Chapter 9

~~Modeling~~ multiword expressions in a parallel Bulgarian-English newsmedia corpus

Petya Osenova

Linguistic Modelling Department, IICT-BAS

Kiril Simov

Linguistic Modelling Department, IICT-BAS

The paper focuses on the ~~modeling~~ of multiword expressions (MWE) in Bulgarian-English parallel news corpora (SETimes; CSLI dataset and PennTreebank dataset). Observations were made on alignments in which at least one multiword expression was used per language. The multiword expressions were classified with respect to the PARSEME lexicon-based (WG1) and treebank-based (WG4) classifications. The non-MWE counterparts of MWEs are also considered. Our approach is data-driven because the data of this study was retrieved from parallel corpora and not from bilingual dictionaries. The survey shows that the predominant translation relation between Bulgarian and English is *MWE-to-word*, and that this relation does not exclude other translation options. To formalize our observations, a catenae-based ~~modeling~~ of the parallel pairs is proposed.

1 Introduction

This work proposes a catenae-based ~~modeling~~ of aligned pairs in parallel Bulgarian-English news corpora. A representation is suggested that handles bilingual pairs comprising at least one MWE. Our main aim is to offer a representation that deals equally well with cross-language symmetries and asymmetries.

In each language, MWEs were annotated independently from the alignments in the corpus. Then, using the alignments, we examined how MWEs were trans-

Petya Osenova & Kiril Simov. 2018. ~~Modeling~~ multiword expressions in a parallel Bulgarian-English newsmedia corpus. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 249–274. Berlin: Language Science Press. DOI:10.5281/zenodo.1182603

lated between the two languages. The following general alignment types of examples are considered: MWE-to-MWE; MWE-to-word; MWE-to-phrases. This general typology is not exhaustive since, and in most of the cases, another translation option could have been used. Thus, it is interesting to observe the lexical choices actually made in the parallel data.

In our work we refer to the classifications of MWEs developed within PARSEME (PARSing and Multiword Expressions)[1] in Working Groups 1 and 4 – WG1: Lexicon-Grammar Interface and WG4: Annotating MWEs in Treebanks. The first one focuses on the linguistic properties of MWEs (structure, reflexes to alternations such as passivisation, etc.) and is more detailed, while the second one is treebank-related and thus focuses on a different set of MWE features such as the structural correspondences among MWEs across languages and the distributions observed in corpora.

The results from the empirical study highlight at least the following issues: (1) realization options of different MWE types in two languages with different morphological complexity and word order; (2) a data-driven typology of alignment possibilities among various types of MWEs; (3) ~~modeling~~ the bilingual data with a catenae-based approach.

The paper is structured as follows: §2 outlines the related work; §3 introduces catenae in a more formal way and also describes the main operations that can be applied on them; §4 presents bilingual catenae; §5 describes the parallel data and its classification; and §6 concludes the paper.

2 Related work

This section comprises two parts: a discussion on MWE classification and a presentation of catenae. Concerning the former, there is extensive literature regarding the study of MWEs within a language and across languages, theoretical issues on MWE modelling, etc. Here only some of them will be mentioned. To the best of our knowledge, this is the first attempt to use catenae for modelling bilingual or multilingual MWE correspondences.

2.1 MWE classifications

There is no widely accepted classification of MWEs (Villavicencio & Kordoni 2012). For the task of automatic recognition of MWEs in Bulgarian Stoyanova

[1] PARSEME is an interdisciplinary scientific network devoted to the role of multiword expressions in parsing – IC1207 COST Action.

(2010) adopts the classification of Baldwin et al. (2003). This classification could be characterized as a semantically oriented division, since the MWEs are classified as non-decomposable by meaning, idiosyncratically decomposable and simple decomposable.

In Sag et al. (2002) another classification is proposed. The MWEs are divided into lexicalized phrases and institutionalized phrases. Here we do not consider institutionalized phrases (being semantically and syntactically compositional, but statistically idiosyncratic) as a distinct group. Lexicalised phrases are further subdivided into fixed expressions, semi-fixed expressions and syntactically flexible expressions. Fixed expressions are said to be fully lexicalized and undergoing neither morphosyntactic variation nor internal modification. Semi-fixed expressions have a fixed word order, but "undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection," Sag et al. (2002: 4) including non-decomposable idioms and proper names. Syntactically flexible expressions allow for some variation in their word order (light verb constructions, decomposable idioms).

On the multilinguality front, there are various approaches to different MWE-related problems. For example, in Rácz et al. (2014) the multilingual annotation of light verb constructions is discussed for English, Spanish, German and Hungarian. The specific annotation properties of these elements are described for each language. Another popular task is the construction of bi- or multilingual MWE lexicons on the base of parallel or comparable corpora. In Seo et al. (2014) a context-oriented method is proposed for French and Korean.

The WG4 classification was specially tailored to reflect the typology of MWEs in syntactically annotated corpora (treebanks). It divides MWEs into the following groups on the basis of the parts-of-speech (PoS) of the head word:

1. Nominal MWEs
2. Verbal MWEs
3. Prepositional MWEs
4. Adjectival MWEs
5. MWEs of other categories
6. Proverbs

Some of these groups are further subdivided into subtypes: *Nominal MWEs* including named entities (NEs), nominal compounds as well as other nominal

MWEs and *verbal MWEs* including phrasal verbs, light verb constructions, VP idioms and other verb MWEs. Thus, the WG4 classification is syntax-based.

WP1 classification elaborates the typology by studying idiomaticity and flexibility on the basis of a large set of morphosyntactic diagnostics. With respect to flexibility, the WG1 approach differs from Sag et al. (2002) in providing a coarser division between semi-flexible and flexible MWEs. With respect to idiomaticity, the classification is based on Baldwin & Kim (2010). It handles five types: lexical, syntactic, semantic, pragmatic and statistical idiomaticity. Our work deals with the syntactic and semantic idiomaticity in a bilingual context.

