Stable Coordinated Pairs in Text Processing *

Igor A. Bolshakov[1], Alexander Gelbukh[1,2], and Sofía N. Galicia-Haro[1]

[1] Center for Computing Research (CIC),
National Polytechnic Institute (IPN), Mexico City, Mexico
{igor,sofia,gelbukh}@cic.ipn.mx, www.Gelbukh.com
[2] Computer Science Department, Chung-Ang University, Korea

Abstract. Stable coordinated pairs (SCPs), e.g., *comments and suggestions*, *far and near*, *sooner or later* occur rather frequently in various European languages, whereas there is only a few thousands of them. We argue that a dictionary of SCPs of the given language supplied with their characteristics of lexical, morphological, syntactic, semantic, and pragmatic nature can help in such important natural language processing tasks and applications as word sense disambiguation, parsing, as well as detection and correction of semantic errors.

1 Introduction

Coordinated constructions (CCs) such as *define the notion and show the result*; *teaching and research assistant*; *efficient, robust and portable*; *Canada and Mexico* are rather frequent in different genres of European languages. E.g., on average each fifth sentence of the news at the Russian web site *Gazeta.ru* contains a CC.

We call a CC stable coordinated pair (SCP) is the mutual information of its components is greater than one. A significant part of CCs are SCPs, e.g., *comments and suggestions*, *air and space*, *far and near*, *sooner or later*. It texts of *Gazeta.ru* nearly each third CC is a SCP. A SCP as a whole can play the syntactic role of any major part of speech—noun, adjective, verb, or adverb.

A SCP is a contiguous segment of text, though one or more words in it can vary in their morphological form depending on the outer context, especially in highly inflectional languages like Russian. The words in SCP are lexically fixed: replacement of any of them even by a synonym converts the SCP to a (non-stable) CC, sometimes awkward. E.g., the Russian *ran′še i teper′* is a SCP, but *ran′še i sejčas* is not, though *teper′* and *sejčas* are synonyms, both CCs meaning 'earlier and today' [15].

All CCs are rather difficult for automatic processing, in analysis, generation, and error correction. In Meaning–Text theory (MTT) [12, 13] a structural description of CCs was done for Russian [12], English [16], and Spanish [2]. CCs were also described in HPSG [18]. We follow the MTT (dependency syntax) approach below, but this affects only the description of syntactical structure of a CC and of a SCP.

While the number of CCs in a language is infinite, there are only few thousands of SCPs. Thus, a complete dictionary of SCPs can be compiled, with all their characteristics

* Work done under partial support of Mexican Government (CONACyT, SNI) and CGPI. The second author is currently on Sabbatical leave at Chung-Ang University.

V. Matoušek and P. Mautner (Eds.): TSD 2003, LNAI 2807, pp. 27–34, 2003.

of lexical, morphological, syntactic, semantic, and pragmatic nature. Here we argue that such a dictionary can help in such important natural language processing (NLP) tasks and applications as word sense disambiguation, parsing, detection and correction of semantic errors, detection of coreferences, and text segmentation into paragraphs. Specifically, our goals in the paper are:

- To expose the most important categorization parameters of SCPs;
- To show the statistics on SCPs in terms of the parameters introduced;
- To overview the dependency substructures corresponding to various SCPs;
- To compare SCPs with another type of stable and lexically restricted word combinations—collocations;
- To review how to use the information on SCPs for the mentioned NLP tasks.

Our study is based on collections of Russian, Spanish, and English SCPs. The Russian collection [3] contains more than 3200 entries. The Spanish and English ones contain only 620 and 340 items; however, for the majority of them we have found Russian analogues, which allows us to consider them together. We give examples in English (marked by ***E***), Spanish (***S***), and Russian (***R***).

2 Categorization Parameters of Coordinated Pairs

Part of Speech. Part of speech (POS) of an SCP is determined by its role in a sentence: it can be a noun (NG), adjective (AjG), adverb (AvG), or verb (VG) group. E.g., ***E*** *coffee and cakes* is a NG, *effective and far-reaching* is a AjG, *divide and rule* is a VG, *sooner or later* is a AvG. A prepositional group can play the role of both AjG and AvG, e.g., ***E*** *in form and in substance* is AjG when modifying *unconstitutionality* (≈ 'formal and substantial') and AvG when modifying *verify* (≈ 'formally and substantially'). We consider such SCPs homonymous.

