PARSEME Survey on MWE Resources

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, Johanna Monti

PARSEME Survey on MWE Resources

Gyri Smørdal Losnegaard[1], Federico Sangati[2], Carla Parra Escartín[3], Agata Savary[4], Sascha Bargmann[5], Johanna Monti[6]

[1]University of Bergen, Norway; [2]FBK Trento, Italy; [3]Hermes Traducciones, Spain;
[4]Université François Rabelais Tours, France; [5]Goethe University Frankfurt, Germany; [6]University of Sassari, Italy

Gyri.Losnegaard@uib.no, federico.sangati@gmail.com, carla.parra@hermestrans.com,
agata.savary@univ-tours.fr, bargmann@em.uni-frankfurt.de, jmonti@uniss.it

Abstract

This paper summarizes the preliminary results of an ongoing survey on multiword resources carried out within the IC1207 Cost Action PARSEME (PARSing and Multi-word Expressions). Despite the availability of language resource catalogs and the inventory of multiword datasets on the SIGLEX-MWE website, multiword resources are scattered and difficult to find. In many cases, language resources such as corpora, treebanks, or lexical databases include multiwords as part of their data or take them into account in their annotations. However, these resources need to be centralized to make them accessible. The aim of this survey is to create a portal where researchers can easily find multiword(-aware) language resources for their research. We report on the design of the survey and analyze the data gathered so far. We also discuss the problems we have detected upon examination of the data as well as possible ways of enhancing the survey.

Keywords: multiword expressions, language resources, language resource infrastructures

1. Introduction

Despite the ample progress in Natural Language Processing (NLP) in the past decades, the high degree of ambiguity and fuzziness in human languages remains a challenge. Multiword expressions (MWEs), which often show idiosyncratic morphological, syntactic, or semantic properties, strongly contribute to that challenge (Sag et al., 2001).

Current research on MWEs shows that most proposals still concentrate on the creation of MWE lexicons or the automatic recognition of MWEs in text, whereas only very few address the links between MWEs and a comprehensive linguistic analysis of text. The IC1207 COST Action PARSEME (PARSing and Multi-word Expressions)[1], an interdisciplinary scientific network devoted to the role of MWEs in parsing, aims to contribute to the further improvement of NLP in this direction (Savary et al., 2015).

In an effort towards consolidating previous and ongoing research, PARSEME is currently conducting a meta survey of language resources (LRs) containing MWEs. Examples of such LRs are monolingual and multilingual lists of MWEs, MWE dictionaries and lexicons, corpora and treebanks with MWE annotations, and any other type of lexical or linguistic resource with MWEs as part of its inventory. The aim is to provide a portal that gives researchers access to previously unavailable or newly created MWE(-aware) resources.

In this paper, we present the preliminary results from this ongoing survey. Section 2 reports on how the survey was designed and conducted. Sections 3 and 4 analyze the collected data. Whereas Section 3 offers a statistical analysis, Section 4 is devoted to a qualitative one, focusing on each LR type separately. Section 5 discusses our preliminary findings and planned future work, and Section 6 concludes the paper.

[1] http://www.parseme.eu/index.php/the-action

2. Methodology

2.1. Survey Design

The survey (PARSEME, 2014c) was designed using Google Forms and consists of two main sections. In the first section, the user is asked to provide general information about the LR:

- name
- link to the LR
- type of the LR
- contact information
- language(s)
- size of the LR
- maximum length of MWEs
- whether non-contiguous expressions are present
- license and accessibility policies

The second section is dedicated to more advanced descriptions of the LR:

- relevant publications
- special MWE features
- grammatical or lexical formalism (if any)

The design of the survey is a trade-off between basic cataloging (i.e. providing an overview of as many relevant LRs as possible) and usability for end users (i.e. meeting user requirements for detailed information and accurate descriptions of each LR).

Google Forms is an efficient crowdsourcing tool. With the aid of optional fields, contributors are encouraged to register as much information as possible, while the limited number of required fields keeps the overall time required to complete the form to a minimum. In order to lower the threshold for adding new LRs even further, they can also be registered by sending an email with the basic information to a mailing list.[2]

[2] parseme-survey@nlp.ipipan.waw.pl

2.2. Registration of LRs

There are already quite a few catalog entries of LRs that include MWEs in international infrastructures such as the CLARIN and META-SHARE repositories. However, they are not always easy to find since the information about MWEs in these repositories is often scarce, non-uniform, or non-explicit. The current survey includes relevant LRs from three major LR inventories:[3]

- META-SHARE: the ILSP managing node
- ELRA: European Language Resources Association
- SIGLEX-MWE: the MWE community website

The LRs were extracted by searching for the strings *mwe*, *mwu*, *multi word*, *multiword*, and *multi-word*. At a later stage, LRs will also be harvested from other infrastructures, and the search will be extended to include further search expressions, such as *collocation*.

