Multiword Expressions: Observations from a Parallel Bulgarian-English Newsmedia Corpus

Petya Osenova, Aleksandar Popov, Kiril Simov (IICT-BAS)
{petya|alex.popov|kivs}@bultreebank.org

Introduction

This work describes an empirical study on aligned multiword expressions (MWEs) in a parallel Bulgarian-English corpus. We assume the classification of MWEs developed within PARSEME WP1 and WG4. The MWEs in each language are annotated independently from the alignments in the corpus. Then, using the alignments, we compare the ways in which MWEs are translated between the two languages. The following types of examples are considered: MWE-to-MWE; MWE-to-Word; MWE-to-phrases.

The results from the empirical study highlight at least the following issues: (1) realizations of different MWE types in two languages with different morphological complexity and word order; (2) a typology of alignment possibilities among various types of MWEs.

Data and preliminary observations

The parallel BG-EN 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.

Our aim was to extract various types of alignments, in which at least one member is a MWE. Thus, our data includes the following preliminary types: MWE-to-word; MWE-to-MWE and MWE-to-phrase in both language directions. More details will be given in the long version, if accepted.

Mapping to PARSEME classifications

The WP4 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 POS heads: *Nominal MWEs*, *Verbal MWEs*, *Prepositional MWEs*, *Adjectival MWEs*, *MWEs of other categories*, *Proverbs*. Some of these groups are further subdivided into subtypes: *Nominal MWEs* include NEs, nominal compounds as well as Other nominal MWEs; *Verbal MWEs* include: phrasal verbs, light verb constructions, VP idioms and Other verbal MWEs. WP1 classification elaborates further the typology with respect to various linguistic criteria.

When mapped to these typologies, the data showed the following properties: in the English part the most frequent MWE types are: *Verbal MWEs* (phrasal verbs, light constructions and VP idioms); *Nominal MWEs* (Nominal compounds); *Other categories of MWEs*. In the Bulgarian part the most frequent ones are: *Verbal MWEs* (se-verbs, light constructions, VP idioms); *Nominal MWEs* (other types) and *Other categories*. To present a slightly more detailed analysis of the types of correspondences, we also use the WP1 classification, which emphasizes the internal structure of the MWEs.

Within *MWE-MWE relations* the correspondences are:

- Straightforward (light constructions; VP idioms; Other categories – adverbs, prepositions, etc.).
    - Light verbs in one language often correspond to similar constructions in the other. For instance, “reach a decision: взема решение”, where V NP in English translates to V NP in Bulgarian, or “take effect: влезе в сила” and “take control: влезе във владение”, where V NP translates to V PP.

    - Complex prepositions in English tend to have structurally similar counterparts in Bulgarian. For instance, “with respect to: що се отнася до”.
    - Nominal MWEs of the kind A N seem to be often preserved across the two languages: “tough line: твърда позиция”, “free marker: свободния пазар”, “real estate: недвижимото имущество”.
    - The MWE construction V NP (not mentioned in WP1) also seems to be sometimes preserved in translation: “is drawing fire: привлича критиките”, “haven't got a clue: нямат представа”.
    - Multi-word adverbial constructions: “on the other hand: от друга страна”, “of course: разбира се”, “more and more: все повече и повече”, “in particular: в частност” (here, however, the prepositional complement varies between Adj and Noun in the two languages).
    - Conjunctions composed of multiple words: “as well as: както и”.
- Cross-language specific types (EN phrasal verbs into BG se-verbs; EN nominal compounds into BG Other NP MWEs, mainly *Adjective Noun* or *Noun preposition Noun*).
    - English phrasal verbs often correspond to Bulgarian se-verbs: “give up: се откаже”, “move back: се върнат”.
    - N N compounds in English can map to A N compounds in Bulgarian: “face amount: номинална стойност”.
    - N N in English can also be translated as N PP in Bulgarian, where the meaning carried by the first noun is expressed through a PP: “law enforcement: силите на реда”.
    - N and N constructions in English can apparently translate in similar coordinated constructions in Bulgarian, but constructed out of different POS constituents: “pros and cons: доводи за и против (N and N / N P and P).
    - An idiomatic clausal construction (V NP PP) can be translated with a light verb construction in Bulgarian: “putting pen to paper: предприел действие”.
    - V AP (not in WP1) in English can be translated with minimal changes into V AdvP in Bulgarian: “broke even: са излезли начисто” (two changes in this case: from single-word verb to se-verb and from A to Adv).
    - V PP (not in WP1) in English can be translated as V NP in Bulgarian: “will be priced of a job: ще загубят работата си”. More interesting here is the observation that a passive construction is translated in the active voice.

Conclusions

The frequency count shows the following: the *MWE-to-MWE* and *MWE-to-word* correspondences are more prevalent. In contrast, the *MWE-to-phrase* correspondence has not shown wide distribution. It would be interesting to perform a detailed analysis on more examples, so as to uncover persistent transformations between the two languages. Such knowledge can be used in designing automatic translation systems, as well as when singling out best practices in human translation. Furthermore, these transformations can possibly illuminate the ways the two languages differ in expressing meaning.