COORDINATIVE ELLIPSIS IN RUSSIAN TEXTS: PROBLEMS OF DESCRIPTION AND RESTORATION

Igor A. NOLSHAKOV
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 algorithms are outlined.

INTRODUCTION. The investigations of ellipsis (omission) 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 replies 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:

In many cases of such a sentence to a formalized language, e.g. for automatic updating of taxonomic 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], [Korlakova et al., 1973] had examined the phenomenon from the point of view of synthesis. But when syntactic rules transform a non-empty entity into 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 elliptic sentences with minimal number of elliptic transformation formulae; 4) 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 \( XQ \& Y \rightarrow X \& Y \).
sentence and the links, even ambiguous, within word groups in incomplete segments; preparing them as disjointed "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 (books, articles, abstracts, articles, popular science books, and brochures). The material was arranged in a minimal number of elliptic formulae. A total of 2% formulae was found, but only four of them exceeded the 4% threshold (see Table).

In the Table the arrows show the direction of synthesis; $Q$ stands for an omitted part replaced with a "dash"; $Q'$, $Q''$, are additional omisions possible to the left and to the right of the dash; $P$ and $P'$ are optional segments (modifiers find 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 reciprocity of formulae together account for more than 50%.
- Number $N_0$ of reciprocal pairs and number $N_0$ of omissions in a sentence usually satisfy the formula $N_0 = N_1 < 1$, since reciprocals and omissions are commonly interlacing each other, but it is possible to construct counter-examples.
- Cases with $N_0 > 2$ and/or $N_0 > 2$ amount for less than two per cent.
- Approximately in 7% of cases non-symmetric ("chiastic") formulae occurred. For example, formula $P\rightarrow Q\rightarrow X\rightarrow Y\rightarrow X\rightarrow Y\rightarrow X\rightarrow Y$ may have the following realization: (Some word) $Q_0$ is a modifier $X$ of the sentence $Y_0$ (broadening of his mathematical background).

- 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, e.g., the word combinations observed are not involved in the ellipsis. Reciprocals are divided into three classes: (classification aspects) with an impossible/possible agreement.

The set of ES is subject to updating, so far. Caution should be used however about updating. Indeed, attempts to directly link words arbiter with case-marking of sentence-dependency tree, though eliminating the very notion of ellipsis, lead to superficial 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 length at the matched nodes should be forcibly involved and these are of the following types:

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

- **Morphological.** In most cases lexemes at the matched nodes belonged to the same part of speech, and their word-forming characteristics were in agreement; nouns and numbers expressed by words — in case, adjectives (proclinal included) and participles — in gender, number, and case; personal verb-forms — in gender, number, and person. Only in 20 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., interactivity of lexemes (ENGLISH and CHINESE) in the sentence "a person went to see how the game had progressed". /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 synonymous indicators 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 example, semantic hierarchies are not excluded, either. We have specifically observed: quantitative and cardinal words (20%), synonymy and antonymy (10%), hypernymy and hyponymy (i.e., genus/species, genus/individual) (20%), and other kinds of such similarity (20%).

When semantic similarity within pairs $X/Y$ and $Y/X$ failed, semantic proportion $X/Y = X/Y$ has been sometimes observed.
but algorithmic verification of the proposition is difficult.

**ALGORITHM.** An algorithm we propose for restoration of core-destructive ellipsis (CE) consists of the following main steps:

1. A brute ("calligraphic") algorithm in conveying on symbolic parsing, sentence by sentence. CE-clues are simultaneously derived from the known other clues signalling completeness of the parse tree, co-ordinative or the sentence structure, and presence of a gap operand.

2. In the CE-clues for the sentence there exists an algorithm of clues coordination either up to five hundredth of the incomplete segment and a dual position, $Y$ and $Y'$ elements are established. They should be internally linked via conjunction and, if necessary, excluding conjunction.

3. The $X$ and $X'$ found, that co-represent $X$ and $X'$ in the complete segment and not embedded regarding the known operands. The node-ly nodes matching for the broken structure from the root node, when all node nodes, and should be combined of the root node nodes. $Y'$ function would not be uncovered at the root.

4. The $X$ and $Y$ found, dependency other in the complete segment not been matched, partly in accordance with the sentence and location of broken, broken, co-represent, and conjunction are then revealed.

5. The possible CE-clues in dừngual and omitted word groups are duplicated with the help from the complete segment to the incomplete one. Re-agreement of the restored words is carried out, if necessary, only in number, person, and/or gender for a predicate.

In these cases more than two core-destructive segments in the sentence, then steps 3 to 5 are repeated for every subsequent segment applying preceding (complete or just restored) one.

**APPLICATIONS.** The author wishes to express his great gratitude to I.G. Bladyk, Ye.I. Mihaylov, I.G. Mitinchuk, V.V. Paduchev, and especially to Yu. A. Aseyan for their kind criticism. He is also thankful to R.B. Panin and A.A. Kohman for additional remarks.

**BIBLIOGRAPHY.**

Korolkov, E. N. (1965) Analyzis and synchronic on Russian adjectival morphology. Moscow-Teh-Cheskaya Informatsiya, 11, Moscow: VASEN, pp. 20--66.

Paduchev, V. V. (1975) The ellipsis as zero morphemic sign. Moscow-Madecheka Informatsiya, 27, Moscow: VASEN, pp. 20--5.

Korolkov, E. N. (1975) Transformation in syntactic constructions: reconstruction and ellipsis. Moscow-Madecheka Informatsiya, 27, Moscow: VASEN, pp. 29--56.

Del'chuk, I. A. Experience in the theory of linguistic "Meaning - Text" models (in Russian). Moscow: Reakta, 1975.

### Table: The Formulas of Core-Destructive Ellipsis in Russian

| # | Formulas | Russian example | Word-by-word translation |
|---|----------|----------------|--------------------------|
| 1 | $xy \triangleright x \times y' \times y$ | (без окончания) $x_0$ (при отсутствии дополнительных элементов) $x_0$ (после окончания) $y_0$ (при отсутствии предикатов) $x_0$... | (without rounding of $x_0$ (when adjusting additional) $y_0$ (code) an error will $y_0$ (negative) $x_0$... | (2/3) $x$ (for microcomputers are used) $y$ (in commercial) $x$... |
| 2 | $...$ | (каждая $q_1$ (and) $q_1$ (1/3) $x_0$... | (in socio-technical) $x_0$... |
| 3 | $...$ | (если температура одинакова, то) $x_0$ (шаровая манипуляция) $x_0$ (геометрический) $x_0$ (в простом случае) $x_0$... | (if the temperature is identical, then) $x_0$ (sphere in named) $x_0$ (and) $x_0$ (in opposite case) $x_0$... |
| 4 | $...$ | (в 2002 году $x_0$ (фактор) $y_0$ (глубинная манипуляция) $x_0$ (динамический), $x_0$... | (in 2002, $x_0$ (factor) $y_0$ (depth m) $x_0$... (approximately 40,000,000 $x_0$... |