Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset

Ivana Filipović Petrović
Croatian Academy of Sciences and Arts, Zagreb, Croatia

Miguel López Otal
University of Zaragoza, Aragon Institute for Engineering Research, Zaragoza, Spain

Slobodan Beliga
University of Rijeka, Faculty of Informatics and Digital Technologies, Rijeka, Croatia

ifilipovic@hazu.hr, mlopezotal@unizar.es, sbeliga@inf.uniri.hr

LREC-COLING 2024, Torino, Italia

Outline

1. Introduction
2. Background and Motivation
3. Methods and Results

    3.1. The Initial Dataset
    3.2. Semantic Representation Model and RDF Generation
    3.3. Linking Idioms

4. Conclusion and Future Work

Introduction

- Idioms: conventionalized multiword expressions (MWEs) with figurative meaning– which sometimes is not the sum of its parts.
- Cultural-dependent, with forms based on everyday cultural references.
- However, numerous idioms appear as different language realizations at the level of expressions, but they all convey the same meaning:

piece of cake (EN)

é canja de galinha (PT)

e’ facile come rubare le caramelle a un bambino (IT)

mačji kašalj (HR)

- meaning when some task is effortless.

Introduction

- Understanding the meaning of idioms for non-native speakers is a significant challenge.
- Natural language processing can provide support in overcoming this challenge, especially for low-resourced language as Croatian.
- The aim of this paper is to integrate Croatian idioms into an existing multilingual linked idioms dataset.

Background and motivation

- There is a lack of comprehensive linguistic resources focusing on Croatian idioms in the Linked Open Data (LLOD) format.
- Prominent contributions can be found in Orešković et al. (2018) (CroLLOD) and Škvorc et al. (2022).
- In 2018, Moussallem et al. developed a multilingual linked idioms dataset called LIdioms to support natural language processing applications by linking idioms from various languages.
- LIdioms included idioms from English, German, Italian, Portuguese, and Russian.
- Croatian idioms have recently become accessible through the *Online Dictionary of Croatian Idioms* (Filipović Petrović and Parizoska, 2022) - open access, highly reliable source of lexical data, which is available in XML format.
- However, it presents its contents linearly, limiting its accessibility to human readers, primarily native Croatian speakers.

Methods and Results
The Initial Dataset

Resources: Utilized the Online Dictionary of Croatian Idioms (Filipović Petrović and Parizoska, 2022) for our research.

Version: This is a corpus-driven dictionary, version 1.0 of which was released in October 2022.

Contents: The dictionary currently contains 513 headwords, comprising a total of 1,042 idioms and their variants.

Reliability: It is a highly reliable source of lexical data, meticulously curated by linguists and lexicographers based on real language usage.

Advantages: The dictionary provides pre-evaluated idioms and carefully designed definitions, making them a ready-made resource for this study.

Limitations: While the data is open-source, it lacks the structuring necessary for benefiting NLP applications and its contents are presented linearly, limiting its accessibility to human readers, primarily native Croatian speakers.

Semantic Representation Model and RDF Generation

Goal: convert our XML source file to an RDF Ontolex-Lemon representation.

Challenge: idioms are complex to represent in Ontolex Lemon

Solution: rely on the Ontolex representation for idioms presented by LIdioms (Moussallem et al., 2018):

- Well-established format for representing idioms (a type of MWE) in Ontolex Lemon.
- Potential interoperability of our resource with the existing LIdioms dataset (linked data).

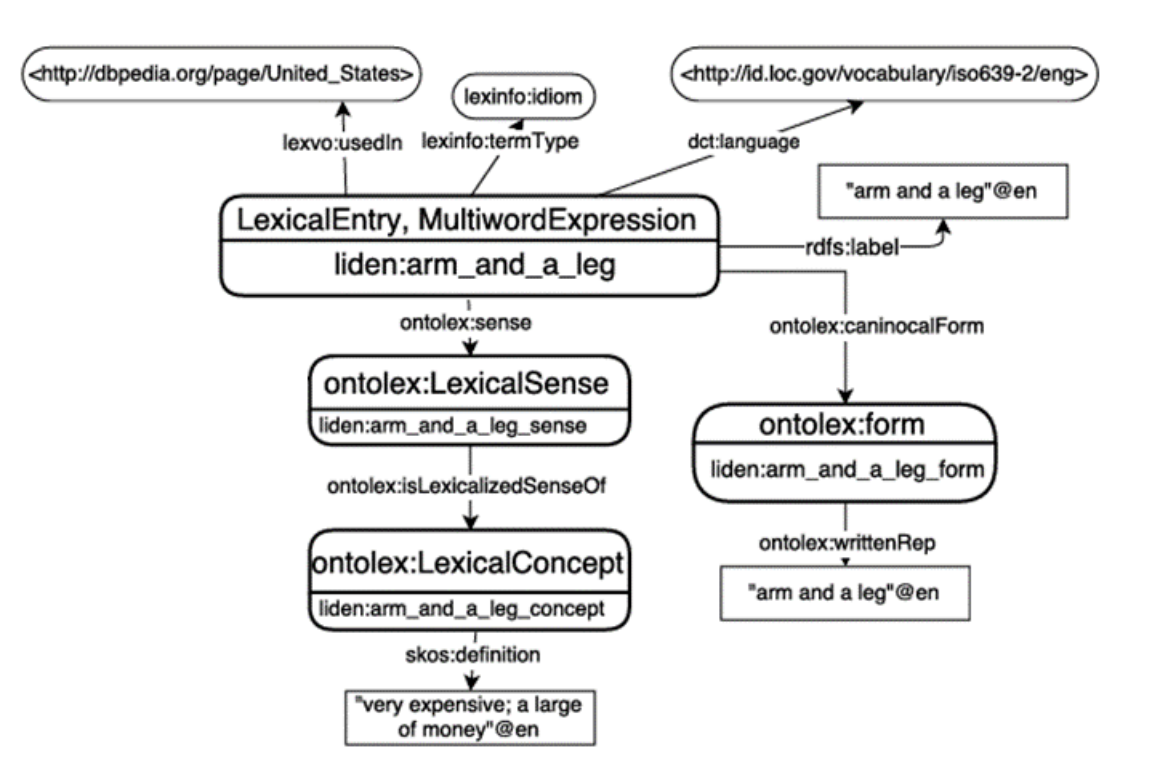


Figure borrowed and adapted from Moussallem et al. (2018), representing an Ontolex Lemon representation for the English idiom ‘arm and a leg’.