2.3. Dissemination

In May and September 2014, requests for contributions were posted to Corpora List and Linguist List, and we also addressed the presenters and participants of the 2014 EACL MWE workshop. The survey will continue to be disseminated to the community at similar future events, and PARSEME members regularly receive reminders.

2.4. Availability

As internally designed by the Google Forms framework, each new entry collected from the survey is automatically appended as a single row in a spreadsheet. An automatic copy of this spreadsheet with anonymized entries has been made publicly accessible (PARSEME, 2014a).
Since the large number of entries and column fields make it a rather difficult repository to consult, we have made use of the Awesome Table Gadget[4] to present its content in a more user-friendly way and to add interactive controls to manipulate the data it displays. The interactive view is now publicly available, too (PARSEME, 2014b).

3. Global Statistics

The LRs gathered through the survey can be grouped into 5 types, as illustrated in Table 1. The types are described in more detail in Section 4. They were established upon reclassification of the original 7 types used in the survey form. This was done by conflating the two types 'MWE dictionary or lexicon' (MWEs only) and 'Dictionary or lexicon with MWEs' (includes but is not limited to MWEs) into the type 'MWE lexicons', and by conflating the two types 'Multilingual list of MWEs' and 'Multilingual parallel list of MWEs' into the type 'Multilingual resources'. On top of that, we manually resolved the categories of the LRs that fell into the 'Others' group, which originally accounted for about 40% of the entries.

As can be observed in Table 1, almost half of the LRs are MWE lexicons. As stated earlier, these are either resources that include MWEs as part of their data or resources consisting of MWEs only. The rest of the LRs are more or less equally divided between the other 4 categories.

Type	Count	%
MWE lists (4.1.1)	13	12%
MWE lexicons (4.1.2)	48	45%
Treebanks with annotated MWEs (4.1.3)	12	11%
Multilingual resources (4.2)	15	14%
Others (4.3)	19	18%

Table 1: Types of the collected MWE Resources

As regards the properties of the MWEs, only 40 survey entries specify the maximum length of the MWEs in the LR. As shown in Figure 1, the length varies between 2 and 23.[5]

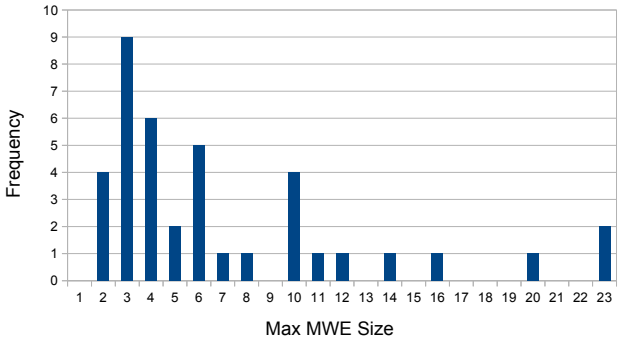

Figure 1: Length of the MWEs

More than half of the LR entries (62) specify whether the LR contains only contiguous MWEs, i.e. MWEs whose components are always adjacent in text occurrences, or also non-contiguous MWEs (cf. Figure 2).

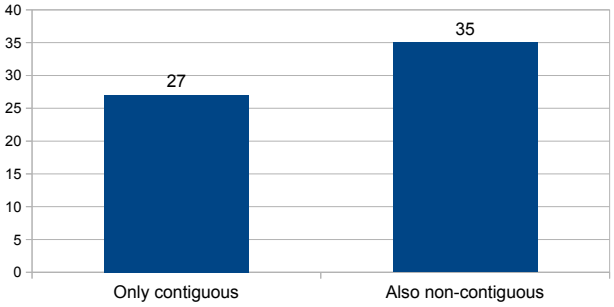

Figure 2: (Non-)Contiguity among the MWEs

End users will also want to know if the LR is available for research, commercial use, etc. All but 8 LRs are reported to be available with either unrestricted or restricted use (Figure 3). Figure 4 provides details on the types of licenses. Figure 5, on the other hand, shows the distribution of the submitted entries over time. The two major peaks correspond to the periods in which submission requests were posted to the mailing lists (cf. Section 2.3).

[3]The list summarizes the infrastructures and inventories searched by October 2015.

[4]https://sites.google.com/site/scriptsexamples/available-web-apps/awesome-tables

[5]The histogram includes all entries except for a single outlier reporting a figure of 57.

