KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing

Nikola Ljubešić,* Tomaž Erjavec,* Darja Fišer†*

*Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si, tomaz.erjavec@ijs.si

†Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 6, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si

Abstract

This paper presents two datasets for supervised learning of terminology extraction. The first is focused on monolingual term extraction and is a lexicon-type dataset of Slovene term candidates labeled by four annotators. The second is focused on extracting and linking terms in different languages which are translations of each other. It contains sentences that satisfy patterns in which terms occur frequently with their translations, with manually annotated terms in English, Slovene and other languages, and links between terms and their translations. For each dataset we set up a baseline approach: for monolingual terminology extraction we train an SVM classifier, while for identifying terms in different languages we train a sequential CRF classifier. The datasets and the described baselines are made freely available.

1. Introduction

In this paper we present two new datasets for training term extraction tools developed in the scope of the Slovene national project KAS, *Slovene scientific texts: resources and description*.

KAS-term is a lexicon-type dataset containing term candidates extracted via morphosyntactic patterns from a selection of PhD theses written in Slovene. Each term candidate is annotated by multiple annotators. The dataset is meant to be used for supervised learning of ranking of term candidates extracted from Slovene texts.

KAS-biterm is a sentence-type dataset consisting of sentences that satisfy some patterns that are typical for terms and their translations into other languages such as "*ekstrakcija terminologije (angl. term extraction)*". These sentences are annotated for terms, partial terms and abbreviations in Slovene, English, or other language. Links between the Slovene terms and their terms or abbreviations in the other languages are encoded as well.

On both datasets baseline approaches are defined and evaluated: for monolingual terminology an instance-level SVM binary classifier is defined which uses various co-occurrence statistics as features, while for bilingual terminology a sequence-level CRF classifier is defined which uses context-based features and annotates each token in a candidate sentence with the respective category.

The rest of this paper is structured as follows: Section 2. gives the related work on terminology extraction and describes the KAS corpus of Slovene academic writing, from which the presented datasets are produced. Section 3. describes in detail the monolingual datasets and the implementation and evaluation of our baseline, while Section 4. does the same for the bilingual case. Finally, Section 5. gives some conclusions and directions for future research.

2. Related work

In this section we give a description of related work in monolingual and multilingual terminology extraction.

2.1. Monolingual terminology extraction

A broad overview of linguistic, statistical and hybrid approaches to automatic terminology extraction (ATE) is given in Pazienza et al. (2005).

The term recognition task is usually formulated as a two-step procedure (Nakagawa and Mori, 2003): candidate term extraction followed by term scoring and ranking. We also follow this approach for monolingual term extraction.

There is a number of ATE datasets already available. Handschuh and QasemiZadeh (2014) present ACL RD-TEC, a dataset for evaluating the extraction and classification of terms from literature in the domain of computational linguistics. The dataset is based on the ACL ARC corpus consisting of papers from the ACL anthology. From that corpus more than 83,000 term candidates are extracted via PoS-based filtering, n-gram-based techniques and noun phrase chunking. They are furthermore annotated either as non-terms, technology terms or non-technology terms. Out of the 84k terms, 22k were annotated as being valid while 62k were annotated as invalid. The authors report an observed agreement of 0.758 and Cohen's κ of 0.517, on a small double-annotated dataset of 250 terms.

A reference dataset for terminology extraction is the GENIA corpus consisting of 2,000 MEDLINE abstracts from scientific publications in biomedical literature that is accompanied by the annotations of 100,000 terms organized in a well-defined ontology (Kim et al., 2003). Another example of a bio-textmining dataset is The Colorado Richly Annotated Full Text Corpus (CRAFT), consisting of 97 articles from the PubMed Central Open Access subset annotated with biomedical concepts (Bada et al., 2012). The authors of the dataset measure weekly inter-annotator agreement (IAA), showing expected improvements through time, as well as an F1 IAA of above 90% after a few weeks / meetings for five out of six tasks. However, the tasks consisted of applying ontologies on text, and not of labeling terms as an open task.

