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

Marek Blahuš̌ ${ }^{1}$, Michal Cukr ${ }^{1}$, Miloš Jakubíček ${ }^{1,2}$, Vojtěch Kovář ${ }^{1}, 2$, Vít Suchomel ${ }^{1,2}$<br>${ }^{1}$ Lexical Computing, Brno, Czechia<br>${ }^{2}$ Faculty of Informatics, Masaryk University, Brno, Czechia<br>E-mail: firstname.lastname@sketchengine.eu


#### Abstract

This paper presents a new generation of terminology extraction grammars for the OneClick Terms system. Unlike previous grammars built using linguistic judgment, the new grammars use rules inspired by patterns frequently observed in existing terminology databases. This evidence-based approach leads to a more accurate coverage of term candidates of lexical structures typical for authentic terms. The internal variety and maximum length of recognized terms have also increased. Due to the use of techniques known from corpus linguistics in their design, the resulting grammars maximize the coverage of terms while keeping a manageable size.


In the paper, we first describe how term grammars are used in OneClick Terms (Baisa et al., 2017) to enable terminology extraction for individual languages. Then we explain the procedure which we use to design next-generation term grammars for seven languages (English, Estonian, French, German, Italian, Portuguese, Spanish). This includes studying the IATE term base (Zorrilla-Agut \& Fontenelle, 2019) to derive information on the typical structure of terms in each language. Eventually, we provide figures concerning the new term grammars and their recall of the IATE terms, and we discuss directions for further development.

Keywords: terminology extraction; evidence-based term grammars; OneClick Terms; IATE

## 1. Introduction

Finding terms in a domain-specific corpus has been a feature of NLP tools for more than a decade (see, e.g., (Aker et al., 2013), (Gojun et al., 2012)). While many of such tools were designed as language-independent, the Sketch Engine corpus management system (Kilgarriff et al. 2014) has pioneered language-aware automatic term extraction for many languages, building on the belief that customization and collaboration with actual speakers of the language lead to higher-quality results (Jakubíček et al., 2014).

Currently, 29 languages are supported in both monolingual or bilingual term extraction. A dedicated web interface called OneClick Terms (Baisa et al., 2017) showcases the essential functionality of Sketch Engine and hides all the background complexity of corpus building, text alignment and the actual term extraction from the eyes of the user. All the user does is upload the document(s) and select the language(s), after which all computation happens seamlessly and the extracted terms are displayed as a result.

For each supported language, OneClick Terms needs a language-specific term grammar. A term grammar is a set of rules which defines the lexical structures, typically noun phrases, which should be included in term extraction. Earliest term grammars for Sketch Engine were prepared for the World Intellectual Property Organisation (WIPO) and typically had the form of a single part-of-speech-based regular expression (e.g. one or more adjectives followed by a noun, for English).

Because these grammars were prepared using linguistic judgment, they could only match term candidates of a limited variety and length. We deem this approach substandard and believe that applying the same principles that are common in corpus linguistics (i.e. statistically exploring large sets of data rather than relying on a linguist's intrinsic knowledge) would provide higher-quality term grammars. The idea is to observe which lexical structures are common in terminological databases, instead of coming up with a selection of our own. Obviously, not all sequences of tokens of an applicable lexical structure are terms, but the existing term extraction algorithm will take care of distinguishing actual terms from mere term candidates.

In this paper, we describe a new generation of terminology extraction grammars for the OneClick Terms system, which we developed with a strictly empirical approach by studying an existing manually curated terms base, namely the IATE (Interactive Terminology for Europe), created by the Translation Centre for the Bodies of the European Union with terms in 24 languages (Zorrilla-Agut \& Fontenelle, 2019). Since the rules are inspired by patterns observed in a terminology database, we call these grammars "evidence-based" term grammars. Our aim is to maximize the term grammar recall of the terms in the terminology database, which serves as the gold standard showing what people actually perceive as terms in the particular language. We have used this new approach to develop evidence-based term grammars for seven languages so far (English, Estonian, French, German, Italian, Portuguese, Spanish) and evaluate them in terms of improvement compared to the existing term grammars and coverage of the terms in IATE.