Inflectionality. A SCP is inflectional when at least one its component changes its morphological form depending on the syntactical governor of the SCP. E.g., ***E*** *cut and paste* has 3 forms: *cut and paste*, *cut and pasted*, *cutting and pasting*;
S *estirar y aflojar* 'to stretch and to loosen' has tens of forms, as any Spanish verb.

Since noun SCPs rarely have both singular and plural forms, we consider ***E*** *mom and dad* and *moms and dads* independently. Grammatical number of a noun SCP is usually plural (except for coreferential SCPs, see below). In Slavic languages noun SCPs change their grammatical case, e.g., ***R*** *armija i flot* 'army and fleet' has six cases. Slavic adjectives change in case, gender, and number (24 combinations but only 12 different forms in Russian).

Use of Conjunctions and Tree Substructures. In overwhelming majority, the coordinated components are jointed by a standard copulative conjunction: ***E*** *and*, ***S*** *y/e*, ***R*** *i*. The single conjunction can be also disjunctive (***E*** *or/but*, ***S*** *o/u*, ***R*** *ili/libo*). There exist in English and Spanish a nonstandard conjunction vs. = versus (***E*** *renting vs. buying*, *evolution vs. creationism*). In English, the sign & can be used instead of *and*.

For all cases of singular conjunction c, the dependency grammars [13] ascribe to CCs with the components P_1 and P_2 the following structure on the surface syntactical level: $P_1 \rightarrow c \rightarrow P_2$. A SCP can include a disjoint pair of conjunctions: ***E*** *both... and*,

either... or, *neither... nor*; ***S*** *y... y*, *ni... ni*, *o... o*, ***R*** *i... i*, *ni... ni*, *ili... ili*; for example, ***R*** *ni tuda ni sjuda* 'neither to nor fro'. The syntactical structure of such SCPs is

$$\overbrace{c_1 \to P_1 \quad c_2 \to P_2}.$$

In the simplest and the most usual case, the components P_1 and P_2 with the conjunction(s) cover the whole SCP: ***E*** *comments and suggestions*, *air and space*, *far and near*, *sooner or later*. When other words participate in the construction, e.g., ***E*** *(Bank for) Reconstruction and Development*, ***R*** *počtovyj i elektronnyj (adresa)* 'postal and electronic addresses', we preserve in the dictionary the outer part (in parentheses) only if a given coordinated pair is used uniquely in this environment. Hence, would we find in texts a noun differing from *Bank* for preceding the pair *Reconstruction and Development*, we preserve only the pair without *Bank for*.

Sphere of Usage. This is semantic parameter. These types seem to suffice:

- Official documentation and mass media cliches, e.g. the titles of well known organizations (***E*** *Bible College and Seminary*, *Bank for Reconstruction and Development*, ***S*** *Hacienda y Crédito Público* 'Treasury and Public Loans');
- Business entities, e.g. the names of common shops and workshops (***R*** *ovošči i frukty* 'vegetables and fruits', ***S*** *tintoreria y lavanderia* 'dry cleaner's and laundry') or store departments (***S*** *carnes y lacteos* 'meat and milk products');
- Cultural and sci-tech terms (***E*** *air and space*, *mechanics and dynamics*);
- Everyday life cliches: ***E*** *mom and dad*.

Semantic Link Between Components. The following values seem to suffice:

- (Quasi-)synonyms and repetitions: ***E*** *fun and games*, *services and offers*; ***R*** *agitacija i propaganda* 'drive and propagation'; ***S*** *desarrollo y consolidacion* 'development and consolidation', *muchos y muchos* 'many and many';
- Co-hyponyms: ***E*** *axioms and theorems*, *Earth and Moon*; ***S*** *maestria y doctorado* 'MS and PhD degrees'; ***R*** *akušerstvo i ginekologija* 'obstetrics and gynecology.'

