CO-ORDINATIVE ELLIPSIS IN RUSSIAN TEXTS: PROBLEMS OF DESCRIPTION AND RESTORATION

Igor A.BOLSHAKOV
VINITI, Academy of Sciences of USSR
Moscow, 125219, USSR

ABSTRACT. Russian elliptic constructions are examined from the point of view of syntactic analysis. Reciprocal elements in a co-ordinative elliptic sentence are exposed and possible types of their similarity are explored. Linear formulae of ellipsis for most textual cases are constructed and statistics of their use is discussed. As a result the main steps of ellipsis restoration algorithm are outlined.

INTRODUCTION. The investigations of ellipsis (gapping) in natural language sentences with structural methods have been carried on for more than 20 years, but algorithms of automatic restoration of omitted words either in coherent Russian texts or in sequences of Man-Machine interaction replicas have not yet been proposed. The problem is still topical. Indeed, at an average each 7th entry in Great Soviet Encyclopedia and 25th abstract in a common Soviet abstract journal contain at least one sentence of a following kind:

В первой базе данных содержится информация по всем отраслям знаний, во второй - только по гуманитарным наукам.

/The first database contains information on all knowledge fields, the second one - only on the Humanities./

Conversion of such a sentence to a formalized language, e.g.for automatic updating of factographic databases, without the ellipsis restoration is impossible. Meanwhile ellipsis in Russian sci-tech texts is very diversified and covers any part of a sentence and most frequently the predicate with adjacent words.

Early Soviet works /Leontieva, 1965/, /Paducheva et al.,1973/, /Korelskaya et al., 1973/ had examined the phenomenon from the point of view of synthesis. But when synthetic rules transform a non-empty entity to an empty one, corresponding analytical rules are not their trivial reversion. We examine elliptic constructions in co-ordinative sentences with orientation to analysis, i.e. to parsing algorithms including restoration of omitted words.

Several important issues should be forced for our purpose: 1) introducing the notion of reciprocal elements in non-omitted parts of elliptic sentences; 2) defining new types of syntagmata for restoration of semi-destroyed links between words within the reciprocals; 3) exploring possible types of reciprocal similarity; 4) describing co-ordinative sentences with minimal number of elliptic transformation formulae; 5) collecting statistics of the formulae use, which implies a search order of a specific formula for a given sentence. Thus a base for outlining the main steps of an ellipsis restoration algorithm is formed.

RECIPROCAL ELEMENTS. There are several kinds of word omission in natural languages. Among them co-ordinative reduction is carried out according to the formula XQ & X*Q => X & X*Q or QX & QX* => QX & X*, where X and X* are different words or word groups, Q is a recurrent group of words, and & stands for a co-ordinative conjunction or just a comma: $(\text{He took})_Q$ $(\text{the bread})_X$ $(\text{and})_{\&}$ $(\text{he took})_Q$ $(\text{the milk})_{X^*}$ => $(\text{He took})_Q$ $(\text{the bread})_X$ $(\text{and})_{\&}$ $(\text{the milk})_{X^*}$.

Co-ordinative ellipsis includes co-ordinative reduction as a subset. Two or more phrase segments co-ordinated in a single sentence have some identical parts not necessarily standing at the borders of the segments. For economy's sake the natural language omits, wholly or partially, the recurrent part of a segment, as a rule, in the second or subsequent one. The omission may be carried out according to the formula XQ_oY & $X^*Q_oY^*$ => XQ_oY & X^* - Y^*, where X and X^* are a pair of somewhat similar, but compared and even semantically opposed elements referred to hereinafter as reciprocal; Y and Y* are another such pair; the co-ordinative segments are XQ_oY and X^* - Y^*. In Russian writing ellipsis is often accompanied with a dash.

The common feature of elliptic omissions is that a connected dependence tree for a sentence cannot be built without their restoration. The parsing algorithm should take into account and properly distinguish the specific phenomena occurring in a given sentence, e.g. ellipsis and zero copula going together.

