EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Skema: A New Tool for Corpus-driven Lexicography

Baisa V.[1], Tiberius C.[3], Ježek E.[2], Colman L.[3], Marini C.[2], Romani E.[2]

[1] *Lexical Computing, Czech Republic*
[2] *University of Pavia, Italy*
[3] *Instituut voor de Nederlandse Taal, The Netherlands*

Abstract

In this paper, we describe the development of Skema and its features. Skema [ˈski:mə] is a new corpus pattern editor system which supports the manual annotation of concordance lines with user-defined labels (each concordance has its own set of labels) and the editing of the corresponding patterns in terms of slots, attributes, examples and other features following the lexicographic technique of Corpus Pattern Analysis. Skema is integrated into the web-based Sketch Engine and can be used by any user for annotating both preloaded and user corpora. Each annotation label is linked to the pattern structure (stored in JSON format) which can be easily customized to individual projects, a generic pattern structure (i.e. a list of user-defined attributes) being available by default. The paper illustrates the use of Skema in three specific projects, i.e. *Woordcombinaties* for Dutch verbs, *Typed Predicate-Argument Structures* for Italian Verbs (T-PAS) and its sister project for Croatian Verbs (CROATPAS).

Keywords: corpus-driven lexicography; editor; pattern dictionary; Sketch Engine; corpus annotation; annotation schema

1 Introduction

Skema [ˈski:mə] is a new corpus pattern editor system. It was implemented to facilitate the management of manual annotations in Sketch Engine (hence the name) (Kilgarriff et al. 2014) allowing to associate word meaning with word use as is advocated by Corpus Pattern Analysis (Hanks 2013).

Corpus Pattern Analysis (CPA) is a lexicographic technique for mapping meaning onto words in text. It is based on the Theory of Norms and Exploitations (Hanks 2004; 2013). This theory distinguishes between normal or prototypical uses of words and their exploitations, like patterns with anomalous collocates or unconventional metaphors. The focus of the analysis is on the prototypical syntagmatic patterns with which words in use are associated. Associating a "meaning" with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns.

In this paper, we describe the development of Skema and its features, and we illustrate the use of Skema in a number of projects, i.e. *Woordcombinaties* for Dutch verbs (Colman & Tiberius 2018), *Typed Predicate-Argument Structures* for Italian verbs (T-PAS; Ježek et al. 2014) and its sister project for Croatian CROATPAS (Marini & Ježek 2019).

2 The Skema editor

Sketch Engine has supported the manual annotation of corpora (concordance lines) for quite a long time (Baisa et al. 2015), but the management of annotations and the integration in Sketch Engine was clumsy and unstable, since different systems on different servers were involved. Therefore, a new interface with a modern look was implemented, which - more importantly - is much more stable and easier to maintain than the previous system. Below, we describe the two main components of Skema: the manual annotation of concordance lines (section 2.1) and the editing of the patterns (section 2.2).

Figure 1: From the left: annotating concordance lines, word sketch collocates and the annotation menu in the Sketch Engine interface

2.1 Manual annotation

Using the CPA technique, the lexicographer starts with analyzing and annotating a sample (usually 250 lines or more) of concordance lines with labels using the annotation feature ( ) in the Sketch Engine interface. Patterns are identified manually by carefully examining similar concordance lines. The labels are chosen from a small pop-up menu (see Figure 1). It is common practice to use numerical labels for the patterns, but any string can be used; labels with a dot (e.g. 1.a) are treated as sublabels and are grouped together under the main label.
It is important to note that each annotated concordance has its own set of labels. In the projects described below, the concordances are based on verb headwords, but any concordance (a query result) can be stored for later annotation.
The set of labels is maintained in the Annotation menu (Figure 1, on the right) which provides an overview of the labels together with the number of concordance lines annotated with that label and allows the user to add new labels, to sort the whole concordance by the labels or to show only lines annotated with a specific label. The user can also go to the Annotation manager, which is on a separate page (Figure 2).
The annotation is not only available in Concordance, but also in Word Sketch (Figure 1 in the center). Annotating per collocate can significantly speed up the process, since the label is assigned to all concordance lines containing the headword-collocate pair.