We left the degree of the meaning intersection between quasi-synonyms or co-hyponyms rather vague, since it in no way affects their belonging to SCP.

- Antonyms, quasi-antonyms, and opposite notions: ***R*** *bednye i bogatye* 'the poor and the rich'; ***E*** *to and from*, *more or less*, *missiles and antimissiles*; ***S*** *material y espiritual* 'material and spiritual'.
- Co-participants of a situation: ***E*** *management and budget*, *globe and mail*; ***S*** *productos y servicios* 'products and services'.

The latter type is most complicated semantically. The situation type can be:

- In ***E*** *management and budget*, it is primarily determined by *management* while *budget* is a resource of *management*.
- The pair can reflect a logical sequence: the first coordinated component P_1 contains an antecedent, while the second, P_2, its logical consequence.
- In ***E*** *dead and buried*, ***R*** *pojti i ne vernut'sja* 'to go and not to come back', there is a time sequence of actions, with the time of P_1 preceding that of P_2.
- In ***E*** *mother and child*, there is a physical cause-consequence link: the mother gave birth to the child. Another example: ***E*** *war and destruction*.
- In ***E*** *newspapers and other mass media*, there is a genus-species link: P_1 is a species and P_2 contains its genus.

Idiomacity. The SCP is considered an idiom if its meaning is not the sum of its components' meanings. Idioms whose meaning contains that of P_1 or P_2 are semiphrasemes [14], e.g., ***R*** colloq. *deševo i serdito* 'cheaply and impressively' (lit. 'cheaply and angrily'). Otherwise it is a complete phraseme [14], e.g., ***R*** *ni Bogu svečka, ni čertu kočerga* 'absolutely useless' (lit. 'neither a candle for the God nor a poker for the devil'),***S*** *a diestra y siniestra* 'recklessly' (lit. 'to the right and to the left'), *entre ceja y ceja* 'constantly in the brain' (lit. 'between brow and brow'). Idioms whose meaning contains that of the main component plus other semantic elements are quasi-phrasemes [14], e.g.: ***E*** globe and mail 'globe-wide world and mail linking its parts'.

To represent the semantics of idioms, their meaning is to be explicitly specified in the dictionary. Luckily, the majority of SCPs are not idioms in the sense of [14].

Reversibility. Some SCP can appear in texts in both orders, e.g., ***E*** *coffee and cakes*. We store them as separate SCPs, without indication of prevalent order.

The irreversible pairs (called irreversible binomials in [11]) often contain temporal, logical, or causative sequence mentioned above: ***E*** *mother and child*, *sooner or later*, *cause and effect*, *prepare and submit*;***S*** *fabricar y comercializar* 'produce and sell'.

Lexical Peculiarity. This means that at least one component cannot be used out of the SCPs: ***E*** *to and from*, ***R*** *i tam i sjam* 'here and there'; ***S*** *dimes y diretes* 'tittle-tattle'. The dictionary entry for such words should consist only of the reference to its SCP.

Inclusion of Proper Names. A component of SCP can be a proper name, maybe geographic: *Asia and Africa*, ***S*** *America Latina y el Caribe*. Each proper name should enter to a basic thesaurus of encyclopedic type.

Coreferentiality. In very rare cases P_1 and P_2 refer to the same person:***S*** *esposa y amiga* 'wife and friend'; ***R*** *muž i povelitel'* 'husband and sovereign.' This parameter is semantic and determines syntactical agreement in number: such SCPs are always singular.

Style. This is a pragmatic parameter: the speaker addresses a specific audience. We consider the following style levels: elevated (very rare:***S*** *alpha y omega*), neutral (standard in speech and texts and without any labels in dictionaries), colloquial (used by everybody for everyone; very frequent in everyday speech; given in dictionaries with the *colloq* label), and coarse colloquial (commonly used by men for men; not so rare in speech but unusual for dictionaries).