2.2 Catena

The notion of catena 'chain' was introduced in O'Grady (1998: 284) as a mechanism for representing the syntactic structure of idioms. He shows that for this task there is need for a definition of syntactic patterns not coinciding with constituents. A variant of this definition was offered by Osborne (2006):

> The words A, B, and C (order irrelevant) form a chain if and only if A immediately dominates B and C, or if and only if A immediately dominates B and B immediately dominates C. (Osborne 2006: 258)

In recent years the notion of catena revived again and was applied to dependency representations. Catenae have been used successfully for the modelling of problematic language phenomena. Gross (2010) presents the morphological and syntactic problems that have led to the introduction of the subconstituent catena level. Constituency-based analysis has to deal with non-constituent structures in ellipsis, idioms, and verb complexes.

Apart from the linguistic modelling of language phenomena, catenae have been used in a number of NLP applications. Maxwell et al. (2013), for example, present an approach to Information Retrieval based on catenae. The authors consider the catena as a mechanism for semantic encoding which overcomes the problems of long-distance paths and elliptical sentences. Also, Sanguinetti et al. (2014) present a catena-related approach for syntactic alignments in multilingual treebanks. In translation research, catenae are best known as "treelets" (Quirk & Menezes 2006). We employ catenae, which have already been used in NLP applications, to model the interface between the treebank and the lexicon.

A first attempt to formalise MWE information with catenae is discussed in Simov & Osenova (2015). In the next section we present the main notions of our proposal.

3 Definition of catena. Operations on catenae

We follow the definition of catena provided by O'Grady (1998) and Gross (2010): a CATENA is a word or a combination of words directly connected with dominance relations. In fact, in the domain of dependency trees, this definition is equivalent to a subtree definition. Figure 1 shows a complete dependency tree and some of its catenae. Notice that the complete tree is also a catena. Individual words are catenae, too. With "root$_C$" we mark the root of the catena that might be identical with the root of the complete tree, but it also might be different as in the case of *John* and *an apple* in Figure 1.

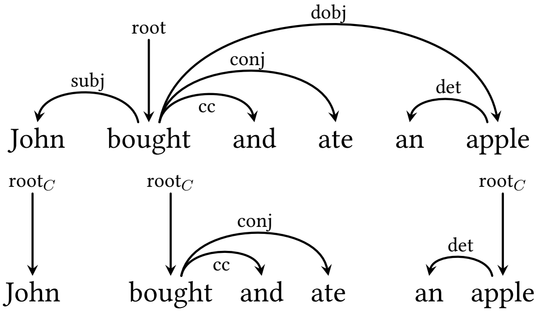


Figure 1: A complete dependency tree and some of its catenae.

A catena as an object on its own is a tree in which the nodes are decorated with various labels including word forms, lemmas, and parts-of-speech; the grammatical features and the arcs are augmented with dependency labels. The labeling function is partial. Thus, some nodes or arcs remain non-decorated in the catena and allow for different mappings to dependency trees. When the catenae are not mapped on dependency trees, they are considered part of the lexicon or the grammar of a given language.

We call the mapping of a catena onto a given dependency tree the *realization of the catena in the tree*. We consider the realization of the catena as a fully specified subtree including all the nodes and arc labels. Each realization of a catena has to agree with its labeling outside of the dependency tree. For example, the catena for *(to) spill the beans* will allow for any realization of the verb form like in: *they spilled the beans* and *he spills the beans*. Thus, the catena in the lexicon will be underspecified with respect to the grammatical features and word forms for the corresponding lexical items.

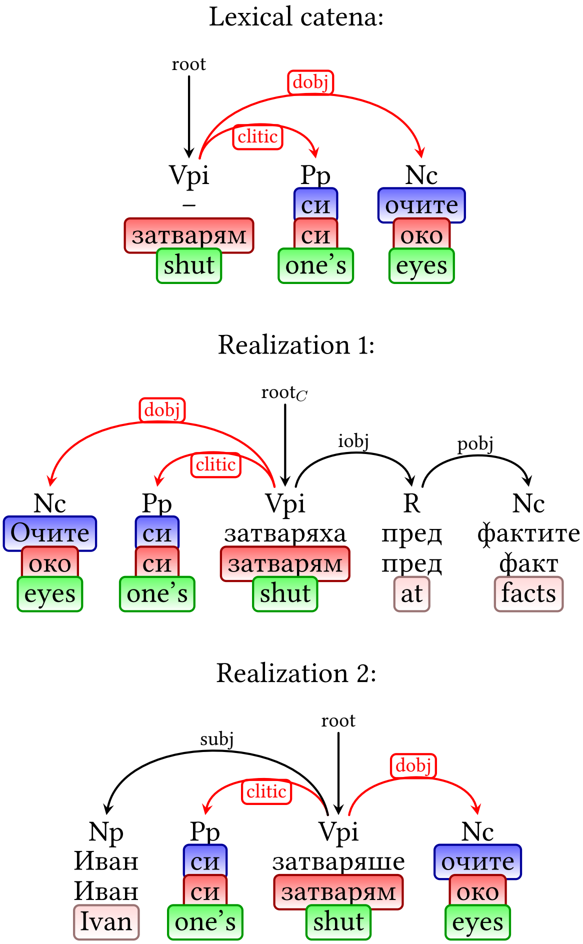


Figure 2: Catena realization.

In this paper, the underspecified catena is called a *lexicon catena* (LC) and it is stored in the lexical entries. Figure 2 shows a lexical catena for the idiom затваря-м си оч-и-те, zatvarya-m si ochite, close-PRS.1SG REFL eye-PL-DEF, lit. 'shut one's eyes' and two of its realizations. Catenae in the lexicon do not specify any particular word order.[2] The word order of the catena realization reflects the rules of the grammar, therefore, the realisation of the same catena in different dependency trees could materialise with different word orders.

[2] Formalisation of the word order within the catena remains an open question for future work.