Semantic Representation Model and RDF Generation

- Difficulties with our source XML file:

Semantic Representation Model and RDF Generation

- Difficulties with our source XML file:
    - Source file consisting of a collection of `<entry>` XML tags, each representing a different idiom, and roughly corresponding to an Ontolex `LexicalEntry`. ✓

Semantic Representation Model and RDF Generation

- Difficulties with our source XML file:

  - Source file consisting of a collection of `<entry>` XML tags, each representing a different idiom, and roughly corresponding to an Ontolex `LexicalEntry`. ✓
  - Additional linguistic information encoded in individual XML entries (e.g. alternative forms and definitions) being difficult to represent in standard Ontolex Lemon. ✗

Semantic Representation Model and RDF Generation

```
<entry>: 'naći se u škripcu'                    <!--'to be in a difficult position'-->
    <other_form>'biti u škripcu'</other_form>                <!--'be in a corner'-->
    <other_form>'izvući se iz škripca'</other_form>   <!-- 'get out of a corner'-->
</entry>
```

Semantic Representation Model and RDF Generation

```
<entry>: ‘naći se u škripcu’                    <!--‘to be in a difficult position’-->
    <other_form>‘biti u škripcu’</other_form>                <!--‘be in a corner’-->
    <other_form>‘izvući se iz škripca’</other_form>   <!-- ‘get out of a corner’-->
</entry>
```

- In order to be represented in LIdioms scheme. Several options:
  a) List all forms under a single `LexicalEntry`? How do we model these forms under this entry?
  b) Each form is to have its own `LexicalEntry`? How do we represent several individual `LexicalEntry` being part of a same base idiom form?

Semantic Representation Model and RDF Generation

```
<entry>: ‘naći se u škripcu’                    <!--‘to be in a difficult position’-->
    <other_form>‘biti u škripcu’</other_form>                <!--‘be in a corner’-->
    <other_form>‘izvući se iz škripca’</other_form>   <!-- ‘get out of a corner’-->
</entry>
```

- In order to be represented in LIdioms scheme. Several options:
  a) List all forms under a single `LexicalEntry`? How do we model these forms under this entry?
  b) Each form is to have its own `LexicalEntry`? How do we represent several individual `LexicalEntry` being part of a same base idiom form?

<entry>: ‘naći se u škripcu’ <!--‘to be in a difficult position’-->

<other_form>‘biti u škripcu’</other_form> <!--‘be in a corner’-->

<other_form>‘izvući se iz škripca’</other_form> <!-- ‘get out of a corner’-->

</entry>

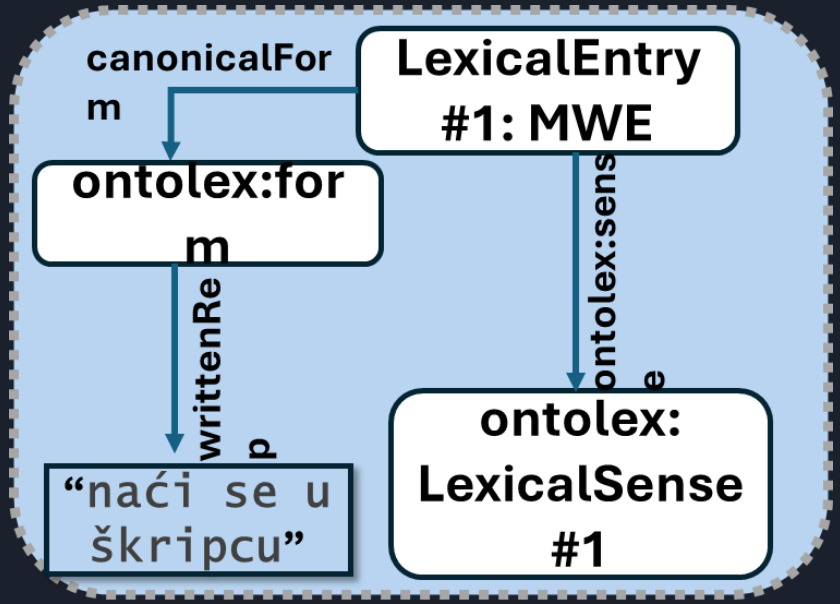
<entry>: ‘naći se u škripcu’ <!--‘to be in a difficult position’-->

<other_form>‘biti u škripcu’</other_form> <!--‘be in a corner’-->

<other_form>‘izvući se iz škripca’</other_form> <!-- ‘get out of a corner’-->

</entry>

“Main” idiom form



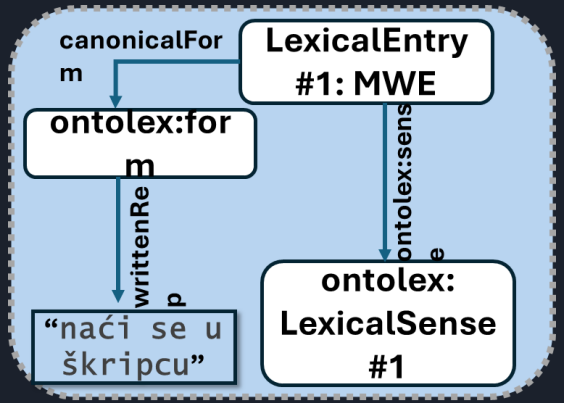
<entry>: ‘naći se u škripcu’ <!--‘to be in a difficult position’-->

<other_form>‘biti u škripcu’</other_form> <!--‘be in a corner’-->

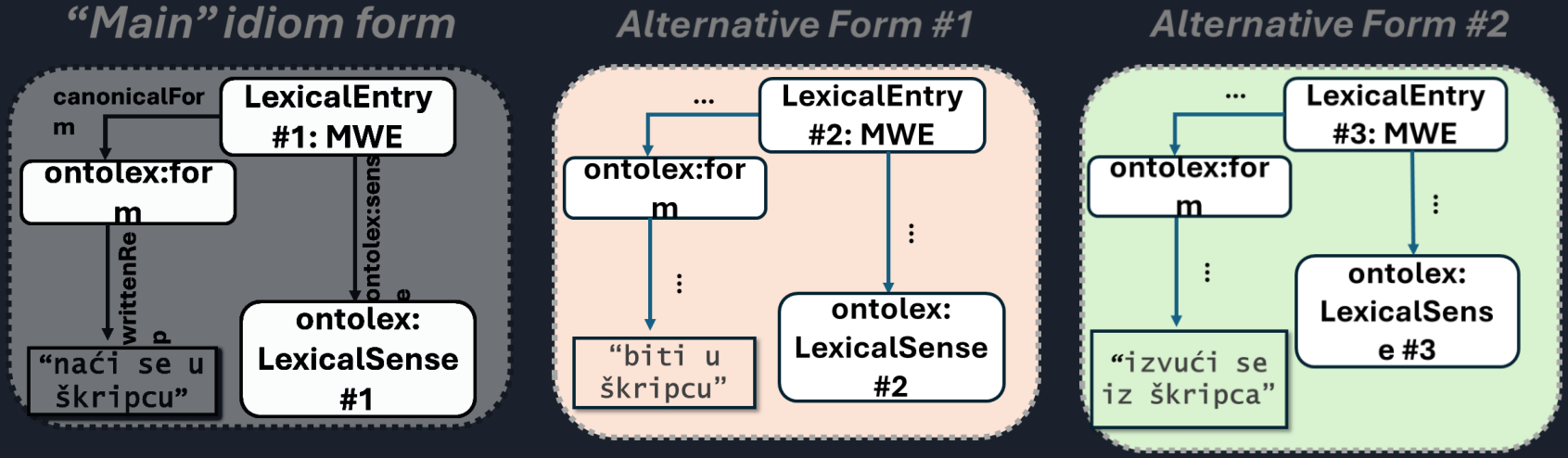
<other_form>‘izvući se iz škripca’</other_form> <!-- ‘get out of a corner’-->