Another reference dataset is a corpus for the evaluation of term extraction in the domain of automotive engineering (Bernier-Colborne and Drouin, 2014). The authors annotate running text, but allow for evaluation of extracted lists of term candidates.

Combining various statistical predictors in a supervised learning setting is a well known approach in natural language processing and has been also applied to the problem of automatic term extraction. Loukachevitch (2012) combines 16 features, and with their logistic regression combination improves the best single result by removing 30-50% of error, depending on the domain. Similarly, Conrado et al. (2013) show on three domain corpora of Portuguese that a combination of 19 features significantly outperform separate well known statistics for ATE.

A very similar problem to ATE is collocation extraction where Pecina and Schlesinger (2006) obtain 21.53% relative improvement when combining 82 association measures with respect to the best individual measure. They also show that feature selection can bring the number of features down to 17 without a significant loss in the evaluation metric.

2.2. Bilingual terminology extraction

Bilingual terminology extraction is typically performed on parallel data (Daille et al., 1994; Vintar, 2010). Another popular line of research is multilingual term extraction from semi-structured multilingual knowledge banks, such as Wikipedia, relying on explicitly encoded cross-lingual links (Gupta et al., 2008; Erdmann et al., 2008). However, since (extensive) parallel corpora and other types of multilingual knowledge sources are difficult to obtain for a lot of specialized domains, researchers are increasingly proposing approaches that extract terms from partially translated (Nagata et al., 2001) or comparable (Tanaka and Iwasaki, 1996) data, where they extract terms for each language separately and then perform post-hoc term pairing.

In this paper we take a different approach, identifying patterns that are used to express the Slovene term and its translation equivalent into English or another foreign language in largely monolingual scientific texts, thereby considering the task to be a classical sequence annotation task. A similar approach has been proposed by Bond (2008) who used a small set of manually defined patterns to extract bilingual term pairs from the web. Abekawa and Kageura (2009) and Abekawa and Kageura (2011) proposed an extension of this basic approach in which they first extract seed bilingual terms from the available parallel glossaries and then use the seed term pairs to identify typical patterns that are used between them, which then serve as the basis of the large-scale bilingual term extraction from the web.

2.3. The corpus

The KAS corpus (Erjavec et al., 2016) was collected via the Open Science Slovenia aggregator (Ojsteršek et al., 2014) which harvests the (meta)data of the digital libraries of Slovene universities and other research institutions. The corpus contains mainly Bachelors, Masters and Doctoral theses and comprises almost 1 billion tokens. The texts were extracted from PDF files, and, after some filtering and cleaning, were tagged with morphosyntactic descriptions (MSDs) and lemmatised with reldi-tagger[1] (Ljubešić and Erjavec, 2016) using its model for Slovene. Each text in the corpus is accompanied with extensive meta-data, containing also classificatory information, such as CERIF (Common European Research Information Format) keywords.

The current, preliminary, version of the KAS corpus contains 700 PhD theses (40 million tokens) from a large range of disciplines[2] and it is this subcorpus that was used as the textual basis for the experimental datasets presented in this paper.

3. Monolingual term extraction

3.1. The dataset

For the term extraction experiments presented here we focused on three fields: Chemistry, Computer Science, and Political Science, which we selected by matching them with their CERIF keywords, thus obtaining 48 PhD theses form Chemistry, 105 from Computer Science, and 23 from Political Science.

From these three subcorpora we sampled 5 PhD theses per area and automatically extracted term candidates, using the CollTerm tool (Pinnis et al., 2012) given a set of manually defined term-indicative MSD patterns. These patterns were initially developed for the Sketch Engine (Kilgarriff et al., 2014) terminology extraction module, and are in detail described in Fišer et al. (2016). For the present experiments we used only 31 nominal patterns, from unigrams and up to 4-grams, e.g. `Nc.*,S.*,Nc.*,Nc.*g.*` which finds sequences of *common noun, preposition, common noun, common noun in the genitive case*, such as *adheziv na osnovi topil* (*adhesive on basis (of) solvents = solvent-based adhesive*).