3 Stable Coordinated Pairs *vs*. Collocations

Stable word combinations are numerous in any language. They are divided into idioms of various types (semiphrasemes, complete phrasemes, quasi-phrasemes) and free combinations, depending on the correspondence between the meaning of components and the whole [4, 14]. Different authors apply the term *collocation* to all of them or only to semiphrasemes, such as lexical functions of syntagmatic type (which have a standard correspondence between the two syntactically linked parts). Collocation collections exist for Russian [4] (on-line) and English [1, 17] (paper). For numerous applications they should include not only idioms but also free combinations [6].

In SCPs the components are not syntagmatic lexical functions, though most SCPs are lexically restricted and some of them include lexical functions of paradigmatic type

(synonyms, antonyms, hyperonyms, etc.). The difference is implied by the type of involved syntactic links. In collocations, the dependency links are of subordinate type, while in coordinated pairs the upper level links are of coordinative type (cf. Section 2) and only inner links within coordinated word groups are subordinate. Hence, SCPs are different linguistic objects, poorly studied in applied linguistics.

However, the property of stability and of mutual lexical restrictions permits us to use SCP collections in the same applications as collocations (cf. Section 5.)

4 Some Statistics

At present, the largest sets of SCPs we possess is Russian (3228 entries) and Spanish (623), the Russian set being near to saturation.

The distribution of part-of-speech roles of SCPs is shown in Table 1. In both languages the overwhelming majority are nouns. The higher percentage of adverbs in Russian can be partially explained by that the predicative phrases like *It is hot and stuffy* in Russian include adverbs, not adjectives: žarko i dušno lit. 'hotly and stuffily.'

Table 1. Part of speech distribution.

Part of speech	Russian	Spanish
Nouns	71%	82%
Adjectives	14%	6%
Adverbs	10%	7%
Verbs	5%	5%

Table 2. Semantic links between components.

Semantic link	Russian	Spanish
Synonyms	10%	2%
Antonyms	25%	7%
Co-hyponyms	58%	86%
Co-participants	7%	5%

The distribution of the types of semantic links between the components is shown in Table 2. The majority are co-hyponyms, the antonyms having the second rank. In Russian, we found only 69 (2.3%) SCPs with proper names and 80 (2.7%) with rarer conjunctions: *ili* 'or,' *i*. . . *i* 'both. . . and,' *ni*. . . *ni* 'neither. . . nor,' *da* 'and'. Thus the majority of SCPs are noun groups coordinated by the standard copulative conjunction (*and*), of neural style, with co-hyponymous non-coreferential links between components, without lexical peculiarities or proper names.

5 Use of SCPs in Language Processing

5.1 Word Sense Disambiguation

Out of context, a component of a SCP can have different meanings; e.g., in Russian there are about 15% of SCPs with at least one ambiguous word. Nearly all SCPs resolve the ambiguity, selecting only one sense: e.g.,***S*** *luz y fuerza* 'electric light and power' while *fuerza* has the sense 'electric power' among many others; other examples: ***S*** *especie*$_2$ *y familia*$_5$ 'species and family (in a classification),' *apoyo*$_2$ *y asistencia*$_3$ 'support and assistance'; in ***R*** *vojna i mir* 'war and peace' the noun *mir* has two senses, *mir*$_1$ 'world' and *mir*$_2$ 'peace', and the SCP selects the second one.

For word sense disambiguation, each specific SCP in the dictionary should be labeled by senses of its components, e.g., by EuroWordNet synset labels [19].

5.2 Parsing

If the collection of SCPs contains, for each entry, its dependency subtree with all nodes supplied with the corresponding lexeme labels and morphological characteristics, the local parsing procedure is rather trivial: the parser only needs to find the sequence of words corresponding to the SCP and copy its dependency sub-tree from the dictionary to the sentence tree being constructed. For highly inflectional languages, matching includes stemming. For example, ***S*** adjectival pair *sanas y salvas* 'sound$_{\mathrm{FEM,PLUR}}$ and safe$_{\mathrm{FEM,PLUR}}$' should be reduced to *sano y salvo* 'sound$_{\mathrm{MASC,SING}}$ and safe$_{\mathrm{MASC,SING}}$' (its dictionary form).

