Year: 2014

Towards machine translation of Russian aspect

Sonnenhauser, Barbara ; Zangenfeind, Robert

Towards machine translation of Russian aspect

Barbara Sonnenhauser (1) and Robert Zangenfeind (2)

(1) Institute for Slavic Philology – Ludwig-Maximilians-Universität München
Geschwister-Scholl-Platz 1, 80539 München
basonne@lmu.de
(2) Center for Information and Language Processing – Ludwig-Maximilians-
Universität München
Oettingenstr. 67, 80538 München
r.zangenfeind@lmu.de

Abstract

One of the most intricate problems for machine translation are grammatical categories which are present in the source language but not in the target language. This problem is further complicated if the category in question is highly polysemous. Grammatical aspect in Russian is one example for such a category. On the categorical level, it has only approximate equivalents in non-Slavic languages (such as the progressive form in English). In addition, language-internally, its semantics and interpretation cannot be sufficiently captured with one specific characteristic feature. This paper aims at establishing a basis for the machine translation of the Russian aspect. To do so, the interaction of verb and aspect semantics has to be described in a systematic way. Moreover, the further lexical components contributing to the meaning computation have to be annotated for the aspectual information they provide. This allows for the formulation of rules for machine translation into target languages either lacking grammatical aspect or having a different aspect system.

Keywords

Grammatical aspect, machine translation, ambiguity, semantic features, syntactic features, Russian, German, English, Turkish

1   The problem of aspect

Grammatical aspect in Russian is problematic for machine translation in at least two respects: 1) it is a highly polysemous category, 2) it does not have categorical equivalents in some languages, and if it does, the equivalence is hardly one to one.

1.1 Polysemy and ambiguity

Traditionally, the interpretations that are possible for both aspects are given in the form of lists (e.g. Zaliznjak & Šmelev, 1997: 15-36). This is definitely useful for didactic intentions; other purposes, however, require a different systematization. One case in point is translation, especially machine translation from Russian into other languages.[1]

One possible way of systematizing aspect interpretations is provided by the analysis developed in Sonnenhauser (2004, 2006), based on the combination of a selection-theoretic (Bickel, 1996) and time-relational (Klein, 1995) account. According to this analysis, aspect operators select, and thereby assert, specific part(s) of the event structure encoded by the verb. Assuming a tripartite event-structure (Moens & Steedman, 1988), verbs may encode dynamic phases 'φ_{dyn}' (preparatory processes), boundaries 'τ' (culmination points) and static phases 'φ_{stat}' (consequent states), depending on the eventuality they refer to. By selecting and asserting some part of the coded event structure, aspect establishes a relation between the topic time interval I(TT) as the time the assertion is about and the event time interval I(e) comprising that part of the run time of the denoted event that is selected by the aspect operator. This yields the relations given in (1) and (2), i.e. the inclusion of the boundary in the topic time for the perfective (= pf) aspect and a general overlap relation for the imperfective (= ipf) aspect (a more detailed account is provided in Sonnenhauser, 2006, 2009):

(1)   pf $I(TT) \supset I(\tau)$

(2)   ipf $I(TT) \bigcirc I(e)$

Both relations are specified in the course of interpretation. For the pf aspect, this mainly concerns the specification of the boundaries of I(TT): the interval may be closed to both sides, i.e. the initial and final points are part of the interval, it may be open to the right or open to the left, i.e. the initial point is part of the interval whereas the final point is excluded and vice versa.[2] This is illustrated with the example in (3a), which can be interpreted in three ways and thus be translated into English as in (3b–d):

(3)   a.   Ja emu *dala* knigu.

        b.   I *gave* him the book [and then …]   I(TT) closed

        c.   I *have given* him the book [and now …]   I(TT) open to the right

        d.   [After] I *had given* him the book   I(TT) open to the left

---

[1]   The other direction of machine translation again has its own specific requirements, cf. Mel'čuk & Wanner (2008).

