Handling Non-compositionality in Multilingual CNLs

Ramona Enache, Inari Listenmaa, and Prasanth Kolachina

University of Gothenburg, Sweden
{ramona.enache,inari.listenmaa,prasanth.kolachina}@cse.gu.se

Abstract. In this paper, we describe methods for handling *multilingual* non-compositional constructions in the framework of GF. We specifically look at methods to detect and extract non-compositional phrases from parallel texts and propose methods to handle such constructions in GF grammars. We expect that the methods to handle non-compositional constructions will enrich CNLs by providing more flexibility in the design of controlled languages. We look at two specific use cases of non-compositional constructions: a general-purpose method to detect and extract multilingual multiword expressions and a procedure to identify nominal compounds in German. We evaluate our procedure for multiword expressions by performing a qualitative analysis of the results. For the experiments on nominal compounds, we incorporate the detected compounds in a full SMT pipeline and evaluate the impact of our method in machine translation process.

1 Introduction

The work describes a series of methods used to enrich multilingual CNLs written in the grammar formalism GF (Grammatical Framework)[20] with multilingual multiword expressions (MMWEs). This aims to give a better separation between compositional and non-compositional constructions in GF applications and a better understanding on representing MMWEs in GF. We present two new GF modules: one for constructions in a multilingual setting, and one specifically for German compound nouns.

We are targeting cases where translation equivalents have different syntactic structure: this covers pairs such as English–French (*apple juice*, *jus de pommes* 'juice of apples') and English–Finnish (*kick the bucket*, *heittää henkensä* 'throw one's life'). Only the latter pair contains a monolingually non-compositional structure, i.e. having an interpretation that cannot be inferred from the components, but we consider both of them as MMWEs, due to the non-compositionality of translation.

We propose a solution to this, that relies on prior analysis of the domain, since GF applications are normally developed starting from positive examples covering the domain [22]. We start from a parallel corpus describing the scope of the grammar and identify MMWEs in order to add them to the grammar as special constructions.

A special case of MMWEs, which we treat separately is that of nominal compounds in German. The need for a multilingual lexicon of such compounds and their translations originated from the use of GF in machine translation [10], [11]. This use case is of particular interest, since it is easier to evaluate—both in terms of precision and recall of the method, and in terms of impact on the machine translation process.

This paper is structured as follows: Section 2 describes the background and related work; Section 3 describes the implementation of the general MWE detection and compound detection methods; Section 4 describes a preliminary evaluation, and finally Section 5 describes future work.

2   Background and Related Work

2.1   Grammatical Framework

GF (Grammatical Framework) is a grammar formalism particularly fit for multilingual natural language applications. In the recent years, it has been used extensively for developing (multilingual) CNLs, such as the in-house implementation of Attempto Controlled English [21], domain-specific applications for mathematical exercises [25], [27], [26], speech-based user interfaces [14], tourist phrases [23], business models [8] and cultural heritage artifacts [6], [7].

Applications written in GF are represented by their *abstract syntax*, which models the semantics of the domain in a language-independent fashion, and a number of *concrete syntaxes*, mapping the semantics to a number of target languages, most commonly natural languages.

The difficulty when dealing with compositional and non-compositional constructs in GF arises, in fact, from the multilingual character of the applications. It is of particular difficulty to design the abstract syntax in a way that accommodates all the concrete syntaxes, without the need for further change. As a potential solution to this, there has been work done on deriving the abstract syntax from an existing ontology [2] or FrameNet [13], [12]. However, such resources are not always available.

2.2   Multiword Expressions

There is a significant body of research on MWEs, ranging from classification [4], linguistic analysis [24] to methods to detect MWEs (for both monolingual [15], [18] and multilingual settings [29], [5], [28]) and evaluation measures for these methods [19].

Following the MWE taxonomy from [4] into fixed, semi-fixed and syntactically flexible expressions, we note that applying the same scale to MMWEs, it is the semi-fixed and syntactically flexible constructions that are most effectively representable in GF. The reason is that GF allows for generalisations in terms of arguments (for relational MWEs, such as transitive verb phrases), declension forms and topicalisation in the sentence.