Since the syntactical role of the SCP is known beforehand, it is necessary to search the word within a sentence that could govern this SCP. In the previous example, a word like *muchachas* 'girls,' *mujeres* 'women', *son* 'are' or the like is searched for ***S*** *sanas y salvas*. The morphological agreements should be checked at this stage. Sometimes the nodes subordinated to the SCP root are also revealed, but they do not change the inner structure of the SCP.

In many cases, this resolves morphological and lexical homonymy. For example, ***S*** *entre el cielo y la tierra* 'between the heaven and the earth' contains entre than can be a form of the verb *entrar* 'to enter', so that the sequence permits the false interpretation 'should enter the heaven and the earth'. The fact that the word chain is found in the SCP dictionary usually resolves all such ambiguities.

5.3 Detection and Correction of Semantic Errors

Semantic (real-word) errors often violate neither orthography nor grammar of a text. Consisting of correct words inappropriate in a given context or of grammatical phrases contradicting to common sense, they break text understanding.

A type of semantic (real-word) errors is malapropism, i.e., a blunder in which one word is replaced by another word existing in the given language, similar in sound but different in meaning, e.g., *travel around the word* (for *world*).

Two methods were proposed for correcting malapropisms. In [9, 10] the detection and correction of malapropisms rely on semantic text anomalies (words distant in WordNet from all other words in the context). Here the distance is based on paradigmatic relations (synonyms, hyponyms, hyperonyms) between words (mainly nouns).

In [7], we rely on syntactic and semantic text anomalies (words not forming sensible and stable word combinations—collocations—with other words in the sentence) supposing that a malapropism destroys collocation(s). E.g., *travel around the word* is coherent syntactically but is not a collocation. In [7] the dependency links between words are of subordinate type; we propose to use coordinative links in the same way.

Suppose we have an SCP collection with the ***S*** adjective SCP *sano y salvo* 'sound and safe' and find in the text a sequence *sano y saldo* 'sound and paid'. The word *saldo* can indicate malapropism since it does not form a SCP in this sentence, while the word salvo similar in sound does, according to the SCP collection. With this, the program can not only indicate the error but also propose the correction.

Similarly, given ***S*** SCP *pobres y ricos* 'the poor and the rich', the program can correct a malapropism *pobres y rizos* 'the poor and the curled.' Loosening the similarity measure to two-letter difference, we can use this method for ***S*** *único e invisible* 'unique and invisible' given the correct ***S*** *único e indivisible* 'unique and indivisible'.

This method can be generalized to real-word errors other than malapropisms, when the second coordinated component in the text is similar to the intended true word not in sound but in something else; e.g., when the initial part of a SCP coincides with that of the chain in the text, while its second part is a synonym or quasi-synonym of the second word in the text. E.g., in ***R*** *ran′še i teper′* erroneously written as *ran′še i sejčas*: synonymy dictionary would recognize *teper′* and *sejčas* as synonyms.

5.4 Cohesion Tests for Detecting Hidden Co-reference and Text Segmentation

The text *John was eating quickly, the donuts were tasty* is cohesive: the reader infers that John was eating just these donuts (food article). This is hidden co-reference, when both parts of utterance refer to the same entity, but in an indirect manner [8].

Numerous SCPs have just the same property: they contain hidden co-references. For example, ***E*** SCP *mother and child* in fact means 'mother and her child'. Other examples: *law and order* standing for 'law and the order implied by it'; *food and agriculture* standing for 'food and the agriculture that produces it'.

Thus, for the semantic representation of texts including SCPs, it is necessary to specify with each SCP its meaning, for both idioms and for pairs including coreferences. This will make the semantic representation cohesive.

There is another application of CSPs related to text cohesion. In [5] we proposed a method of automatic text segmentation into paragraphs. We measured cohesion at a running word as density of semantic and syntactic links the word is in. The paragraph breaks were set after local minimums of such cohesion function. In [5] the syntactic links were only of subordinate type; here we add coordinative links.

6 Conclusion and Future Work

We have defined the notion of a *stable coordinated pair* and described its main features. Since the inventory of SCPs in a language is limited (about 3200 for Russian) and so is the supplementary information of morphological, syntactic, semantic, and pragmatic nature necessary for applications, we propose to create the SCP dictionaries and to use them for several important applications, such as word sense disambiguation, parsing, and semantic error detection and correction.

