Stable Coordinate Pairs in Spanish: Statistical and Structural Description*

Igor A. Bolshakov[1] and Sofia N. Galicia-Haro[2]

[1] Center for Computing Research (CIC),
National Polytechnic Institute (IPN), Mexico City, Mexico
igor@cic.ipn.mx
[2] Faculty of Sciences,
National Autonomous University of Mexico (UNAM),
Mexico City, Mexico
sngh@fciencias.unam.mx

Abstract. Stable coordinate pairs (SCP) like *comentarios y sugerencias* 'comments and suggestions' or *sano y salvo* 'safe and sound' are rather frequent in texts in Spanish, though there are only few thousands of them in language. We characterize SCPs statistically by a numerical Stable Connection Index and reveal its unimodal distribution. We also propose lexical, morphologic, syntactic, and semantic categories for SCP structural description — for both a whole SCP and its components. It is argued that database containing a set of categorized SCPs facilitates several tasks of automatic NLP.. The research is based on a set of ca. 2200 Spanish coordinate pairs.

1 Introduction

In all European languages, coordinate constructions are rather frequent in common texts. For example, in the Internet version of a Mexican newspaper *La Jornada* a coordinated construction is on an average in each fourth sentence. We name word combination of two content words (or content word compounds) linked by a coordinative conjunction Stable Coordinate Pair (SCP), if the occurrence rates of the whole entity and its components satisfy a statistical criterion introduced below. One component in a SCP more or less predicts another. In other words, one component restricts the other both lexically and semantically: *café y té* 'coffee and tea', *guerra y paz* 'war and peace', *ida y vuelta* 'roundtrip'.

Notwithstanding frequent occurrence of SCPs, general linguistics gave them scant attention [1, 7]. Our works [4, 5] seem insufficient either.

The objective of this paper is to describe SCPs in more detail. To characterize them statistically, we propose Stable Connection Index similar to Mutual Information Index well known in statistics [8]. To categorize both whole SCPs and their components, we introduce parameters of lexical, morphologic, syntactic, semantic, and pragmatic nature. It is argued that gathering a set of fully characterized SCPs into a database facili-

* Work done under partial support of Mexican Government (CONACyT, SNI) and CGEPI-IPN, Mexico.

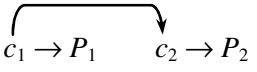
M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 489 – 497, 2005.

tates a variety of NLP applications. The research is based on ca. 2300 Spanish coordinate pairs (2165 stable ones after testing).

2 Stability of Coordinate Pairs

SCP as a whole plays the syntactic role of any major part of speech: noun, adjective, verb, or adverb. SCP occurs in a text as a contiguous segment, with its components that vary morphologically depending on the intra-sentence context. The surface syntactic structure of SCPs can be of two shapes [9]:

- In frequent cases the structure is $P_1 \rightarrow C \rightarrow P_2$ where components P_1 and P_2 are linked with unique conjunction C equal to *y/e* 'and', *o* 'or', or *pero* 'but.'
- In rarer cases the structure contains disjoint conjunctions *y* ... *y* 'both ... and', *ni* .... *ni* 'neither ... nor', *o bien* ... *o bien* 'either ... or':



During the recent years we have gathered a set of ca. 2300 Mexican coordinate pairs intuitively considered stable. Then the problem arose to formally define and to numerically test their stability, in order to filter off the scratch set. We did not take the criterion based only on frequencies of the entire would-be SCPs met in some corpus, since these frequencies depend on the corpus size S while the frequencies of the components P_i taken apart are not considered. A possible solution is to involve Mutual Information well known in statistics [8]

$$\mathrm{MI}(P_1, P_2) \equiv \log \frac{\mathrm{S} \times \mathrm{N}(P_1, P_2)}{\mathrm{N}(P_1) \times \mathrm{N}(P_2)},$$

where N() is frequency of the entity in parentheses met through the corpus. Regrettably, only a limited part of our set proved to be in the text corpus compiled by us from Mexican newspapers [6].