## 2. Background

The term extraction in OneClick Terms is based on a corpus-based contrastive technology involving two key steps: (1) applying the rules in the term grammar to a corpus to generate a list of term candidates (2) scoring the candidates by comparing their frequencies in the uploaded document(s) (which form a focus corpus) to their frequencies in general language represented by an extensive reference corpus (Jakubíček et al., 2014). OneClick Terms uses the corpora of the TenTen Corpus Family (Jakubíček et al. (2013), target size $10^{10}$ words) as reference. A later extension to the system allows for bilingual terminology extraction from aligned documents (Kovář et al., 2016) based on co-occurrences in aligned segments being ranked using the logDice association score (Rychlỳ, 2008). Recently, the support for bilingual extraction from non-aligned documents was added (Baisa et al., 2015).

We used Sketch Engine to build a single-purpose term corpus, consisting of multi-word terms from the current version (September 2021) of the IATE term base (all domains, all collections, only the "term" term type, any evaluation, any reliability), cleaned by removing any HTML markup (e.g. $<i>$ ), quotation marks, text in brackets, and even full entries if they look like a complex chemical formula (e.g. 6 -chloro-N-ethyl-N-(propan-2-yl)-1,3,5-triazine-2,4-diamine), a list of multiple terms (e.g. period of driving time, driving
period) or an incomplete term (e.g. to inform ... of ...). Each term is represented as a separate paragraph and the corpus is processed using the standard processing pipeline for the particular language, which includes part-of-speech tagging, lemmatization and morphological annotation.

A report (see Figure 1) is then generated using the Sketch Engine API showing the frequency distribution of part-of-speech combinations (e.g. adjective + noun) in the terms (paragraphs) in the term corpus, ordered by descending frequency. For each such combination, a second-level frequency distribution is computed on the morphological tags, revealing that, for instance, in languages with gender agreement, the combination masculine singular + masculine singular is much more frequent than masculine singular + feminine singular; the latter being either the result of incorrect tagging, or a random grouping of words (if found in a regular running text corpus) which do not form any lexical structure. For each part-of-speech and morphology combination, a list of one hundred random examples of matching terms is displayed to allow for quick inspection during the term grammar design process.

The imposed order within the report makes it possible for the term grammar author to focus on the most frequent patterns and provides hints at probable grammatically incorrect readings and other rare cases unworthy of attention. As a rule of thumb, only items with a relative frequency of at least $0.15 \%$ were considered for inclusion in the term grammar. At the same time, collaboration with a speaker of the language makes it easier to understand the observed patterns and generalize them where useful (e.g. enforcing an overall agreement in gender and number instead of running in the risk of omitting some less-frequent cases such as with the plural). On the other hand, some constraints need not be reflected in the rules, such as grammatical case governed by a preposition, because false positives seem to be rare and by not demanding a particular case we allow for possible incorrectly tagged terms to be included and the term grammar to be simpler.

Generalization, compromising and application of linguistic knowledge contribute to shortening the length of the resulting term grammar (i.e. lowering the number of rules), making the internal structure of the grammar easier to understand and also making the computation quicker. It is assumed that a breakdown of the gold-standard terms into part-of-speech and morphological tags is sufficient for the creation of term grammar rules. If, during the rule design or during later evaluation, it is observed that some constituent of a rule should be specified in more detail, it is possible to further limit the accepted words to certain lemmas or word forms (e.g. in most Romance languages, only a limited set of adjectives is permitted to appear in front of the noun they relate to), to enforce additional relationships between two constituents of a rule, or to limit the acceptable context (adjacent words) of a valid term candidate within running text.

## 3. Term Grammars

A term grammar is a carefully crafted set of rules (expressed in CQL, the Corpus Query Language (Jakubíček et al., 2010) describing the lexical structures, typically noun phrases, which should be included in term extraction. Noun phrases are manifested by the presence of a head noun, but their internal morphosyntactic structure is variable and by far not all sequences of words that include a noun are terms.

[^0]
## 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 • modal numbers • non-recurring expenses • numeric data • outdated data • peritrichous flagella - photo axes • polar latitudes • preliminary surveys • psychomotor activities • repetitive duties - residual stocks • self-supporting elements • settlable solids • short-horned grasshoppers • social dynamics • speculative damages • structural arrangements • super singles • tamping ties - toothed whales • undercover activities • urban centres • white pages...