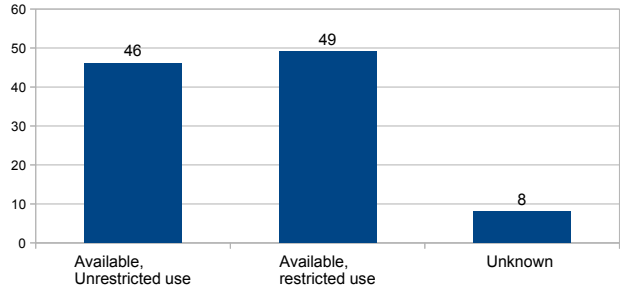

Figure 3: Availability of the MWE Resources

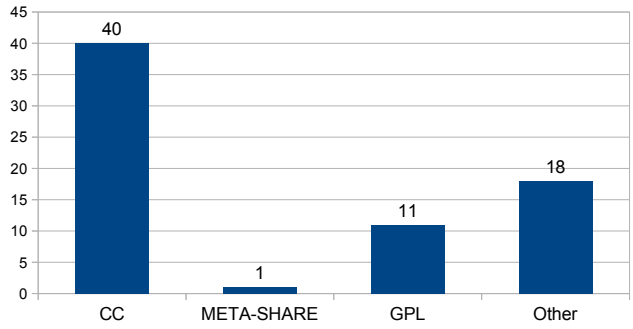

Figure 4: License Types of the MWE Resources: Creative Commons, META-SHARE, GNU General Public License

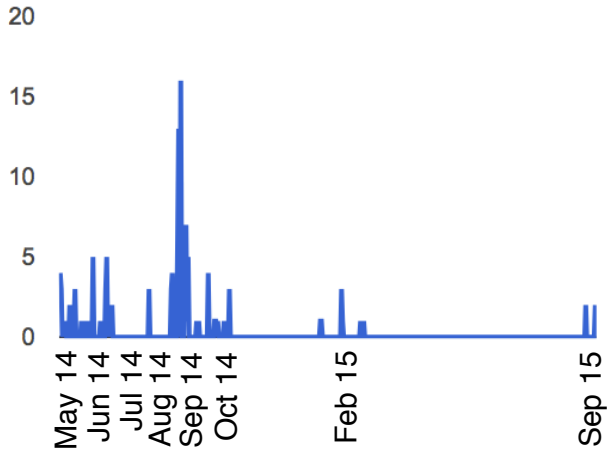

Figure 5: Entry Times of the MWE Resources

4. Results

In this section, we present the data obtained for each of the five MWE resource types distinguished in Table 1. The data were gathered from the responses directly and unaltered, i.e. the information provided in the forms was not enriched. As explained in Section 6, this will be done in future work. In some cases, however, the collected data required an adjustment of the resource type.

The remainder of this section is broken down into three subsections: 4.1 is about *mono*lingual, 4.2 about *multi*lingual MWE resources. Subsection 4.3 focuses on other types of input, such as NLP applications and tools.

4.1. Monolingual MWE Resources

Here, we report on all those LRs that were indicated to be monolingual. They are further subdivided into MWE lists (4.1.1), MWE lexicons (4.1.2), and treebanks containing annotated MWEs (4.1.3).

4.1.1. MWE Lists

The survey contains two categories for MWE lists: monolingual and multilingual.[6] Of the compiled LRs, 13 can be classified as monolingual lists. A few of these (3) are datasets already included in the LR inventories mentioned in section 4.3. Most (7) of the 10 remaining lists are pure MWE resources, i.e. they contain nothing but MWEs.

Half of the 10 LRs had been registered as monolingual lists right from the start. They include 2 English wiktionaries with phrasal verbs and idioms, 2 lists of English and Hungarian light verb constructions, and 1 resource of English MWEs mined from Wikipedia and annotated with information about lemma, POS, frequency, source, sense, etc.

The other half of the MWE lists had originally and erroneously been registered in the category 'Others'. This half includes 2 lists for English, 1 for Croatian, 1 for Greek, and 1 for Portuguese. 3 of them include MWEs as part of their data. They consist of 1 term bank, 1 list of stopwords, and 1 list of word forms with IPA representations. The other 2 lists consist of MWEs only. One is a database with detailed linguistic descriptions, the other a list with human-annotated semantic compositionality scores.

Of the lists providing information about size, the largest one has more than 350,000 entries (all MWEs). The smaller lists only have a few hundred entries, but these tend to have more annotations. Some of the lists that are not exclusively MWE resources are also large, but they do not provide the number of MWE entries.