The Web search engines are incomparably richer, but they deliver statistics on queried words and word combinations measured in Web-pages. We can re-conceptualize N() as numbers of relevant pages, and S as the page total managed by the engine. However, now N()/S are not empirical probabilities of occurrences: the same words occurring in a page are counted only once, while the same page is counted repeatedly for each word included. Thus, MI is not now a strictly grounded statistical measure for words. Since *MI* depends on $\mathrm{N}(P_1,P_2)$ and $\mathrm{N}(P_1) \times \mathrm{N}(P_2)$, we may construe other similar criteria from the same 'building blocks.' Among those we have preferred Stable Connection Index

$$\mathrm{SCI}(P_1, P_2) \equiv 16 + \log_2 \frac{\mathrm{N}(P_1, P_2)}{\sqrt{\mathrm{N}(P_1) \times \mathrm{N}(P_2)}},$$

where the constant 16 and the logarithmic base 2 are chosen quite empirically: we tried to allocate a majority of SCI values in the interval [0...16]. To calculate SCI, we do not need to know the steadily increasing total volume S under the search engine's control. SCI reaches its maximally possible value 16 when P_1 and P_2 always go to-

gether. It retains its value when N(P_1), N(P_2), and N(P_1,P_2) change proportionally. This is important since all measured values fluctuate quasi-synchronously in time.

Computing SCI values for the available set by means of Google, we have plotted a unimodal (= single-peaked) statistical distribution with the mean value $M = 7.2$ and standard deviation $D = 2.8$ (Fig. 1). While dividing the SCP set into three groups, the lower (SCI < $M - D$), the middle ($M - D \leq$ SCI < $M + D$), and the upper one (SCI ≥ $M + D$), their relative proportions are 23:57:21.

Hereafter a coordinate pair is considered stable if the following formula is valid:

$$\text{SCI} \geq 0.$$

Taking into account the shape of the distribution and the negligible number of the pairs that did not pass the test on positivity, the threshold seems adequate.

Examples of SCPs with maximal possible SCI values are given in Table 1. One can see than they are of three types: idioms (*a diestra y siniestra* 'to the right and to the left'); usually inseparable geographic names (*América Latina y el Caribe* 'Latin America and the Caribbean') or office names (*Hacienda y Crédito Público* 'Treasury and Public Credit'); fixed everyday-life expressions, also rather idiomatic (*un antes y un después* 'somewhat before and somewhat after', *a tontas y a locas* 'without thinking or reasoning' (lit. 'to idiots and crazies'). The non-idiomatic pairs (*pequeño y mediano* 'small and medium,' *lavadoras y secadoras* 'washing machines and dryers,' *términos y condiciones* 'terms and conditions,' etc.) are rather rare within the upper group. Except for the proper names and the fixed formulas like *una de cal y otra de arena* 'changing one's mind' (lit. 'one of lime and other of sand'), these SCPs can be also used in the inverse order, but with significantly lower SCI values (cf. the figures after '/' sign in the middle column).

The most numerous middle group is illustrated by the following SCPs with SCI in the interval 7.0 to 8.0: *trabajadores y sindicatos* 'workers and trade unions,' *normas y políticas* 'norms and policies,' *casa y jardín* 'house and garden,' *previsible y evitable*