### 2.3. JJ NP (940 terms, 0.15\%)

African Eve • Argentine Republic • Bosnian Serbs • Chocolate Point • Dedicated Target • Euro-Mediterranean Partnership • Euromediterranean Bank • European Commission • European Union • Feminist Initiative • Focal Point • Governmental Committee • Honourable Member • Injured Party • Legislative Council • MAb-based ELISA• Molecular Engineering • Neutral Red • Norwegian Trench • Permanent Secretary • Spanish Constitution • Standby Mode • Test-ban Treaty • Transatlantic Forum • Wet Sump • active NFE • anatomical MRI • asynchronous TTY-terminal • climate-neutral Union • competent Court • depletion-mode FET - dideoxy sequencing • east Berlin • far IR • flash EAROM• free-floating e-scooter • glacé ki - helical CT • lazy Susan • multilayer TVS • nationwide ISDN • postgenomic biosciences • prokaryotic promotor • regional programm • serial DAS • soft Brexit • total AMS • underfloor wheel-lathe • visual CAPTCHA • Hal Qormi...

Figure 1: Part of the report for English IATE data: The adjective + noun pattern is the second most common in multi-word terms. Majority of these terms are tagged JJ NN (i.e. adjective followed by a singular or mass noun) in the corpus. In some such terms, the noun is in the plural (JJ NNS). A few terms consist of the adjective followed by a proper noun, what is sometimes the result of inaccuracies in tagging due to the use of title case (e.g. Governmental Committee or Standby Mode) or due to the fact that acronyms such as $M R I, N F E$, or $C T$ are tagged as NP. Used tagset is the English TreeTagger PoS tagset with Sketch Engine modifications (see https://www.sketchengine.eu/tagsets/english-par t-of-speech-tagset//)

It should be noted that the full internal structure of a term candidate is usually not visible in Sketch Engine, because only shallow parsing is performed and the exact dependencies within a complex noun phrase may remain ambiguous. Such cases require our attention, because some isomorphic syntactic structures might erroneously be discarded if rule conditions have been set too tight (with only the prevailing structure in mind). For example, in French noun phrases of the type noun + preposition + noun + adjective, imposing gender and number agreement between the last two words (e.g. in gestion des exploitations agricoles, i.e. management of agricultural exploitations) is wrong, because the adjective can as well link to the noun in the first position (e.g. pardon des péchés obtenu, i.e. obtained forgiveness of sins).

Besides the CQL query that a sequence of words must match to produce a term candidate, each term grammar rule ensures that the term is represented in its canonical (citation) form. The tradition differs across languages, but it usually includes using the lemma for the head noun and its optional modifiers (Gojun et al., 2012). For many Romance and Slavic languages, the lemma used for adjectives must be gender-respecting (e.g. nuée ardente instead of nuée ardent in French, see Jakubíček et al. (2014)). The rules are even more complex in German (with its capitalized nouns and adjectives ending in suffixes corresponding to the gender of the related noun).

Full implementation of such rules may rely on special attributes present in the corpus. Examples of attributes that had to be added or modified include: corpus attributes for the comparison of the agreement in number and case, context-based disambiguation of non-conclusive gender and number in the output of the FreeLing tagger ${ }^{2}$, or an extension of gender-respecting lemmas to the past participle (while the past participle behaves like adjectives and appears within terms, its lemma used to be the verb infinitive).

In the formula describing the citation form, individual matched tokens are referred to by their labels (numbers) in the CQL query. For convenience, the numbering of tokens in the query is chosen so as to provide an idea about the syntactic structure of the noun phrase, starting from number 1 for the head noun (with necessary limitations, due to the fact that a single consecutive row of integers is used). In theory, tokens may be present in a different order in the citation form, but we have not found a need for this in any of the languages we have worked with. Sometimes, tokens from the query may be missing in the citation form, usually when they are used only as negative filters, e.g. to ensure that another noun does not follow a matched sequence of nouns, so that Centro Robert Schuman is considered a term candidate, but not its substring Centro Robert. Such negative restrictions are typically put in place only during the evaluation of a term grammar draft, because the term corpus itself does not contain such incomplete terms.