2.2 Editing patterns

Once the annotation is done, the lexicographer starts editing the patterns in the Annotation manager. Here, lexicographers have an overview of the list of all headwords (Figure 2). They can search the list and per headword, they can maintain the list of labels identifying the patterns (Figure 2 at the bottom). Labels can be renamed, added, removed and reordered in the Annotation manager, and these operations are synchronized with Sketch Engine.

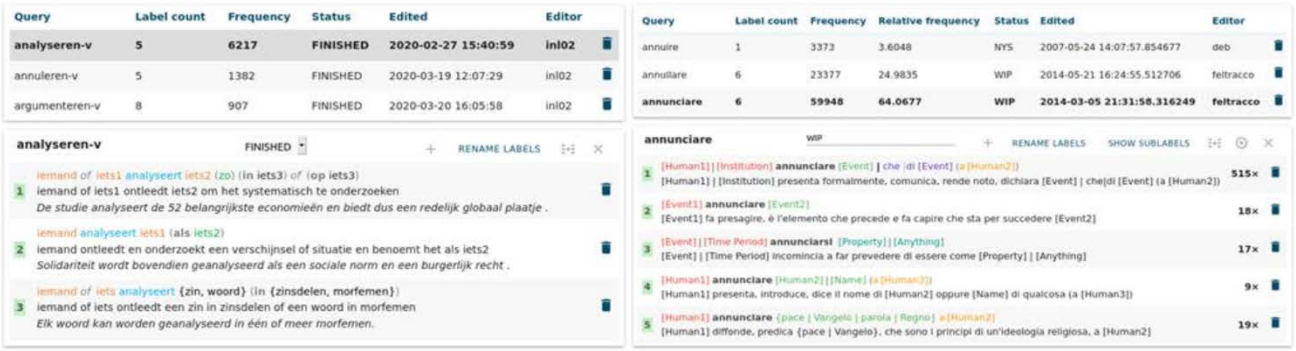

Figure 2: The list of concordances and a list of labels with different pattern visualization (Dutch labels and patterns for *analyseren* 'to analyze' and Italian labels and patterns for *annunciare* 'to announce')

When a pattern has been edited, a preview of the pattern is shown next to the label. Each pattern corresponds to one of the labels used in the manual annotation. The visualization of patterns is also project-specific and can be customized. By clicking on one of the labels, the pattern (structured information) opens up and can be edited (see Figure 3).