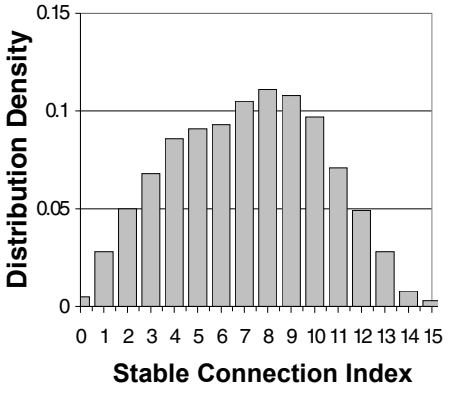


Fig. 1. Distribution of SCI values for the SCP set

Table 1. Several SCPs of the upper group

Spanish SCP	*SCI* (dir/inv)	Translation
a tontas y a locas	**16.8/7.3**	without thinking or reasoning
monedas y billetes	**16.1/11.5**	coins and currency
ayudas y subvenciones	**16.0/11.9**	aid and subventions
un antes y un después	**15.8/4.1**	somewhat before and somewhat after
pequeño y mediano	**15.3/7.7**	small and medium
una de cal y otra de arena	**15.2/-**	changing one's mind
todos para uno y uno para todos	**14.9/11.7**	all for one and one for all
en las buenas y en las malas	**14.9/9.8**	for better or worse
las mil y una noches	**14.9/1.3**	thousand and one night
a diestra y siniestra	**14.8/4.9**	hobnob
lavadoras y secadoras	**14.5/2.9**	washing machines and dryers
escuelas y universidades	**14.4/8.7**	schools and universities
bebés y niños	**14.4/9.6**	babies and children
imagen y sonido	**14.3/10.4**	image and sound
lo público y lo privado	**14.3/11.6**	public and private domains
carteles y edictos	**14.2/-**	posters and proclamations
comentarios y sugerencias	**14.1/11.3**	commentaries and suggestions
Hacienda y Crédito Publico	**14.1/-**	Treasury and Public Credit
términos y condiciones	**14.0/8.7**	terms and conditions
América Latina y el Caribe	**14.0/3.8**	Latin America and the Caribbean

'foreseeable and evitable,' *autobuses y tractores* 'buses and tractors,' *negocios y comercios* 'shops and services,' *cartón y cartoncillo* 'board and chipboard.' Nearly all of them are non-idiomatic with commonly used words as components.

SCPs with the lowest positive SCI values can be illustrated as follows: *servicio y equipo* 'service and equipment,' *noticias y foros de opinión* 'news and opinion polls,' *señores y niños* 'gentlemen and children,' *concentrados y sabores* 'concentrates and flavors,' *granito y concreto* 'granite and concrete.' Mostly, these are commonly used non-idiomatic expressions with components occurring apart much more frequently than the components of the middle group pairs.

The pairs with negative SCI values (ca. 6%) were removed from the initial set so that the total of actual set is now ca. 2200. Most of them were morphological variants of the same SCPs. For example, the pair *acción y proyecto* 'action and project' has negative SCI, while its plural *acciones y proyectos* 'actions and projects' has the SCI value 7.4.

3 External Categorization of SCPs

As a whole, SCP can be characterized by the following categories.

Part of speech is determined by its syntactic role of a SCP in a sentence: SCP can be a noun group (NG, 85% of our set), adjective group (AjG, 7%), adverb group (AvG, 5%), or verb group (VG, 3%). E.g., *embajadas y consulados* 'Embassies and Consulates' is NG, *infantil y juvenil* 'infant and juvenile' is AjG, *comer y beber* 'to eat and drink' is VG, *por arriba y por abajo* 'by above and below' is AvG. Some

prepositional groups can play the role of both AjG and AvG, e.g., *en cuerpo y alma* 'in body and soul' is AjG when modifying *hermosa* 'beautiful' or is AvG when modifying *apoyar* 'to help.' We consider these roles as separate SCPs.

Number is relevant only for NGs. Usually it is plural, independently of number of the components P_1 and P_2. Indeed, both *padre y madre* 'father and mother' and *padres y madres* 'fathers and mothers' can be substituted by *they*. However, in cases when P_1 and P_2 refer to the same person, the SCP is externally singular (cf. Sect. 4).

Sphere of usage can be subdivided as follows, without strict borders between the branches:

- Official documentation entries and mass media clichés, including the names of well known organizations: *pérdidas y ganancias* 'losses and gains'; *mayoreo y menudeo* 'wholesale and retail'; *Hacienda y Crédito Público* 'Treasury and Public Credit';
- Common business notions including the names of common shops, workshops or store departments: *frutas y verduras* 'fruits and vegetables'; *vinos y licores* 'wines and liquors';
- Everyday life clichés: *dimes y diretes* '*squabble*'; *noche y día* 'night and day';
- Sci-tech terms: *temperatura y presión* 'temperature and pressure'; *hidráulica y mecánica* 'hydraulics and mechanics'; *álgebra y geometría* 'algebra and geometry'
- Cultural and religious terms: *Sansón y Dalila* 'Samson and Delilah'; *Adán y Eva* 'Adam and Eva';
- Official geographic names: *Bosnia y Herzegovina, Trinidad y Tobago.*