In most languages, the citation form of terms traditionally uses lower-case letters only. This is convenient in order to reconcile differences in letter case in the word forms (e.g. when a phrase is sometimes spelled in the corpus in Title Case) and to cope with the fact that the built-in lemmatization for some languages returns lower-case output only. Another peculiarity is that term grammar rules currently cannot enforce use of the plural for the headword of a citation form, although this is customary in some contexts. As such cases are difficult to recognize, this difference is disregarded and all terms' headwords are rendered in the singular, in turn producing occasional incorrect citation forms (e.g. united

[^1]state of america). We believe that a future addition to the OneClick Terms algorithm might improve the quality of citation forms generation, by taking advantage of their surface form frequency in order to generate correctly capitalized output in the correct number (e.g. United States of America).

```
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
```

Figure 2: Simplified example of a rule from the new Spanish term grammar, along with definitions of the used macros. The head noun in position 1 is output as lemma, the noun and adjective in positions 3 and 4 must agree in gender and number. The shown example term means "reduction of red eyes"

When writing a term grammar, we have found it useful to divide the rules into blocks, depending on the number of tokens in the produced term candidates (note that single tokens are not considered terms, but keywords). Within each such block of same-length rules, interactions among the rules are possible, which may lead to overlaps and possibilities to generalize. We try to order the rules within a block by decreasing frequency, although this constraint is sometimes broken in favour of similar rules (such as all starting with a noun) being listed next to each other. For the processing of the term grammar in OneClick Terms, the order of rules in the term grammar, as well as their possible overlaps are irrelevant.

To make orientation in the term grammar and the editing thereof easier, we make use of macros in the rule definitions and show example terms next to each rule. Macros such as noun (instead of [tag="NN"]) or modif (meaning noun or adjective) have been used also in the existing term grammars, ever since the adoption of the m4 macro language for term grammars has enabled this, but with the increased complexity of terms recognized by the next-generation term grammars, their usefulness and variety has risen substantially. One and sometimes more examples of terms matched by a rule are included as comments in the term grammar file and provide the possibility of noting that a noun phrase of a certain morphological structure may correspond to two or more syntactic structures, as already explained above.

Many times, incorrect tagging comes into play too, because some rules may partially or fully match terms that have been assigned incorrect part-of-speech or morphological tagging in the corpus. If this is the case, we note this fact in the term grammar by providing an extra example marked as such, but we do not feel obliged to cover all such cases, for the inconvenience of doing so and for the belief that in such cases the respective taggers should be improved instead.

## 4. Development

The initial design consisting of writing rules corresponding to the most frequently observed patterns in the term corpus is followed by testing the resulting term grammar draft against an actual focus corpus and a reference corpus. We have asked the collaborating speakers to come up with a domain-specific focus corpus of their own preference, expecting that subject knowledge can lead to better results. These focus corpora have varied in size from about 700,000 to $2,000,000$ tokens and most were built specifically for this purpose using WebBootCaT (Baroni et al. 2006). To speed up the iterative evaluation process (i.e. each change in the term grammar requires the terms to be recompiled for both the focus corpus and the reference corpus), we did not use the full standard reference corpus (i.e. one of the TenTen corpora), but a downsized sample thereof instead (approximately 200 million tokens) as a sufficient approximation.

Since for each processed language, there had been an existing term grammar before and our aim was to improve it, we did not stop at generating a list of terms in the focus corpus with the new term grammar, but we also ran term extraction from the same focus corpus with the old term grammar. Then we could visualize the differences by putting the two lists side by side and marking for each item in each list whether it is present in the other list or not, and if it is in both, then how much did its ranking (i.e. position in the list) change. See Figure 3 for an example of such comparison. A term's ranking can easily have changed due to factors such as inflection or incorrect tagging when different tokens (or differently tagged tokens) share the same citation form. For example, the old term grammar could only match the term in the nominative, while the new term grammar matches it in all cases (and outputs it in a lemmatized form, i.e. the nominative, thus increasing the term's frequency and therefore ranking).

It is natural that some term grammar rules produce more terms than others, and some terms may have been contributed to by multiple rules. In the regular list of extracted terms, it is impossible to make such distinctions. In order to evaluate each rule performance separately, we split the created term grammar into a set of single-rule mini-grammars and run term grammar extraction separately with each of them. This process is time-consuming (tens of minutes for longer term grammars), but it provides useful data not available in a different way. The term lists generated in this way can be combined to form the full grammar term list, with the extra information on which rule(s) produced each term. With such per-rule lists, it is also easy to spot when some rule does not produce any terms at all, which means it should be either fixed or discarded.

