THE CASE OF RUSSIAN SUBJECT PRO IN MACHINE TRANSLATION SYSTEM

Bogdanov A. V. (abogdanov@abbyy.com)[1],
Gorbunova I. M. (igorbunova@abbyy.com)[1, 2]

[1]ABBYY, Moscow, Russia
[2]Russian State University for the Humanities, Moscow, Russia

This paper concerns a problem of Russian floating quantifiers (also known as semipredicatives) in machine translation. Floating quantifiers in Russian (such as *оба* 'both', *один* 'alone', *сам* 'on one's own' etc) are inclined for case, number and gender and agree in those categories with the subject of the minimal (finite) clause containing them. However, the case of a floating quantifier in an infinitive clause varies according to the type of PRO control applied and some other structural characteristics of the infinitive clause. This poses a problem for rule-based machine translation, to choose the correct case for the quantifier at synthesis, or to link it correctly to its antecedent at analysis. A model-based machine translation system, such as ABBYY Compreno, can handle the case choice problem, as this paper is to show.

Key words: subject, PRO, case marking, floating quantifiers, semipradicatives, machine translation, Russian

Introduction

This article deals with the problem of floating quantifiers in Russian from the perspective of ABBYY Compreno, a universal text analysis technology. Recently there has been a number of presentations concerning the information extraction features of the technology ([Anisimovich et al. 2012; Starostin et al. 2014; Bogdanov et al. 2014]). This article, however, deals more with the machine translation benefits that arise from the complete semantic-syntactic analysis of an input text, a task solved by Compreno.

1. Floating quantifiers and the case of Russian subject PRO

According to Babby, floating quantifiers are "adjectives that adjoin to VP and agree in case, gender and number with the subject of the minimal clause containing them". This can be illustrated by examples (1), (2), (3) and (4) (floating quantifier is marked nominative in (1), (3) and (4), but dative in (2); singular in (1–3), but plural in (4); masculine in (1–2), but feminine in (3), all in accordance to the features of the subject)

(1) *Я пришел сам*

(2) *Мне прийти самому?*

(3) *Она пришла сама*

(4) *Они пришли сами*

In Compreno syntactic parcer the notion of agreement is narrowed to a relation between two directly bound nodes in a tree and therefore in order to support the agreement floating quantifier is considered as moved from within subject NP, as in figure 1 (a syntactic tree in Compreno parcer for ex. (1))

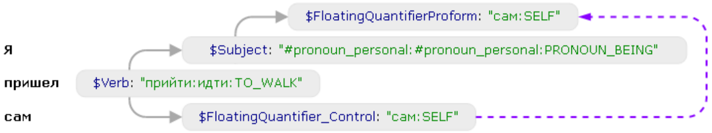
$FloatingQuantifierProform: "сам:SELF"
я
$Subject: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
пришел
$Verb: "прийти:идти:TO_WALK"
сам
$FloatingQuantifier_Control: "сам:SELF"

Fig. 1. Я пришел сам

The subject of infinitive clause in Russian is normally dative, as in (2). Indeed, in dependent infinitive clause with a PRO subject a dative floating quantifier can also be found, as in (5). However, the nominative case is sometimes the only option for a floating quantifier in infinitive clause, as in (6).

(5) *Он приказал нам прийти самим*

(6) *Он хочет прийти сам*

Amongst the numerous works on the subject of so-called second dative one can point out three main hypothesis about the nature of nominative and dative of floating quantifier in infinitive clause. Before turning to the description of the mechanism applied for the case choice in Compreno, we will give a brief account for those three hypothesis: universally local agreement ([Comrie 1974]), long-distance agreement for nominative and default dative assignment ([Franks 1990, 1995; Greenberg and Franks 1991]), as well as direct predication in subject control PRO constructions ([Babby 1998]).

Local agreement hypothesis appears in [Comrie 1974], the first paper to consider the problem of second dative. The idea is that the PRO of infinitive clause, whether lexically controlled or not, is assigned one of the cases—nominative or dative. Nominative is restricted to subject control PRO constructions and dative is a default case for the subject of infinitive clause. The syntactic structures for (5–6) are proposed as in (7–8):

(7) [*Он*] [*приказал нам*$_{i}$ [*PRO*$_{i,\ DAT}$ *прийти самим*$_{i,DAT}$]]

(8) [*Он*$_{i}$] [*хочет* [*PRO*$_{i,NOM}$ *прийти сам*$_{i,NOM}$]]

Long-distance agreement hypothesis was proposed by S. Franks. He claims that only a subject of a tensed CP infinitive clause can be assigned dative case, and PRO is essentially caseless. This claim is supported by the fact that most of the infinitives with an overt dative subject can take a tense auxiliary for future and past, as in (9) and its counterpart (10).