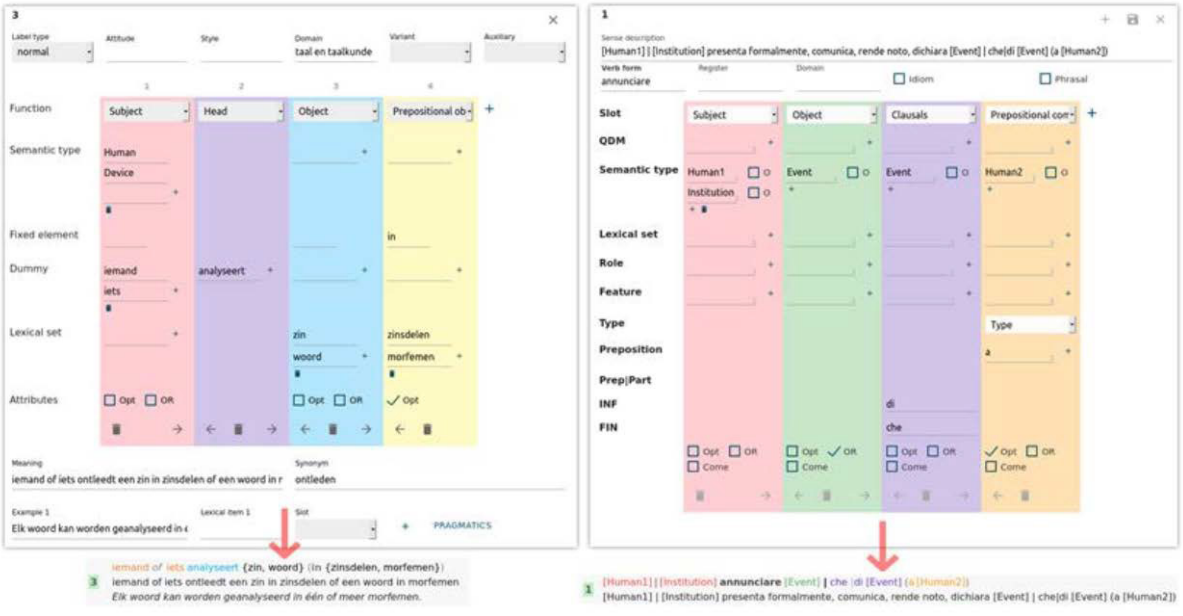

Figure 3: Pattern editors for Dutch and Italian; the generated patterns at the bottom can be customized with colors, typography etc.

Editing is done by selecting the right number of slots for a specific pattern and completing the information for the relevant features for each slot (e.g. syntactic function, semantic type, lexical set). Each project can define its own features for the

slots. Each project can thus have a different information structure linked to the slots and the different projects currently using Skema do indeed take advantage of the possibility to fine-tune the information in the slots to the particular needs of the project.

2.3 Skema: the technical solution

In the old version (called CPA editor; Baisa et al. 2015), the structured information was stored in a PostgreSQL database in several tables and every structural change in the pattern structure had to be reflected in the DB schema. Due to frequent changes, the schema became clumsy and hard to maintain. In Skema, the whole pattern structure is saved in JSON format in a SQLite database (one DB per corpus and project), so that changes are easily done at the level of the Skema pattern editor without a need to change the DB schema.

To use a customized pattern editor, Sketch Engine users need to be assigned to a specific project by Sketch Engine administrators. Otherwise, users will see a generic pattern editor with an option to save an arbitrary list of attribute-value pairs.

The system is currently not well-suited for parallel annotation by several annotators. Even though multiple annotators can work on different queries (stored concordances), the situation when more annotators (within one project and in one specific corpus) are editing the same concordance and changing the labels is not treated well at the moment and might lead to inconsistencies and conflicts. In the future, these situations can be treated by locking the annotation temporarily.

In each project, the selected queries and their labels can be published as a read-only single-page website. At the moment, only two of the projects using Skema have such access: the English Pattern Dictionary of English Verbs[1] and the Italian T-PAS.[2] Currently, it is not possible for users to publish their own data, but Sketch Engine administrators can set up a new single-page website with a unified user interface (with simple customization options) similar to the two examples above.

3 Projects using Skema

In this section, we illustrate the use of Skema in three ongoing projects.

3.1 Dutch *Woordcombinaties* project ('Word combinations')

Woordcombinaties is a new online lexicographic resource from the Dutch Language Institute,[3] which merges a pattern dictionary of Dutch verbs, following the example of the Pattern Dictionary of English Verbs, with a collocation application, following the example of Sketch Engine for Language Learning (SkELL).[4] A demo[5] of the resource has recently been released describing the combinatorics of a selection of 150 verbs taken from a list of high frequency verbs for advanced learners of Dutch as a second language. For all verbs, example sentences and a kind of word sketch are provided, and for a subset, a pattern description is also available. The project is based on a corpus of approx. 230 million tokens consisting of newspaper material and domain specific texts from the Netherlands and Belgium, in order to reflect language variety in Dutch.

In the editorial process, Skema is used for editing patterns, whereas an in-house system (Tiberius et al. 2014) is used for editing the example sentences and word sketches. The data from both editors is integrated in the online *Woordcombinaties* application. The basic setup of Skema for *Woordcombinaties* is fairly similar to the other projects. A pattern consists of a number of slots and each slot has a number of features and attributes attached to it. Specific to the Dutch project are the features 'fixed element' and 'dummy'. Dummies (such as *iemand* 'someone', *iets* 'something') are used in addition to semantic types for the sake of readability. This practice is inspired by E-VALBU,[6] where complements in the patterns are also embedded in dummies so that semantic roles are more or less implicitly recognizable. In the patterns in the online application, only dummies are shown, not the semantic types. The fixed element was introduced to have a placeholder for prepositions and conjunctions separate from the dummy in prepositional complements and predicative modifiers as is illustrated in the pattern below, where there are two optional prepositional complements, one introduced by *in* ('in') and the other by *op* ('on').