Importantly, per-rule lists allow us to quickly review the top-scored terms for each rule with the aim of making sure that no rule produces invalid terms with scores high enough so that they risk spoiling the overall list of terms. The presence of invalid terms is common due to noise in corpora (typos, foreign words, broken language etc.) and inaccuracies in the processing (incorrect tagging, inherently ambiguous rules, incorrectly created citation forms etc.) and we limit our effort to making sure that the top terms produced by each rule are correct. If a rule produces problematic output and all of it is low-scored (compared to the top scores found in the full list), it can be considered for deletion, because its removal is not going to substantially change the overall results of term extraction. All in all, the effort spent at fixing a rule should be proportional to the score of the terms it generates. The full list of all term candidates, produced by all rules as a whole, may contain tens of thousands of items and is never used in practice, because it is the normalized-frequency scoring which

1．pasta sfoglia
2．secondo piatto 수
3．primo piatto $\boldsymbol{\Lambda}-11$
4．ricetta facile＋1
5．pasta fillo－1
6．forno vegetariana $\uparrow$－3
7．tempi di cottura－
8．verdure in padella－
9．prossimo commento－-2
10．cookie salvi－
11．ricette antipasti－
12．torta in padella $\uparrow-54$
13．verdure miste－
14．cottura in padella $\boldsymbol{4}-17$
15．maria bonaccorso－
16．cottura in forno $\boldsymbol{\wedge}-2$
17．forno statico $\boldsymbol{\uparrow}$
18．padella antiaderente $\uparrow-2$
19．email necessario © -2
20．indirizzo email necessario $\boldsymbol{\text {－}}-2$
informazioni di profilo－
22．informazioni di profilo pubbliche－
3．profilo pubbliche－
ricette di antipasti－
pasta fredda $\boldsymbol{T}^{-8}$
piatto unico $\boldsymbol{4}-11$
campi obbligatori－
social login－
ultime ricette－
tempo di cottura +25
nostra ricetta－
tue impostazioni－
ricette vegetariane－
mio consenso－
patate in padella－
peperoni ripieni－
carta forno－-9
ricetta vegetariana $\boldsymbol{\square}+30$
pollo in padella T $_{-15}$
ricetta semplice $\boldsymbol{H}_{-3}$
antipasti veloci－
domus spa－
editoriale domus spa－
g．mazzocchi－
r．e．a．di milano－
cookie completa－
informativa cookie completa－
proprietà di maria bonaccorso －
．$x$ fonte－
．proprietà di maria－

```
pasta al forno +
```

pasta al forno +
pasta sfoglia +1
pasta sfoglia +1
ricetta facile 苗-1
ricetta facile 苗-1
secondo piatto +2
secondo piatto +2
5. tempo di cottura ©-2.
5. tempo di cottura ©-2.
6. pasta fillo + +1
6. pasta fillo + +1
.verdura al forno +
.verdura al forno +
ricetta vegetariana 苂-30
ricetta vegetariana 苂-30
forno vegetariana +3
forno vegetariana +3
cookie salvo +
cookie salvo +
prossimo commento +2
prossimo commento +2
antipasto veloce ©-90
antipasto veloce ©-90
pasta al forno vegetariana +
pasta al forno vegetariana +
primo piatto +11
primo piatto +11
torta salata -4-124
torta salata -4-124
6. verdura in padella <4-4641
6. verdura in padella <4-4641
17. antipasto sfizioso +-35
17. antipasto sfizioso +-35
18. cottura in forno + +2
18. cottura in forno + +2
19. forno statico
19. forno statico
20. padella antiaderente \#+2
20. padella antiaderente \#+2
21. email necessario \+ +2
21. email necessario \+ +2
2. indirizzo email necessario
2. indirizzo email necessario
informazione di profilo +
informazione di profilo +
informazione di profilo pubbliche
informazione di profilo pubbliche
+
+
. informazione di profilo pubbliche
. informazione di profilo pubbliche
fornite +
fornite +
network scelto in base +
network scelto in base +
profilo pubblica fornita +
profilo pubblica fornita +
network scelto +
network scelto +
social network scelto +
social network scelto +
profilo pubblica +
profilo pubblica +
cottura in padella +17
cottura in padella +17
verdura mista +
verdura mista +
pasta fredda + +8
pasta fredda + +8
impostazione sulla privacy +
impostazione sulla privacy +
cottura della pasta +
cottura della pasta +
verdura cotta - -506
verdura cotta - -506
piatto unico +11
piatto unico +11
filo d' olio +
filo d' olio +
ricetta antipasti +
ricetta antipasti +
patata al forno +
patata al forno +
dieta vegetariana \-41
dieta vegetariana \-41
cottura al forno +
cottura al forno +
ricetta semplice + +3
ricetta semplice + +3
patata in padella +
patata in padella +
carta da forno +
carta da forno +
carta forno \# +9
carta forno \# +9
peperone ripieno +
peperone ripieno +
metodo di cottura \ -28
metodo di cottura \ -28
campo obbligatorio +
campo obbligatorio +
50. antipasto vegetariano ©-181

```
50. antipasto vegetariano ©-181
```