The upper part of the image in Figure 2 represents the lexicon catena for the idiom. It determines the fixed elements of the catena: arcs and their labels as well as nodes and their labels. More precisely, the following information is included: extended part of speech (PoS),[3] word forms, and lemmas.[4] The translations of the word form are presented, too. A dash (–) under a node indicates that the corresponding element is not defined for the given node. In Figure 2, the dash represents the fact that the word form for the verb node is underspecified, therefore the idiom can be marked with a variety of tense, person and other values.

In the two realizations, the fixed elements of the catena are represented as in the lexicon catena. Thus, the lemmas are the same as the word forms, the parts-of-speech and the grammatical features for the direct object and for the clitic are also the same. The realizations are different from the lexicon catenae with respect to the word forms and the grammatical features of the verb node: in both examples the verb is in past tense while in the first realization it is in plural and in the second in singular number. The word order in the two realizations is different. Thus, the underspecified catenae representation allows for various levels of morphosyntactic and semantic flexibility within the multiword expressions.

The catena representation of the lexical items explicitly denotes their properties that constrain their interaction. We proceed to show how we model the selectional restriction of a given lexical unit with respect to a catena in a sentence. The main operation for ~~modeling~~ the interactions among the catenae is called COMPOSITION. For example, let us assume that the verb *to read* requires that its subject denotes a human and that its object denotes an information object. In Figure 3 we present how the catena for *I read* is combined with the catena *a book* in order to form the catena *I read a book*. The figure represents the level of word forms and the level of semantics (specified only for the node, on which the composition is performed). The catena for *I read …* specifies that the unknown direct object has the semantics of an *Information Object* (InfObj). The catena for *a book* represents the fact that the book is an Information Object. Thus the two catenae are composed on the two nodes marked as InfObj. The result is represented at the lower part of Figure 3. We have defined the composition operation for catenae that agree with each other on one node; the operation can be defined on more agreeing nodes.

[3] The extended parts of speech are defined as prefixes of the tags in the BulTreeBank tagset: http://www.bultreebank.org/TechRep/BTB-TR03.pdf

[4] In some examples we give the important information only, thus, some of these rows are missing. In some examples new rows are used to introduce additional information.

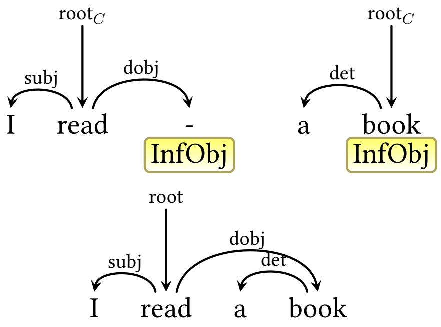


Figure 3: Composition of catenae.

In Figure 4 the structure of the lexical entry for the verb бяга-м, byaga-m, run-PRS.SG, 'run' is presented in the sense 'run away from facts'. The verb selects an indirect object in the form of a prepositional phrase introduced with the preposition *от*, ot, 'from'. In Figure 5 we give the catena for the synonymous MWE затварям си очите, zatvaryam si ochite, close.PRS.1SG REFL eye.PL, 'I close my eyes'.

The lexical entry of a MWE uses the format: a **lexicon-catena**, **semantics** and **valency**.[5] Lexicon-catenae for the MWEs are stored in their canonical form. The semantics part of a lexical entry is represented with a logical formula comprising elementary predicates. The role of possible modifiers has to be specified in the lexicon-catena, if modification of the MWE is possible, for instance when structures with modifiers of the noun can be attested in the data. For example, the MWE затварям си очите, zatvaryam si ochite, close.PRS.1SG REFL eye.PL.DEF, which is synonymous to the verb бягам, byagam, run.PRS.1SG, is presented in Figure 5.[6] The valency level is built as follows: the root of the valency catena is marked with the identifier of the node in the lexical catena for which the particular valency representation is applicable. In Figure 5 the valency representation is applicable to the root node CNo1 of the lexical catena. The two catenae are composed on this node. The composition is applied to the semantics of the lexical catena and of the valency catena. Note that the nodes No1 and No2 are different from the nodes CNo1 and CNo2.

We use catenae to represent both single words and MWEs because single words are also catenae by definition.

[5] The corresponding fields in the lexical entry (rows in the tables below) are marked as: LC, SM, Fr (for valency frames).

[6] The grammatical features are: 'poss' for possessive pronoun, 'plur' for plural number and 'def' for definite noun.

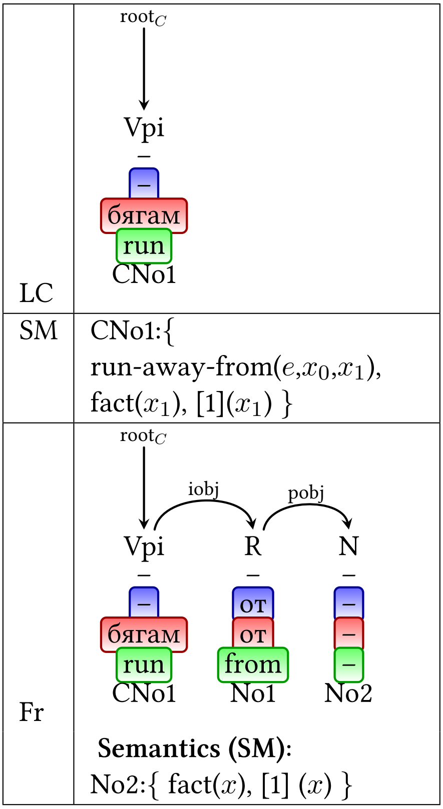


Figure 4: Lexical entry for the verb *run*.

