Parallel Corpora in Aspectual Studies of Non-Aspect Languages

Maria Stambolieva
Laboratory for Language Technologies, New Bulgarian University
mstambolieva@nbu.bg

Abstract

The paper presents the first results, for Bulgarian and English, of a multilingual Trans-Verba project in progress at the NBU Laboratory for Language Technologies. The project explores the possibility to use Bulgarian translation equivalents in parallel corpora and translation memories as a metalanguage in assigning aspectual values to "non-aspect" language equivalents. The resulting subcorpora of Perfective Aspect and Imperfective Aspect units are then quantitatively analysed and concordanced to obtain parameters of aspectual build-up.

1 Aims of the investigation

At the time of the appearance of the first studies on Aspect in the early 20th century, this term (a calque from the all-Slavonic "vid"), was solely used for the description of a category typologically characterising Slavonic languages, setting them apart from "non-aspect" languages. After a century of aspectual studies, the term has undergone considerable widening of meaning and forms part, in modern linguistics, of the grammatical description of languages of different groups. Thus, "aspectual classes" are set out for Romance and Germanic languages; the English opposition "non-progressive-progressive" is called "Aspect"; even the category of Correlation is often described as a "Perfect Aspect".

Far from supporting cross-language investigations, foreign language teaching and translation, the shoving of different language phenomena in the same Aspect-bag is nothing but misleading and problem-raising. Bulgarian teachers of English who have tried to draw a parallel between Bulgarian and English Aspect to their pupils are well aware of the unsatisfactory results. Translators from Bulgarian to English and back, and their editors, point to Aspect as a major pitfall. Aspect is, again, the category where systems for automatic translation seem to offer the least help – Cf. the translation equivalents provided by Google Translate for a few English sentences:

1. He sang the song. – Toy izpya pesenta.(Perfective Aspect, Aorist)
2. He sang for an hour – ?Toy peeshe za edin chas. (Imperfective Aspect, Imperfect Tense)
3. They ate the sandwich. – *Te yade sandvich. (Imperfective Aspect, Aorist/Present?)
4. Did you eat the sandwich? – *Znaete li, yade sandvich? (???)

In what follows, I will try to:

- define the essence of Slavonic aspect and, in particular, aspect as expressed in the Bulgarian language – in an attempt to demonstrate why, and with respect to the expression of what semantic oppositions, Bulgarian can be used as a metalanguage in aspectual studies;
- contrast Bulgarian aspect to the aspectual system of English;
- demonstrate the possibilities of using parallel corpora and translation memories in the cross-language study of aspect and present first quantitative results of the computational analysis of the data, with parameters of English aspect construal.

2 The Slavonic Category of Aspect

Slavonic aspect is an ***equipollent*** lexico-grammatical category covering the entire verbal system and unambiguously defined in the Lexicon. The semantic basis of the opposition is the presence or absence of a bound ([+Bound] / [-Bound]) in the topological structure of a situation or, in other words, the +Event/-Event nature of the situation. Events and non-events in Slavonic languages define a small set of Situation Types, which, after lexical filling, result in a large number of 'Action Modes'.

Depending on their situation type, eventive verbs may mark one-bound or two-bound situations: *zapeya* ('start to sing') / *izpeya* ('sing from beginning to end'). One-bound verbs mark either the beginning of a situation or its end phase - compare *zapeya* above:

..... [.....

Fig. 1. One-bound situations with initial bound

and *dopeya* ('finish singing'):

Proceedings of Recent Advances in Natural Language Processing, pages 39–42,
Hissar, Bulgaria, 12-14 September 2011.

.....].....

Fig. 2. One-bound situations with final bound

Two-bound situations can be minimal – *namigna* ('wink'), *padna* ('fall'), *otlepya* ('unglue'):

.....[X].....

Fig. 3. Two-bound minimal situations

or extended: *procheta* ('read through'), *prepluvam* (swim through), *pospya* ('sleep a while'):

.....[XXX..].....

Fig. 4. Two-bound extended situations

Non-Eventive verbs may mark s i m p l e non-bounded situations of the Action Modes Statal: *haresvam* ('like'), *imam* ('have'), *izglezhdam* ('seem', 'appear'), *cherveneya* ('be red), *mladeya* ('appear young') or Processual - *ticham* ('run'), *zreya* ('ripen'):

.....]XXX..[.....

Fig. 5. Simple noun-bounded situations

or else ***complex*** non-bounded situations: preparative situations, i.e. processes preceding an event - *zapyavam* ('be about to start singing'):

.....] [[X(XX..)].....

Fig. 6. Complex non-bounded situations: Preparatives

and iterative situations, i.e. series of similar events – *kiham* ('sneeze'), *izpyavam* ('repeatedly sing'):

.....][X] [X] ([X]...)[.....