Of the lists that were supplied with information about contiguity, 2 (the term bank and the Wikipedia resource) contain only contiguous MWEs, while 4 lists (the Greek database, the 2 lists of light verb constructions, and the list of MWEs with semantic compositionality scores) were reported to include also non-contiguous MWEs.

Most of the lists (6) are for English. The others (4) are for Hungarian, Croatian, Greek, and Portuguese.

Half of the 10 lists are available with unrestricted use and 2 with restricted use. For 3 of the lists, their availability is unknown. However, not all lists seem to have a proper license: 2 have a license from the European Language Resources Association (ELRA), of which 1 is a license that makes the list freely available for research, whereas the other is restrictive (academic but non-commercial). 2 other lists have a Creative Commons attribution, share-alike license (CC BY-SA), and 1 has its own license allowing for unrestricted use. For the rest, no proper license seems to have been established.

Finally, it is important to point out the great diversity of LRs gathered under this category. This in part reflects the fact that no definition of 'list' was provided in the survey guidelines. If the term is taken to mean a compilation of unannotated MWEs and MWEs that have been classified

[6] For multilingual lists, see Section 4.2.

according to a single feature or type, only 8 of the LRs qualify as lists. In some LRs, such as the English Wikipedia and the linguistically rich Greek resource, the MWEs (or lexical entries) are annotated according to more than one parameter. These LRs are perhaps more accurately classified as datasets or databases. These observations will be used to improve the design of the survey in the future.

4.1.2. MWE Lexicons

MWE lexicons (or more precisely e-dictionaries and lexicons dedicated to MWEs or containing MWEs) account for almost half (45%) of the LRs documented in the survey. As explained in Section 3, this category came into existence by the unification of two separate categories listed in the original survey: 'MWE dictionary or lexicon' (MWEs only) and 'Dictionary or lexicon with MWEs' (includes but is not limited to MWEs).

The lexicons gathered in our survey concern 19 languages: Arabic, Bulgarian, Croatian, Czech, Dutch, English, Estonian, French, German, Hebrew, Italian, Macedonian, Polish, Portuguese, Russian, Serbian, Slovak, Slovene, and Swedish. The best-represented languages in this category are English (12) and Italian (7).

The MWE lexicons can be classified according to various criteria. Firstly, the MWEs contained in them divide into two classes:

- contiguous MWEs, i.e. those whose components are adjacent in text occurrences: compound nouns, including proper names (Savary et al., 2009), adverbs, adjectives, pronouns, prepositions (Litkowski, 2014), post-positions, and conjunctions (30% of the lexicons contain contiguous MWEs only)
- possibly non-contiguous MWEs: support verb constructions, phrasal verbs, verbal idioms, particle verbs[7], noun-verb expressions, etc. (contained in 34% of the lexicons), e.g. (Odijk, 2013; Borin et al., 2013a; Gantar et al., 2013)

The survey entries for the remaining 36% of the lexicons do not contain the relevant data.

Secondly, most lexicons contain MWEs from the general language register, but some are restricted to a specialized domain: medicine[8], economy (Savary et al., 2012), environment, transportation, etc.

Thirdly, most are meant for NLP applications. Only very few are dedicated to human users (e.g. language learners).

Finally, the lexicon construction methodology ranges from a) totally automatic extraction from corpora or Wikipedias (Quochi et al., 2012; Fadida et al., 2013) over b) automatic extraction with human post-processing to c) mostly manual descriptions of MWEs in traditional dictionaries and lists.

The nature of the linguistic description of the MWEs is also highly variable in these LRs and includes:

- raw lists of MWEs linked with their morphosyntactic variants (e.g. acronyms)
- intensional morphosyntactic descriptions of MWE lemmas including variation patterns (Al-Haj et al., 2014), which possibly allow for the automatic generation of the extensional descriptions, e.g. all inflected forms (Krstev et al., 2013; Czerepowicka and Savary, 2015)
- valency frames, where verbs with their subcategorization (compulsory arguments, not necessarily lexicalized) can be seen as objects on the frontier between single words and MWEs (Borin et al., 2013b; Žabokrtský and Lopatková, 2007; Urešová et al., 2014)
- data related to corpus occurrences of the MWEs, such as frequencies, concordances, or links to treebank nodes, which allow the user to retrieve the morphological, syntactic, and semantic annotations of these occurrences (Bejček and Straňák, 2010)
- diachronic information (e.g. on neologisms)
- links to semantic networks

The sizes of the lexicons are difficult to compare since they are not uniformly documented in the survey, and for some lexicons, the size specification is missing altogether. The largest lexicons for which the relevant data are available contain 9,000 to 140,000 MWE base forms or 70,000 to 300,000 inflected forms and variants.