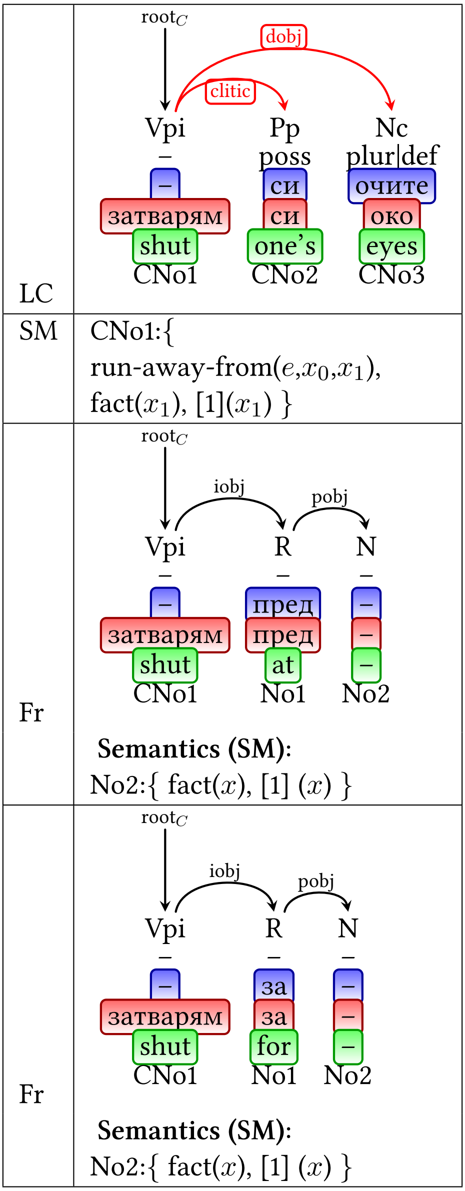


Figure 5: Lexical entry for *I close my eyes*.

We can specify all the grammatical features of a lexical item using the formal definition of catena given above. The semantics defined in the lexical entry can be attached to each node in the lexicon-catena. In Figure 4 there is just one node of the lexicon-catena. In this paper, we present only the set of elementary predicates rather than providing their full semantic structures because we focus on the principles of the representation.[7] In Figure 4 the verb introduces three elementary predicates: *run-away-from*(e, x_0, x_1), *fact*(x_1), [1](x_1). The predicate *run-away-from*(e, x_0, x_1) represents the event and its main participants: x_0, x_1. The predicate *fact*(x_1) is part of the meaning of the verb in the sense that the agent represented by x_0 will run away from some (unpleasant) situation. The underspecified predicate [1](x_1) has to be compatible with the predicate *fact*(x_1). This predicate is used for incorporating the meaning of the indirect object *at something* in the frame *shut one's eyes at something*. The valency frame is given as a set of valency elements defined as a catena with a semantic description. The catena describes the basic structure of the valency element including the necessary lexical information, grammatical features, and the syntactic relation to the main lexical item. The semantic description determines the main semantic contribution of the frame element and is incorporated in the semantics of the whole lexical item with structural sharing. In Figure 4 there is only one frame element. It is introduced with the preposition от, ot, 'from'. The semantics originates in the dependent noun that has to be compatible with the predicate *fact*(x) and in the underspecified predicate [1](x_1), that may introduce a specific predicate. Via the structure sharing index [1], this specific predicate is copied on the semantics of the main lexical item.

The lexical entry in Figure 5 is similar to the one shown in Figure 4. The main differences are: the lexicon-catena represents a MWE and not a single word. The semantics is the same, because the verb and the MWE are synonyms. The valency frame contains two alternative elements for indirect object introduced by two different prepositions. The conclusion that the two descriptions are alternatives follows from the fact that the verb has only a free indirect object slot. If a direct object slot was free as well then the valency set would contain elements to fill also this slot; however, in the MWE presented, the direct object slot is occupied by a fixed element.

In a nutshell, catenae are an appropriate mechanism for the representation of MWEs because they adequately encode the grammatical flexibility of some elements within the MWEs and also allow for the informative representation of single words.

[7] For a full semantic representation we employ *Minimal Recursion Semantics*, introduced by Copestake et al. (2005).

In the rest of the paper we extend the above lexicon model in order to handle correspondences among translation pairs with at least one MWE as a member.

4 Bilingual catena modelling

In this section we show the treatment of the following bilingual types of pairs in Bulgarian and English: MWE-to-MWE and MWE-to-word. Our survey is corpus-driven and we have chosen to discuss the most frequent pairs in our data (see next section for data statistics).

4.1 MWE-to-MWE

Let us consider the example:

(1) Example RD:[8] 'reach a decision' взема решение, vzema reshenie, take.PRS. 1SG decision.

The two MWEs are flexible in several ways. First, the verb *reach* (and the corresponding one in Bulgarian взема, vzema, 'take, get') ~~allow~~ for morphological variation, including tense, person, etc. The noun *decision* allows for pre- and post-modifiers as in: *we reached an important decision* or *they will reach a decision about us tomorrow*. The Bulgarian MWE presents the same behavior. Figure 6 shows the lexical entry for the parallel MWEs that are modeled as catenae. In the lexical entries we can see the catenae for both MWEs. In the next row, the semantics of the parallel MWEs is represented with a set of elementary predicates coupled with a coindexation strategy between the semantics of the MWE and its frame semantics.

In Figure 6, the indices [1] and [2] represent the unknown semantics of the modifying nouns. If no modification phrases exist, these predicates are assumed to express the most general one, namely *everything(x)*. Thus, the set {take-decision(e,x_0,x_1), decision(x_1,x_2), [1](x_1), problem(x_2), [2](x_2)} represents the meaning of the MWE[9]: event "take-decision" e with two participants x_1 and x_2. The participant x_0 is the agent who takes the decision. The participant x_1 is the main argument of the predicate for the relational noun *decision* that, being a two-argument predicate, introduces a third participant in the event, namely the

[8] We use a special notation after each example: RD, IG and CH for ensuring the correct connection with the corresponding pictures in Figures 6, 7, and 8.

[9] The examples present light verb constructions that are translational equivalents between Bulgarian and English.