Figure 3：Top of a side－by－side comparison of terms generated from an Italian vegetarian cuisine corpus using the old（left）and the new（right）Italian term grammar：Newly identified terms are marked with a plus sign，discarded terms are marked with a minus sign． Each type is further highlighted in a corresponding color（green and red，respectively）． For terms generated by both grammars，the difference in their ranking across the two sets is marked with an up arrow or a down arrow，followed by the change of ranking expressed as a signed integer．
makes term extraction in OneClick Terms so powerful, as it helps to distinguish actual terms from mere term candidates. Because of this, during the development, we only strive to have the first few hundred items in the list as clean as possible, increasingly tolerating noise further down the list.

## 5. Evaluation

In order to estimate the coverage of terms in IATE by the produced term grammars, we ran each rule's CQL query on the term corpus (with a restriction that the full paragraph/term must be matched) and calculated the number of unique matches in the output. When compared with the total number of terms in the term corpus, this says what portion of IATE terms is recalled by our term grammar. We ran the same calculation also with the old term grammar to be able to observe if there has been progress. Results for each language are shown in Table 1

| Language | IATE <br> 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 ${ }^{a}$ | 378,133 | 176,836 | $46.8 \%$ | 277,246 | $73.3 \%$ |
| Por- <br> tuguese | 302,843 | 176,836 | $58.4 \%$ | 277,246 | $91.5 \%$ |
| Spanish | 365,066 | 201,990 | $55.3 \%$ | 265,435 | $72.7 \%$ |

Table 1: Recall of multi-word terms in IATE by old and new term grammars

[^2]Please note that when performing these calculations, we did not consider in any way the selection bias of terms found in IATE, which might over-represent terms from a particular domain or of a particular lexical structure and thus make the results less applicable to general terminology extraction. The calculated numbers are also representative only of the term corpus, i.e. recognized isolated terms. More authentic results would be achieved if we were to search for these terms inside running text, in which they would be used in sentence context and possibly inflected.

The figures in Table 1 demonstrate that we have managed to achieve our goal, namely that we have improved the coverage of actual terms by OneClick Term's term extraction grammars. The observed differences of recall rank from $17.4 \%$ for Spanish to $49.3 \%$ for French. Except for Estonian, whose dataset in IATE is smaller by an order of magnitude, all other languages have more than $72 \%$ of the multi-word IATE terms covered by the
newly developed term grammars. Importantly, recall has risen from $57.8 \%$ to $79.5 \%$ of multi-word IATE terms for English, which is the most requested language by OneClick Term users.

Some factors that contribute to the recall not being $100 \%$ are:

- Ambiguous or incorrect tagging which hides important information that could be used to identify a term candidate
- Ambiguity in language and lack of information on syntax which makes it impossible to distinguish actual lexical structures from mere token sequences that span across syntactic borders
- Low-frequent patterns in term candidate structure that are ignored to reduce term grammar complexity
- Terms longer than the longest rule in the term grammar (e.g. $8.1 \%$ of the English IATE terms are longer than 5 tokens and $1.8 \%$ of terms are 10 tokens or longer, e.g. communal estate of husband and wife comprising only property acquired after their marriage)
- Terms of type deliberately not supported by the term grammar (verbal terms, e.g. Italian fare click - "to click" - constitute approximately $1 \%$ of IATE data but their inclusion in term extraction is questionable)