References

1. Benson, M. *et al*: *The BBI combinatory dictionary of English*. John Benjamin, 1989.
2. Bolshakov, I. A.: Surface Syntactic Relations in Spanish. In: A. Gelbukh (Ed.): Proc. CICLing-2002, Computational Linguistics and Intelligent Text Processing. *Lecture Notes in Computer Science*, No. 2276, Springer, 2002, p. 210–219.
3. Bolshakov, I. A., Gaysinsky, A. N.: A Dictionary of Stable Coordinated Pairs in Russian. *Nauchnaya i Tekhnicheskaya Informatsiya*. Ser. 2, No. 4, 1993, p. 28–33.
4. Bolshakov, I. A., Gelbukh, A. F.: A Very Large Database of Collocations and Semantic Links. In: NLDB-2000, Natural Language Processing and Information Systems. *Lecture Notes in Computer Science* 1959, Springer, 2001, p. 103–114.
5. Bolshakov, I. A., Gelbukh, A. F.: Text Segmentation into Paragraphs Based on Local Text Cohesion. In: Matousek V. et al. (Eds.): TSD-2001: Text, Speech and Dialogue. *Lecture Notes in Artificial Intelligence*, No. 2166, Springer, 2001, p. 158–166.

6. Bolshakov, I. A., Gelbukh A. F.: Word Combinations as an Important Part of Modern Electronic Dictionaries. *Procesamiento de Lenguaje Natural* 29, 2002, p. 47–54.
7. Bolshakov, I. A., Gelbukh A. F.: On Detection of Malapropisms by Multistage Collocation Testing. In Proc. NLDB-2003, Applications of Natural Language to Information Systems. *Lecture Notes in Computer Science*, Springer, 2003, to appear.
8. Gelbukh, A., Sidorov, G., Bolshakov, I. A.: Dictionary-based Method for Coherence Maintenance in Man-Machine Dialogue with Indirect Antecedents and Ellipses. In: Sojka, P. et al. (Eds.): Proc. TSD-2000, Text, Speech and Dialogue. *Lecture Notes in Artificial Intelligence*, No. 1902, Springer, 2000, p. 357–352.
9. Hirst, G., St-Onge, D.: Lexical Chains as Representation of Context for Detection and Corrections of Malapropisms. In: Fellbaum K.(Ed.): *WordNet: An Electronic Lexical Database*. The MIT Press, 1998, p. 305–332.
10. Hirst, G., Budanitsky, A.: Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion. *Computational Linguistics* (to appear).
11. Malkiel, Y.: Studies in Irreversible Binomials. *Lingua*, v. 8, 1959, p. 113–160.
12. Mel'čuk, I.: *An Experience on the Theory of Linguistic Models Meaning–Text*. Semantics, Syntax (in Russian). Moscow, Nauka Publ., 1974.
13. Mel'čuk, I.: *Dependency Syntax: Theory and Practice*. SONY Press, NY, 1988.
14. Mel'čuk, I.: Phrasemes in Language and Phraseology in Linguistics. Everaert, M. et al (Eds.): *Structural and psychological perspectives*. Lawrence Erlbaum Associates, p. 169–252.
15. Mel'čuk, I.: "Sejčas" i "teper'" v sovremennom russkom jazyke (In Russian). In: Mel'čuk, I. A.: *The Russian language in the Meaning–Text perspective*, Wiener Slawistischer Almanach. Sonderband 39, Moskau–Wien, 1995, p. 55–76.
16. Mel'čuk, I., Pertsov, N.: *Surface Syntax of English: A Formal Model in the Meaning–Text Framework*. Amsterdam/Philadelphia: Benjamins. 1987.
17. *Oxford Collocation Dictionary for Students of English*. Oxford Univ. Press. 2003.
18. Sag, I. A., Wasow, T.: *Syntactic Theory: A Formal Introduction*. CSLI Publ., 1997.
19. Vossen, P. (Ed.): *EuroWordNet General Document*, `www.hum.uva.nl/~ewn`.