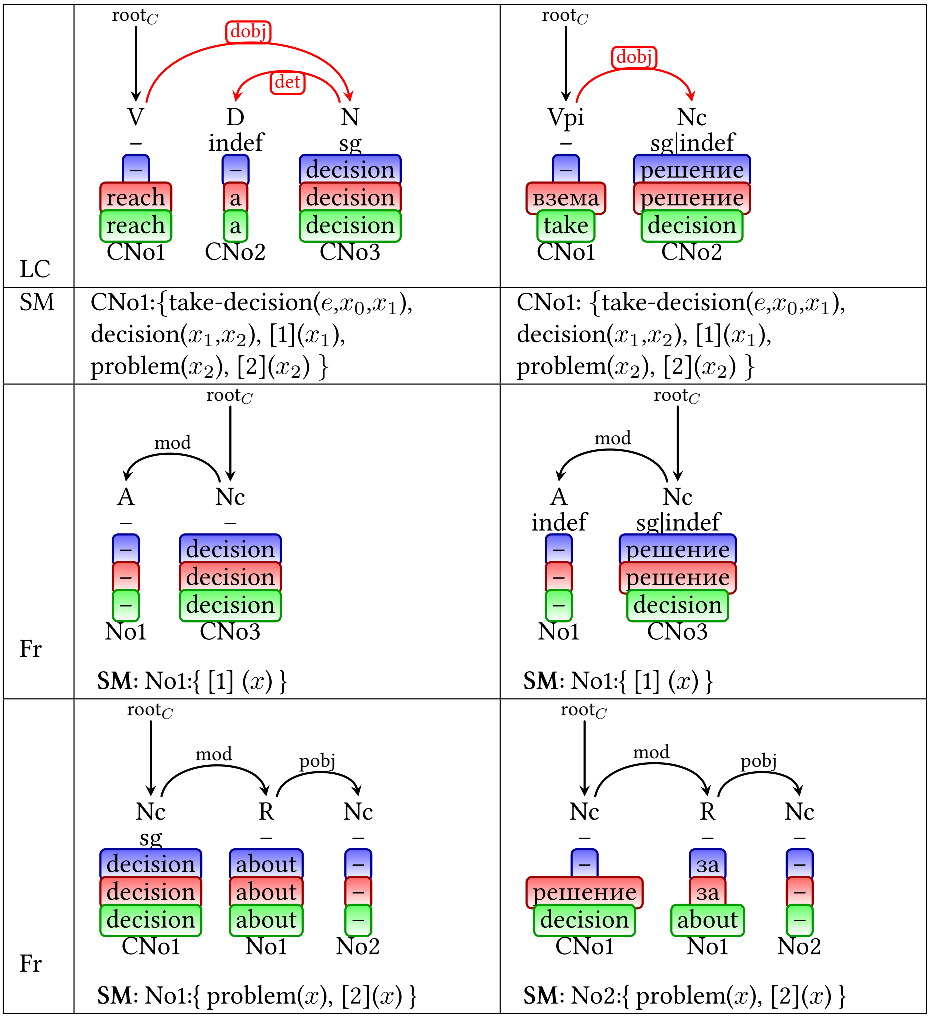


Figure 6: Parallel Lexical Entries for the parallel MWEs: **example RD**.

problem that the decision is about, denoted with the variable x_2. If along with the lexicon catena the frame catena is also realized in the sentence, then the new predicates introduced by the corresponding nouns are added to the semantics of the new bigger catena. This mechanism of representing bilingual lexicon entries is suitable for the processing of the bilingual information including the shared representation of the semantics and correspondences between the grammatical features of the parallel realisations of the catenae in the different languages.

In some cases the lexical entry of the parallel MWEs might be quite simple, as in the following example:

(2) Example IG: ‘in general’, като цяло, kato tsyalo, as whole.

In Figure 7 the adverbials share the same semantics. They do not have frames and they allow for no modification. Only the PoS assigned to their elements may be different.

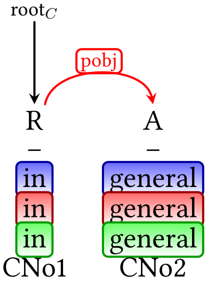
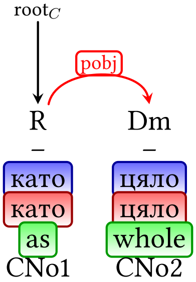
LC	 root$_C$ pobj R A – – in in in general general general CNo1 CNo2	 root$_C$ pobj R Dm – – като като as цяло цяло whole CNo1 CNo2
SM	CNo1:{ generally(e,e_1) }	CNo1:{ generally(e,e_1) }
Fr		

Figure 7: Parallel Lexical Entries for the parallel MWEs: **example IG**.

4.2 MWE-to-word

Concerning the relation MWE-to-word irrespectively of the language direction, two main cases can be observed. The first one relates to functional PoS, such as the English preposition *after* and the Bulgarian complementiser след като, sled kato, after when, that are translational equivalents and have identical semantics but differ in PoS and some selectional properties.

A challenging problem occurs when non-functional counterparts are considered. For example, the term

(3) Example CH: the English term *chemicals* translates into the Bulgarian MWE химическ-и продукт-и, himichesk-i produkt-i, chemical-PL product-PL, chemical products.

Both expressions might be modified by adjectives, PPs or clauses: ***dangerous*** *chemicals*, *chemicals* ***from airplanes***, and *chemicals* ***that are used by the pharmaceutical industry***. We find similar examples in Bulgarian like отровни химически продукти, otrovni himicheski produkti, poisonous.PL chemical.PL product.PL, poisonous chemical products.