3   Methods for MMWE Extraction

3.1   General MMWE Candidate Extraction

The algorithm for general MMWE extraction parses a pair of sentences (X, Y) with a wide-coverage GF grammar, often resulting in multiple parse trees for each sentence. Then it compares all pairs of trees $\{(x, y) \mid x \in parse(X), y \in parse(Y)\}$, and if no identical trees are found, the phrases are candidates for containing BMWEs.

```
weather_adjCl : AP -> Cl ;          -- it is warm / il fait chaud (Fre)
n_units_AP    : Card -> CN -> A -> AP ;  -- x inches long
glass_of_CN   : NP -> CN ;           -- glass of water / lasillinen vettä (Fin)
where_go_QCl  : NP -> QCl ;          -- where did X go / vart gick X (Swe)
```

Fig. 1. Example of constructions

Part of the test material was not parsed by the regular GF grammar. To add robustness, we used a new chunking grammar[1] for the language pair English–Swedish. French and German didn't have the chunking grammar implemented, so for pairs including them, we used robust parsing in GF [1], [3]. With the chunking grammar, the trees kept their local structure better, whereas the robust parser resulted in flatter structure, making the distance to any well-formed tree high. Thus these sentences were always reported as BMWE candidates. For our small test set, this wasn't a problem, but for future work, a fallback for partial trees should be considered, e.g. one that translates the sentences both ways and calculates the word error rate.

We used material from two sources. First, we took 246 sentences from the Wikitravel phrase collection[2] in English, German, French and Swedish. The material consists of sentences such as asking for direction or expressing needs, in various language pairs of which other is English. For another type of text, we took the 61–sentence short story "Where is James?", from the website UniLang[3], which contains free material for language learning. In total our test set was 307 sentences, functioning mostly as a proof of concept.

After running the experiments, we found various MMWE candidates in all language pairs. We added relevant new findings to the GF multilingual dictionary, some replacing the old translations, some as new lexical items. However, the majority of the candidates were predicates that span over a larger structure, and couldn't be covered just by lexicon—instead, we added them to a new module, called *Construction* (see Figure 1).

The module is, in the spirit of construction grammar, between syntax and lexicon. Instead of applying to categories in general, most of the functions in the module are about particular predicates which are found to work differently in different languages. The purpose of the module is hence not so much to widen the scope of string recognition, but to provide trees that are abstract enough to yield correct translations. It is being developed incrementally, but we envision being able to develop the module in a more systematic manner by employing data-driven methods, such as extracting constructions from a treebank.

3.2 GF Lexicon of Compound Words

A substantial part of the work on MWEs involved the detection and representation of compound words in GF. The motivation for this lies in the need to improve GF-

---

[1] https://github.com/GrammaticalFramework/GF/blob/master/lib/
   src/experimental/Chunk.gf
[2] http://wikitravel.org/en/List_of_phrasebooks
[3] http://www.unilang.org/

```
fun ConsNomCN : N -> CN -> CN ;
fun Cons_sCN : N -> CN -> CN ;
fun Cons_enCN : N -> CN -> CN ;
```

Fig. 2. Example of compounds

driven machine translation from English into German, especially in the bio-medical domain [9].

The goal is to extract pairs consisting of German compound words and their English translations from parallel corpus, to syntactically analyse the compound and to build a GF representation of the pair, which will be added to a compound lexicon. Because the most frequent such compound words are nominals [4], we consider them as the use case of our method.

The method relies on a GF resource describing rules for nominal compounding. The following rules describe three types of compounding: first one with the modifier in nominative, second one with the morpheme 's' in the end (*Lebensmittel* 'life-s-means') and third one with the ending 'en' (*Krankenwagen* 'sick-en-vehicle').

The basic procedure is the following:

– we extract candidate pairs, which fulfil the following criteria:
  • their probability is above a confidence threshold
  • the English part parses as an NP in GF
  • the German part is composed of one word
– we employ a greedy algorithm to split the German word into a number of lexical items from the German monolingual dictionary from GF (based on Wiktionary), based on the German compound grammar described above; we select the split which employs the least number of tokens
– we add the pair of GF trees to a lexical resource for compounds

In our experiments, phrase translations extracted from a English-German parallel corpus [17] are used to detect possible nominal compounds in German. For practical reasons, we restrict the set of possible phrase translations to phrases determined to be *constituents* in the parse tree for the English sentence by a constituency parser [16]. This restricts the amount of noise in the translation memories, where *noise* is defined as a pair of random sequence of words in English and German that are seen together in the translations. Furthermore, we restrict our interest to entries that are labelled as noun phrases by the parser.

4   Evaluation

4.1   Evaluation of General MMWE Extraction

As a tentative evaluation for the general MMWE extraction method, we used the results of the language pair English–Swedish and did qualitative analysis of the findings. We chose Swedish, because it had the best grammar coverage out of the languages we

tested; the results for French and German were poorer, due to the flat structure of trees from robust parsing. The chunking grammar made it possible to compare trees even when one has a complete parse and other not, since the well-formed sentence can also be expressed as chunks.

Table 1. General MMWE extraction

Not MWE candidates	92
MWE candidates	215
False positives	44
Lexical MWEs	29
Predicates	142
All sentences	**307**

Table 1 shows the results of the analysis. Of the 307 sentences in English and Swedish, we found 215 candidates, of which 44 were considered false positives, due to parsing problems. For the algorithm to recognise two sentences as identical, it needs to have parsed them properly, so we did not get false negatives.

