland • complementary medicine • concurrent liability • critical assembly • dental floss • environmental effectiveness • ever-married survivor • express request • ferrous iron • fragmented mechanization • governmental aid • hedge period • hybrid selection • little plover • louvred fitting • mass effect • medical cannabis • mizzen sail • natural recovery • non-motorized vessel • on-line separation • political instability • poor soil • posterior kidney • preformed joint • private shareholder • public procurement • radiant density • random choice • reverse calf • sealed ampoule • semi-scale brewing • single licence • standard tare • straight lease • synthetic fluid • terminal bar • top performer • two-price system • unobservable variable • up-to-date inventory • variable pad • written assessment…

2.2. JJ NNS (8613 terms, 1.35%)

Introductory Notes • Physical contingencies • administrative courts • adverse consequences • algebraic parentheses • ancillary restrictions • beneficial contracts • calcareous algae • collective arrangements • cumulative grounds • descriptive markings • discouraged people • error-free seconds • essential workers • executive powers • fine seeds • hazardous substances • high-speed data • industrial trucks • interest-induced shifts • journey-related variables • locked points • major effects • mass properties • military mails • minor repairs • missing plants •

Writing a Term Grammar

- Compromising & generalization for length & simplicity
  - more attention paid to more frequent patterns
  - threshold for inclusion (0.15%)
  - native speaker's introspection (e.g. agreement)
  - deliberate ommission of some constraints (e.g. case government)
- Citation form for output
  - lemma, gender-respecting lemma, or word
  - typically lower case
- Rules grouped by number of tokens
- Example term for each rule

Rule Example

```
define(‘common_noun’, ‘[tag="NC.*"]’)
define(‘preposition’, ‘[lc="a|al|con|de|del|en|entre|para|por|sin|sobre"]’)
define(‘adjective’, ‘[tag="A.*" | tag="VMP.*"]’)
define(‘agree’, ‘$1.gender=$2.gender & $1.number=$2.number’)

*COLLOC "%(1.lemma) %(2.lc) %(3.lc) %(4.lc)"
1:common_noun 2:preposition 3:common_noun 4:adjective & agree(3, 4)
# example:  reducción de ojos rojos
```

Advanced Rule Design

- imperfect input
  - incorrectly tagged tokens
  - crossing noun-phrase boundaries (e.g. conjunctions)
- imperfect output
  - incomplete lexical structures (e.g. **Centro Robert ~~Schuman~~*)
  - plural-only terms (e.g. **foreign affair*, **United State of America*)
- occasional corpus research
  - prepositive adjectives
  - noun noun
- modification of corpus processing pipelines(!)

Results: Performance Comparison

1. pasta sfoglia ⬆ -1
2. secondo piatto ⬆ -2
3. primo piatto ⬆ -11
4. ricetta facile ⬇ +1
5. pasta fillo ⬆ -1
6. forno vegetariana ⬆ -3
7. **tempi di cottura —**
8. **verdure in padella —**
9. prossimo commento ⬆ -2
10. **cookie salvi —**
11. **ricette antipasti —**
12. torta in padella ⬆ -54
13. **verdure miste —**
14. cottura in padella ⬆ -17
15. **maria bonaccorso —**
16. cottura in forno ⬆ -2
17. forno statico ⬆ -2
18. padella antiaderente ⬆ -2
19. email necessario ⬆ -2
20. indirizzo email necessario ⬆ -2
21. **informazioni di profilo —**
22. **informazioni di profilo pubbliche —**
23. **profilo pubbliche —**
24. **ricette di antipasti —**

1. **pasta al forno +**
2. pasta sfoglia ⬇ +1
3. ricetta facile ⬆ -1
4. secondo piatto ⬇ +2
5. tempo di cottura ⬆ -25
6. pasta fillo ⬇ +1
7. **verdura al forno +**
8. ricetta vegetariana ⬆ -30
9. forno vegetariana ⬇ +3
10. **cookie salvo +**
11. prossimo commento ⬇ +2
12. antipasto veloce ⬆ -90
13. **pasta al forno vegetariana +**
14. primo piatto ⬇ +11
15. torta salata ⬆ -124
16. verdura in padella ⬆ -4641
17. antipasto sfizioso ⬆ -35
18. cottura in forno ⬇ +2
19. forno statico ⬇ +2
20. padella antiaderente ⬇ +2
21. email necessario ⬇ +2
22. indirizzo email necessario ⬇ +2
23. **informazione di profilo +**
24. **informazione di profilo pubbliche +**

IATE Recall

Language	IATE terms	Old grammar		New grammar	
English	635,700	367,693	57.8%	505,431	79.5%
Estonian	37,485	7,624	20.3%	24,884	66.4%
French	585,112	136,783	23.4%	425,133	72.7%
German	227,652	110,418	48.5%	169,558	74.5%
Italian	378,133	176,836	46.8%	277,246	73.3%
Portuguese	302,843	176,836	58.4%	277,246	91.5%
Spanish	365,066	201,990	55.3%	265,435	72.7%

Table: Recall of multi-word terms in IATE by old and new term grammars

Results: Term Grammar Size

Language	Number of rules	Maximum term length
English	21	5
Estonian	61	5
French	47	8
German	73	6
Italian	40	7
Portuguese	64	9
Spanish	52	8

Table: Number of rules and maximum supported length of terms (in tokens) in the new term grammars

Finalization

- Optimization of rules
    - Use of macros
    - Combining similar rules
- Testing
    - Different domains and corpus sizes
    - User feedback
- Deployment
    - Installation in Sketch Engine
    - CC BY-NC license

Future Work

- New & evidence-based term grammars for more languages
    - All 24 IATE languages and beyond
    - Ukrainian, Arabic, ...
- Learning on running texts rather than isolated terms
    - Higher tagging accuracy
    - Non-canonical forms