[2]   This is based on a set theoretic definition of intervals as proposed, e.g., by Guentchéva & Desclés (1982), cf. also Sonnenhauser (2006, 115–118). According to this conception intervals consist of sets of points. Unbounded intervals are not segmented, bounded intervals are segmented. With open-bounded intervals I_o the initial (a) and the final point (b) out of the set of points (x) are not part of the interval, i.e. $I_o = \{x, a < x < b\}$, with closed-bounded intervals I_c both are part of the interval, i.e. $I_c = \{x, a \leq x \leq b\}$.

The basic relation for the ipf aspect is much more general, since contrary to the pf aspect it is not restricted as regards the selected part of the event – this may be a phase or the event as a whole. Closer inspection reveals three specific relations, which are given in (4), with one example each illustrating the relations:

(4)   a. $I(TT) \subset I(\varphi_{dyn})$

Kogda on vošel, ona *čitala* knigu. 'When he came in, she *was reading* a book.'

$(I(\varphi_{dyn}))$: the time interval of her reading the book, covering only this process excluding beginning or end; $I(TT)$ is included in the reading-process and specified by the moment when he came in)

        b. $I(TT) = I(e)$

Ona *rabotala* v universitete. 'She *worked* at the university.' [= She was employed there.]

$(I(e))$: the time interval when she was employed at the university; $I(TT)$ runs exactly parallel to the time interval of her working at the university)

        c. $I(TT) \supset I(e)$

Ona uže *rasskazyvala* emu ėtu istoriju. 'She *has* already *told* him this story.'

$(I(e))$: the time interval of her telling the story; $I(TT)$ includes the complete story-telling event)

The outlined analysis in terms of specific relations assumes them to be clearly distinct, which in turn suggests ambiguity. Each of these distinct relations may give rise to a specific range of interpretations. For the purposes of machine translation, only the factor of ambiguity is decisive; both the structures underlying the representations and the specific interpretations can be neglected.

Having proposed a solution for the systematization of the manifold interpretations possible for the Russian pf and ipf aspect, the second problem can be addressed: the cross-linguistic similarities and divergences.

1.2   Language comparison

The justification for postulating the three specifications for the pf aspect is provided not only on language internal grounds, but also by the fact that these relations can be morphologically coded in other languages, which render it mainly in terms of temporal distinctions. Table 1 illustrates this for Russian, English and German, with the brackets indicating the boundedness-characteristics of the intervals. Note that these correlations hold for the past tense; with morphological present in Russian, the relation 'I(TT) closed' yields a future interpretation.

semantics	interpretation	Russian	English	German
group I_{pf} TT closed: [---τ---]	eventive	pf	simple past	imperfect / perfect[3]
group II_{pf} TT right open: [---τ---[	perfect (existential, current relevance, extended now, etc.)	pf	perfect	perfect
group III_{pf} TT left open:]--- τ---]	pluperfect	pf	pluperfect	pluperfect

Table 1: Ambiguity of pf aspect

Likewise, the cross-linguistic validity of assuming three basic ipf configurations is suggested by two facts: the three configurations may be coded morphologically in other languages in terms of aspect distinctions, and if coded, they give rise to a similar range of interpretations. This is illustrated in table 2, comparing 'imperfective' grammemes in Russian, English and Turkish (for more details cf. Sonnenhauser 2006).[4] This indicates that even though aspect is grammaticalized in all three languages, they are by no means equivalent as regards the semantic range of the respective grammemes.

semantics	interpretation	Russian	English	Turkish
group I_{ipf} $TT \subset \varphi_{dyn}$	processual, conative	ipf	progressive	-iyordu -mekteydi
group II_{ipf} $TT = e$	habitual, non-actual, potential, permanent, atemporal	ipf	simple form	-irdi
group III_{ipf} $TT \supset e$	general-factive, durative	ipf	simple form	-di

Table 2: Ambiguity of ipf aspect

The ambiguity of the Russian aspects and the cross-linguistic validity of the possible disambiguated configurations are crucial for the question of machine translation in that this provides the basis for stating clearly formulated rules.

1.3 Disambiguation

Having pointed out the advantages of assuming a basic ambiguity of the semantics of aspect as regards Russian and the cross-linguistic perspective, the next question we address is that of disambiguation. In natural language communication, interpreting an utterance requires the resolution of the aspect-ambiguity; disambiguation is also the first step towards machine translation.

---

[3] The distinction between imperfect and perfect is getting blurred in German. It is intact in Northern varieties but has been completely lost in Southern varieties, where the perfect has taken over narrative functions.

[4] The comparison in table 2 is confined to the past, since group III_{ipf} is not possible for the other tenses. Accordingly, the Turkish forms are specified with the past tense morpheme *-di*. Note that being opposed to *-iordu* and *-irdi*, simple *-di* can be interpreted as the morphologically and semantically unmarked element in the past tense aspect system of Turkish (cf. Sonnenhauser 2006).

Disambiguation is achieved by specifying I(TT) in terms of its boundedness-features and – for the ipf aspect – by specifying the relevant part of the Aktionsart that is selected and related to this interval. In Russian, this specification is possible mainly by lexical and syntactic means: as regards the ipf aspect, adverbs like *medlenno* 'slowly' or *postepenno* 'gradually' specify I(TT) as open-bounded, adverbs like *ran'še* 'formerly' as unbounded, particles like *uže* 'already' as closed-bounded, and hence the interpretation as belonging to group I_{ipf}, II_{ipf}, or III_{ipf} respectively. Concerning the pf aspect, conjunctions like *i* 'and [then]' disambiguate eventive (group I_{pf}) from perfect (group II_{pf}) interpretations, adverbials specifying a point in time suggest the pluperfect interpretation (group III_{pf}), etc. This is due to the fact that here the consequent state following the selected boundary does not hold at the time of utterance (which would yield the perfect interpretation) but at the time specified by the temporal adverbial, which is prior to the time of utterance.

As can be seen from tables 1 and 2, for machine translation from Russian to English, German or Turkish it is enough to solve these basic ambiguities. What is rendered by means of the perfect in English or German has the same interpretational range as the 'perfect' / group II_{pf} specification of the Russian pf aspect, what is rendered by means of the *-irdi* suffix in Turkish may give rise to the same variety of interpretations as group II_{ipf} of the Russian ipf aspect. The same reasoning applies to the other ambiguities.

For an automatic disambiguation, the relevant lexical and syntactic means have to be annotated in the lexical entries of lexemes as regards the aspectual information they contribute to the meaning computation. The computation may then proceed in the form of 'if-then' statements along the lines proposed by Vazov (1999), which is also used by Mel'čuk & Wanner (2008) for aspect-establishing rules in the process of German-Russian translation.

One way to provide the necessary kind of information for the relevant lexemes can be to annotate these lexemes with appropriate semantic features. ETAP-3 (cf. section 2) is a machine translation system that uses dictionaries with semantic and syntactic features.

2 Aspectually relevant semantic and syntactic features in the dictionaries of ETAP

The machine translation system ETAP-3 (for an earlier version cf. Apresjan et al. 1992) provides a lot of information for lexemes that can be useful for the interpretation of aspect. Until now, this information is mainly used for syntactic analysis and is given in the dictionaries of ETAP in terms of semantic and syntactic features. Some of the semantic features that are potentially relevant for the interpretation of aspect are 'VREMJA' (to characterize temporal lexemes[5]), 'DEJSTVIE' (for nouns and verbs that denote an action which develops in time and which is initiated by an active subject), 'PROCESS' (for nouns and verbs that denote a process which develops in time and which is initiated by a passive subject) etc. Some important syntactic features for the interpretation of aspect are 'DLIT'

---

[5] The descriptions of this and the following features are corresponding to the Russian help manual for ETAP-3 in version 3.1.91 from the year 2008. This is part of ETAP-3 system which we are very grateful that Leonid L. Iomdin back then placed at our disposal at the Center for Information and Language Processing in Munich.

which characterizes a period of time or 'NEOPR' for indefinite pronouns, to name just two of them.

Another valuable instrument for our purpose is the classification of predicates by Apresjan (2006). This classification includes 17 classes. Some of them exclude certain disambiguation possibilities and/or make others highly probable. For 'dejatel'nosti' ('activities')[6], such as *torgovat'* 'to trade', *upravljat'* 'to rule', for instance, the actual-processual and the general-factual readings are ruled out, whereas a durative interpretation is most likely. For these verbs, therefore, the information 'group II_{ipf}' can already be assigned in its lexical (semantic) information. Other classes, such as 'dejstvija', i.e. 'actions', are less explicit and allow for all possible interpretations. For their disambiguation, further information provided by aspectually relevant components of the sentence have to be taken into account.

These relevant components are realized by adverbials, particles and conjunctions.[7] So, besides information about aspect and class of the predicate in a sentence, aspectually relevant information must be provided in form of semantic (and syntactic) features in the lexical entries of these parts of speech. Another crucial bit of information is provided by tense. Present tense, for instance, excludes ipf interpretations out of group III_{ipf} and all pf interpretations except for the future interpretation (cf. section 1.2). The combination of all this kind of information can be the basis for the "calculation" of a temporal and aspectual interpretation of the whole sentence.[8] The next section will show the problems of such a calculation and steps towards a possible solution.

3   Towards a solution

An example to illustrate which information in a sentence is relevant is given in (5):

(5)   Ran'še ja po večeram prodelyval èti gimnastičeskie upražnenja po pjat' raz.[9]

         lit. 'formerly I in evenings do.PAST.ipf these gymnastic exercises each five times'

Lexemes and phrases that are important for our interpretation are the following: *ran'še* 'formerly', *po večeram* 'in the evenings', *prodelyvat'* '[to] do', *upražnenie* 'exercise' and *po pjat' raz* 'five times each'.

---

[6]   The English terms for classes of predicates are taken from Apresjan (2005).

[7]   These components correspond to the contextual clues (imperfective and perfective triggers) of Mel'čuk & Wanner (2008).

[8]   This is by no means to say that there have not been any compositional approaches to aspect before. These approaches (for a basic overview cf. Verkyul 2012) are concerned with the modeling of aspectual composition in order to arrive at a principled syntactic description and the mapping of composition onto syntactic structure, not with possible implementations into NLP. Moreover, they mostly lack a distinction between lexical information and aspect semantics.

[9]   Example from Bendixen et al. (2005–2012).

The dictionary entries of ETAP provide the following information about these lexemes amongst other characteristics[10]:

- *ran'še* 'formerly' has the syntactic feature 'VREM' which characterizes temporal adverbs

- *večer* 'evening' has the syntactic features 'DLIT' (to characterize a period of time) and 'VREM' (here to characterize a point of time or a period of time) as well as the semantic feature 'VREMJA' (temporal lexeme)

- *prodelyvat'* '[to] do' has the semantic features 'FAKT' (event) and 'DEJSTVIE' (action, i.e. a situation which develops in time and which is initiated by an active subject)

- *upražnenie* 'exercise' has the semantic features 'FAKT' and 'DEJSTVIE', cf. above.

For our purposes, this information should be enriched by the following:

- *ran'še* 'formerly' is temporally and referentially (as concerns reference to event) indefinite and thus excludes group I_{ipf} interpretations; appropriate semantic features in addition to ETAP's features could be 'temporally indefinite' and 'referentially indefinite'[11]

- *po [večeram]* 'in [the evenings]' is a preposition that – when governing a temporal lexeme, i.e *po16*[12] – expresses regularity. An adverbial phrase like *po večeram* 'in the evenings' can be annotated by labeling the preposition *po16* with the feature 'regularity', and thus excludes group I_{ipf} and group III_{ipf} interpretations

- *prodelyvat'* '[to] do' is used as a support verb; i.e. it has no semantics, only aspectual information is relevant (here: ipf); semantic information must be provided by the predicative noun in the sentence (here: *upražnenie* 'exercise')

- *upražnenie* 'exercise' is the semantic predicate in the sentence and can be labeled as 'zanjatie' ('occupation') according to Apresjan (2006: 83, 86f.), i.e. an action whose

---

[10] We only cite here semantic and syntactic features that seem to be relevant for aspect interpretation.