With regard to availability, 21 of the lexicons are available and unrestricted in their use, 20 are available but restricted, 3 entries indicate availability to be unknown, and 4 lack availability specifications altogether.

Overall, 37 of the entries show information on the lexicons' licenses: 6 have their own, the remaining 31 use standard licenses. 5 of the latter belong to the GNU family: 3 are General Public Licenses (GNU GPLs), the other 2 Lesser General Public Licenses (GNU LGPLs).

Of the remaining licenses, 5 are from ELRA, and the others belong to the Creative Commons family: 7 only require attribution (CC BY), 5 require attribution and are shared alike (CC BY-SA), 7 require attribution and prohibit commercial usage (CC BY-NC), 2 require attribution, are shared alike, and prohibit any commercial usage (CC BY-NC-SA), and 1 requires attribution and prohibits both commercial usage and redistribution (CC BY-NC-ND).

4.1.3. Treebanks with annotated MWEs

The documented LRs include 10 treebanks with MWE annotations and 2 grammars. The treebanks are for Polish (1), English (2), Dutch (2), Turkish (2), Hungarian (1), Italian (1), and Slovene (1), while both grammars are for Polish.[9]

A few of the reported LRs are not really treebanks, though, but rather annotation layers on top of treebanks.

There is a group of treebanks that provide information about the grammatical framework involved. This group consists of a dependency treebank, a hybrid treebank (constituency and dependency), and a treebank using EAGLES for morphosyntactic annotation and ISST (based on FAME) for functional annotation.

The types of annotated MWEs are reported for 5 treebanks and include compound and MWE named entities;

[7] e.g. the LINGO resources or the Estonian verbal MWEs

[8] e.g. The SPECIALIST lexicon

[9] Some of these treebanks are also described in a separate qualitative survey about the representation of MWEs in treebanks, carried out by PARSEME's WG4 (Rosén et al., 2015).

phrasal verbs, light verb constructions, and verbal expressions; named entities, idiomatic expressions, and expressions in foreign languages; light verb constructions; compounds, support verb constructions, and idioms; place names, names of organizations, names of persons, and proper names. Most treebanks have annotations of non-contiguous MWEs. The numbers of annotated MWEs in the treebanks range from 20,000 to 2,704, but they are not really comparable.

All documented treebanks are available, 7 for restricted and 5 for unrestricted use. 3 of the treebanks have a Creative Commons license. While 2 of them are CC-BY-SA (attribution, share-alike), 1 also has the non-commercial (NC) restriction. All Polish LRs have a General Public License (GNU GPL). One of the English treebanks is licensed through the Language Data Consortium (LDC), and the Italian treebank is licensed through ELRA. In both cases, the type of license depends on the type of end user.

One of the Turkish treebanks also has a dual license (academic + commercial), and 2 LRS (the other Turkish and the Hungarian treebank) are only available for academic purposes. Finally, 1 of the Dutch treebanks (the Alpino treebank)[10] is freely available but seems to lack any type of licensing.

4.2. Multilingual MWE Resources

Overall, 15 LRs were classified as bilingual or multilingual. However, upon closer inspection, two had to be moved to the 'Lexical resources' type. In one case, the LR was actually monolingual and had simply been wrongly classified. In the other, the data were multilingual, but there were no links between the different languages (i.e. no pointers to translations). For this reason, we decided to consider it monolingual rather than multilingual.

The 13 remaining multilingual resources can be divided into two main categories: multilingual MWE lexicons, which represent the majority of the data gathered in the survey, and multilingual MWE lists, which represent a significantly smaller percentage, as shown in Table 2.

Type	Count	%
Multilingual MWE lists (4.1.1)	4	26%
Multilingual MWE lexicons (4.1.2)	10	66%
Others (4.3)	1	0.6%

Table 2: Types of Multilingual MWE Resources

Out of these 13 LRs, 5 contain only MWEs, whereas the other 8 contain MWEs as part of their overall data. The LRs containing only MWEs each focus on a particular type of MWE: collocations (2), light verb constructions (1), named entities (1), or phrasal verbs (1).

The multilingual MWE resources vary in terms of the number of languages covered. While one of them includes dozens of languages (the JRC-Names database by Steinberger et al. (2011)), the remaining ones focus on bilingual (Hungarian-English; English-French; Arabic-French) or trilingual (French-Romanian-German; Chinese-English-French; French-Portuguese-Spanish; Korean-English-French) data. One of the LRs, the Multilingual Collocation Dictionary system Centre Tesniere (MultiCoDiCT)[11], could be considered to include 4 LRs, as it contains four different lexicons.