External semantic correspondences of an entire SCP are usually their synonyms in the form of:

- A single word: *padre y madre* = *padres* (father and mother = parents);
- The same SCP given in the reverse order. In the examples *hospitales y clínicas* (SCI = 11.8) = *clínicas y hospitales* (SCI = 11.6) 'clinics and hospitals'; *industrial y comercial* (10.2) = *comercial e industrial* (10.3) 'industrial and commercial' SCI values are comparable, and there is no reasons to prefer any order except of a mere habit. There are also cases when the opposite order changes communicative organization of the expression: *México y el mundo* means approximately 'México and its relations with the world', whereas *el mundo y México* means 'the world and its relations with México.' Such oppositions do not seem fully synonymous, even if their SCI values are close to each other.
- A SCP with components synonymous to the corresponding components of the source SCP (one component may be the same). Such SCPs can have comparable SCI values: *colegios e institutos* (12.0) ≈ *escuelas y universidades* (14.4) 'schools and universities'; *astronomía y física del espacio* (11.7) ≈ *astronomía y ciencias del cosmo* (11.6) 'astronomy and space science'. In the case when the options differ in SCI more significantly (*ida y vuelta* (12.4) = *ida y venida* (8.4) 'go and return'; *docentes y estudiantes* (9.6) ≈ *maestros y discípulos* (6.1) 'teachers and pupils') it is recommendable to prefer more stable synonym.

Style is a pragmatic parameter for us: the speaker addresses the given expression to a specific audience. The style can be elevated (very rare: *alfa y omega* 'alpha and omega'), neutral (standard in speech and texts and without any labels in dictionaries), colloquial (used by everybody addressing everyone, very frequent in everyday speech, and given in dictionaries with the *colloq* label), and coarse colloquial (commonly used by men addressing men, not so rare in speech but rarely represented in dictionaries).

4 Internal Categorization of SCPs

Internally, SCPs can be characterized as follows.

Inflectionality. A SCP is inflectional if at least one its component changes its morphologic form depending on the syntactic governor of the SCP and/or of its semantically-induced morphologic characteristics (like tense of verb).

Noun SCPs that at the first glance have both singular and plural forms frequently do not correspond to each other as usual grammatical numbers. We consider each number as a separate SCP, e.g., *bar y cantina* 'bar and canteen' vs. *bares y cantinas* 'bars and canteens'.

Concerning the articles, the situation is different. We adopt the modern Mel'čuk's point of view [10] that the pairs *bar y cantina* and *el bar y la cantina* are grammatical forms of the same pair with variants differing in definiteness. The fact that the purely grammatical feature is represented by a separate auxiliary word is irrelevant for us. Indeed, in some other languages (e.g., Romanian, Bulgarian or Swedish) the definite article is suffixal part of the corresponding noun. So we compute SCI separately for each member of 'morphological' paradigms {*bar y cantina* 'bar and canteen,' *el bar y la cantina* 'the bar and the canteen'}, {*bares y cantinas* 'bars and canteens,' *los bares y las cantinas* 'the bars and the canteens'} and take the maximal SCP in a paradigm to characterize it as a whole. Conventionally, the paradigm may be represented by the version without articles. Note that if there occur in texts also indefinite form of a given coordinate pair like *un bar y una cantina* 'a bar and a canteen', the third member is added to the paradigm proposed, and the maximum is searched among the three variants.

Spanish adjectives change in gender (masculine and feminine) and in number (singular and plural), having totally four combinations. We compute SCI values for each member of the morphologic paradigm, e.g., {*activo y saludable, activos y saludables, activa y saludable, activas y saludables*} ('active and healthful') and then take the maximal value to characterize the whole. By usual convention, the {masculine, singular} form is taken as dictionary representation of the whole paradigm.

Hence we initially had ca. 5400 various forms of coordinate pairs, and after evaluations and unifications the total has reduced to ca. 2300 SCPs.