Each found term candidate was extracted in the form of its lemma sequence and the most frequent inflected phrase, keeping those that appear at least three times in a doctoral thesis. For manual annotation the candidates were first alphabetically sorted, in order to remove bias coming from frequency or statistical significance of co-occurrence, both types of information being provided by the CollTerm tool.

We produced a separate list of term candidates for each doctoral thesis. These lists were then annotated by four annotators. Annotators, who were graduate students of the three fields in focus, were asked to label each potential term with one of the five labels:

- *in-domain*: words and phrases that represent an in-domain term, i.e. one from the focus field;
- *out-domain*: words and phrases that represent a term from a field other than the one in focus;

[1] `https://github.com/clarinsi/reldi-tagger`

[2] The body parts of the KAS corpus and the KAS-Dr (PhD theses only) corpus are available for exploring through the concordancer at CLARIN.SI: KonText (`http://www.clarin.si/kontext/`) and noSketch Engine (`http://www.clarin.si/noske/`).

```
{
"document_id": "kas-845894",
"area": "Kemija",
"annotation_round": 1,
"lemmas": "gradient magneten polje",
"wordforms": "gradientom magnetnega polja",
"pattern": "Nc.*|A.*g.*|Nc.*g.*",
"length": 3,
"annotator_1": "t_termin",
"annotator_2": "t_termin",
"annotator_3": "n_nerelevantno",
"annotator_4": "n_nerelevantno",
"frequency": 7,
"tfidf": 0.11325,
"chisq": 0.79361,
"dice": 0.11079,
"ll": 0.25669,
"mi": 0.55263,
"tscore": 0.1324,
"cvalue": 11.09473
}
```


Figure 1: JSON encoded monolingual dataset entry

- *general*: vocabulary that is typical for academic discourse;

- *irrelevant*: words and phrases that belong to the general vocabulary, foreign-language expressions, definitions, fragments of terminology;

- *discuss*: borderline cases that needed to be discussed and resolved. These do not occur in the final dataset.

The instances of the dataset are thus term candidates annotated with the above categories, and various frequency and co-occurrence statistics. The final dataset consists of 22,950 such instances.

As illustrated in Figure 1, the fields of each instance are the thesis identifier, the scientific field, annotation round, lemma sequence, its most frequent surface form, morphosyntactic pattern, length in words and the manual annotations by annotator number 1 – 4. We also encode seven statistics calculated with the CollTerm tool during the term candidate extraction. These statistics are the frequency of the term candidate, and its tf-idf, χ^2, dice, log-likelihood point-wise mutual information and t-score values. Due to its popularity we also give the C-value (Frantzi et al., 2000), although this statistic is not based on co-occurrence, but the frequency of the term candidate and the frequency and number of other candidate terms containing that term candidate.

We distribute this dataset both in JSON and CSV formats. It is available from the CLARIN.SI repository (Erjavec et al., 2018b).

3.2. Baseline method

We set up a baseline for the task of predicting whether a candidate is a term or not given the variables available in the prepared dataset. We build the baseline as an SVM classifier with `scikit-learn` (Pedregosa et al., 2011)[3].

Given that we have four labels present in our dataset, we defined two mappings (inclusive and exclusive) of the four labels to a binary system of positive and negative classes. Both the inclusive and exclusive mappings take the irrelevant terms as instances of the negative class, but the inclusive mapping considers out-of-domain terms and academic vocabulary to be instances of the positive class, while the exclusive mapping considers them to be negative class instances. In the remainder of the paper we experiment with the more strict, exclusive mapping.

The explanatory variables we have at our disposal are the already mentioned frequency and seven co-occurrence statistics: *frequency*, *dice*, *chisq*, *ll*, *mi*, *tscore*, *tfidf*, and *cvalue*.

We consider the response variable to be the rounded average of the human responses, i.e., if three annotators claim an instance to be a term, and one annotator the opposite, the gold response for this term will be 1, i.e., the positive class. In (infrequent) cases where the average is 0.5, it is rounded up to 1.

We separate the prediction of multi-word terms (MWT) and single-word terms (SWT) as for single-word terms the only available variables are the frequency and the tf-idf statistic. For MWTs of all lengths all the seven variables are available.

We give the results on using single statistics, as well as the SVM classifier combining all the statistics in ranking multi-word term instances in a receiver-operating-characteristic (ROC) curve analysis in Figure 2. The ROC curve shows for each separate statistic to be surprisingly close to the random baseline (*baseline*), but that combining all these statistics in a supervised fashion (*all*) significantly improves the ranking of term candidates. If we quantify each ranking as an area under curve (AUC), our supervised baseline achieves a value of 0.736, while ranking by specific statistics achieves AUC scores between 0.505 (*tscore*) and 0.590 (*dice*).

For SWT ranking, where we have only two statistics at our disposal, namely *freq* and *tf-idf*, we calculated AUC scores for each separate statistic, as well as the ranking obtained through supervised learning on the two explanatory variables. The *freq* variable obtains an AUC of 0.523, *tfidf* performs much better with AUC of 0.703, while the combination of these two variables achieves an AUC of 0.613. Therefore as our baseline for SWT ranking we propose the *tfidf* statistic.

4. Bilingual term extraction

4.1. The dataset

The bilingual term extraction dataset contains complete sentences selected from all the PhD theses from the KAS corpus. We chose only sentences that have a high chance of containing the term in the original language and its translation into Slovene. The sentences were extracted using noSketch Engine via queries in its Corpus Query Language (CQL). After experimenting with various queries we

[3]The code of the baseline is published on `https://github.com/clarinsi/kas-term`

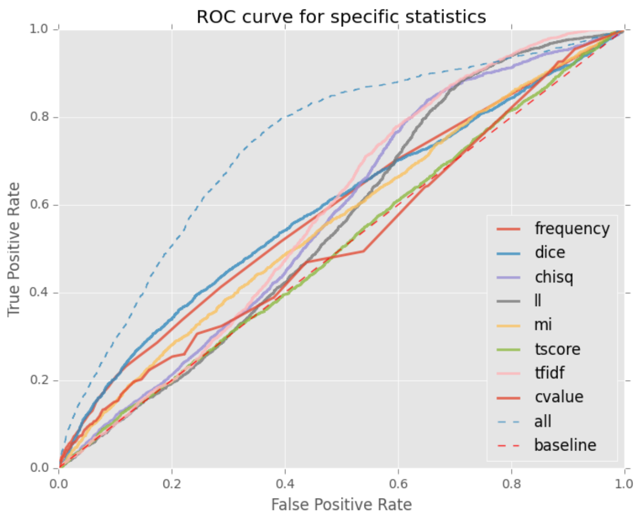

Figure 2: ROC curves for each of the variables in ranking multi-word term candidates, and for their combination *all*. The *baseline* is a random baseline.

then extracted the sentences with the following three CQL queries:[4]

1. `"\(" ".*"? "an\.|ang\.|angl\.|angleš.+" [tag="U"]? [word!="\)"]+ "\)"`

2. `[tag!="Nj"] "\(" "."? [tag="Nj" & word="...+"] [word!="\)" & word!="[0-9,]+"]* "\)"`

3. `"[a-zA-ZščžŠČŽ]+" "ali" "[\'\"]"`

The sentences retrieved by the queries were the basis for the manually annotated corpus. We first randomly sampled the results of the queries and then imported, for each query result separately, the sample into WebAnno (Yimam et al., 2013), a tool for Web-based manual annotation of corpora. Even though not all the annotations were used in the current baseline experiment, we, for the sake of completeness and possible further use, annotated the samples on the following levels:

- Type of term (*full term, partial term, abbreviation*): this distinction was made as the sample showed that the sentences often contain not only complete terms, but also terms which only partially cover its corresponding translation or original. Furthermore, the context of many terms or their translations also contains their abbreviation.

- Language of the term (*Slovene, English, Other*): even though our focus was on Slovene-English pairs, some found terms were also in other languages. We chose a middle road between ignoring these terms and marking them with their actual language, by assigning them all the *Other* language.