Two of the LRs, the INCYTA[12] and the THAMUS dictionaries[13], are actually collections of bilingual dictionaries (English-Spanish, English-Italian, and German-Italian) across different domains (data processing, economics, medicine, computer science, telecommunications, etc.).[14]

The remaining 6 LRs cover a wide variety of language pairs, including Bulgarian, Romanian, Greek, Polish, Russian, and up to the 50 languages covered in BabelNet (Navigli and Ponzetto, 2012).

The information provided by the informants is not complete, and comparisons are hard to make, which is also due to the diverging nature of the LRs as well as their differences in size and languages covered. Only 7 LRs document whether the MWEs they contain are only contiguous (3) or also non-contiguous (4), and only 9 LRs include licensing information. 2 LRs have different ELRA licenses with a fee that depends on the type of the intended use (academic + commercial). 5 LRs have a Creative Commons (CC) license: 3 CC-BY-NC (attribution, non-commercial); 1 CC-BY (attribution only); 1 CC-BY-SA (attribution, share-alike). The remaining 2 licenses are not standard. One establishes that the LR is only for academic use, the other makes the LR mostly freely available.

The data gathered in the survey proves that multilingual MWE resources are hard to find. It is particularly revealing that most of the LRs in this category are not MWE resources, but rather lexical resources containing MWEs as part of their data. This lack of large bilingual and multilingual MWE lexicons may hamper research both on the translation of MWEs across languages and NLP involving two or more languages.

4.3. Others

The category 'Others' includes 19 out of the 107 survey entries (17.8%). It encompasses all those LRs that, according to the respondents, did not fit into any of the four resource types mentioned above and, therefore, is quite diverse. It includes MWE (extraction) tools (7), corpora with annotated MWEs (3), lists of MWE resources (2), a dataset with MWEs, an annotated text, a pattern dictionary of verbs, a list of domains, a WordNet with MWEs, a Web service, and an ontology. Of these 19 LRs, 10 are concerned with just one language; the others deal with at least 2 languages (5) or are language-independent tools (4).

The most prominent languages are English (4), Bulgarian (3), and Serbian (3). The majority of the entries (11) show no indication in terms of the size of the respective LR. The ones that do, vary in the way this kind of information is

[10] http://www.let.rug.nl/~vannoord/trees/

[11] http://tesniere.univ-fcomte.fr/multicodict-en.html

[12] http://metashare.ilsp.gr:8080/repository/search/?q=INCYTA

[13] http://metashare.ilsp.gr:8080/repository/search/?q=thamus

[14] All the dictionaries contained in these two collections have individual ISRLNs. For space reasons they are not listed here.

expressed, so that it is difficult to compare the LRs in this respect. Only 4 entries indicate whether the LRs described by them contain non-contiguous MWEs: 2 do, 2 do not. The majority (11) of the LRs are available but restricted in their use. Only 3 are non-restricted. For the other 5, there is no information on this issue.

Of the 19 'Others' entries, 14 show licensing information. The licenses include 4 Creative Commons (CC) licenses (3 CC-BY-NC, 1 CC-BY-SA), 4 General Public Licenses (GPLs), 1 META-SHARE license (MS-NC-NoReD-ND), 1 ELRA, 1 Apache, and 3 non-standard licenses.

5. Discussion and Future Work

On the basis of our analysis, three major issues are worth discussing: metadata on MWE resources, licensing issues, and information on LRs containing MWEs in existing LR cataloging initiatives.

5.1. Metadata Issues

As reported in the previous sections, some of the data from our survey are difficult to analyze and compare because they are non-uniform or incomplete. There are two reasons for this: incomplete descriptions from respondents and the use of pragmatic solutions in the design of the survey. For instance, many fields allow the respondents to enter free text, instead of forcing them to choose from a predefined list of values. Such solutions are used for types of information that do not easily fit into well-defined categories.

Moreover, in some cases, a free-text field 'Other' is offered to allow respondents to freely formulate their replies when none of the options listed match their case. This holds for instance for 'LR type', where a list of options is offered but also a free-text field 'Other', in case neither of the listed options applies. Free-text fields are also used for 'LR size' and 'language(s)', for which it is difficult to define a limited and at the same time-exhaustive set of values. The type of size unit and the size intervals will vary greatly, while language descriptions may get complex for multilingual resources. It also proved difficult to integrate the ISO language inventory of languages into the Google form as a predefined list.