In Figure 8 a part of the parallel lexical entry for this example is presented. It can be seen that in the English part of the lexical entry there is a catena for a single word while in the Bulgarian part there is a catena for a noun phrase of type adjectival modifier - head noun. The catena for the Bulgarian MWE is underspecified for the word form and the grammatical features because the whole phrase might be definite: химически**те** продукти, himicheski**te** produkti, '**the** chemicals'. The English and the Bulgarian entries are specified for the same semantics. In the frame part of the lexical entries all possible modifications have to be defined (in the example just one of them is given, namely left modification with adjectives; however, modification with PPs has been encountered in the data, etc.). The important point here is that the lexicon catenae for the two languages have to contain appropriate correspondences of the frames in order to be proper translations of each other. The correspondences of the frames have to be established on semantic grounds—the corresponding frames in the English and the Bulgarian part have to define the same semantic contributions to the lexical catenae.

The frame catena in Figure 8 marks the fact that the lexical catena can be modified by an adjectival modifier. The realization of such a modifier is additional to the realization of the adjectival modifier *химическ-и*, himichesk-i, chemical-PL that is a fixed part of the MWE. In the frame catena we mark only the nominal head of the MWE.

Note that we do not aim at an exhaustive analysis of all the bilingual pairs. Our aim is to present a mechanism which would deal with both—symmetric (MWE-to-MWE) and asymmetric (MWE-to-word) relations in translations. Our hypothesis is that the correspondences between the two languages in the lexicon have to be governed by the semantics of the lexical catenae and the semantic contribution of the possible frames. A consequence of this hypothesis is that, in the lexicon, we have to allow for correspondences not only between MWEs, but also between MWEs and words, and between words/MWEs in one of the languages and compositional phrases in the other.

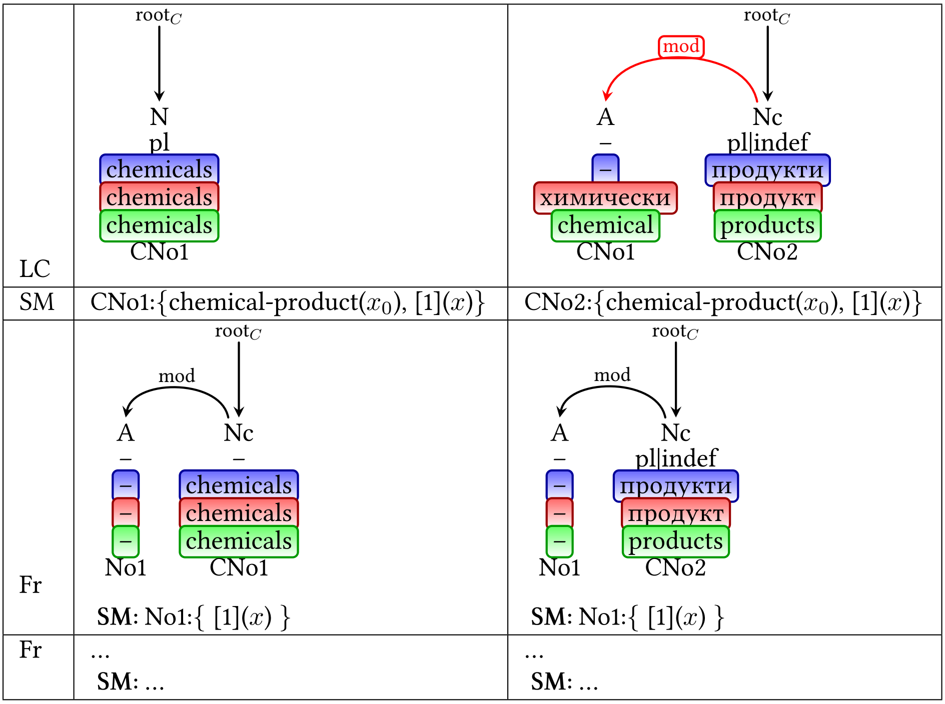


Figure 8: Parallel Lexical Entries for the parallel MWEs: **example CH**.

5 Classification of the parallel data

In this section we provide a classification of parallel pairs that consist of two MWEs or an MWE and a word. For each class of correspondences the minimum information to be included in the lexical entries has been specified. The parallel Bulgarian-English newsmedia corpus consists of two parts: SETimes plus CSLI dataset (920 sentences, or 9308 tokens); PenTreebank dataset (838 sentences, or 21949 tokens). Thus, our final dataset consists of: 1758 sentences or 31 257 tokens.

The data was aligned according to Simov et al. (2011). However, the alignments did not mark the MWEs. For that reason, additional annotation was performed for detecting the alignments with MWEs in at least one of the two languages.

Our aim was to extract various types of alignments with at least one MWE as a member. Thus, our data included the following general types: MWE-to-word; MWE-to-MWE and MWE-to-compositional phrase in both language directions.

As shown in Table 1, ~~636~~ occurrences of MWEs were detected within these data. 370 MWEs of these occurrences are of type MWE-to-word (for example

Table 1: General Classification.

	Occurrences
MWE-to-MWE	126
MWE-to-word	370
MWE-to-phrase	14
Total	~~636~~

Table 2: MWE-to-Word classification.

	MWE-to-word
Bulgarian MWE	220
English MWE	150
Total	370

the English *within* is translated as *в рамките на*, v ramkite na, in frame.PL of, within the framework of); 126 MWEs are of type MWE-to-MWE (for example the English *with respect to* is translated as *що се отнася до*, shto se otnasya do, as far as relate.PRS.3SG to), and 14 MWEs are of type MWE-to-phrase (for example, the English *take-it-or-leave it* is translated as *приемаш или се отказваш*, priemash ili se otkazvash, accept.PRS.2SG or refuse.PRS.2SG).

Table 2 shows the distribution of MWEs in the largest set, namely the set of the type MWE-to-word: 220 Bulgarian and 150 English MWEs were detected.

Two types of classification are applied. First, the aligned pairs are classified into three groups: MWE-to-MWE, MWE-to-word and MWE-to-phrase (see Tables 1 and 2). This classification offers a coarse picture of the bilingual situation. Then, the classification methods developed in PARSEME WG1 and WG4 are applied. These classifications draw on the structural and the semantic features of MWEs.