Semantic link between components can be of the following types:

- Synonyms, quasi-synonyms, and mere repetitions (2% in our set): *presidente y director* 'president and director'; *cine y artes audiovisuales* 'movies and audiovisual arts'; *más y más* 'more and more';

- Co-hyponyms in an unspecified genus–species hierarchy (86%): *maestría y doctorado* 'magister and doctorate degrees'; *axiomas y teoremas* 'axioms and theorems'; *ginecología y obstetricia* 'gynecology and obstetrics.' The degree of the meaning intersection between quasi-synonyms or co-hyponyms is rather vague.
- Antonyms, quasi-antonyms, conversives, and opposite notions (7%): *material y espiritual* 'material and spiritual'; *más o menos* 'more or less'; *a dios y al diablo* 'to God and to devil'; *compra y venta* 'purchase and sale'; *frío y caliente* 'cold and hot'.
- Co-participants or actions in a situation (5%): *gerencia y presupuesto* 'management and budget'; *productos y servicios* 'products and services'.

The latter type is the most complicated semantically. Some subtypes of the situation are as follows:

- In *muerto y enterrado* 'died and buried' there is a time sequence of actions, with the time of P_1 preceding that of P_2.
- In *fabricación y venta* 'manufacturing and sale', *crimen y castigo* 'crime and punishment', *arbitraje y mediación* 'arbitration and mediation', there is a material cause-consequence link: fmanufacturring brings about a product to sell, crime leads to official punishment, and arbitration entails mediation.

Idiomaticity. A SCP is called idiom if its meaning is not just a sum of its components' meanings. Idioms whose meaning does not contain meanings of any of their component are complete phrasemes [10], e.g., *una de cal y otra de arena* 'changing one's mind' (lit. 'one of lime and other of sand'); *ni con melón ni con sandía* 'neither pro nor contra' (lit. 'neither with melon nor with watermelon'). The majority of SCPs are non idiomatic.

Irreversibility of SCP components could be induced by a temporal or causative sequence mentioned above. However many SCPs are reversible, maybe with change of SCI (cf. Sections 2 and 3).

Lexical peculiarity means that at least one component is not used separately. For example, in *toma y daca* 'give and take' word *daca* is peculiar (compare with the word *fro* in *to and fro*).

Coreferentiality. In very rare cases, P_1 and P_2 co-refer to the same person: *padre y esposo* 'father and husband', *madre y amiga* 'mother and wife'. This parameter determines morpho-syntactic agreement in number: such NGs are considered singular.

5 Stable Coordinate Pairs in Natural Language Processing

If we supply each SCP of the available set with all parameters introduced above, including the corresponding syntactic subtree, semantic interpretation (for idioms), and SCI value, the resulting SCP dictionary becomes a database very useful for various applications. Let us give their synopsis.

Referencing in text editing is needed while preparing a new text or editing an already existing one. Indeed, even a native speaker can feel uneasiness while selecting a

convenient expression for a given case, SCPs being among such expressions. The appropriate SCP could be found by means of any its component or a one-word synonym of the whole SCP (if any).

Learning foreign language is greatly facilitated with the SCP database. A student should know that *ida y vuelta* is more preferable than *ida y venida* (both are 'round-trip') and *docentes y estudiantes* is much more preferable than *maestros y discípulos* (both are 'teachers and pupils').

Word sense disambiguation. Out of context, a component of a SCP can have different meanings. In our set about 20% of SCPs contain at least one ambiguous word. Nearly all SCPs resolve this ambiguity, selecting only one sense. E..g., in *centros y departamentos* 'centers and departments', the noun *center* has at least two senses: 'midpoint' and 'institution', and the SCP selects the second one; in *pacientes y familiares* 'patients and relatives', the noun *pacientes* ('sick person' or 'object of an action') resolves to the first sense. We suppose that all SCP components in the database are labeled by their sense numbers.