Besides the relevant information on the size, license(s), or language(s) included in the LR, researchers working with MWEs (either from a theoretical or an applied point of view), require other information to be available for resources they would consider using. Initiatives such as CLARIN and META-SHARE have worked on the standardization of the metadata description of LRs (Gavrilidou and Desypri, 2003; Gavrilidou et al., 2011; Borin and Lindh, 2011; Broeder et al., 2012). Despite these efforts, to our knowledge, the particular metadata that would be relevant for describing MWE(-aware) LRs have not been thoroughly researched.

Also, even though these metadata could potentially be useful for classifying the data in our survey, the nature of the data we collected requires additional work on the elements needed to describe MWE resources and the inclusion of additional metadata in the metadata schemata used to described LRs. Information on the types of MWEs gathered and the types of information offered for each MWE in the dataset are highly relevant metadata that are currently missing. We tried to address these issues in the survey by allowing LR developers and owners to describe their resources as detailed as they wished and by including specific questions about relevant issues related to MWEs.

On the basis of the preliminary results presented here, we have planned a qualitative analysis of all entries to clean up the data, determine closed-vocabularies, add missing information wherever possible, and make the results accessible in a user-friendly manner. This qualitative analysis will be complemented with the input from two upcoming events organized by the PARSEME Cost Action: the PARSEME/ENeL joint workshop on MWE e-lexicons[15] and the PARSEME 6th General Meeting[16].

One of the planned outcomes of the workshop will be a wish list of features that an ideal MWE lexicon should contain to be maximally NLP-applicable. A specific time slot of the General Meeting has been designated to the discussion of a potential taxonomy of MWE resources and what metadata are particularly needed to describe such LRs.

The results of the qualitative analysis together with the outcomes of the two upcoming PARSEME events will result in a proposal for a metadata schema to describe resources containing MWEs. This schema will be further discussed and developed within the work carried out by the Working Groups of the PARSEME Cost Action.

5.2. Licensing Issues

Although most of the LRs gathered in our survey included information about their availability for research (only 8 showed 'Unknown' for availability, cf. Table 3), not all of them had standardized licenses (or any license at all). In fact, some LRs are publicly available because they are offered for download on the websites of research institutions, but they are not part of any large infrastructure and lack a proper license. In other cases, the LR seems to have a license, but it is not at all standardized.

Large research infrastructures such as CLARIN[17] or LT-Observe[18] have fostered the depositing and licensing of LRs to ensure their reusability and curation. Although this was not one of our initial goals when creating the survey, it is worth exploring the option of joining forces with research infrastructures to make sure that all LRs in our survey that are not deposited in any catalog can be properly licensed and stored. This effort will enhance the reusability of such resources in the long run, as well as the replicability of experiments using such resources.

[15] http://www.parseme.eu/index.php/2-general/135-enel-parseme-workshop-on-mwe-lexicons

[16] http://www.parseme.eu/index.php/2-general/130-6th-general-meeting-spring-2015-struga-fyr-macedo

[17] https://www.clarin.eu/content/license-categories

[18] http://www.lt-innovate.org/lt-observe/resources-list

5.3. Cataloging Issues

As stated in Section 2, we attempted to manually collect LRs from existing LR infrastructures. This effort proved to be a challenge, however, as this type of information is not consistently provided in the metadata of such initiatives, and one depends on MWEs being mentioned in any of the free-text fields describing the LRs. Potentially, many LRs including MWEs were not included in our survey results because they lack specific reference to them.

This points out the need to make both LR developers and research infrastructures aware of the benefits of properly specifying in the description of an LR whether it contains MWEs or not. Having this information available would not only help researchers working with MWEs to find relevant resources for their research, it would also foster the use of already-existing LRs for new research.

6. Conclusion

In this paper, we have presented and discussed the preliminary results of an ongoing survey on multiword resources carried out within the IC1207 Cost Action PARSEME. These results provide a basic overview of more than 100 LRs of different types, representing a range of different languages. We have also discussed the design of our survey and the main issues that we have encountered, as well as potential ways of addressing them.

The findings presented here will be used to normalize the existing data and to improve the survey design in order to create a meta resource that is maximally useful to end users. Based on this first analysis, it seems necessary to revise or redefine the types of LRs used in the survey and to consider whether new categories are justified.

The overall aim of the survey is to create a meta resource that is maximally useful to end users. The planned work towards drafting a suitable metadata schema for LRs dedicated to or containing MWEs will also help to represent the diversity of LRs in a standardized form and display them in a proper way to end users.