Fig. 7. Complex non-bounded situations: Iteratives.

Preparative and iterative situations are generally expressed by verbs which are derivatively (prefixally or suffixally) formed out of perfective verbs marking momentary or extended events.

The grammaticalisation of the opposition Non-Event/Event is a typological feature of Slavonic languages which sets them apart from languages of the Germanic and Romance groups. Further, in Slavonic languages the expression of aspectual information is concentrated in the verb. Hence, the presence of a perfective or imperfective verb defines unambiguously the aspectual value of the sentence.

Bulgarian stands out among Slavonic languages in that it manifests the Perfective-Imperfective opposition to the highest degree of regularity and grammaticalisation within the language group. As Yu. Maslov points out (Maslov 1984, p.97):

'It should not be thought that the principle of the positive suffixal expression of the Imperfective Aspect and the negative, null expression of the Perfective Aspect forms an exclusive feature of the Bulgarian language area [...] However, it is precisely in the Bulgarian language area that this principle has found its fullest and most consistent development. The specifics of the Bulgarian system in this respect [..] is not in the deviation from the Slavonic language type, but in the fullest expression of the developmental tendencies built in the Slavonic grammatical system [...].'

It is the regular, systematic character of the expression of the eventive/non-eventive nature of a situation in the verb and the richess of lexical verb types that defines the possibility to use Bulgarian as a metalanguage *sui generis* in aspectual studies.

3 Aspect Studies for the English Language

Even though Aspect forms part of the verbal categories claimed by English grammar, little -- if any -- of the defining features of the Slavonic category can be said to be applicable to the English data.

In harmony with the analysis of other non-aspect languages, aspectual studies of the English verb start with Verb Classes. A proliferation of classifications of these is in circulation, ranging from Aristotle's tripartition through Vendler (1957), Kenny (1963), Mourelatos (1981), Smith (1997), to name but a notable few. Surprisingly, not one of these classifications parallels the grammaticalised Slavonic opposition in distinguishing, first and foremost, events from non-events. Quite the reverse, the first line is, as a rule, drawn between states and processes. J.-P. Descles (1990) even goes as far as to claim a *topological* distinction between states as non-bounded situations against processes and events as bounded situations. Such verb classifications are not very helpful in event construal and cannot form the basis of cross-language parallels with Aspect languages.

Unlike other non-Aspect languages, the grammatical system of English does, in fact, incorporate an opposition of an aspectual type - the so-called "Progressive Aspect'. This is a *privative* opposition between an unmarked form and a marked form expressing non-boundedness, plus a large number of other components of meaning of a non-topological nature -- such as limited duration, irritation and other emotional colouring, increasing or decreasing activity, etc. The non-progressive form in the English "aspectual" opposition is *unmarked with respect to boundedness*. In other words, the English non-progressive verb cannot unambiguously define a situation as eventive or not. Seeing that, on average, English non-progressive forms occur approximately 20 times oftener than progressive ones in an English narrative text, this means that *English verbs are, largely, unmarked for boundedness*.

In his 1972 dissertation, Henk Verkuyl tried to demonstrated that in non-Aspect languages such as English, *events are construed*, i.e. boundedness obtains at VP and Sentence level as a result of the combination of verbs belonging to particular verb classes with quantified or unquantified complement or subject NPs. About the same time and independently of Verkuyl, M. Ridjanovic (1969) and A.Danchev, B. Alexieva (1974) in their English-Serbo-Croatian and English-Bulgarian contrastive studies, respectively, arrived at similar results, namely: aspect markers in English occupy a large stretch of the discourse. While Ridjanovic concentrated on the articled/non-articled noun phrases as major markers of Aspect, Danchev/Alexieva, processing a large parallel corpus (20 000 file-cards of English Simple Past Tense sentences and their Bulgarian equivalents!) arrived at a much greater variety of contextual markers. The authors ranked these as follows: adverbial phrases, verb semantics, subject phrase semantics, object quantification.

4 Parallel Corpora in the Aspectual Study of English

In view of the abundance of English-Bulgarian or Bulgarian-English parallel texts, (mainly in the form of TRADOS or Wordfast translation memories, but also simply aligned -- whether with tools for automatic alignment such as WinAlign or computer-assisted aligners such as MIX), the idea of using translation units and the aspectual values of the Bulgarian verbs to assign aspectual values to English sentences seems to make sense. While a wider-scope study based on a set of registers from a balanced corpus is the ultimate task of this project, the data presented below are drawn from a smaller parallel corpus of fiction texts. Even this corpus, however, clearly pinpoints lines of investigation and possibilities for applications of the approach.