When mapped to the PARSEME WG1/WG4 typologies, both languages showed very similar MWE properties. Thus, the most frequent MWE types in both languages are: *verbal MWEs*; *noun MWEs*; *other categories of MWEs*. The language specific features are evident in the subtypes. Thus, phrasal verbs and reflexive (formally or semantically) ce-verbs, se-verbs, seem to be the most frequently used verb MWEs in the English and Bulgarian data respectively. Both languages feature light verb constructions and VP idioms. Lastly, compounds are the most frequent type of noun MWEs in English while adjective-noun phrases are in Bulgarian.

To present a slightly more detailed analysis of the correspondence type MWE-to-MWE, we use the WG1 classification (predominantly the syntactic and semantic dimensions), that focuses on the internal structure of the MWEs.

Within the set of the *MWE-to-MWE pairs*, correspondences are grouped to straightforward mappings and to cross-language specific types. A presentation of these two groups follows.

5.1 Straightforward mappings

The class of straightforward mappings includes: verb MWEs (light verb constructions, VP idioms) and other categories (adverbs, prepositions), etc.

In this group of translation equivalent, two main classes of Bulgarian–English MWE pairs are identified: pairs with cross-lingual variance that have to be considered in the lexicons, and MWEs with no cross-lingual variance that are trivially handled in the lexicon. In the first case, the grammatical behavior of the MWE elements in both languages has to be taken into account, such as the possibility of inflection for number, or of accepting modifiers. In the second case, the MWE elements hardly undergo inflection or modification, so the translational equivalents are registered in the lexicon without further elaboration on the behavior of their elements.

The first case includes verb and noun MWEs and the second one complex PoS and non-inflecting MWEs.

Examples for the first group are given below:

- Light verbs in one language often correspond to similar constructions in the other. For instance,

  ‘reach a decision’, взем-а решение, vzem-a reshenie, take-PRS.3SG decision

  where V NP in English translates to V NP in Bulgarian,

  ‘take effect’, влез-е в сила, vlez-e v sila, enter-PRS.3SG in power,

  ‘take control’, влез-е във владение, vlez-e vav vladenie, enter-PRS.3SG in possession

  where V NP translates to V PP.

  In this group the MWEs are assigned identical semantics, but they might differ in the elements and in valence selection.

- Noun MWEs of the type A N that are translational equivalents, often are literal translations of each other:

  ‘tough line’, твърда позиция, tvarda pozitsiya, tough position,

  ‘free market’, свободни-я пазар, svobodni-ya pazar, free-DEF market,

  ‘real estate’, недвижимо-то имущество, nedvizhimo-to imushtestvo, nonmoving-DEF property.

  The MWEs in this group share the same semantics and the same modification mechanisms.

- The structure V NP tends to characterise both members in pairs consisting of verb MWE translational equivalents:

  ‘is drawing fire’, привлич-а критик-и-те, privlich-a kritik-i-te, attract-PRS.3SG critic-PL-DEF,

  ‘haven’t got a clue’, няма-т представа, nyama-t predstava, not.have.PRS.3PL idea.

  The MWEs in this group are assigned the same semantics, but vary in their elements and valence selection.

Examples for the second group are given below:

- Multiword adverbial constructions:

  ‘on the other hand’, от друга страна, ot druga strana, from other side,

  ‘of course’, разбир-а се, razbir-a se, understand-PRS.3SG REFL,

  ‘more and more’, все повече и повече, vse poveche i poveche, even more and more,

  ‘in particular’, в частност, v chastnost, in detail.
  Here, however, the prepositional complement varies in the PoS across the two languages. For example, in the last translational equivalent the English prepositional complement is the adjective *particular*, while in Bulgarian it is the noun частност, chastnost.

  The MWEs in this group are assigned the same semantics, but may vary in the elements. However, this difference is not taken into consideration, because the elements hardly inflect and do not allow for insertion of additional elements.

- Complex prepositions in English tend to have structurally similar counterparts in Bulgarian. For instance,

  ‘with respect to’, ~~що се отнас-я до~~, ~~shto se otnas-ya do,~~ ~~as far as relates PRS.3SG to.~~

  The MWEs in this group are assigned the same semantics, but since, presumably, they are assigned the same PoS and do not inflect, the element variance is not relevant.

- Conjunctions composed of multiple words:

  ‘as well as’, както и, kakto i, as and.

Like the complex preposition group, this group also contains MWEs that are assigned the same semantics and the same PoS; these MWEs do not inflect, therefore the element variance is not relevant.

5.2 Cross-language specific types

Here we include English phrasal verbs having Bulgarian reflexive ce-verbs, se-verbs, as translational equivalents and English nominal compounds having Bulgarian

other NP MWEs, mainly adjective-noun or noun-preposition-noun, as translational equivalents. In this group, translational equivalents are assigned the same semantics, but they may present systematic structural differences due to language specific constructions. The elements in the MWEs always differ across languages.

- English phrasal verbs often correspond to Bulgarian ce-verbs, se-verbs:

  ‘give up’, се откаж-е, se otkazh-e, REFL decline-PRS.3SG,

  ‘move back’, се върна-т, se varna-t, REFL return-PRS.3SG.

  Bulgarian and English MWEs in this group may differ in valency and in the way meaning is constructed. Thus, Bulgarian uses the lexical aspect and the reflexive ‘ce’, se, to construct MWE meanings, while English uses the verb in combination with the phrasal affix.

- English N N compounds can map to A N compounds in Bulgarian:

  ‘face amount’, номинална стойност, nominalna stoynost, nominal value.

The MWEs in this group differ in the PoS of the modifier of the head noun: with Bulgarian A N MWEs the head noun is modified by an adjective and with English N N MWEs by a noun.