- Link between the term and its translation or between the term and its abbreviation (*link*): as the final goal is to automatically link terms and translations, the manual annotation of the link between the two is essential.

[4]Rather than explaining each query, we give links for returning the shuffled results of the three queries, in order: `http://hdl.handle.net/11346/clarin.si-ZNAN`, `http://hdl.handle.net/11346/clarin.si-7GRN`, `http://hdl.handle.net/11346/clarin.si-WHDX`.

Each sentence was annotated by two annotators and then the differences in annotation were resolved by the curator. Table 1 gives the statistics over the dataset, by query and in total. The numbers of sentences and tokens show that the queries had a significantly different yield, while the ”Marked” column gives the number of sentences in which something was annotated, i.e. they contained either a term or abbreviation with its translation; the last query thus not only returned the least sentences, but even the ones returned were typically not marked. The next three columns give the distribution by the type of the entity marked: in all cases, complete terms predominate, with abbreviations being about one tenth as frequent, and partial terms even less. Finally, the last three columns give the distribution by language: naturally, the Slovene and English items are quite similar in size, with other languages representing a very small minority.

The dataset was exported from WebAnno and merged with the source TEI encoding of the corpus as illustrated bin Figure 3. Here, the type of term is distinguished by the name of the element (`abbr` or `term`) and, in the case of terms, its `@type` attribute (`complete` or `partial`)q, while the language is distinguished by the value of the standard `@xml:id` attribute. Furthermore, the value of the `@subtype` gives the tag as it was used in WebAnno. The linkings are made via the `@corresp` attribute, which points to the value of the `@xml:id` attribute of the relevant term(s) or abbreviation(s). It should be noted that all the pointers are two-way.

The dataset is freely available in the scope of the CLARIN.SI repository (Erjavec et al., 2018a).

4.2. Baseline

Given that in this task we have running text instances annotated per token with term information, we frame this task as a sequence labeling task. Similar as with the task of monolingual term prediction, we use the traditional method applicable given the type of data: we use CRF, in particular the CRFSuite implementation (Okazaki, 2007). The baseline is published on `https://github.com/clarinsi/kas-biterm`.

Since the first pattern is the most productive one, as well as having a much higher precision than the remaining two patterns, we run the baseline experiments only on the 1,000 sentences following that pattern. The goal of the defined baselines is, namely, not only to set the stage for future experiments, but also to produce systems that will be easily applicable to various datasets, starting with the full KAS corpus. We split the available instances 80:20 into a training and a testing set.

We experimented with various features and, given the results of our experiments, we kept the following ones:

Query	Tokens	Sents	Marked	Complete	Partial	Abbrev	sl	en	und
q1	36,716	1,000	864	2,134	141	299	1,159	1,392	23
q2	34,773	787	427	1,324	51	169	696	707	141
q3	7,002	165	36	81	1	1	40	39	4
Σ	78,491	1,952	1,327	3,539	193	469	1,895	2,138	168

Table 1: Statistics over the BiTerm dataset

```
<abbr xml:id="patt1-001.abbr.2"
      xml:lang="en"
      corresp="#patt1-001.term.9
               #patt1-001.term.8"
      subtype="4AbbrEng">
  <w lemma="msd" ana="msd:Ncmsn">MSD</w>
</abbr>
<c> </c>
<term xml:id="patt1-001.term.8"
      xml:lang="sl"
      type="complete"
      corresp="#patt1-001.term.9
               #patt1-001.abbr.2"
      subtype="2TermSlv">
  <w lemma="oblikoskladenjski"
     ana="msd:Agpfsg">oblikoskladenjske</w>
  <c> </c>
  <w lemma="oznaka"
     ana="msd:Ncfsg">oznake</w>
</term>
<c> </c>
<pc ana="msd:Z">(</pc>
<w lemma="angl." ana="msd:Y">angl.</w>
<c> </c>
<term xml:id="patt1-001.term.9"
      xml:lang="en"
      type="complete"
      corresp="#patt1-001.abbr.2
               #patt1-001.term.8"
      subtype="1TermEng">
  <w lemma="Morpho"
     ana="msd:Npmsn">Morpho</w>
  <c> </c>
  <w lemma="Syntactic"
     ana="msd:Npmsn">Syntactic</w>
  <c> </c>
  <w lemma="Description"
     ana="msd:Npmsn">Description</w>
</term>
<pc ana="msd:Z">)</pc>
```

Figure 3: Example of a TEI bilingual term annotation for the segment *MSD oblikoskladenjske oznake (angl. Morpho Syntactic Description*)
.

- focus token: lowercased token for which features are currently extracted
- focus MSD: morphosyntactic description of the focus token
- focus PoS: part-of-speech of the focus token (first two letters of the morphosyntactic description tag)
- focus token length: number of characters in the focus token
- focus token case (lower, upper, title)
- lower cased tokens in a -3...3 window
- PoS tags in a -3...3 window

While performing baseline experiments, we calculated the informativeness of each feature set by performing ablation experiments. We ablated specific features, but also the set of features based on the focus token and the set of features based on the context window. We present the results of the ablation experiments in Table 2.

The results show that the most relevant feature sets are those of the focus token's context window, with the largest loss being when all window features are removed (8.89% relative loss), followed by the setup where all focus token features are removed (2.65% relative loss). Removing specific features generates a relative loss ranging between 1% and 0.1%.

We also experimented with other features, but they decreased our results. These are the features with their relative loss when added to the optimal feature set:

- focus token character 5-grams (best performing length), extended with a initial and ending character (loss of 0.2%)
- MSDs in a -3...3 window (loss of 0.3%)
- 100 embedding dimensions learnt from the slWaC corpus with fasttext using the skipgram model (loss of 0.3%)

The most surprising among the negative results is the loss when word embedding features are added to the sequential classifier. This result can probably be explained with the sensitivity of the CRF classifier to irrelevant features as most of the embedding dimensions do not hold any relevant information for the task at hand.

The full results of our best performing system (comparable to the system in ablation experiments with no ablated features) are presented in Table 3. As expected, the SL-ABBR class performs the worse as the number of tokens annotated with this label is by far the lowest. The class EN-TERM is better predicted as the class SL-TERM, which is also not surprising as identifying the borders of an English term in Slovene text is much easier than the borders of a Slovene term. Regarding the balance between precision and recall, there are no surprises with a good overall balance.

Ablated features	0	SL-TERM	SL-ABBR	EN-TERM	EN-ABBR	weighted
support	6177	601	10	527	66	7381
none	0.969	0.789	0.000	0.896	0.683	0.945
focus token	0.968	0.778	0.000	0.896	0.634	0.943
focus MSD	0.968	0.776	0.000	0.892	0.710	0.943
focus PoS	0.966	0.755	0.000	0.890	0.708	0.940
focus length	0.968	0.773	0.000	0.894	0.698	0.943
focus case	0.969	0.778	0.000	0.895	0.650	0.944
all focus token	0.957	0.702	0.000	0.815	0.452	0.920
tokens in window	0.964	0.733	0.000	0.894	0.625	0.936
PoS in window	0.968	0.771	0.000	0.896	0.672	0.943
all window	0.924	0.289	0.000	0.845	0.370	0.861

Table 2: Ablation experiments over the feature sets used for bilingual term extraction. The labels the results are given for are: O (other), SL-TERM (Slovene term), SL-ABBR (Slovene abbreviation), EN-TERM (English term), EN-ABBR (English abbreviation)

Metric	0	SL-TERM	SL-ABBR	EN-TERM	EN-ABBR	weighted
precision	0.965	0.839	0.000	0.872	0.737	0.945
recall	0.974	0.745	0.000	0.920	0.636	0.947
F1	0.969	0.789	0.000	0.896	0.683	0.945

Table 3: Final experiment on bilingual term extraction

5. Conclusions

We presented two newly developed manually annotated datasets for Slovene: the KAS-term dataset for learning monolingual term extraction and the KAS-biterm dataset for learning bilingual term extraction.

We set up baseline approaches with good, far from random results. However, we strongly believe that these results can further be improved and encourage other researchers and NLP practitioners to improve over these baselines and share their results.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful comments. We are also indebted to Špela Arhar Holdt and Maja Bitenc for conducting the annotation campaigns, to the the students who participated in the annotation process, and to their supervisors, Urban Bren, Marko Robnik Šikonja, and Boštjan Udovič. The research described in the paper was supported by the project ARRS J6-7094 “Slovene scientific texts: resources and description”.

6. References

Takeshi Abekawa and Kyo Kageura. 2009. Qrpotato: A system that exhaustively collects bilingual technical term pairs from the web. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 115–119. ACM.

Takeshi Abekawa and Kyo Kageura. 2011. Using seed terms for crawling bilingual terminology lists on the web. *Trans. Comp*.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner Jr., K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13:161.

Gabriel Bernier-Colborne and Patrick Drouin. 2014. Creating a test corpus for term extractors through term annotation. *Terminology*, 20:50–73.

Francis Bond. 2008. Extracting bilingual terms from mainly monolingual data. In *14th Annual Meeting of the Association for Natural Language Processing*, Tokyo.

Merley Conrado, Thiago Pardo, and Solange Rezende. 2013. A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 16–23.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics – Volume 1*, pages 515–521. Association for Computational Linguistics.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An approach for extracting bilingual terminology from Wikipedia. In *International Conference on Database Systems for Advanced Applications*, pages 380–392. Springer.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, and Milan Ojsteršek. 2016. Slovenska znanstvena besedila: prototipni korpus in načrt analiz (Slovene Scientific Texts: Prototype Corpus and Research Plan). In *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana University Press, Faculty of Arts.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, and Maja Bitenc. 2018a. *Bilingual terminology extraction dataset KAS-biterm 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1199.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, and Špela Arhar Holdt. 2018b. *Terminology identification dataset KAS-term 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1198.

Darja Fišer, Vit Suchomel, and Miloš Jakubiček. 2016. Terminology extraction for academic Slovene using Sketch Engine. In *RASLAN 2016: Recent Advances in Slavonic Natural Language Processing*, pages 135–141.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, Aug.

Anand Gupta, Akhil Goyal, Aman Bindal, and Ankuj Gupta. 2008. Meliorated approach for extracting bilingual terminology from Wikipedia. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on*, pages 560–565. IEEE.

Siegfried Handschuh and Behrang QasemiZadeh. 2014. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In *COLING 2014: 4th International Workshop on Computational Terminology*.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36, Jul.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*, pages 180–182.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Natalia V. Loukachevitch. 2012. Automatic term recognition needs multiple evidence. In *Proceedings of the Eighth Conference on Language Resources and Evaluation, LREC 2012*, pages 2401–2407.

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the web as a bilingual dictionary. In *Proceedings of the workshop on Data-driven methods in machine translation-Volume 14*, pages 1–8. Association for Computational Linguistics.

Hiroshi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

Milan Ojsteršek, Mojca Kotar, Marko Ferme, Goran Hrovat, Mladen Borovič, Albin Bregant, Jan Bezget, and Janez Brezovnik. 2014. Vzpostavitev repozitorijev slovenskih univerz in nacionalnega portala odprte znanosti (The Set-Up of the Repository of Slovene Universities and the National Portal of Open Science). *Knjižnica*, 58(3).

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Maria Pazienza, Marco Pennacchiotti, and Fabio Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. *Knowledge mining*, pages 255–279.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL*, pages 651–658, Stroudsburg, PA, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*.

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 580–585. Association for Computational Linguistics.

Špela Vintar. 2010. Luščenje terminologije iz angleškoslovenskih vzporednih in primerljivih korpusov (Terminology mining from English-Slovene parallel and comparable corpora). In Špela Vintar, editor, *Slovenske korpusne raziskave*, pages 37–53. Znanstvena založba Filozofske fakultete, Ljubljana.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August. Association for Computational Linguistics.