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

Table 2: Number of rules and maximum supported length of terms (in tokens) in the new term grammars

The size of each term grammar (expressed in the total number of rules in it), as listed in Table 2. depends on several factors:

- Precision with which rules were written (less strict rules often mean tolerance to small errors in tagging and lead to less complex term grammars while letting in no or very little extra noise)
- Level of detail in the used tagset
- Maximum term length defined in the term grammar (which itself is influenced by the following factor:)
- Variety of the language's morphology and syntax (e.g. Romance languages typically chain nouns by means of a preposition like $d e$ and possibly an article, so their
terms tend to be longer than English terms which often expresses the same with adjectives or noun juxtaposition)

In general, we strived to keep the number of rules a two-digit number in order to keep the term grammar friendly to a human editor and the computation of term candidates fast enough (each extra rule means an extra query that has to be executed on the corpus). The number of rules can be somewhat reduced during final optimization of the term grammar, e.g. by creating macros that combine conditions which are often both applicable in a context like having a macro meaning adjective or past participle, or by relaxing some rules in order to merge them with other similar rules without causing any actual damage by such generalization.

During finalization, each produced term grammar was tested with several other focus corpora, including different domains and one rather small corpus, to ensure that it performs reasonably well in real-life situations. The final term grammars are made available under the Creative Commons Attribution NonCommercial license. All the new grammars are already installed in OneClick Terms at the time of writing and can be used also in Sketch Engine. Feedback received from both creators and users of these tools suggests that the change has been to their satisfaction and that the quality of term extraction for these languages has noticeably improved in their opinion.

## 6. Future Work

The fact that we work with isolated terms is a source of inconvenience, both in the design stage and during the evaluation of a term grammar. In authentic use cases, terms are extracted from running text, composed of full sentences. In running text, terms can appear nested within more complex syntactic structures and possibly inflected. The collaborating speaker's linguistic knowledge is likely to mitigate this issue to some extent because of forethought. For instance, rules can be written with all grammatical cases in consideration, even if in the studied list of terms, only the nominative is used. However, if we were able to look up the IATE terms and their possible inflected forms inside full sentences, we could produce a performance estimate that would be more representative of real-life situations. Sentences containing the IATE terms in use could possibly be found and extracted from large corpora, such as those of the TenTen Corpus Family.

More strikingly, the inconvenience of using isolated terms manifests in the term corpus which we use as a gold standard. Although morphological taggers should in theory be able to handle non-sentences such as titles or list items and process them correctly, this is not always the case. For instance, the FreeLing tagger for Spanish had the tendency to sometimes mark nouns at the start of a term as verbs: e.g. in aduana de primera entrada ("customs office of first entry"), the first word is asserted to start with the third person singular of the verb aduanar, i.e. "(he) pays the customs", rather than the correct noun meaning "customs" or "customs office". Similarly, capital letters in proper names at the start of terms would get confused for sentence-start capitals, possibly influencing the tag assigned to the word (the FreeLing tagset distinguishes common and proper nouns).

In an effort to prevent these problems, we experimented with enclosing the terms into sentence frames before the tagging and removing these frames afterwards. For instance, English terms could be prefixed with the words I know the or Spanish terms with the
word Hay ("There is/are"), creating a full sentence in which the term constituents get tagged more accurately. It is, however, not always possible to come up with such universal sentence frame in a language which would work with all or almost all terms; many times, such a frame would need to be differentiated in form by the grammatical number or gender of the term that follows it, which is information unknown to us at the time and not easy to derive. Our research so far has been inconclusive in whether the creation and usage of such sentence frames is desirable and worth the effort.

There are also some intended deficiencies in the produced term grammars, due to situations we could not handle without letting in too much noise. Many terms that include a conjunction, mainly "and" or "or", are not covered by the new term grammars because these conjunctions are frequently used to join lexical structures and even sentences and therefore most of the output generated by rules that feature a conjunction would in fact be spanning across these syntactic borders and not represent an actual lexical structure. In rare cases, we could allow conjunctions in rules with confidence due to it clearly being situated inside, rather than possibly at the edge of a lexical structure. An example is the French système de séparation et de tri ("separation and sorting system"), in which the conjunction et ("and") is followed by the preposition de ("of"), indicating that it is joining the two attributes of the preceding headword (système).