[11] The semantic feature 'temporally indefinite' should indicate that there is just a vague temporal specification in terms of localization on the time axis; this feature should also characterize adverbs like *skoro* 'soon', *togda* 'then', *vsegda* 'always' etc. (the semantic feature 'temporally definite', on the other hand, should indicate a more precise temporal specification; e.g. for adverbs like *teper'* 'now', *segodnja* 'today' etc.). The lists of adverbs with these and other semantic features, of course, still must be thoroughly examined (the need for a list of such triggers is pointed out also by Melčuk & Wanner 2008: 141). 'Referentially indefinite' concerns the selection and assertion of a specific part of the event structure carried out by aspect (cf. section 1.1): adverbs like *ran'še* indicate that there is no specific part of the event structure selected by aspect (contrary to 'group I_{ipf}' interpretations, where the dynamic phase is selected by the ipf aspect and asserted to hold within I(TT)). Particles like e.g. *uže* 'already', on the other hand, should be annotated with the features 'referentially definite' and at the same time 'temporally indefinite'.

[12] cf. Slovar' russkogo jazyka 1983.

immediate object is just to accomplish this action[13]; in combination with an ipf support verb such as *prodelyvat'*, it allows for group I_{ipf}, II_{ipf} and III_{ipf} interpretations

- *po [пять раз]* '[five times] each' is a preposition that – when governing a noun that can have a numeral as syntactic dependent, i.e. *po20*[14] – expresses distributivity of the verbal complement and allows for group I_{ipf}, II_{ipf}, III_{ipf} interpretations[15]; the appropriate feature for the preposition *po20* could be 'distributive'.

Based on the newly added information, the aspectual information given in (5) can be disambiguated as belonging to group II_{ipf}. This is calculated as follows: The predication *prodelyvat' èti upražnenija po pjat' raz* 'do.ipf these exercises five times each' is ambiguous between all three groups. This range of possibilities becomes restricted by the contribution of past tense and the adverbials *po večeram* 'in the evenings' and *ran'še* 'formerly'. The decisive information is provided by *po večeram* which excludes two of the three possible specifications and thus overrides the less specific information given by *ran'še*, which excludes only one specification. Based on this specification, the verbal lexeme should be rendered by the simple form in the English translation.[16]

To sum up, this interpretation could be formalized as conditional instructions ('if-then') in the following way:

(6)   for language-internal disambiguation:

IF predicate has feature 'occupation'

AND IF aspect = ipf

AND IF tense = past

AND IF adverb 'group II_{ipf}'

THEN 'group II_{ipf}' interpretation

---

[13] This means that the information from ETAP ('action') for *upražnenie* 'exercise' can be further specified by the feature 'occupation'.

[14] cf. Slovar' russkogo jazyka 1983.

[15] Cf. Mehlig (2008) on the hybrid nature of distributive predications. These predications may be ipf and receive an actual-processual interpretation even though their complement is bounded, as is *upražnenija* by *po pjat' raz* in (5). Note the crucial role of the secondary imperfective *prodelyvat'* here. As Filip (2008 : 247) points out, "[p]redications with secondary imperfectives [...] have sets of partially ordered events in their denotation, due to the contribution of the prefix, but the imperfective suffix on the verb explicitly suspends the requirement that the verb only has maximal events in its denotation [...]".

[16] The most adequate translation would be with the habitual construction 'used to'. This specification can be solved by means of language-internal paraphrasing rules and is not necessarily an immediate concern of translation.

(7)   for translation:

        IF 'group II$_{ipf}$' interpretation

        THEN 'simple form' in English

Formal descriptions like these can be the basis for an implementation in a machine translation system like e.g. ETAP.[17]

4   Conclusion

We have argued that, based on the combination of a selection-theoretic and time-relational account, it is possible to systematize the semantics of the Russian verbal aspect and its interpretations. This systematization comprises several groups specifying the relation between topic time interval and event time interval possible for the pf and ipf aspect. These groups may have morphological counterparts in the tense-aspect systems of other languages. In order to choose the right morphological means when translating the Russian verbal aspect it is necessary to disambiguate its semantics. Disambiguation is made possible by annotating all relevant lexemes with specific, aspectually relevant information. This is the starting point for a possible computational implementation of aspect interpretation. The system of semantic and syntactic features as used by the machine translation system ETAP is a workable basis for this implementation. Enriching this system with information taken from Apresjan's classification of predicates and with additional, more detailed semantic features, we illustrated the problems of a "calculation" of aspect interpretation and presented steps towards a possible solution.