(9) *Куда нам поставить этот ящик?*

(10) *Куда нам было поставить этот ящик?*

Therefore, according to Franks, all the constructions where a subject of infinitive clause is overt and undoubtedly dative, are CPs, whereas dependent infinitives are IPs (as they are tenseless) and cannot assign dative to its subject. The nature of nominative and dative cases of floating quantifiers then are essentially different. Nominative case restricted to subject control PRO infinitive constructions is "transmitted" to the floating quantifier from the understood antecedent, with a long-distant agreement arising. Dative case on the other hand is a default case assigned to the sister to I' (the same rule as for the dative subject, but unlike the subject, a quantifier can be assigned the case directly without restriction to CPs).

Direct predication hypothesis was introduced in [Babby 1998]. Subject control PRO infinitive complement is viewed as a bare VP without a PRO, whereas other infinitives have a PRO which is assigned dative as a default case for the subject of infinitive. Floating quantifier thus receives the case form the nearest subject by agreement. The corresponding structures for (5–6) in this theory are (11–12)

(11) [*Он*] [*приказал нам*$_{i}$ [*PRO*$_{i,DAT}$ *прийти самим*$_{i, DAT}$]]

(12) [*Он*$_{i}$] [*хочет* [*прийти сам*$_{i.NOM}$]]

However, there is data that comes to conflict with each of the theories. First, it can be easily demonstrated that overt dative subject of infinitive clause is not bound to the tensed constructions. In (2) above one cannot add a tense auxiliary for future or past (unlike examples (9–10) borrowed from [Franks 1990]), and therefore one cannot argue that the construction in (2) is tensed. It contradicts the main argument of [Franks] that only a tensed CP can assign dative to the subject, and thus there is no consistent argument against the local agreement hypothesis[1].

Second, direct predication of [Babby 1998] can only account for infinitival sentential actants, but not for adjunct CPs such as that of (13), yet in (14) a nominative

[1] The same refers to the claim of [Fleisher 2006] that the overt dative NP in an infinitive clause is not subject to the infinitive clause but rather a subject to copula construction with infinitive complement.

case is acceptible (if not preferable) for the floating quantifier. As there can be no direct predication in (14), this hypothesis fails to explain this kind of nominative.

(13) *Я купил машину, чтобы ездить на работу самому*

(14) *Андрей слишком труслив, чтобы прийти сам*

This last pair of examples also pose a problem for the long-distance agreement hypothesis, as the infinitive clause in (14) is surely a CP (as it contains a conjunction in C), and therefore its subject must be assigned dative case, while nominative cannot be "transmitted" from above (CP must be blocking such a transmission).

Third, despite the claims of most of the cited authors, there are other cases accessible to the floating quantifiers in an infinitive clause, cf (15).

(15) *Меня просят прийти самого*

A problem therefore arises for the local agreement theory, because to explain examples like (15) one has to agree that the subject PRO of an infinitive must have a choice of three cases instead of two. When it comes to machine translation system, however, this latest problem appears to be the least of them all, as we will show in section 3.

2. PRO control in Compreno

In theory the control of infinitive PRO in Russian is dependent on theta-roles. Namely, the choice of the controller follows the hierarchy of [Jackendoff 1972: 43]

Patient > Addressee > Agent

This makes it difficult to build the control link without semantic analysis of the input text. For Compreno, however, this problem can be solved. As was already stated above, Compreno transfigures an input text into a semantic-syntactic tree, where each node is a notion given a package of grammatical information and diathesis description. Therefore, if some node is a parent to an infinitive clause, given all the information about the model of the lexical item in this node we can predict, what kind of control will be applied in the particular construction. For instance, consider (16–19).