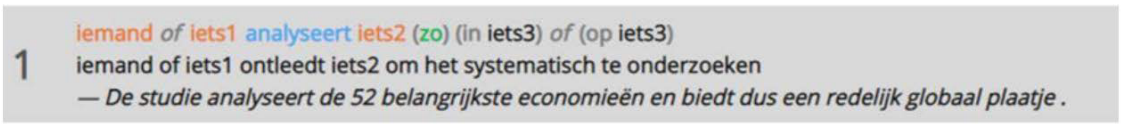

Figure 4: Pattern 1 of Dutch *analyseren* ('to analyse')

The visual rendering of the patterns in Skema has been customized completely to the *Woordcombinaties* project and is very similar to the layout and color-coding used in the online application of the project, providing a WYSIWYG preview to the lexicographer. For instance, the OR feature is displayed as the Dutch word *of* ('or') in the pattern (see Figure 4), and different syntactic functions are marked by different colors. Note also that the pattern uses the inflected form of the verb.

[1] pdev.org.uk [04/05/2020]
[2] tpas.sketchengine.eu [04/05/2020]; The T-PAS project is currently using this feature only internally and enriching it with good examples (GDEX) from the corpus for each label. The results will be made public by the end of 2020.
[3] ivdnt.org [04/05/2020]
[4] www.sketchengine.eu/skell/ [04/05/2020]
[5] woordcombinaties.ivdnt.org/ [04/05/2020]
[6] grammis.ids-mannheim.de/verbvalenz [04/05/2020]

In addition to the pattern, a definition and a general example are given.
In *Woordcombinaties*, sublabels are used to distinguish subpatterns from main patterns. Subordinate clauses, idioms, proverbs and formulas are normally considered as subpatterns. For instance, the pattern in Figure 4 has two subpatterns: one where the object slot of the main pattern is realized by a subordinate clause introduced by *of* ('or') or a wh-word, and one where the object is realized by a quote.
In the *Woordcombinaties* version of Skema, patterns are complemented by two types of example sentences, a general example and selection of examples of lexical items instantiating a particular slot. The slots in a pattern are numbered and the active slot is highlighted so that the (GDEX sorted) selected examples are automatically linked to this slot (see slot 6 in Figure 5). This numbering of the slots is especially important if a slot with a particular syntactic function occurs more than once in a pattern, as in the example below.

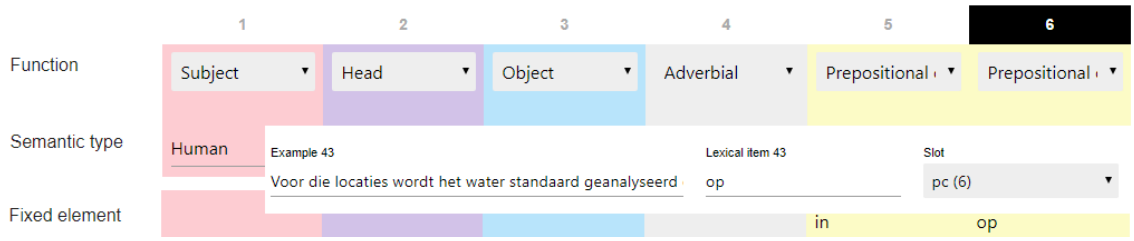

Figure 5: Pattern for Dutch verb *analyseren* 'to analyze' illustrating automatic linking of slots to example sentences

Both types of example sentences are stored in the JSON file which can be downloaded from Skema. In addition to this, the full set of annotated concordances in the corpus is extracted through the Sketch Engine API and shown on demand to the user in the online version of *Woordcombinaties*.