Our future work will be to develop a refined system of semantic features for verbs (and predicative nouns), adverbials, particles and conjunctions, based on ETAP's features and Apresjan's classification of predicates. With these tools at hand, it is our aim to implement rules for aspect translation in a machine translation system like ETAP. Besides the practical utility, an implementation in a rule based system has the great virtue to verify the linguistic theory in practice and, with that, to enable to improve the theory.

Bibliography

Apresjan, Ju.D. 2005. Prolegomena to systemic lexicography. In Apresjan, Ju.D. & L.L. Iomdin (eds.). *East West encounter: second international conference on Meaning ↔ Text Theory*. Moscow, 20–29.

Apresjan, Ju.D. 2006. Fundamental'naja klassifikacija predikatov. In Apresjan, Ju.D. (ed.). *Jazykovaja kartina mira i sistemnaja leksikografija*. Moskva, 75–110.

---

[17]   An implementation of rules like these in ETAP's already much elaborated translation system from Russian to English could be a first step. But, of course, similar rules can be established for the translation into other languages. For a fragment of the translation from Russian to German as an extension of ETAP cf. Zangenfeind (2011).

Apresjan, Ju.D. et al. 1992. *Lingvističeskij processor dlja složnych informacionnych sistem.* Moskva.

Bendixen, B. et al. 2005–2012. *Russisch aktuell.* Wiesbaden.

Bickel, B. 1996. *Aspect, mood and time in Belhare.* Zürich.

Filip, H. 2008. Events and maximalization: the case of telicity and perfectivity. In Rothstein, S. (ed.). *Theoretical and crosslinguistic approaches to the semantics of aspect.* Amsterdam, Philadelphia, 217–256.

Guentchéva, Z. & J.-P. Desclés. 1982. A la recherché d'une valeur fondamentale du parfait bulgare, *Săpostavitelno ezikoznanie* 7/1–2, 44–56.

Klein, W. 1995. A time-relational analysis of Russian aspect, *Language* 71(4): 669–695.

Mehlig, H.-R. 2008. Aspect and bounded quantity complements in Russian. In Rothstein, S. (ed.). *Theoretical and crosslinguistic approaches to the semantics of aspect.* Amsterdam, Philadelphia, 257–290.

Mel'čuk, I.A. & L. Wanner. 2008: Morphological Mismatches in Machine Translation. In *Machine Translation*, 22: 101–152.

Moens, M. & M. Steedman. 1988. Temporal ontology and temporal reference, *Computational Linguistics* 14(2): 15–28.

Slovar' russkogo jazyka 1983. Tom III, Moskva.

Sonnenhauser, B. 2004. Underspecification of 'meaning': the case of Russian imperfective aspect. In *Proceedings of the ACL-04 workshop on 'Text Meaning and Interpretation'.* Barcelona, 89–96.

Sonnenhauser, B. 2006. *Yet there's method in it. Semantics, pragmatics, and the interpretation of the Russian imperfective aspect.* München.

Sonnenhauser, B. 2009. Definiteness and specificity of verbal referents. In Birzer, S., Finkelstein, M. & Mendoza, I. (eds.). *Proceedings of the second international Perspectives on Slavistics conference (Regensburg 2006).* München, 115–126.

Vazov, N. 1999. Context-scanning strategy in temporal reasoning, *Modeling and using context. Lecture notes in computer science* 1688: 389–402.

Verkuyl, H. 2012. Compositionality. In Binnick, R. (ed.). 2012. *The Oxford handbook of tense and aspect.* Oxford, 563–585.

Zaliznjak, A.A. & A.D. Šmelev. 1997. *Lekcii po russkoj aspektologii.* München.

Zangenfeind, R. 2011. Transfer of Russian actantial syntactic relations into German, In Boguslavsky, I. & L. Wanner (eds.). *Meaning – Text Theory 2011.* Barcelona, 306–315.