(16) *Я пришел починить трубу*

(17) *Вы сказали мне починить трубу*

(18) *Меня прислали вам починить трубу*

(19) *Я был прислан вам починить трубу*

Compreno correctly coindexes the first person pronoun with the subject position of the infinitive clause in all the cases, cf fig. 2–5 respectively:

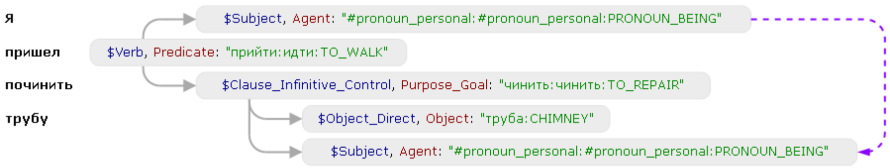
Я
пришел
починить
трубу
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Verb, Predicate: "прийти:идти:TO_WALK"
$Clause_Infinitive_Control, Purpose_Goal: "чинить:чинить:TO_REPAIR"
$Object_Direct, Object: "труба:CHIMNEY"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"

Fig. 2. Я пришел починить трубу

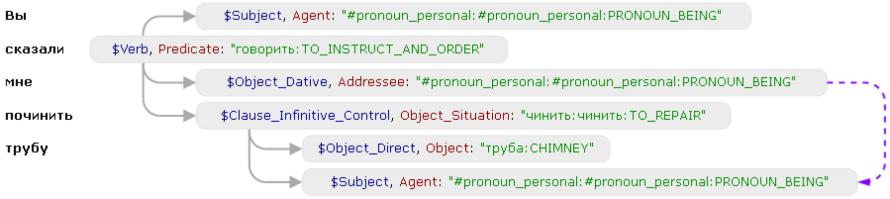
Вы
сказали
мне
починить
трубу
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Verb, Predicate: "говорить:TO_INSTRUCT_AND_ORDER"
$Object_Dative, Addressee: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Clause_Infinitive_Control, Object_Situation: "чинить:чинить:TO_REPAIR"
$Object_Direct, Object: "труба:CHIMNEY"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"

Fig. 3. Вы сказали мне починить трубу

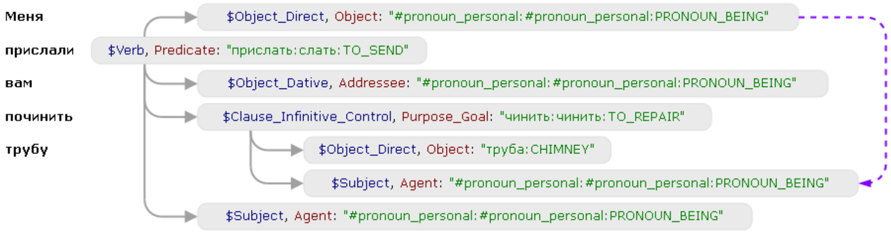
Меня
прислали
вам
починить
трубу
$Object_Direct, Object: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Verb, Predicate: "прислать:слать:TO_SEND"
$Object_Dative, Addressee: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Clause_Infinitive_Control, Purpose_Goal: "чинить:чинить:TO_REPAIR"
$Object_Direct, Object: "труба:CHIMNEY"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"

Fig. 4. Меня прислали вам починить трубу

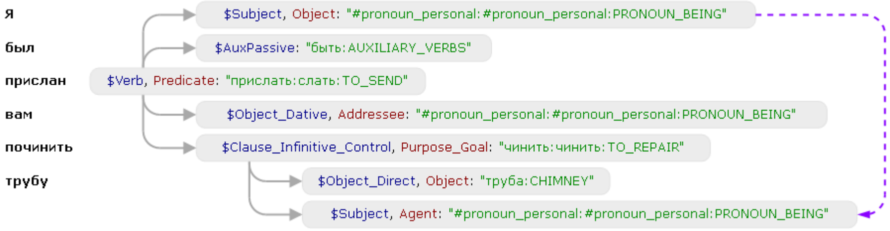
Я
был
прислан
вам
починить
трубу
$Subject, Object: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$AuxPassive: "быть:AUXILIARY_VERBS"
$Verb, Predicate: "прислать:слать:TO_SEND"
$Object_Dative, Addressee: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Clause_Infinitive_Control, Purpose_Goal: "чинить:чинить:TO_REPAIR"
$Object_Direct, Object: "труба:CHIMNEY"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"

Fig. 5. Я был прислан вам починить трубу

As one can see, the non-tree links follow strictly the hierarchy rule mentioned above[2].

Technically this mechanism is comprised of two separate tools. First, each verbal lexical class is marked with a classifying flag that encodes the information about what type of control this verb can have when attaching infinitive (like SubjectControl, DirectObjectControl, DativeObjectControl etc.); second, when choosing the control type

[2] Object slot in Compreno system roughly corresponds to Patient role, referring to a general undergoer of the situation

not only those classifying flags are taken into consideration, but also the voice of the verbal node (like Active or Passive).

In the examples (18–19) above the verb прислать 'to send' is marked with a "DirectObjectControl" flag. For a verb with such classifying flag the control rule has but two options—the first is to build a link between its direct object and infinitive PRO if and only if the diathesis is active; the second is to build a link between its subject and infinitive PRO if and only if the diathesis is passive.

Having enumerated all possible combinations of classifying flags and information of the chosen diathesis it is not difficult to build a system of such non-tree rules as will be able to cope with all the PRO control infinitive constructions. It should be taken into consideration that there are verbs that can choose different adjuncts for antecedent in Active voice, cf.:

(20) *просить мальчика сделать что-то*

(21) *просить у мальчика сделать что-то*

Although the number of combinations is therefore quite large, this system in general is nevertheless pretty straightforward.

3. Floating quantifiers and the case choice in Compreno

Floating quantifiers, as already shown in section 1 (fig.1, 6), are moved from within NP in Compreno syntactic structure and agree in case, number and gender with the parent node before movement. So for the floating quantifier to be marked with case X the subject of the minimal clause containing it has to be assigned the same case X. This is consistent with the local agreement hypothesis rather than any other.

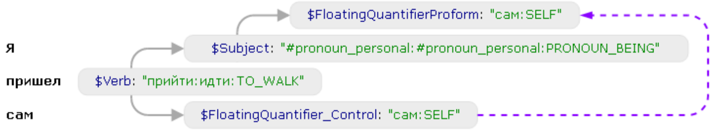
я
пришел
сам
$FloatingQuantifierProform: "сам:SELF"
$Subject: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Verb: "прийти:идти:TO_WALK"
$FloatingQuantifier_Control: "сам:SELF"

Fig. 6. Я пришел сам

This implies that a subject of an infinitive clause can be nominative (as in (22)), dative (as in (23)) or accusative (as in (24)).

(22) *Я хочу починить трубу сам*

(23) *Мне сказали починить трубу самому*

(24) *Меня просят починить трубу самого*

Moreover, there is a number of cases where the floating quantifier is unacceptable or at least dubious. Such are the instances where the PRO of the infinitive clause is coindexed with instrumental NP, as in (25).

(25) *Правительством планируется восстановить
*разрушенные территории *само/*самим*

In Compreno this problem is solved as follows. Every infinitive node of the tree (after the tree is built and all the non-tree links are established) is assigned a special flag that encodes information for the type of control applied in the particular structure. Let us call it TypeOfPRO flag. Restricting the subject-predicate relation, which is represented as one arc in the tree, we assign a certain case to the subject according to the flag of the parent node. Consider semantic-syntactic trees for (22–23), figures 7–8 respectively.

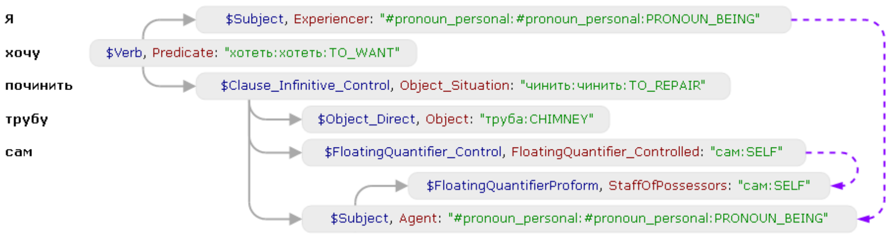
я
хочу
починить
трубу
сам
$Subject, Experiencer: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Verb, Predicate: "хотеть:хотеть:TO_WANT"
$Clause_Infinitive_Control, Object_Situation: "чинить:чинить:TO_REPAIR"
$Object_Direct, Object: "труба:CHIMNEY"
$FloatingQuantifier_Control, FloatingQuantifier_Controlled: "сам:SELF"
$FloatingQuantifierProform, StaffOfPossessors: "сам:SELF"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"

Fig. 7. Я хочу починить трубу сам

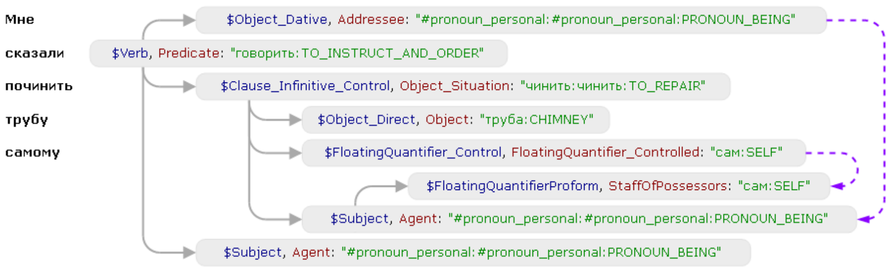
Мне
сказали
починить
трубу
самому
$Object_Dative, Addressee: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Verb, Predicate: "говорить:TO_INSTRUCT_AND_ORDER"
$Clause_Infinitive_Control, Object_Situation: "чинить:чинить:TO_REPAIR"
$Object_Direct, Object: "труба:CHIMNEY"
$FloatingQuantifier_Control, FloatingQuantifier_Controlled: "сам:SELF"
$FloatingQuantifierProform, StaffOfPossessors: "сам:SELF"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"
$Subject, Agent: "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"

Fig. 8. Мне сказали починить трубу самому

In (22), figure 7, the PRO of infinitive clause is controlled by the subject of the matrix predicate. Due to it, the infinitive node bares the NominativePRO flag and its PRO is assigned nominative case. It transmits nominative to the floating quantifier before movement, so that the moved quantifier is also marked nominative. This makes example (26) with dative case invalid.

(26) **Я хочу починить трубу самому*

In (23), figure 8, the PRO of infinitive clause is controlled by the dative object of the matrix predicate, so the infinitive node bares DativePRO flag. Its PRO is assigned dative case and transmits dative to the floating quantifier by agreement. Thus (27) with nominative is also analyzed as invalid

(27) **Мне сказали починить трубу сам*

By the same principle object control PRO as in (24) can be assigned accusative and allow for the floating quantifier to be accusative via agreement. As for (25), the infinitive node is marked with MarginalPRO flag, which means that its PRO is controlled in such a way that some syntactic transformations are blocked inside this clause. Floating quantifier movement is one of those blocked transformations, hence unacceptability of (25).

4. Further applications of the TypeOfPRO flag

The mechanism illustrated above has several other applications apart from the case choice for the floating quantifiers. It has been noticed before, that subject control PRO infinitive constructions can take a short form adjective as a complement (28), whereas object control PRO infinitive constructions cannot (29)

(28) *Я должен/хочу быть красив/красивым*

(29) *Мне хочется быть *красив/красивым*

As there are lexical items without full form it is crucial for the machine translation to choose a synonymous lexical item for constructions with object control PRO infinitive, cf (30–31).

(30) *Я должен быть рад*

(31) **Мне хочется быть рад*

In Compreno it is simply done by applying the TypeOfPRO flag on the infinitive node to restrict form choice in the complement node. It is according to this test that the PRO in constructions such as (25) are assigned dative in Compreno: short form of adjective is unacceptable as a complement in such constructions, cf. (32)

(32) *Правительством планируется быть *компетентно/компетентным*

5. Conclusion

Although a model-based approach to machine translation is known to be relatively labour-intensive, it looks more promising when interpreting and translating such complex structures as those with floating quantifiers in infinitive clauses. For the analysis of those constructions it seems more reasonable to follow the local agreement hypothesis and assign case to PROs, however intuitively dubious that may be.

References

1. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii Dialog”], Bekasovo, pp. 90–103.
2. *Babby, L. H.* (1998), Subject control as direct predication: Evidence fromRussian, Annual Workshop on Formal Approaches to Slavic Linguistics: The Connecticut Meeting, pp. 17–37.
3. *Bogdanov A. V., Dzhumaev S. S., Skorinkin D. A., Starostin A. S.* (2014), Anaphora analysis based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”], Bekasovo, pp. 89–102.
4. *Comrie B.* (1974), The second dative: A transformational approach, Slavic Transformational Syntax, No 10, pp. 123–150
5. *Fleisher N.* (2006), Russian dative subjects, case, and control, Ms., University of California, Berkley
6. *Franks S.* (1990), Case, Configuration and Argumenthood: Reflections on the Second Dative’, Russian Linguistics, Vol. 14, pp.231–254.
7. *Franks S.* (1995), Parameters of Slavic Morphosyntax, Oxford University Press, Oxford.
8. *Greenberg G. A., Franks S.* (1991), A parametric approach to dative subjects and the second dative in Slavic, Slavic and East European Journal, Vol 35, pp. 71–97.
9. *Jackendoff R* (1972), Semantic Interpretation in Generative Grammar, MIT Press, Cambridge, MA.
10. *Starostin A. S., Smurov I. M., Stepanova M. E.* (2014), A production system for information extraction based on complete syntactic-semantic analysis, available at: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/StarostinAS.full.pdf.