Out of the remaining 171 candidates, we classified 29 to be lexical MWEs, such as English *locker* vs. Swedish *låsbart skåp* 'lockable closet', or *hide from* vs. *gömma sig för* 'hide REFL for'. Not all of them were one-to-many; in 11 cases it was just a question of similar words, such as *little* and *small* used in the parallel sentences.

142 candidates were predicates that span over a larger structure. The expressions could be classified to the following subcategories: a) greetings; b) weather expressions; c) time expressions; d) money; e) units of measurement, containers; f) spatial deixis.

These expressions are non-compositional due to different factors: e.g. greetings and weather expressions are highly idiomatic, fixed phrases. Other cases, such as units, are less rigid: a certain semantic class of words appears in structures like *glass of NP*, which work differently in different languages. For example, Swedish uses no preposition, Finnish uses a special form *glassful*. Since adding a general rule for *NP of NP* would be overgenerating, we added these constructions separately for each container word (e.g. *glass, bottle, cup, bucket*).

An example of spatial deixis is the correspondence of direction adverbs between languages: e.g. the word *where* in the sentence *where did X go* should be translated in German to *wohin* 'where to' instead of *wo* 'where in'; same with *here* and *there*. We added these constructions as combinations of a motion verb and a direction adverb.

Finally, a number of the 142 phrases were correctly recognised as containing a differing subtree, but we judged the difference not to be general enough to be added as a construction. For example, sentence (1) from the short story has the auxiliary verb *can* in the English version and not in the Swedish, and the adverb *tydligt* means 'clearly, distinctly'. While not general enough for the construction module, results like this could still be useful for some kind of application grammar; the method correctly recognises them, as long as the sentences are fully parsed.

(1)   Hon hör   det tydligt nu   (Swe)
      'she hears it   clearly now'

      She can hear it well now (Eng)

4.2   Evaluation of German Nominal Compounds

We evaluated the German nominal compounds detected by our algorithm based on their utility in the task of machine translation. In this experiment, we provided the detected nominals as possible dictionary items to an SMT pipeline and extracted a translation memory from a news domain corpora augmented with the nominal compounds. We evaluated the improvements in translation quality after augmenting the translation memories with these nominal compounds. Translation quality is evaluated in terms of BLEU score, a standard metric used in evaluating performance of MT systems. Table 2 shows the BLEU scores obtained from two different SMT systems, a baseline system and the same system using the translation memory augmented with nominal compounds. The BLEU scores are reported on standard test datasets used in the evaluation of SMT systems.[4]

The improvement gained by using this simple method suggests that a proper handling of MWEs could improve the BLEU scores in an even more significant manner, by taking advantage of the full power of the GF representations, mainly by aligning all declension forms of MWEs and adding them to the translation memories.

Table 2. BLEU scores obtained from the SMT systems

SMT system	newstest2011	newstest2012
Baseline	11.71	11.64
+Compounds	11.83	11.96

5   Future Work

As GF has proven to be a reliable environment for writing multilingual CNLs and compositionality is a known problem of such applications, our method to isolate non-compositional constructions would be a great aid for the development of GF grammars, if it were applied on more domains and language sets. In this manner, one could also asses the generality of the method, both in terms of languages and types of constructions, more clearly.

For the purpose of aiding the development of GF domain grammars, we are also considering a combination between our method and the related efforts of constructing multilingual FrameNet-based grammars [13], [12].

---

[4] The datasets can be found at
`http://www.statmt.org/wmt14/translation-task.html`. We use the new-stest2011 and newstest2012 datasets in our experiments.

Regarding the use of MWE in machine translation, one can consider integrating the GF resources in a more meaningful manner, by not just aligning the basic forms, but also the declension forms. The MWE resources could also be helped to improve the existing GF-driven hybrid translation systems [9].

Last, but not least, as our initial experiments have shown a rather large number of false positives, we aim to develop specific pre-processing methods to address this issue. A boost in accuracy would lead to a decrease in the size of the initial resources that are automatically created and reduce the effort for evaluation. A possible solution would be comparing the shape of the parse trees, in order to asses differences in the constructions.

In conclusion, our work represents the first step in handling non-compositional constructions in multilingual GF applications. The methods are still under development, but they still highlight the significant advantages that the feature brings, both to general CNLs written in GF and to large translation systems.