3.2 The Italian T-PAS project

The T-PAS resource is a corpus-derived inventory of semantic structures for Italian verbs to be used for linguistic analysis, language teaching, and computational applications. It is developed at the Department of Humanities of the University of Pavia, in collaboration with the Fondazione Bruno Kessler (FBK, Trento) and the technical support of Lexical Computing ltd. It is based on the Generative Lexicon theory of compositionality (Pustejovsky 1995) and on the corpus pattern lexical analysis proposed in Hanks (2004, 2013). It currently consists of 1160 analyzed verbs for about 8000 patterns, and ca. 190,000 annotated concordances. The verb sample of T-PAS was selected according to two criteria: a random sample of average polysemy verbs from the Sabatini Coletti 2008 dictionary (10% of 2 sense verbs, 60% of 3-5 sense verbs, 30% of 6-11 sense verbs), and coverage of the fundamental verb lemmas ("lemmi fondamentali") from De Mauro 2000.[7] The corpus used to extract the patterns is the itWaC reduced (935,698,409 tokens), a wide corpus gathered by crawling texts from the Italian domain in the web using medium frequency vocabulary as seeds (Baroni et al. 2009). The resource includes a repository of patterns, a hierarchically organized system of semantic types to classify the semantic properties of the verbal arguments, and a corpus of annotated concordances with pattern numbers, that represent instantiations of the corresponding patterns.
The T-PAS System of Semantic Types (Jezek 2018) is a hierarchy of general semantic categories obtained from manual clustering of lexical items found in the argument positions of verbal structures in the corpus. The System currently contains 180 semantic types that are organized hierarchically based on the "is a" (subsumption) relation (e.g., [Human] is an [Animate]). The System of Semantic Types, together with definitions and examples for each type, is made accessible to lexicographers through a customized function of Skema (Figure 6), so that they can easily and instantly consult it while editing the patterns.

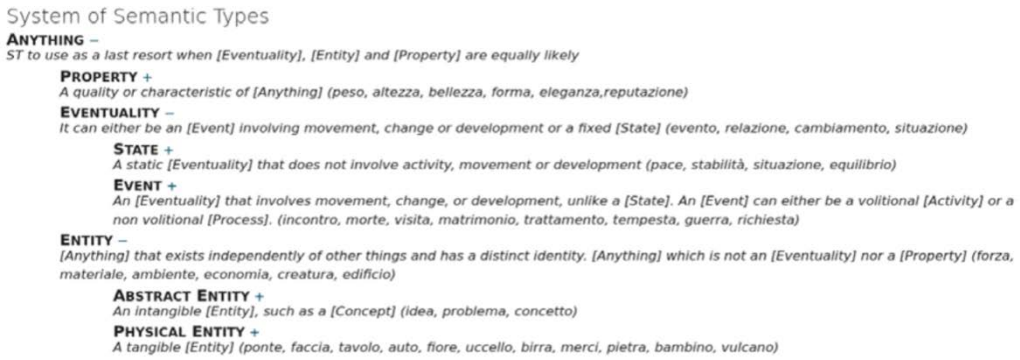

Figure 6: A selection of the top level of the hierarchy for the T-PAS System of Semantic Types in Skema.

Importantly, pattern strings in the T-PAS customized version of Skema only show semantic and lexical information, that

[7] In De Mauro's classification, fundamental lemmas are the words that in all languages tend to cover on average about 90 percent of the occurrences of words in texts and discourse.

is, they include the verb, the semantic type of the arguments, a selection of the best examples of lexical items for the types, the role played by the arguments (i.e. Athlete, Doctor), the features (i.e. Female, Visible) associated to the types, and a preposition or a complementizer (*a*, *per*, *di*), should they be present in the pattern. The syntactic information is encoded in Skema, but it is deliberately not made visible in the pattern string, neither in Skema nor in the online version. This is because the resource is intended primarily as a semantic resource. The syntactic features available for pattern encoding by the lexicographer are: subject, object, prepositional complement (this includes indirect objects), adverbial, clausal, predicative complement, and QDM (quantifier, determiner, modifier) - the latter for argument slots with rigid syntax regarding these features, for example arguments that must be introduced by a determiner. One important feature of T-PAS is that it allows for syntactic alternation within the same pattern, as in Figure 7:

Figure 7: Syntactic alternation in T-PAS for the verb *finire* 'to finish', which allows for both a direct object and a clausal argument introduced by *di* to express the semantic selection [Activity] for the second argument.

Another main feature of T-PAS is that it encodes metonymic shifts on the arguments (Pustejovsky & Jezek 2008). The idea of registering metonymic shifts in the patterns emerged from the need of addressing the divergence between the frequency of metonymic instances in the corpus and the lack of a proper way to record this kind of information in the resource. Therefore, we implemented Skema with the addition of a specific sublabel, .m (where ".m" stands for metonymic); metonymic sublabels are linked to their main label and reflect their syntactic structure, as well as the sense of the verb, which does not change. The metonymic sublabel encodes the new semantic type(s); the shift between the type in the label and the metonymic type is also registered, see the second line of sublabel 1.m in Figure 8. The metonymic sublabel has been applied to a preliminary sample of 30 verbs in Romani (2020); we are interested in extending the number of verbs annotated for metonymies in their arguments.

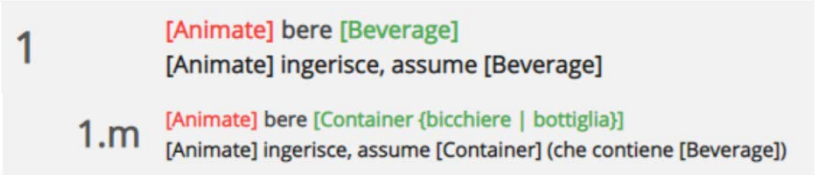

Figure 8: Metonymic sublabel in T-PAS for the verb bere 'to drink', where the semantic shift between the type [Beverage] and the metonymic type [Container] is registered.

Finally, in developing and customizing Skema for T-PAS, we devoted attention to the graphic layout and visualization of the patterns, in order to make them easily-readable and user-friendly. We used as few symbols as possible and conceived a system to properly combine lexical and semantic information (as in the metonymic sublabel in Figure 8).

3.3 CROATPAS

The CROATian Typed Predicate Argument Structure resource (CROATPAS, Marini & Ježek 2019) is the Croatian sister project of the T-PAS resource (see section 3.2). The two projects share the same corpus-based lexicographic methodology and a number of common features, such as the focus on metonymic shifts taking place within argument structures. The reference corpus linked to the resource is the Croatian Web as Corpus (Ljubešić & Erjavec 2011), which contains over 1.2 billion tokens. CROATPAS's first release is scheduled for the end of 2020. At present, its inventory consists of 101 verb entries, 457 patterns, 106 metonymic subpatterns and 22,052 annotated corpus lines (Marini & Ježek 2020). Being a Slavic language, Croatian posed a certain number of issues which had to be tackled when Skema was implemented for CROATPAS, such as the graphical rendering of case inflection in pattern strings. The Croatian case system consists of seven cases, namely nominative, genitive, dative, accusative, vocative, locative and instrumental (Barić et al. 1997: 101). Since it is mainly noun endings that express the grammatical relations between sentence components, we soon realized that – if we planned to translate Croatian valency structures into CROATPAS patterns using non-inflected Semantic Types – we had to find an effective way to convey the morpho-syntactic information usually provided by case, since we could not even rely on fixed word order nor on an extensive inventory of prepositions. In addition to color-coding the different argument slots, the solution was adding case markings as bottom-right indexes to the Semantic Types in the pattern strings, as in the example portrayed in the figure below.

Figure 9: CROATPAS pattern 1 of the verb *preporučivati* 'to recommend'

Another Croatian-specific feature to be taken into account when setting up Skema was verbal aspect. Croatian verbs usually come in pairs featuring both a perfective and an imperfective lexical variant, thus allowing language users to choose between two different options according to the temporal constituency of the given event. Therefore, in CROATPAS each aspectual variant is treated as an independent verb entry.

4 Conclusion

This paper introduced Skema, a new corpus pattern editor system. Skema is a web-based editor integrated into Sketch

Engine which combines two new features: annotating concordance lines with labels for patterns and management of these labels with the possibility of storing arbitrarily structured information for each pattern label.
In this paper, we described three projects which employ the technique of Corpus Pattern Analysis, all of which are using Skema. Another project which has been recently moved to Skema is the Pattern Dictionary of English Verbs (Hanks & Pustejovsky 2005). Since all four projects use a very similar pattern structure, inter-language linking of patterns (verb meanings) should be relatively easy and the resulting dictionary (as envisaged in Baisa et al. 2016) of verb valencies would form a valuable resource for both researchers and language learners.
Skema is being actively developed and new features are expected to be added, such as user-customizable pattern structures, reliable collaborative annotation, support for online self-publishing of the data and an export function.

5 References

Baisa, V., El Maarouf, I., Rychlý, P. & Rambousek, A. (2015). Software and Data for Corpus Pattern Analysis. In *RASLAN*. pp. 75-86.

Baisa, V., Može, S. & Renau, I. (2016). Multilingual CPA: Linking Verb Patterns across Languages. In *Proceedings of the XVII Euralex International Congress,* pp. 410-417.

Barić, E., Lončarić, M., Malić, D. Pavešić, S., Peti M., Zenčević V. & Znika M. (1997). *Hrvatska gramatika*. Zagreb: Školska knjiga.

Baroni, M., Bernardini S., Ferraresi A. & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. In *Language resources and evaluation*, 43(3), pp. 209–226.

Colman, L., Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In *Proceedings of the XVIII EURALEX International Congress*, pp. 233-246.

De Mauro, T. (2000). *Grande dizionario dell'uso* (GRADIT). Torino: UTET.

Hanks, P. (2004). Corpus pattern analysis. In *Proceedings of the 11th EURALEX International Congress*, pp. 87-98.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: The MIT Press.

Hanks, P., Pustejovsky, J. (2005). A pattern dictionary for natural language processing. In *Revue Française de linguistique appliquée*, *10*(2), pp. 63-82.

Ježek, E. (2018). Classi di nomi tra semantica e ontologia. In F. Masini, F. Tamburini (eds.) *CLUB Working Papers in Linguistics*, 2, Bologna: CLUB (Circolo Linguistico dell'Università di Bologna), pp. 117-131.

Ježek E., Magnini B., Feltracco A., Bianchini A. & Popescu, O. (2014). T-PAS: A resource of Typed Predicate Argument Structures for Linguistic Analysis and Semantic Processing. In *Proceedings of LREC 2014*, Reykjavik, Iceland, pp. 890-895.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, 1(1), pp. 7-36.

Marini, C., Ježek E. (2019). CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics* (CliC-it). Bari, Italy.

Marini, C., Ježek, E. (2020). Annotating Croatian Semantic Type Coercions in CROATPAS. In *Proceedings of the Sixteenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (ISA-16). France, Marseille, pp. 50-55.

Ljubešić N., Erjavec, T. (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In I. Habernal, V. Matoušek (eds.) *Text, Speech and Dialogue, Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer-Verlag, pp. 395-402.

Pustejovsky J. (1995). *The Generative Lexicon.* Cambridge MA: The MIT Press.

Pustejovsky J., Jezek E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. In *Rivista di Linguistica (Italian Journal of Linguistics)*, vol. 20, pp. 181-214.

Romani, E. (2020). Searching for Metonymies in Natural Language Texts. A Corpus-based Study on a Resource for Italian Verbs. BA Thesis, University of Pavia, Pavia, Italy.

Sabatini, F., Coletti, V. (2007). *Il Sabatini Coletti 2008. Dizionario della Lingua Italiana*. Milano: RCS Libri S.p.A. https://dizionari.corriere.it/dizionario_italiano/. [04/05/2020]

Tiberius, C., Niestadt. J. & Schoonheim, T. (2014): 'The INL Dictionary Writing System'. In I. Kosem, M. Rundell (eds.) *Slovenščina 2.0: Lexicography*, 2 (2), pp. 72–93.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.