BASIC ASSUMPTIONS. We examined a co-ordinative ellipsis assuming the following:

- A context confined to a standalone elliptic sentence is sufficient for restoration.
- An ellipsis in a sentence is explicated with a dash (along with some implied indications), but omission may occur not where the dash is or not only there. Comparative and other constructions without a dash have not been examined, but most of our statements cover them, too.

Omitted parts, after restoration and, if necessary, morphological re-agreement, exactly restore the original meaning of a given sentence as a whole. Therefore recurrent parts in different segments have had exactly the same length, linear order, and lexical content before omission.

- A dependency tree was selected as a tool for description of links between word-forms in a sentence, and the nature of the links corresponds to those suggested in "Meaning - Text" model /Mel'čul, 1973/.
- Co-ordinatively reduced words should be restored, along with the ellipsis proper, where this provides connectedness of a parse tree and saturation of obligatory valencies.
- There exists an algorithm for syntactic analysis of Russian texts which can cope with parsing any ellipsis-free sentence. It can also parse, consistently and unambiguously, the complete segments of an elliptic

sentence and tie links, even ambiguous, within word groups in incomplete segments, preparing them as disjoited "bushes" for a final parse tree.

STATISTIC OBSERVATIONS. As many as 600 elliptic sentences were extracted from large (more than 1000 pages) corpus of texts in various fields and genres (abstracts, articles, popular science books, and brochures). The material was arranged in a minimal number of elliptic formulae. A total of 24 formulae was found, but only four of them exceeded the 4% threshold (see Table).

In the Table the arrows show the direction of synthesis; Q_0 stands for an omitted part replaced with a dash; Q_{-1}, Q_1 are additional omissions possible to the left and to the right of the dash; P_0 and P_1 are optional segments (modifiers and the like), not involved in the ellipsis.

Available statistics permits to conclude:

- The most widely used formula (single omission between two reciprocals) has exceeded the use total of all other formulae, and four leading formulae together account for more than 80 per cent.
- Number N_1 of reciprocal pairs and number N_0 of omissions in a sentence usually satisfy the formula $| N_1 - N_0 | <= 1$, since reciprocals and omissions are commonly interleaving each other, but it is possible to construct counter-examples.
- Cases with $N_0 > 2$ and/or $N_1 > 2$ amount for less than two per cent.
- Approximately in 7% of cases nonsymmetric ("chiastic") formulae occurred. For example, formula P_0QXY & $X^*QY^* => P_0QXY$ & $X^* \neq Y^*$ may have the following realization: (Совместная работа)$_P$ (требует)$_Q$(от инженера)$_X$(расширения математического кругозора)$_Y$ (,а)$_{\&}$ (от математика)$_{X^*}$ - (владения некоторыми математическими знаниями)$_{Y^*}$. /(A joint work)$_P$ (requires)$_Q$ (from engineer)$_X$ (broadening of his mathematical outlook)$_Y$ (and)$_{\&}$ (from mathematician)$_{X^*}$ - (the possession of some technical background)$_{Y^*}$./
- Approximately four per cent of all cases did not fit into our formulae. Either a possible formula was too complex to be practical, or the shape of the sentence was dubious and even incorrect from the point of view of a human editor.

ENCLOSING SYNTAGMATA. Before searching reciprocals it is necessary to establish syntactic links within word groups in an incomplete segment. The convenient tools for describing these links, the so-called syntagmata, are not always sufficient here, see, e.g. the word combinations: обобщенные экономические/отдельные производственные(показатели) /generalized economic/ some productional (indications)/; 17 старыми/пятью новыми (предприятиями) /by 17 old/by five new (enterprises)/; из капиталистических/из социалистических (стран) /from capitalist/from socialist (countries)/ число имевшихся/число вновь приобретенных (книг) /the number of available/ the number of newly bought (books)/; 80 тыс. публицистических/200 тыс. научных(публикаций) /80,000 publicistic/200,000 scientific (publications/. Reciprocals in the pairs above are divided with slashes, and connected nouns (derived from complete segments) are given in parentheses.

We propose dependencies of a new type, named enclosing syntagmata (ES). These are established either between co-subordinated attributes, linking brother nodes from right to left (обобщенные <= экономические)

or between a number and co-subordinated attribute (17 <= старых) or between a proposition or a noun with predicative properties and subordinate noun's attribute (из <= капиталистических, число <= имевшихся, тыс. <= публицистических).

Several ES constitute the links between a dominating word and its indirectly subordinated one. Via ES the nodes most important for reciprocal matching turned to be within easy distance from their bushes' roots (may be, at the very roots). Similar links can be established in complete segments, too. They transform a relevant subtree to an acyclic graph, which facilitates matching nodes and arcs.

The set of ES is subject to updating so far. Caution should be used however about updating. Indeed, attempts to directly link words arbitrarily distant within a convenient-dependency tree, though eliminating the very notion of ellipsis, lead to superfluous complexity of ES and of a global model of natural language, too.

SIMILARITY FEATURES. Manual segmentation of all available sentences has made clear that antagonists in reciprocal pairs are not in general case mutually isomorphic, i.e. their subtree do not quite coincide. Therefore, the labels at the matched nodes should be forcedly involved and these are of the following types:

Lexical. Lexemes at the roots and/or their direct subordinates in 22 per cent were strictly coincide.

Morphological. In most cases lexemes at the matched nodes belonged to the same part of speech, and their wordforming characteristics were in agreement: nouns and numbers expressed by words - in case; adjectives (pronominal included) and participles - in gender, number, and case; personal verb-forms - in gender, number, and person. Only in 30 per cent of cases the agreement has not been revealed (numbers in digits, abbreviations, etc.).

Syntactic. Some indicators treated in the "Meaning - Text" model as syntactic, might coincide, e.g. interogativity of lexemes КУДА and СКОЛЬКО in the sentence Я спросил, куда идти, а он - сколько времени. /I asked where I should go, and he asked what time it was./

Semantic. If matching labels of the three kinds mentioned failed or at once several nodes in a complete segment were similar to the node in an incomplete one, then coincidence of even one semantic indicator is important. Taxonomy of such indicators is not established yet. Several facets (classification aspects) with admissible intersection of their scopes are expected to suit well, but simple thesaurical hierarchies are not excluded, either. We have specifically observed: quantitative and cardinal words (28%), synonymy and antonymy (10%), hyponymy and hyperonymy (i.e. genus/species, part/whole) (3%), and other kinds of such similarity (2%).

When semantic similarity within pairs X/X* and Y/Y* failed, semantic proportion X/Y ≈ X*/Y* has been sometimes observed,

but algorithmic verification of the proposition is difficult.

DRAFT ALGORITHM. An algorithm we propose for restoration of co-ordinative ellipsis (CE) consists of the following main steps:

1) A basic ("ellipsis-free") algorithm is carrying on syntactic parsing, sentence by sentence. CE-flag is simultaneously derived from the three other flags signalling connectedness of the parse tree, co-ordinativity of the sentence structure, and presense of a zero copula.

2) If the CE-flag is set for the sentence then algorithm of ellipsis restoration starts up. Given borders of the incomplete segment and a dash position, X° and Y° elements are extracted. They should be inwardly linked via convenient and, if necessary, enclosing syntagmata.

3) The X° and Y° found, their reciprocals X and Y in the complete segment are searched starting from its borders inwards. The node-by-node matching for the bushes starts from their roots, uses all node labels, and should be continued at lower tree level, if function words are encountered at the roots.

4) The X and Y found, dependency links in the complete segment are traced, mainly in ascending order. The number and location of borders between reciprocals and omissions are thus revealed.

5) The specific CE-formula is diagnosed and omitted word groups are duplicated with its help from the complete segment to the incomplete one. Re-agreement of the restored words is carried out, if necessary, e.g. in number, person, and/or gender for a predicate.

If there are more than two co-ordinative segments in the sentence, then steps 3 to 5 are repeated for every subsequent segment applying preceding (complete or just restored) one.

ACKNOWLEDGEMENTS. The author wishes to express his great gratitude to I.G.Bider, L.L.Iomdin, L.G.Mitiushin, E.V.Paducheva, and especially to Yu.D.Apresian for their keen critique. He is also thankful to M.S. Frudkin and A.A.Ruskina for editorial remarks.

REFERENCES

Iomdiev, N.N. (1965) 'Analysis and synthesis of Russian elliptic sentences.' Nauchno-tekhnicheskaya informatsiya, 11, Moscow: VINITI, pp. 41-46.

Paducheva, E.V. & Dashchenko, T.K. (1973) 'Ellipsis as zero anaphoric sign.' Nauchno-tekhnicheskaya informatsiya, Ser. 2, 5, Moscow: VINITI, pp. 29-31.

Korelskaya, T.D. & Paducheva E.V. (1973) 'Transformation in symmetric constructions: coordination and ellipsis.' Nauchno-tekhnicheskaya informatsiya, Ser. 2, 9, Moscow: VINITI, pp. 29-38.

Mel'čuk, I.A. Experience in the theory of linguistic "Meaning - Text" models (in Russian). Moscow: Nauka, 1973.

TABLE. The four most frequent ellipsis formulae

#	%	Formula	Russian example	Word-by-word translation
1	61.7	$[P_c]XQ_0Y$ & $X^\circ Q_0 Y^\circ$ => $[P_0]XQ_0Y$ & X° – Y°	(Без округления)$_{P_0}$ (при сдвиге дополнительных)$_X$ (кодов ошибка будет)$_{Q_0}$ (отрицательной)$_Y$ (,а)$_{\&}$ (при сдвиге обратных)$_{X^\circ}$ – (положительной)$_{Y^\circ}$	(Without rounding of)$_{P_0}$ (when shifting additional)$_X$ (codes an error will be)$_{Q_0}$ (negative)$_Y$ (,and)$_{\&}$ (when shifting reverse)$_{X^\circ}$ – (positive)$_{Y^\circ}$
2	8.0	$[P_c]XQ_0YQ_1$ & $X^\circ Q_0 Y^\circ Q_1$ => $[P_0]XQ_0YQ_1$ & X° – Y°	(2/3)$_X$ (микроЭВМ используются)$_{Q_0}$ (в коммерческой)$_Y$ (сфере)$_{Q_1}$ (,а)$_{\&}$ (1/3)$_{X^\circ}$ – (в научно-технической)$_{Y^\circ}$	(2/3)$_X$ (of microcomputers are used)$_{Q_0}$ (in commercial)$_Y$ (field)$_{Q_1}$ (and)$_{\&}$ (1/3)$_{X^\circ}$ – (in sci-tech)$_{Y^\circ}$
3	6.7	$[P_c]Q_0X$ & $P_1Q_0X^\circ$ => $[P_0]Q_0X$ & P_1 – X°	(Если температура одинакова, то)$_{P_0}$ (плазма называется)$_{Q_0}$ (изотермической)$_X$ (,а)$_{\&}$ (в противном случае)$_{P_1}$ – (неизотермической)$_{X^\circ}$	(If the temperature is identical then)$_{P_0}$ (plasma is named)$_{Q_0}$ (isotermic)$_X$ (and)$_{\&}$ (in opposite case)$_{P_1}$ – (noniso-termic)$_{X^\circ}$
4	3.7	$[P_c]Q_{-1}XQ_0Y$ & $Q_{-1}X^\circ Q_0 Y^\circ$ => $[P_0]Q_{-1}XQ_0Y$ & X° – Y°	(В 1982 г.)$_{P_0}$ (фонды)$_{Q_{-1}}$ (университетских)$_X$ (библиотек составили)$_{Q_0}$ (600 тыс. т)$_Y$ (,а)$_{\&}$ (специальных)$_{X^\circ}$ – (порядка 40 тыс. тт)$_{Y^\circ}$	(In 1982)$_{P_0}$ (the resources)$_{Q_{-1}}$ (of university)$_X$ (libraries account for)$_{Q_0}$ (600,000 vol.)$_Y$ (and)$_{\&}$ (of specialized)$_{X^\circ}$ – (approximately 40,000 vol.)$_{Y^\circ}$