References

1. Angelov, K.: The Mechanics of the Grammatical Framework. PhD thesis, Chalmers University of Technology (2011)
2. Angelov, K., Enache, R.: Typeful Ontologies with Direct Multilingual Verbalization. In: Rosner, M., Fuchs, N.E. (eds.) CNL 2010. LNCS, vol. 7175, pp. 1–20. Springer, Heidelberg (2012)
3. Angelov, K., Ljunglöf, P.: fast statistical parsing with parallel multiple context-free grammars. In: European Chapter of the Association for Computational Linguistics, Gothenburg (2014)
4. Baldwin, T., Kim, S.N.: Multiword expressions. In: Handbook of Natural Language Processing, 2nd edn. (2010)
5. Bouamor, D., Semmar, N., Zweigenbaum, P.: Identifying bilingual multi-word expressions for statistical machine translation. In: Calzolari, N., Choukri, K., Declerck, T., Doan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, European Language Resources Association (ELRA) (May 2012)
6. Dannélls, D., Damova, M., Enache, R., Chechev, M.: A framework for improved access to museum databases in the semantic web. In: Recent Advances in Natural Language Processing (RANLP) (2011)
7. Dannélls, D., Enache, R., Damova, M., Chechev, M.: Multilingual online generation from semantic web ontologies. In: WWW 2012, EU projects track (2012)
8. Davis, B., Enache, R., van Grondelle, J., Pretorius, L.: Multilingual Verbalisation of Modular Ontologies using GF and Lemon. In: Kuhn, T., Fuchs, N.E. (eds.) CNL 2012. LNCS, vol. 7427, pp. 167–184. Springer, Heidelberg (2012)

9. Enache, R.: Frontiers of Multilingual Grammar Development. PhD thesis, University of Gothenburg (2013)
10. Enache, R., España-Bonet, C., Ranta, A., Màrquez, L.: A hybrid system for patent translation. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012), Trento, Italy, pp. 269–276 (2012)
11. España-Bonet, C., Enache, R., Angelov, K., Virk, S., Galgóczy, E., Gonzàlez, M., Ranta, A., Màrquez, L.: WP5 final report: Statistical and robust machine translation (D 5.3) (2013)
12. Grūzītis, N., Dannélls, D.: Extracting a bilingual semantic grammar from FrameNet-annotated corpora (2014)
13. Gruzitis, N., Paikens, P., Barzdins, G.: FrameNet Resource Grammar Library for GF. In: Kuhn, T., Fuchs, N.E. (eds.) CNL 2012. LNCS, vol. 7427, pp. 121–137. Springer, Heidelberg (2012)
14. Kaljurand, K., Alumäe, T.: Controlled natural language in speech recognition based user interfaces. In: Kuhn, T., Fuchs, N.E. (eds.) CNL 2012. LNCS, vol. 7427, pp. 79–94. Springer, Heidelberg (2012)
15. Kiela, D., Clark, S.: Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In: EMNLP, pp. 1427–1432. ACL (2013)
16. Klein, D., Manning, C.D.: Accurate Unlexicalized Parsing. In: Proceedings of ACL (2003)
17. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit (2005)
18. Korkontzelos, I.: Unsupervised Learning of Multiword Expressions. PhD thesis, University of York (2010)
19. Ramisch, C., De Araujo, V., Villavicencio, A.: A broad evaluation of techniques for automatic acquisition of multiword expressions. In: Proceedings of ACL 2012 Student Research Workshop, ACL 2012, Stroudsburg, PA, USA, pp. 1–6. Association for Computational Linguistics (2012)
20. Ranta, A.: Grammatical Framework: Programming with Multilingual Grammars. CSLI Publications (2011)
21. Angelov, K., Ranta, A.: Implementing Controlled Languages in GF. In: Fuchs, N.E. (ed.) CNL 2009. LNCS, vol. 5972, pp. 82–101. Springer, Heidelberg (2010)
22. Ranta, A., Camilleri, J., Détrez, G., Enache, R., Hallgren, T.: Grammar tool manual and best practices (D 2.3) (2012)
23. Ranta, A., Enache, R., Détrez, G.: Controlled language for everyday use: The MOLTO phrasebook. In: Rosner, M., Fuchs, N.E. (eds.) CNL 2010. LNCS, vol. 7175, pp. 115–136. Springer, Heidelberg (2012)
24. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002)
25. Saludes, J., Xambó, S., The, G.F.: Mathematics Library. In: Proceedings of First Workshop on CTP Components for Educational Software, THedu 2011 (2011)
26. Saludes, J., Xambó, S.: Proceedings of EACA 2012, TODO (2012)
27. Saludes, J., Xambó, S.: Multilingual Sage. Tbilisi Mathematical Journal (2012)
28. Tsvetkov, Y., Wintner, S.: Extraction of multi-word expressions from small parallel corpora. In: Huang, C.-R., Jurafsky, D. (eds.) COLING (Posters), pp. 1256–1264. Chinese Information Processing Society of China (2010)
29. Villada Moirón, B., Tiedemann, J.: Identifying idiomatic expressions using automatic word alignment. In: Proceedings of the EACL 2006 Workshop on Multiword Expressions (2006)