</entry>

“Main” idiom form



```
<entry>: ‘naći se u škripcu’                    <!--‘to be in a difficult position’-->
    <other_form>‘biti u škripcu’</other_form>          <!--‘be in a corner’-->
    <other_form>‘izvući se iz škripca’</other_form>    <!-- ‘get out of a corner’-->
</entry>
```

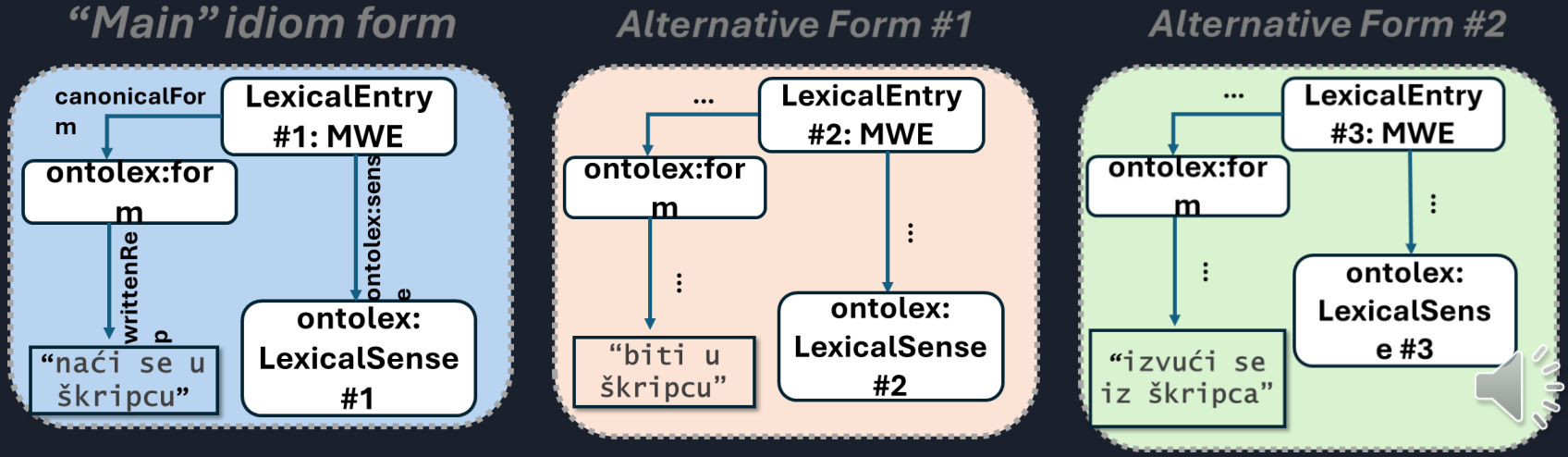


<entry>: ‘naći se u škripcu’ <!--‘to be in a difficult position’-->

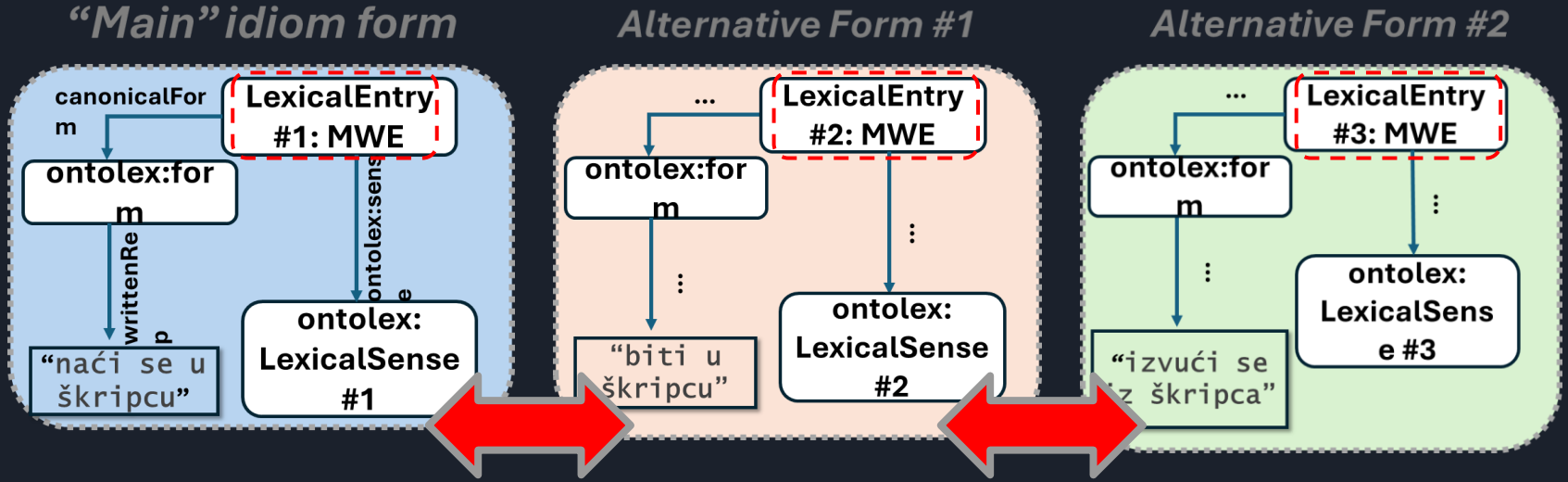
<other_form>‘biti u škripcu’</other_form> <!--‘be in a corner’-->

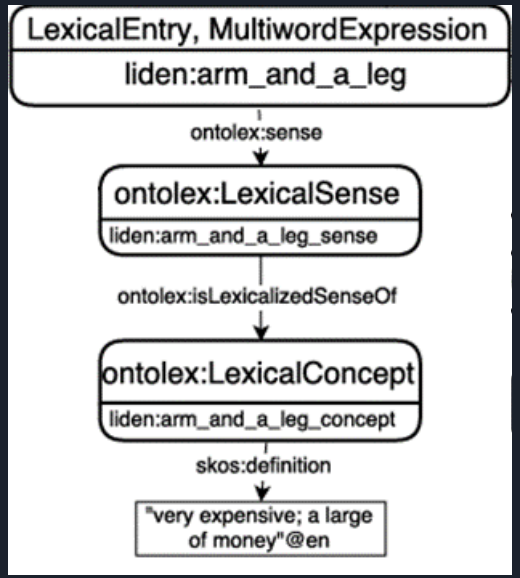
<other_form>‘izvući se iz škripca’</other_form> <!-- ‘get out of a corner’-->