The IATE term base is a unique, large and freely accessible source of terms in multiple languages, but an alternative needs to be identified when writing term grammars for languages not present in IATE. Our ongoing effort at developing a term grammar for the Ukrainian language has shown that resources similar to IATE are scarce and it might be necessary to adopt a different approach and start identifying terms where they are highlighted in running texts rather than collected in ready-made term bases.

## 7. Conclusion

We have designed a procedure for the creation of a new generation of term extraction grammars, which are inspired primarily not by someone's linguistic judgement, but by an existing term base such as IATE, which serves as a model of what lexical structures are likely to be considered terms by end users. The existing term base, which serves as a gold standard, also provides a way of evaluating the quality of the new term grammars. The development of each new term grammar happens in a standardized process, in the cooperation of a computer linguist with a speaker of the respective language. In the article, we have described possible challenges during term grammar design presented by specific languages or linked to cases of inaccuracies or ambiguities, along with recommendations of how they should be handled.

By the time of writing, we had produced such next-generation term grammars for seven European languages (English, Estonian, French, German, Italian, Portuguese, Spanish). Evaluation showed that recall indeed increased after the new grammars had been designed with IATE in mind, as on average three fourths of the (cleaned, multi-word) IATE terms can now be detected during term extraction. Most of these new evidence-based term grammars have been already installed in OneClick Terms and Sketch Engine and positive feedback from users confirms that they are actually getting higher-quality results than with the old term grammars. Lack of negative comments suggests that, while the number and versatility of term extraction rules increased, we managed to avoid polluting the term
extraction results with incorrect terms, or, more specifically, with sequences of words which are matched by some of the new rules and would be lifted high enough by OneClick Term scoring algorithm, but which would not be considered proper terms by the user.

We plan to produce term grammars for more languages using the described method in the future, including languages not represented in IATE. For other languages than the 24 covered by IATE, another similar term base or another approach at gold standard compilation will need to be identified.

## 8. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

We would like to thank Eleri Aedmaa, Hanna Barabakh, Kristína Koláčková, Ondřej Matuška, Merily Plado and Emma Romani for assistance with particular languages.

## 9. References

Aker, A., Paramita, M.L. \& Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 402-411.
Baisa, V., Michelfeit, J. \& Matuška, O. (2017). Simplifying terminology extraction: OneClick Terms. The 9th International Corpus Linguistics Conference.
Baisa, V., Ulipová, B. \& Cukr, M. (2015). Bilingual Terminology Extraction in Sketch Engine. In RASLAN. pp. 61-67.
Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlỳ, P. et al. (2006). WebBootCaT: a web tool for instant corpora. In Proceeding of the EuraLex Conference, volume 1. pp. 123-132.
Gojun, A., Heid, U., Weissbach, B., Loth, C. \& Mingers, I. (2012). Adapting and evaluating a generic term extraction tool. In LREC. pp. 651-656.
Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P. \& Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 53-56.
Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. \& Suchomel, V. (2013). The TenTen Corpus Family. In 7th International Corpus Linguistics Conference CL 2013. Lancaster, pp. 125-127. URL http://ucrel.lancs.ac.uk/cl2013/.
Jakubíček, M., Kilgarriff, A., McCarthy, D. \& Rychlý, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. PACLIC, pp. 741-47.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. \& Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1, pp. 7-36.
Kovář, V., Baisa, V. \& Jakubíček, M. (2016). Sketch engine for bilingual lexicography. International Journal of Lexicography, 29(3), pp. 339-352.
Rychlỳ, P. (2008). A Lexicographer-Friendly Association Score. In RASLAN. pp. 6-9.
Zorrilla-Agut, P. \& Fontenelle, T. (2019). IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. Terminology, 25(2), pp. 146-174.


[^0]:    ${ }^{1}$ https://www.sketchengine.eu/documentation/api-documentation/

[^1]:    2 https://nlp.lsi.upc.edu/freeling/

[^2]:    ${ }^{a}$ The existing Italian term grammar used the TreeTagger tagset, but because Sketch Engine was switching to FreeLing for Italian at the time, the new term grammar was written for this tagset. The figure for the old grammar in this table was produced by an unpublished rewrite of the old grammar for the new tagger.