Parsing. Since the DB with SCPs contains their partial parses, the parsing of the embedding sentence is facilitated: the parser finds the sequence of words corresponding to the SCP and copies its dependency subtree from the DB to the dependency tree of the sentence under parsing. For Spanish this operation includes lemmatization. E.g., the textual expression *sanas y salvas* 'safe and sound'$_{\text{FEM,PL}}$ should be reduced to the standard dictionary form *sano y salvo* labeled with FEM,PL. In many cases, the subtree substitution resolves morphological and lexical homonymy. For example, *entre el cielo y la tierra* 'between the heaven and the earth' contains *entre* that can be a form of the verb *entrar* 'enter', so that the sequence permits the false interpretation 'should enter the heaven and the earth'. The finding of the word chain in the SCP dictionary resolves such ambiguities at once.

Detecting and correcting malapropisms. Malapropisms are semantic errors replacing one content word by another, similar in sound but different in meaning. Syntactic links of a malapropos word with its contextual words often remain the same. In [3] semantic text anomalies are detected by noting that malapropisms, as a rule, destroy the collocation(s) that the mutilated word would be in. We can apply the same idea to SCPs. Suppose that the program of malapropism detection and correction finds in a text the syntactically correct coordinate pair *vivito y boleando* 'alive and shoe shinning' with ultimately negative SCI value. By few editing operations on both components, a special subprogram finds the unique similar SCP *vivito y coleando* 'alive and kicking,' thus indicating both the error and its possible correction.

Linguistic steganography is automatic concealment of digital information in rather long orthographically and semantically correct texts. In [2] a steganographic algorithm replaces words by their synonyms, taking into account the context. However, only few SCPs do have synonyms, while the rest permits synonymous paraphrases of neither the whole pair nor its components. This knowledge is quite important for steganography.

6 Conclusion

A convenient numerical measure of stability—Stable Connection Index—is proposed for coordinate pairs and on this ground the notion of a stable coordinate pair is introduced. Various lexical, morphologic, syntactic, semantic, and pragmatic features are proposed, for both entire SCPs and their components. So far, as many as 2200 Spanish SCPs passed the test on positive SCI. Supplied with all categorial information proposed, the set of SCPs forms a useful database. Such DB facilitates several modern applications of NLP. All our examples and calculations were done for Spanish, but our earlier work [5] shows that all our classifications description are applicable also to some other European languages.

References

[1] Bloomfield, L. *Language*. Halt, Rinehart and Winston, 1964.
[2] Bolshakov, I.A. A Method of Linguistic Steganography Based on Collocation-Verified Synonymy. In: J. Fridrich (Ed.) *Information Hiding* (IH 2004), Revised Selected Papers. Lecture Notes in Computer Science, N 3200, Springer, 2004, p. 180–191.
[3] Bolshakov, I.A. An Experiment in Detection and Correction of Malapropisms through the Web. In: A. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing*. (CICLing-2005). Lecture Notes in Computer Science, N 3406, Springer, 2005, p. 803–825.
[4] Bolshakov, I.A., A.N. Gaysinski. Slovar' ustojčivyx sočinennyx par v russkom jazyke (in Russian). *Nauchnaya i Tekhnicheskaya Informatsiya*. Ser. 2, No. 4, 1993, p. 28–33.
[5] Bolshakov, I.A., A. Gelbukh, S.N. Galicia-Haro. Stable Coordinated Pairs in Text Processing. In: V. Matoušek, P. Mautner (Eds.) *Text, Speech and Dialogue* (TSD 2003). Lecture Notes in Artificial Intelligence N 2807, Springer, 2003, p. 27–34.
[6] Galicia-Haro, S. N. Using Electronic Texts for an Annotated Corpus Building. 4th *Mexican International Conference on Computer Science* (ENC-2003), 2003, p. 26–33.
[7] Malkiel, Y. Studies in Irreversible Binomials. *Lingua*, v. 8, 1959, p. 113–160.
[8] Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
[9] Mel'čuk, I. Dependency Syntax: Theory and Practice. SUNY Press, NY, 1988.
[10] Mel'čuk, I. Phrasemes in Language and Phraseology in Linguistics. In: M. Everaert *et al.* (Eds.) *Structural and Psychological Perspectives*. Hillsdale, NJ / Hove, UK: Lawrence Erlbaum Associates Publ., p. 169–252.