</entry>

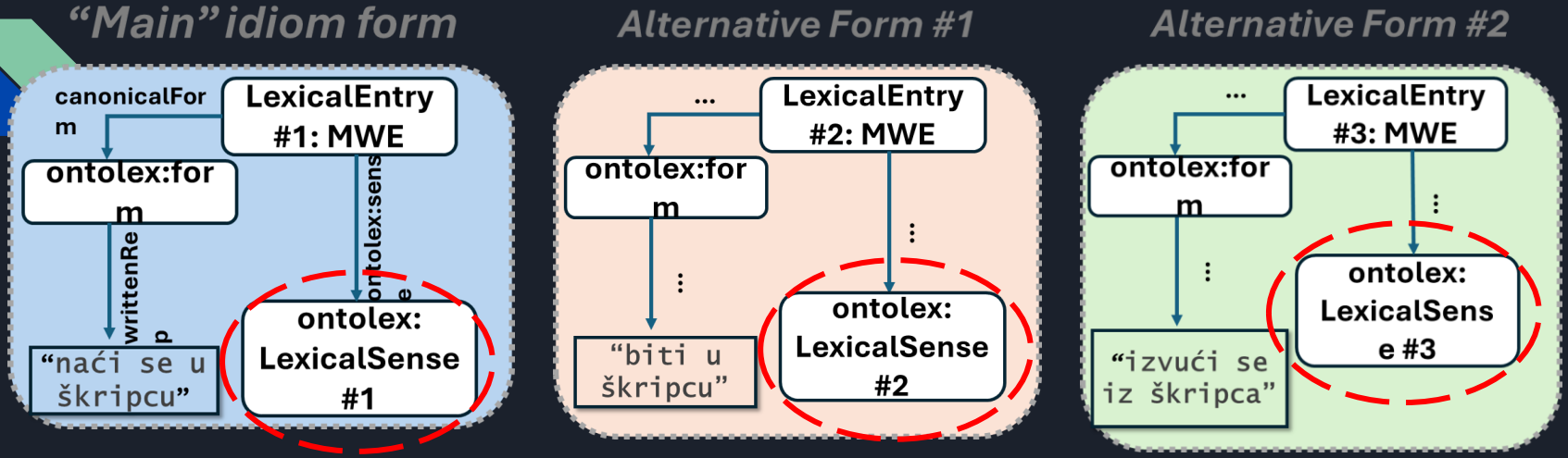


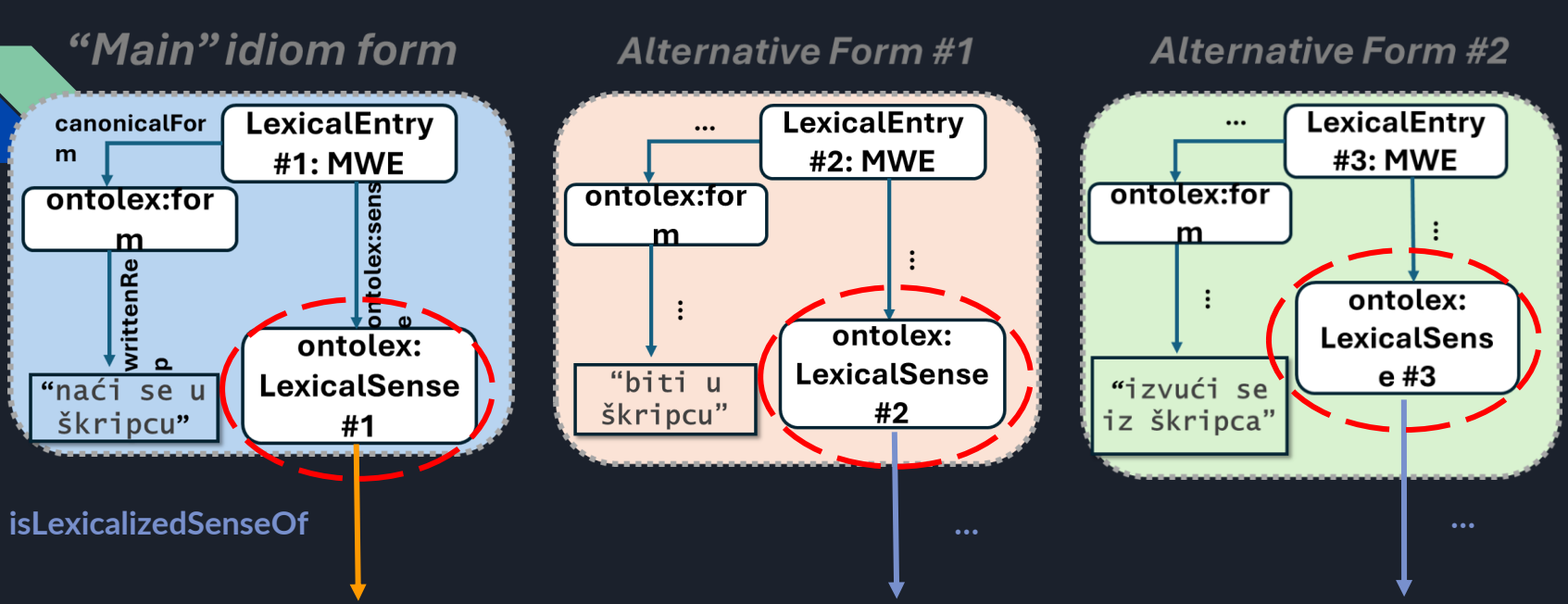
```
<entry>: ‘naći se u škripcu’                    <!--‘to be in a difficult position’-->
    <other_form>‘biti u škripcu’</other_form>          <!--‘be in a corner’-->
    <other_form>‘izvući se iz škripca’</other_form>    <!-- ‘get out of a corner’-->
</entry>
```

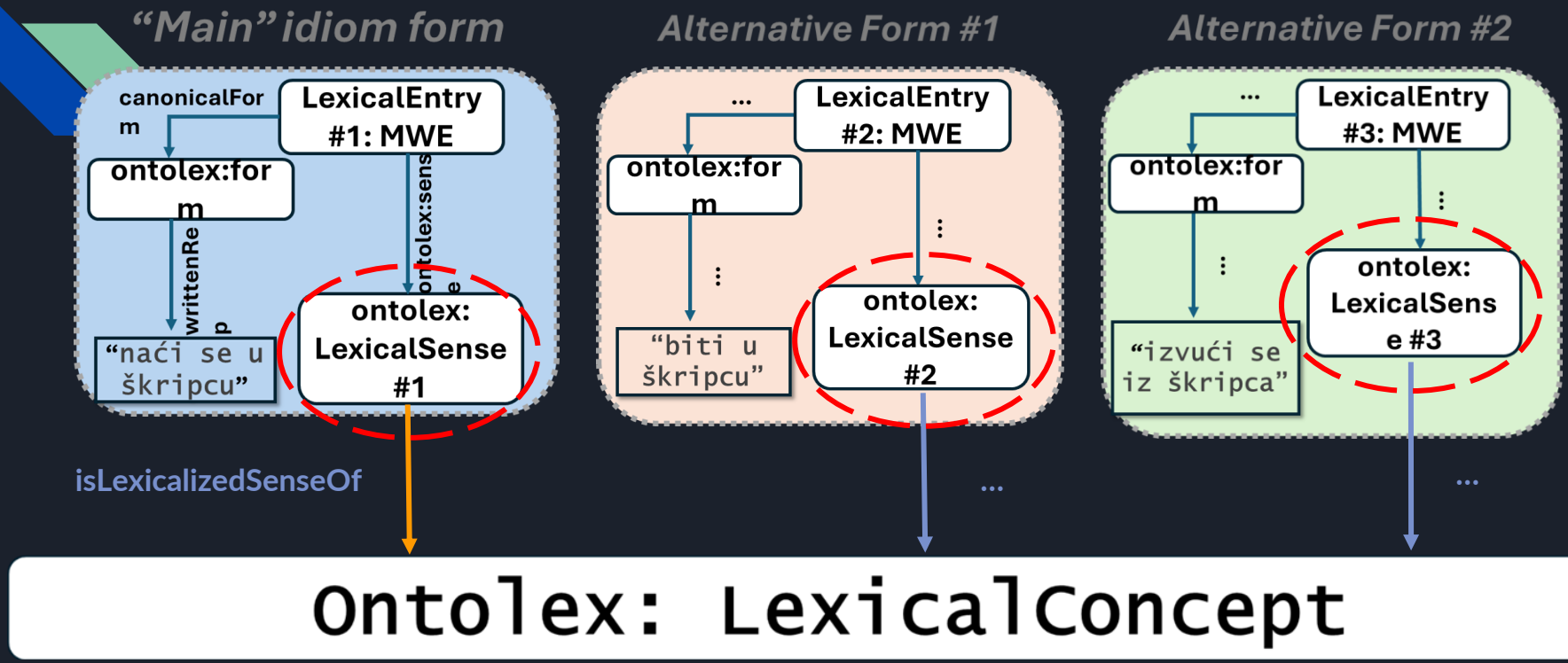


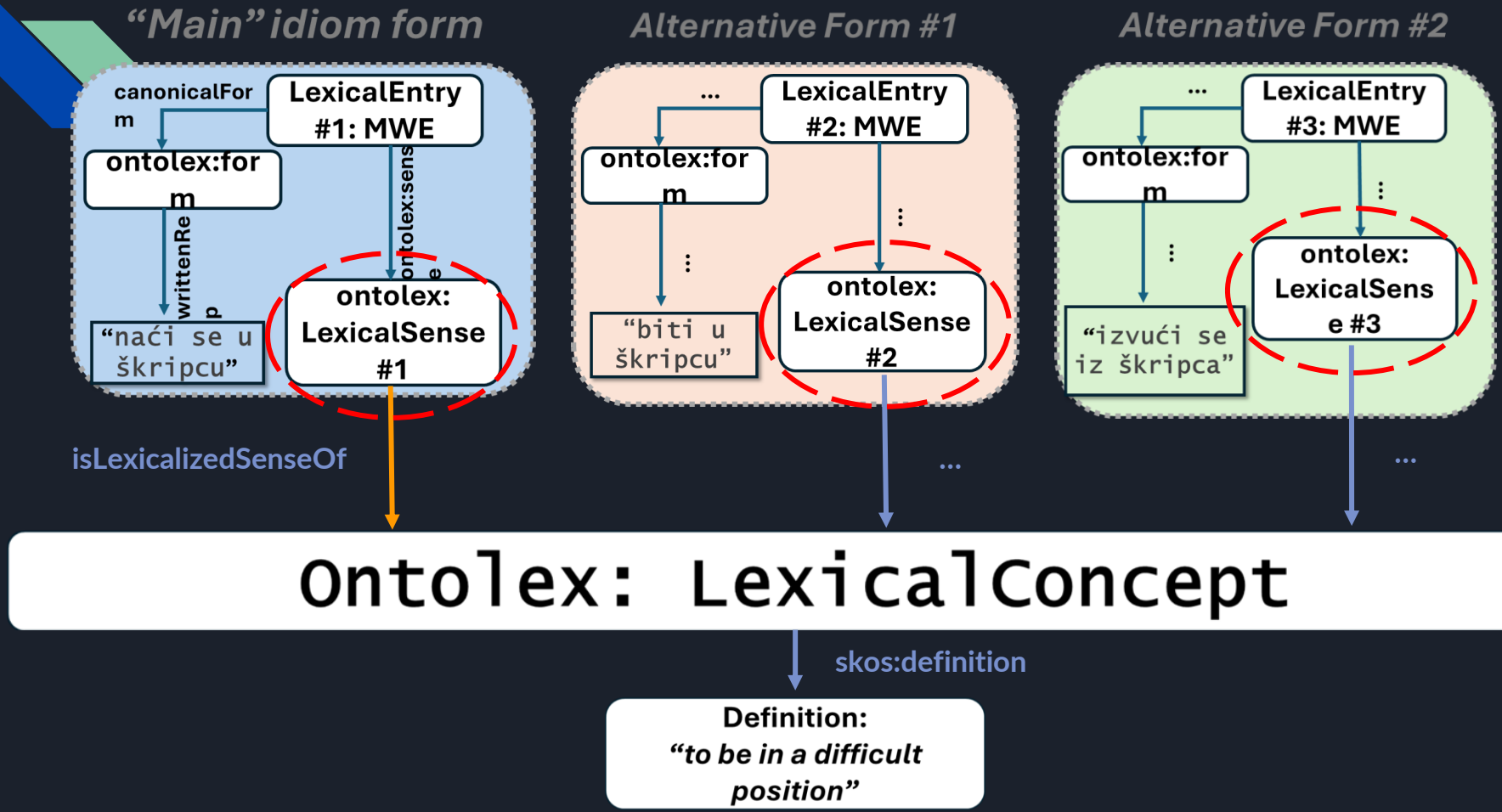


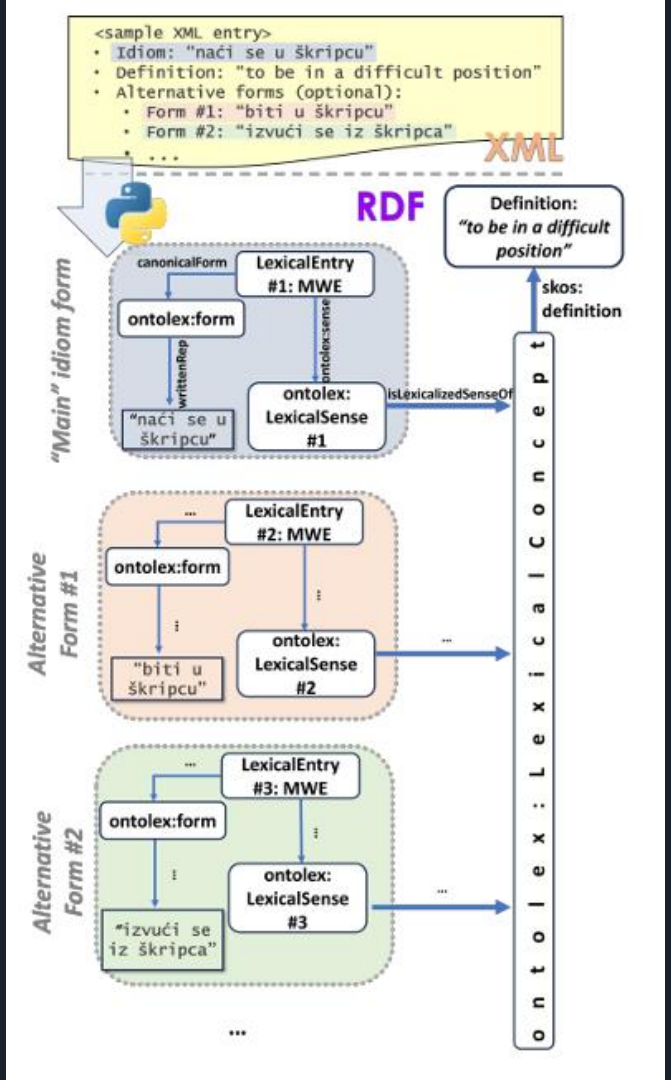
Adapted from Moussallem et al. (2018)


“Main” idiom form
canonicalForm
LexicalEntry #1: MWE
ontolex:form
writtenRep
“naći se u škripcu”
ontolex:sense
ontolex: LexicalSense #1
Alternative Form #1
...
LexicalEntry #2: MWE
ontolex:form
“biti u škripcu”
ontolex: LexicalSense #2
Alternative Form #2
...
LexicalEntry #3: MWE
ontolex:form
“izvući se iz škripca”
ontolex: LexicalSense #3


“Main” idiom form
Alternative Form #1
Alternative Form #2
canonicalForm
LexicalEntry #1: MWE
ontolex:form
writtenRep
“naći se u škripcu”
ontolex:sense
ontolex: LexicalSense #1
isLexicalizedSenseOf
...
LexicalEntry #2: MWE
ontolex:form
“biti u škripcu”
ontolex: LexicalSense #2
...
LexicalEntry #3: MWE
ontolex:form
“izvući se iz škripca”
ontolex: LexicalSense #3
...


“Main” idiom form
Alternative Form #1
Alternative Form #2
canonicalForm
LexicalEntry #1: MWE
ontolex:form
writtenRep
ontolex:sense
“naći se u škripcu”
ontolex: LexicalSense #1
LexicalEntry #2: MWE
ontolex:form
“biti u škripcu”
ontolex: LexicalSense #2
LexicalEntry #3: MWE
ontolex:form
“izvući se iz škripca”
ontolex: LexicalSense #3
isLexicalizedSenseOf
Ontolex: LexicalConcept


"Main" idiom form
Alternative Form #1
Alternative Form #2
canonicalForm
LexicalEntry #1: MWE
ontolex:form
writtenRep
ontolex:sense
"naći se u škripcu"
ontolex: LexicalSense #1
...
LexicalEntry #2: MWE
ontolex:form
"biti u škripcu"
ontolex: LexicalSense #2
LexicalEntry #3: MWE
ontolex:form
"izvući se iz škripca"
ontolex: LexicalSense #3
isLexicalizedSenseOf
Ontolex: LexicalConcept
skos:definition
Definition:
"to be in a difficult position"



RESULTS:

- Ontolex Lemon-compliant
- Compatible with LIdioms
- Representing all encoded information from source XML dictionary
- 17,000 triples

Linking Idioms

- Idioms rarely match word per word between languages, **but some idiomatic expressions share the same meaning across several languages → source for interlinguistic data links**
- Basic idea behind LIdioms (Moussallem et al., 2018):
  - create several monolingual Ontolex-based idioms dictionaries (English, French, German, Russian, Portuguese and Italian), and
  - **link some of the idioms from each dictionary to similar entries in other languages**

Linking Idioms

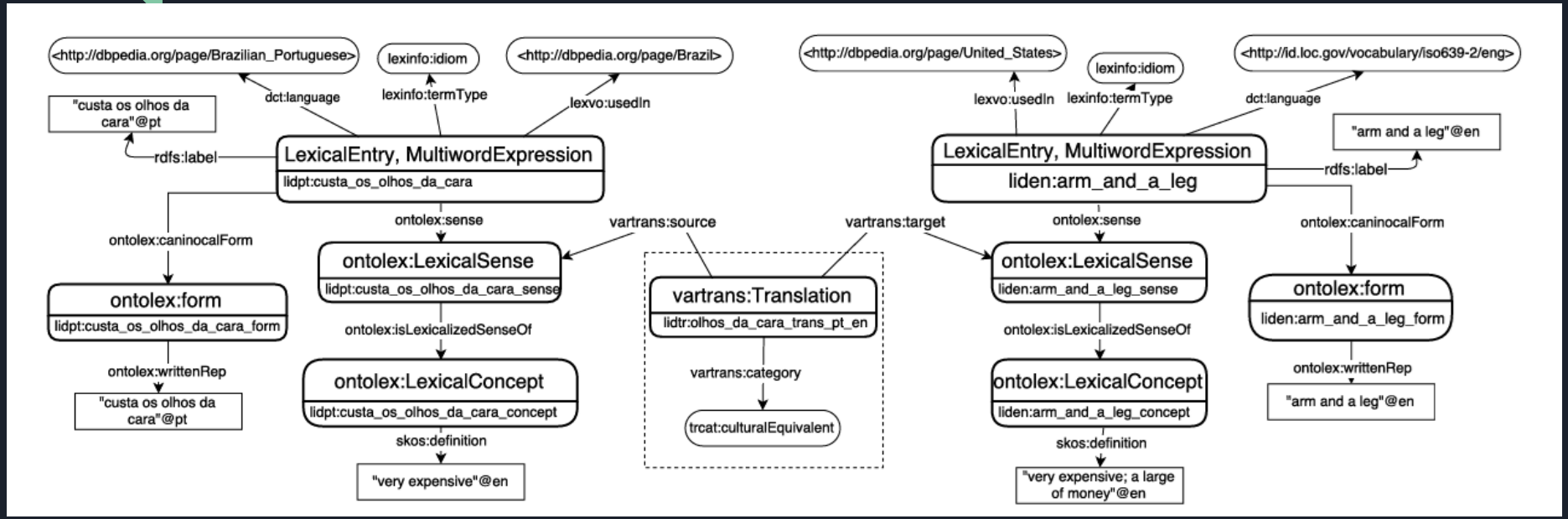


Figure taken from Moussallem et al. (2018), representing the linking convention used between an idiom in Portuguese (‘custa os olhos da cara’) and an equivalent in English (‘arm and a leg’).

Linking Idioms

Croatian idiom

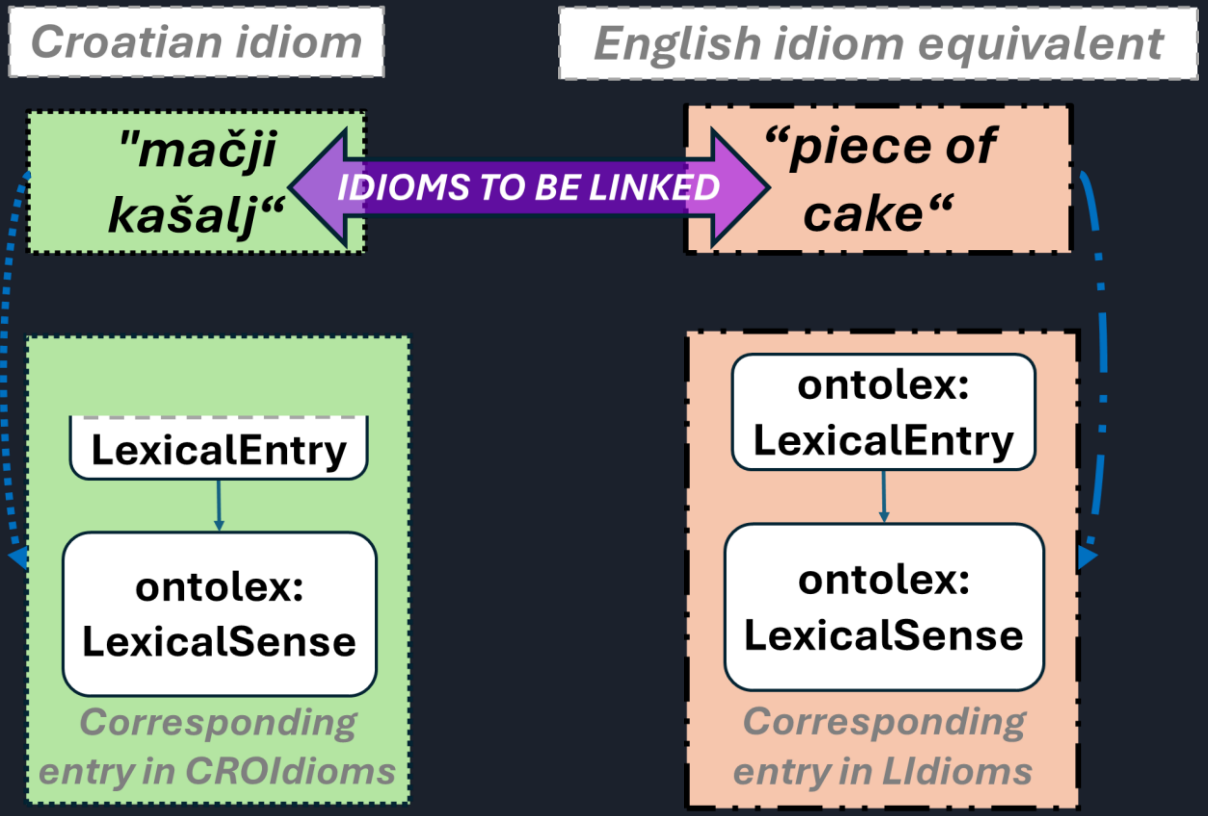
"mačji kašalj“

Linking Idioms

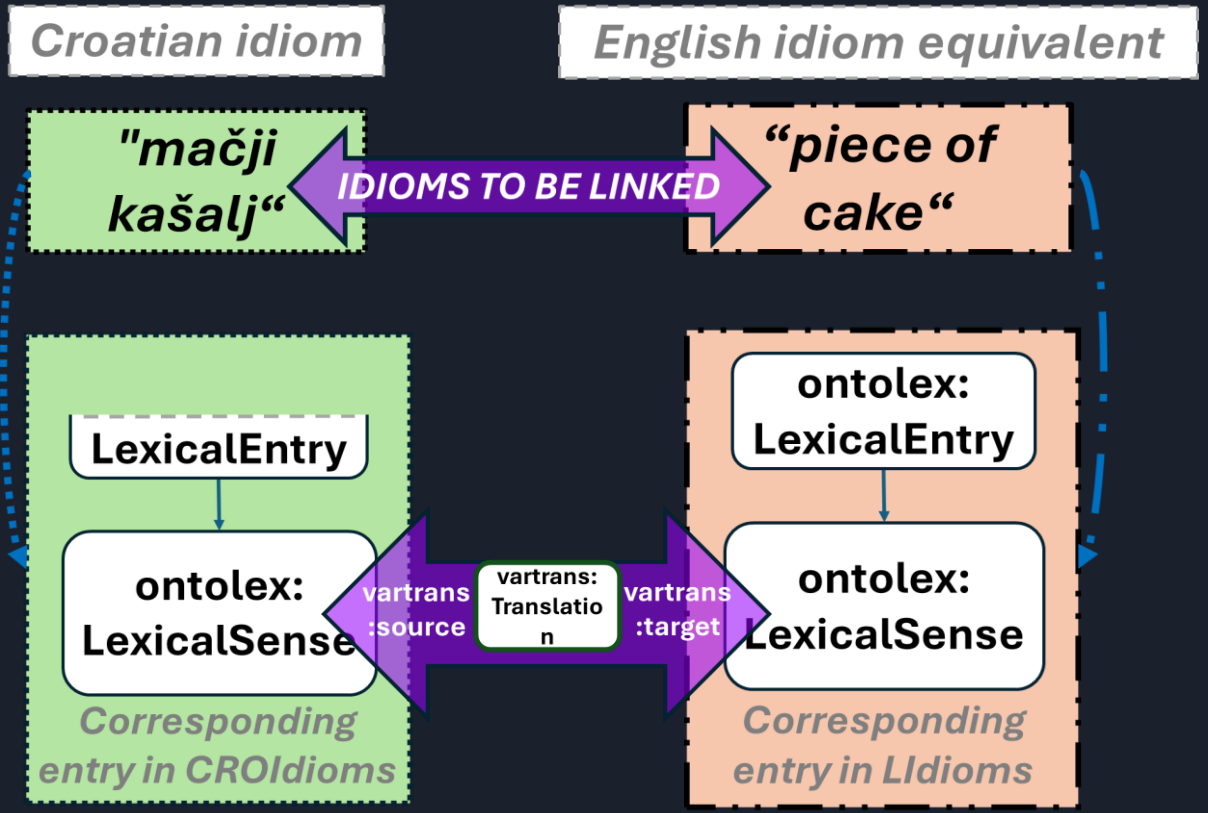
Croatian idiom	English idiom equivalent
"mačji kašalj“	“piece of cake“

Linking Idioms

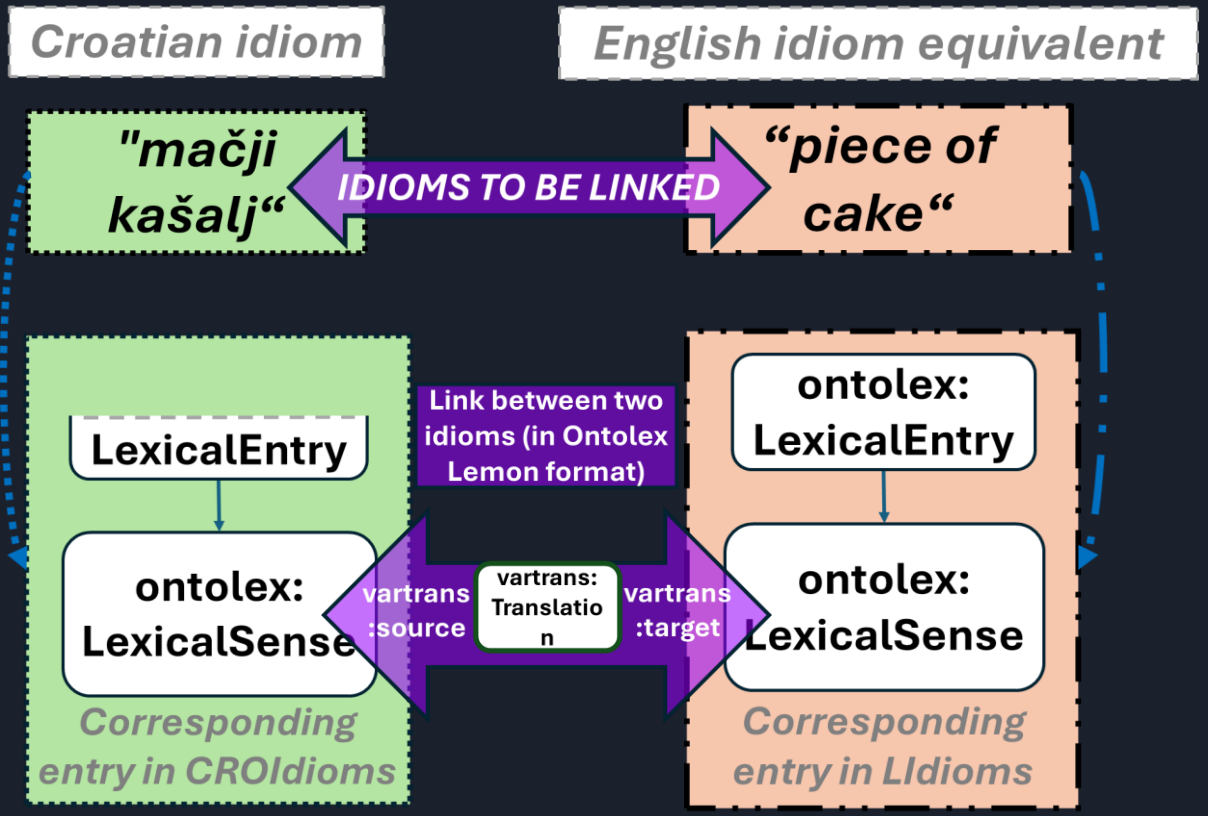
Linking Idioms



Linking Idioms



Linking Idioms



Linking Idioms

- 49 idiom pairs manually identified and linked:
  - 28 English-Croatian pairs
  - 13 Italian-Croatian pairs
  - 8 Portuguese Croatian pairs
- Each corresponding `ontolex:LexicalSense` entry retrieved via SparQL queries and linked with `varTrans` module.
- Additions to existing idiom links in LIdioms dataset

	EN	PT	IT	DE	RU	HR
Idioms	219	114	175	130	105	1,042
Links	192	79	73	60	82	49

Conclusion

This study extends the multilingual linked idiom dataset by adding Croatian idioms to the existing LIdioms dataset.

The aim was to create an enriched resource that can enhance its reusability in language learning and translation and contribute to a broader understanding of universalities and differences between languages and cultures.

Future work and enhancements

- expanding the number of Croatian idioms linked to those of other languages
- enhancements to the CroIdioms dataset can occur in two main directions:
    a. 1. contextual embeddings (like MICE) could be utilized for mining procedures to identify idioms outside our current dataset
    b. 2. building a system for recommending (suggesting) new links between idioms of other languages: calculating semantic similarities between definitions of Croatian idioms and idioms from other languages

Bibliography

Slobodan Beliga and Sanda Martinčić-Ipšić. 2017. Network-enabled keyword extraction for under resourced languages. In Semantic Keyword Based Search on Structured Data Sources, pages 124–135, Cham. Springer International Publishing.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. Linguistic Linked Data: Representation, Generation and Applications, 1st edition. Springer Publishing Company, Incorporated

Ivana Filipović Petrović and Jelena Parizoska. 2022. Frazeološki rječnik hrvatskoga jezika. Hrvatska akademija znanosti i umjetnosti.

Polona Gantar and Simon Krek. 2022. Creating the Lexicon of Multi-Word Expressions for Slovene. Methodology and Structure. In Dictionaries and Society. Proceedings of the XX EURALEX International Congress, pages 549–562, Mannheim. IDS-Verlag.

Jorge Gracia, Elena Montiel-Ponsoda, Daniel Vila Suero, and Guadalupe Aguado-de Cea. 2014. Enabling language resources to expose translations as linked data on the web. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 409–413, Reykjavik, Iceland. European Language Resources Association (ELRA).

A. Langlotz. 2006. Idiomatic Creativity: A cognitive linguistic model of idiom-representation and idiom-variation in English. Human Cognitive Processing. John Benjamins Publishing Company.

Bibliography

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: Development and applications. In Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, in Leiden, Netherlands, pages 587–597. Lexical Computing CZ s.r.o.

Julia Miller. 2018. Research in the pipeline: Where lexicography and phraseology meet. Lexicography, 5:23–33.

Diego Moussallem, Mohamed Ahmed Sherif,Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Marko Orešković, Sandra Lovrenčić, and Mario Es sert. 2018. Croatian Network Lexicon within the Syntactic and Semantic Framework and LLOD Cloud. International Journal of Lexicography, 32(2):207–227.

Jelena Parizoska and Marija Omazic. 2020. Sheme dinamike sile i promjenjivost glagolskih frazema [force-dynamic schemas and variability of verbal idioms]. Jezikoslovlje, 21:179–205.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In Proceedings of the 28th International Conference on Computational Linguistics, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bibliography

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mi titelu, Archna Bhatia, Maja Buljan, Marie Can dito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Ko valevskait˙e, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Es cartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identifi cation of verbal multiword expressions. In Pro ceedings of the Joint Workshop on Linguistic An notation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebe skind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Be jček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Ko valevskaite, Timm Lichte, Nikola Ljubešić, Jo hanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 31–47, Valencia, Spain. Association for Compu tational Linguistics.

Tadej Škvorc, Polona Gantar, and Marko Robnik Šikonja. 2022. Mice: Mining idioms with contex tual embeddings. Knowledge-Based Systems, 235:107606