The Bulgarian verbs in the parallel corpus were aspect-tagged with a choice of PA (perfective aspect) or IA (imperfective aspect) values. Translation units containing one or the other tag were asigned to one of three sub-corpora: an IAcorpus, a PA corpus and a "Mixed" corpus, with sentences containing both perfective and imperfective verbal forms. Each of the subcorpora was processed with the NBU BUILD segmentation programme, yielding quantitative information. At a next stage, concordancing was performed for larger segment identification.

Setting aside some 7% verbless setences, our corpus yielded the following quantitative information: appr. 31 % of the Bulgarian sentences contained Imperfective verbs only; appr. 23% of the sentences contained perfective verbs only; appr, 29% of the sentences contained both perfective and imperfective verbs, in different patterns.

4.1 Analysing the PA subcorpus

The analysis of the PA corpus quantitative data points to the following major PA markers in the English sentences:

Adverbial modifiers of time:

- *when* - upon concordancing, found to present, in about all cases, an instance of the relative adverbial, introducing a time clause;
- *then*, *now*, *now that*, *before*, *as* (=when), *eventually*, *finally*, *in+year* (e.g. *in 1984*), *at lunch*, *to begin with*, *the moment +subject+V*.

Coordination:

- *and* - as a coordinative link between event clauses;
- commas - Cf. above.

Lexical meaning of the verbs:

- communication verbs in the simple past tense, esp. *admitted, announced, insisted, lied, mumbled, prompted, said, thought (to myself), urged*;
- phrasal verbs: *drove away*, *went away*, *sat down*, etc.
- process verbs in the simple past tense.

4.2 Analysing the IA subcorpus

The following were found to be the major IA markers in the corpus:

Adverbial modifiers:

- temporal adverbials, e.g. *still*, *sometimes*, *repeatedly*, *when* (= *whenever*, closely followed by *would*), *as* (= *while*)
- *for*-phrases: e.g. *for a few minutes*;
- *do nothing but*, e.g. *We did nothing but quarrel*.
- adverbial modifiers of time containing NPs with attributes pointing to iterative situations, e.g. *every summer*.

Lexical meaning of the verb:

- link verbs, e.g. *was*, *seemed*, *grew*;
- extended state verbs, e.g. *know*, *hope*, *love*, *remember*.

Subject phrase semantics:

- Subjects semantically characterised as [-Animate], and esp. 'Inalienable property' subjects, e.g. *the symmetrical limbs*, *her expression*, etc. are systematically present in IA clauses.

4.3 Analysing the Mixed subcorpus

The most frequent patterns were found to be: IP (appr.9%), PI (appr. 4,5%), IPP and PPI (appr. 2.5% for each subtype). Typical factors defining the "mixed" status of the sentences are: complex verbal predicates, V + complement clause groups, presence of verbs of communication (typically Perfective), presence of verbs of thinking (typically Imperfective), Frame and Event situations. Conjunctions and complementizers, as markers of coordination and subordination, appear high in the rank list of most "mixed" subgroups.

5 Conclusions

The approach not only yielded results paralleling closely those of Danchev and Alexieva's corpus-based study (op. cit.) and the Stambolieva 2008 system-based one, but also contributed interesting additional information. Thus, coordination/compounding, of which no mention has ever been made in previous work, was found in the present study to occupy an important position in the hierarchy of English contextual PA markers. On the other hand, argument NP quantification was not found to hold the high-rank position predicted by Verkuyl (op. cit. and 1993). Concordancing elements of context occurring in both corpora - such as *when* - allows to arrive at structures which disambiguate them as PA or IA markers. Another important advantage of the approach is the possibility to obtain reliable quantitative information defining the hierarchy of units participating in IA or PA-marked predications. Above all, the specialised corpus thus obtained can be used as valuable translation memory or teaching aid.

References

J.-P. Descles. 1990. '*State, Event, Process and Topology*'. In: General Linguistics, vol. 29. No.3. Pennsylvania, pp. 159-200.

A, Kenny. 1963. *Action, Emotion and Will*. Routledge and Kegan Paul, London and New York.

Maslov 1984. Ю. С. Маслов. *Очерки по аспектологии*. Издательство Ленинградского университета, Ленинград.

Mourelatos A. Mourelatos. 1981. '*Events, Processes and States*'. In: Syntax and Semantics. Tense and Aspect, No.14. Philip Tedeschi, Annie Zaenen (eds.). Walter de Gruyter, Berlin and New York, pp. 191-212.

M. Stambolieva. 2008. *Building Up Aspect*. Peter Lang, Oxford, Bern, New York.

Z. Vendler. 1957. '*Verbs and Times*'. In: The Philosophical Review 66, 143-160.

H. Verkuyl. 1972. *On the Compositional Nature of the Aspects*. Reidel, Dordrecht.

H. Verkuyl. 1993. A Theory of Aspectuality. Cambridge Studies in Linguistics 1964. Cambridge University Press.