Following the public release of the user-friendly version of the survey, we also expect to include new LRs in our database. The results of this iterative process will be published through the PARSEME website and promoted in the SIGLEX-MWE Section. In the long run, we consider publishing the survey data also on other relevant platforms.

7. Acknowledgements

This work has been supported by the IC1207 COST Action PARSEME[19] as part of its scientific program.

The authors wish to thank all the contributors to the PARSEME Survey on MWE Resources.

Carla Parra Escartín is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471.

8. Bibliographical References

Al-Haj, H., Itai, A., and Wintner, S. (2014). Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170, June.

Bejček, E. and Straňák, P. (2010). Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Borin, L. and Lindh, J. (2011). Deliverable D4.1: Metadata descriptions and other interoperability standards. Version 1.0, 2011-05-02. Deliverable in the META-NORD project (CIP 270899).

Borin, L., Forsberg, M., and Lönngren, L. (2013a). Saldo: a touch of yin to wordnet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Borin, L., Forsberg, M., and Lyngfelt, B. (2013b). Close encounters of the fifth kind: Some linguistic and computational aspects of the swedish framenet++ project. *Veredas*, 17(1):28–43.

Broeder, D., Windhouwer, M., van Uytvank, D., Goosen, T., and Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. In Victoria Arranz, et al., editors, *Proceedings of the Eigth Conference on International Language Resources and Evaluation (LREC'12)*, volume Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, pages 8 – 11, Istanbul, Turkey, 22 May 2012. European Language Resources Association (ELRA).

Czerepowicka, M. and Savary, A. (2015). SEJF - a Grammatical Lexicon of Polish Multi-Word Expression. In *Proceedings of Language and Technology Conference (LTC'15), Poznań, Poland*. Wydawnictwo Poznańskie.

Fadida, H., Itai, A., and Wintner, S. (2013). A Hebrew verb–complement dictionary. *Language Resources and Evaluation*, 48(2):249–278.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Kocjančič, P., Grabnar, K., Yerošina, O., Zaranšek, P., and Drstvenšek, N. (2013). Slovene lexical database 1.0. Slovenian language resource repository CLARIN.SI.

Gavrilidou, M. and Desypri, E. (2003). Deliverable D.2.2: Report for the definition of common metadata description for the various types of national LRs, ENABLER project. Deliverable in the ENABLER project.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Speranza, M., Monachini, M., Arranz, V., and Francopoulo, G. (2011). Deliverable D.7.2.1 Specification of Metadata-Based Descriptions for LRs and LTs. Deliverable in the T4ME Project (META-NET).

Krstev, C., Obradovic, I., Stankovic, R., and Vitas, D. (2013). An approach to efficient processing of multiword units. In Adam Przepiórkowski, et al., editors, *Computational Linguistics - Applications*, volume 458 of *Studies in Computational Intelligence*, pages 109–129. Springer.

Litkowski, K. (2014). Pattern Dictionary of English Prepositions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1274–1283, Baltimore, Maryland, June. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-

[19] www.parseme.eu

coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250.

Odijk, J. (2013). Identification and lexical representation of multiword expressions. In *Essential Speech and Language Technology for Dutch*, pages 201–217. Springer.

PARSEME. (2014a). Anonymized table with survey entries. https://goo.gl/P4To2f.

PARSEME. (2014b). Awesome table with survey entries. https://sites.google.com/site/mwesurveytest/home.

PARSEME. (2014c). Mwe survey online form. https://goo.gl/eYz8qL.

Quochi, V., Frontini, F., and Rubino, F. (2012). A MWE acquisition and lexicon builder web service. In Martin Kay et al., editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2291–2306. Indian Institute of Technology Bombay.

Rosén, V., Losnegaard, G. S., De Smedt, K., Bejček, E., Savary, A., Przepiórkowski, A., Osenova, P., and Mitetelu, V. (2015). A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference*, Warsaw, Poland, December.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2001). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

Savary, A., Rabiega-Wiśniewska, J., and Woliński, M. (2009). Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. *Lecture Notes in Computer Science*, 5070:111–141.

Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A., and Makowiecki, F. (2012). SEJFEK - a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 195–214, Mumbai, India, December. The COLING 2012 Organizing Committee.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard, G. S., Parra Escartín, C., Waszczuk, J., Constant, M., Osenova, P., and Sangati, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November.

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., and van der Goot, E. (2011). Jrc-names: A freely available, highly multilingual named entity resource. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 104–110, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Urešová, Z., Štěpánek, J., Hajič, J., Panevova, J., and Mikulová, M. (2014). PDT-vallex: Czech valency lexicon linked to treebanks. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Žabokrtský, Z. and Lopatková, M. (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60.