- English N N can also be translated as N PP in Bulgarian. The first N in the English MWEs and the PP in the Bulgarian MWEs make the same semantic contribution:

  ‘law enforcement’, сил-и-те на ред-а, sil-i-te na red-a, force-PL-DEF of order-SG.DEF.

  The MWEs in this group differ in the PoS of the modifier of the head noun: with Bulgarian N NP MWEs the head noun is modified by a PP and with English N N MWEs by a noun.

- English N and N constructions can apparently be translated with coordinated constructions in Bulgarian; however, the PoS of the coordinated constituents differs across the two languages:

  ‘pros and cons’, доводи за и против, dovodi za i protiv, argument.PL for and against
  (N and N / N p and p).

  The MWEs in this group differ in the head obligatoriness. In Bulgarian the head noun is present, while in English a head noun is only inferred.

- An English idiomatic clausal construction (V NP PP) can be translated with a light verb construction in Bulgarian:

  ‘putting pen to paper’, предприел действие, predpriel deystvie, take. PTSP.3SG action.

  The MWEs in this group differ with respect to modification and selectional properties. The English MWE does not seem to admit any modifiers, while its Bulgarian translational equivalent allows for them (for example, предприел *важно* действие, predpriel vazhno deystvie, taken.PTSP.3SG, important action

- English V AP can be translated in Bulgarian with minimal changes into V AdvP:

  ‘broke even’, са излезли начисто, sa izlezli nachisto, are come.out.PRST.3SG clean.

  The English adjective *even* translates into the Bulgarian adverb *начисто*, nachisto, ‘clean’.

- English V PP can be translated as V NP in Bulgarian:

  ‘will be priced of a job’, ще загубя-т работа-та си, shte zagubya-t rabota-ta si, will lose-PRS.3SG job-DEF.

  It is interesting to observe that an English passive construction can be translated with a Bulgarian active construction. In such cases the valency parts will differ with respect to both the predicate and the participants.

Our work on the Bulgarian-English lexicon aims to provide representations for all these types of correspondence: the representations will be bilingual catena-based lexical entries.

6 Conclusions

The paper has argued that the catena approach can be extended to model pairs of translational equivalents retrieved from parallel English-Bulgarian corpora with at least one MWE as a member. In this way, cross-language asymmetries are handled. Our frequency counts have shown that the *MWE-to-MWE* and *MWE-to-word* correspondences are prevalent. In contrast, the *MWE-to-phrase* correspondence was not found to have a wide distribution. It would be interesting to perform a detailed analysis of more examples in order to uncover persistent correspondences between the two languages. Such knowledge can be used in designing automatic translation systems and in identifying best practices in human translation. Furthermore, these correspondences can possibly illuminate the different ways employed by the two languages to express meaning.

The proposed catena model takes into consideration both flexibility and idiomaticity when representing MWEs and words in the lexicon. These dimensions can be detailed further depending on the available specific subclassifications in a cross-lingual aspect.

Acknowledgments

This research has received support by the EC’s FP7 (FP7/2007-2013) under grant agreement number 610516: “QTLeap: Quality Translation by Deep Language Engineering Approaches” and by European COST Action IC1207: “PARSEME: PARSing and Multiword Expressions. Towards linguistic precision and computational efficiency in natural language processing.”

Abbreviations

check abbr for small-caps

def	definite noun	poss	possessive pronoun
LC	lexicon catena	SM	semantics
POS	part of speech	FR	valency frames
plur	plural number		

References

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka & Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, vol. 18, 89–96. Sapporo.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 2nd edn., 267–292. Boca Raton: CRC Press.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation* 3(4). 281–332.

Gross, Thomas. 2010. Chains in syntax and morphology. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto & Yasunari Harada (eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 143–152. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. http://www.aclweb.org/anthology/Y10-1018.

Maxwell, K. Tamsin, Jon Oberlander & W. Bruce Croft. 2013. Feature-based selection of dependency paths in ad hoc Information Retrieval. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 507–516. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/P13-1050.

O'Grady, William. 1998. The syntax of idioms. *Natural Language and Linguistic Theory* 16. 279–312.

Osborne, Tim. 2006. Beyond the constituent – A Dependency Grammar analysis of chains. In *Folia ~~linguaistica~~*, vol. 39, 251–297.

Quirk, Christopher & Arul Menezes. 2006. Dependency treelet translation: The convergence of statistical and example-based machine-translation? *Machine Translation* 20(1). 43–65. http://dblp.uni-trier.de/db/journals/mt/mt20.html#QuirkM06.

Rácz, Anita, Istvan Nagy T. & Veronika Vincze. 2014. 4FX: Light verb constructions in a multilingual parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. 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)*, 1–15.

Sanguinetti, Manuela, Cristina Bosco & Loredana Cupi. 2014. Exploiting catenae in a parallel treebank alignment. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Seo, Hyeong-Won, Hong-Seok Kwon, Min-ah Cheon & Jae-Hoon Kim. 2014. Constructing bilingual multiword lexicons for a resource-poor language pair. *Advanced Science and Technology Letters* 54. 95–99.

Simov, Kiril & Petya Osenova. 2015. Catena operations for unified dependency analysis. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 320–329. Uppsala, Sweden: Uppsala University, Uppsala, Sweden. http://www.aclweb.org/anthology/W15-2135.

Simov, Kiril, Petya Osenova, Laska Laskova, Aleksandar Savkov & Stanislava Kancheva. 2011. Bulgarian-English parallel treebank: Word and semantic level alignment. In *Proceedings of the Second AEPC Workshop*, 29–38. http://www.aclweb.org/anthology/W11-4305.

Stoyanova, Ivelina. 2010. Factors influencing the performance of some methods for automatic identification of multiword expressions in Bulgarian. In *Proceedings of the 7th FASSBL Conference*, 103–108.

Villavicencio, Aline & Valia Kordoni. 2012. *There is light at the end of the tunnel: Multiword expressions in theory and practice, course materials*. Tech. rep. Technical report, Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT).