Using Selectional Preferences for Extending a Synonymous Paraphrasing Method in Steganography*

Hiram Calvo and Igor A. Bolshakov

Center for Computing Research, CIC
National Polytechnic Institute, IPN
Mexico City, Mexico
hcalvo@sagitario.cic.ipn.mx; igor@cic.ipn.mx

Abstract. Linguistic steganography allows hiding information in a text. The resulting text must be grammatically correct and semantically coherent to be unsuspicious. Among several methods of linguistic steganography we adhere to previous approaches which use synonymous paraphrasing, i.e., substituting content words by their equivalents. Context must be considered to avoid possible substitutions that break coherence (for example *spicy dog* instead of *hot dog*). We base our method on previous work in linguistic steganography that uses collocations for verifying context. We propose using selectional preferences instead of collocations because selectional preferences can be collected automatically from large corpora in a reliable manner, thus allowing our method to be applied for any language. The steganographic algorithm is informally outlined and an example of hiding binary information in a Spanish text fragment is presented, with a rough evaluation of the ratio of hidden information size to the necessary size of the original text.

1 Introduction

Linguistic steganography is a technique for hiding information in a text based on linguistic knowledge. In the approach of synonymous paraphrasing, information is coded by changing an existing text. The changes are determined by the information to be hidden; for example, a specific synonym for a word is selected to encode a bit 0, whereas other synonym is selected to encode a bit 1. However, not all synonyms can be used coherently in all situations. Context must be considered.

This work is based on previous work [3, 4] which uses a manually collected collocations database. Nevertheless, collecting collocations manually is a task that may take many years. For example, a Russian collocations database [2] has taken more than 14 years to be completed. On the other hand, using the Internet for verifying collocations, as in [7], is not adequate for split collocations, e.g. *make an* awful *mistake*, because current web search engines do not allow such searches—the closest search tool is the NEAR operator, which is not precise since it is not restricted to a single sentence.

* Work done under partial support of Mexican Government (CONACyT, SNI, PIFI-IPN, CGEPI-IPN) and RITOS-2. Many thanks to Prof. A. Gelbukh for critical remarks.

In this paper we propose using selectional preferences instead of collocations for the same purpose—to inconspicuously change synonyms along the text in full correspondence with the information to be hidden. In Sections 2 through 4 we present the framework and the considerations for our work; in Section 5 we outline the algorithm. Just as in previous papers on synonymous paraphrasing, our method does not depend on language, but its implementation depends heavily on available language-specific resources. This paper is aimed mainly to Spanish, with a specific steganographic example presented in Section 0. Some English examples are used only for transitory illustrations.

2 Some Definitions

Linguistic steganography is a set of methods and techniques that allow hiding information in a text based on linguistic knowledge. To be effective, the resulting text must have grammatical correctness and semantic cohesion.

There are two main approaches for achieving this: 1) generating text and 2) changing previously written text. To illustrate the first approach, consider a sentence model such as verb-preposition-noun. This model can generate valid sentences such as *go to bed, sing a song*, etc. A non-trivial problem arises when trying to generate a coherent text using these sentences: *John goes to bed, and then John sings a song*. In this way, a non-coherent text is not free of suspicions. As Chapman *et al.* [10] pointed out, the same happens when using more elaborated sentence models extracted from previously written text.

In the second approach, for hiding information some words in the source text are replaced by other words depending on the bit sequence to be hidden. These changes are detectable only at the intended receiver's side. In the best cases, the resulting text keeps the meaning of the source text.

As in works [3, 4, 7], we adhere to the second approach because it is far more realistic; indeed, generating text from scratch needs not only syntactic or semantic information; it also needs pragmatic information.

In this work we do not consider other methods of textual steganography such as text formatting, varying space widths, or other non-linguistic encoding methods. This is because genuine linguistic methods allow a message to be transmitted independently from the medium—linguistic steganography allows transmitting a message over the internet, using the telephone, by a radio broadcast, etc.

To put it in other words, we continue developing the method of linguistic steganography that replaces textual words by their synonyms [3]. This work allows to keep information concealed in unsuspicious texts along with linguistic correctness and meaning of the source text. In addition, we take advantage of existing resources to extend coverage of this method to virtually any language—whenever the required resources exist for such language. This is done by considering an alternative source for the context of words.

3 The Context of a Word

The context of a given word is the words that surround it within a sentence. In many papers just the surrounding words are considered forming collocations with the given word. Other authors consider collocations as *a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components*, as defined by Choueka [11]. Examples that fall within this definition, including idioms, are *hot dog*, *white wine* (actually white wine is yellow) [13], *to kick the bucket* and *to be a piece of cake*.

Furthermore, current usage of the term *collocation* includes also combinations of words which combine their original meaning such as *strong coffee*, but are considered collocations because substituting any of their components by equivalent words yields an understandable but strange-sounding combination, such as *powerful coffee*, *strong rain* (instead of *heavy rain*), or *to do a mistake* (instead of *making a mistake*). Moreover, collocations are not necessarily adjacent words, e. g. *making an* awful *mistake* and may involve subcategorization issues, as we will illustrate below.

The links between components of collocations are syntagmatic. These are, for example, the link between a verb and a noun filling its valence (*made* → *of stone*), or the link between a noun and its adjective modifier (*black* ← *stone*). This kind of relations can be clearly seen in a dependency representation. The **context of a word** is given then by its dependency relations. For example, see Figure 1 for the sentence *Mary read us a fairy tale*.

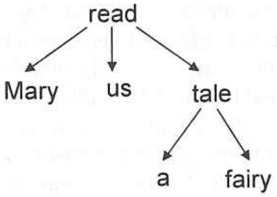

Fig. 1. Simplified dependency representation tree for *Mary read us a fairy tale*

To illustrate the influence of word's context in a sentence, we will substitute some words by their equivalents (i.e. synonyms). For example, among synonyms of *fairy* there are *pixie* and *nymph*. Substituting *fairy* by these equivalents yields, however, a very strange sounding sentence—*Mary read us a pixie tale* or *Mary read us a nymph tale*,—which are possible, but sound odd, since *fairy* depends strongly on *tale*. Another example is substituting *tale* for *yarn*—forgetting about *fairy* for a moment. In this case it sounds odd to say *read us a yarn*, and it would sound much more natural *spin us a yarn*—not considering *spin us a fairy yarn*! This shows the strength between the verb (*read*, *spin*) an one of its arguments (*tale* and *yarn*, respectively).

Subcategorization has an important role when considering collocations. For example, consider the synonyms for *to tell*: *to relate*, *to spin*, and *to say*. If one wanted to change *read* for one of its synonyms, context and structure must be considered to keep the same meaning—and natural sounding—of the sentence. Simply changing *read* for *related* would yield **Mary related us a fairy tale;* for this to be a natural sentence, a different structure (i.e. subcategorization frame) should be used: *Mary related a fairy tale to us.*

In contrast to this latter example, in this work we focus only on synonyms that keep the structure and word order of a sentence, as well as the number of words—counting stable multiwords such as *hot dog* as one unit. We use word combinations to verify that the synonymous paraphrasing results in a natural and coherent text.

4 Sources for Verifying Word Combinations

Our goal is to make synonym paraphrasing considering context. In [3, 4] a previously collected collocation DB was used. Currently only few electronic databases of collocations are readily available. To our knowledge, publicly available electronic databases of collocations did not exist until 1997, when the Advanced Reader's Collocation Searcher (ARCS) for English emerged [0], but it is now inferior to the Oxford Collocation Dictionary [15] in all aspects.

The only project in the recent decade to develop a very large collocation DB available now for local use was dedicated to Russian and it produced an interactive system called CrossLexica [2, 5, 6]. Its core part is a large database of Russian collocations, but it contains also something like a Russian WordNet. Particularly, the WordNet-like part contains a synonymy dictionary and a hyponymy/hyperonymy hierarchy.

A manually collected DB of collocations cannot list every possible pair of words, especially free word combinations such as *big boy*, *walk in the street*, etc. There are several methods for extracting collocations automatically described in [13] and [16]. However, the quality of these automatically obtained collocations is not as good as the quality of those obtained by hand. In addition, as we showed in Section 0, the context of a word is strongly related with the structure of the sentence. Hence, we need linguistic knowledge in addition to purely statistic methods.

Besides that, polysemous words have several synonyms that cannot substitute the original word without changing the meaning of a text, because these are synonyms for other senses of the word. For example *plant* can be substituted by *vegetable* or by *factory*, depending greatly on context.

So far we have identified the following requirements for automatic determination of possible combinations of words: a corpus to learn from, semantic knowledge, and structure of the sentence. The linguistic knowledge that covers semantics and allows determining the structure of the sentence is a set of **selectional preferences**.

4.1. Selectional Preferences for Synonym Paraphrasing

Selectional preferences measure the degree in which a verb *prefers* an argument—a subject, an object or a circumstantial modifier. Selectional preferences principle can be applied also to adjective–noun relations, verb–adverb relations and prepositional phrases, yielding a database of *preferences* that can be regarded as graded collocations with the aid of semantic generalizations. For example, if *harvest plants* appears in a training corpus, and we know that a *harvest* prefer arguments of the kind *flora*, then we can restrict the synonyms of *plant* to only those related with *flora*, excluding those related with the manufacturing process.

In addition, selectional preferences can be an aid to determine the structure of sentences. For example, the syntactic structure of *I see the cat with a telescope* is disambiguated considering that *see with* {*an instrument*} is more frequent than a *cat with a* {*instrument*}. Calvo and Gelbukh present in [9] a method for PP attachment disambiguation, and show in [8] how this information can be used for restricting the sense of a word.

For this work, we use a selectional preferences database based on a corpus of four years of Mexican newspapers with 161 million words as in [9].

5 The Algorithm

The proposed steganographic algorithm has two inputs:

- The information to hide, in the shape of a bit sequence.
- The source text in natural language of the minimal length evaluated as approximately 500 times greater than of the information to hide. The text format can be arbitrary, but the text proper should be orthographically correct, to lessen the probability of unintended corrections during transmission. The corrections can change the number of synonymous words in the text or the conditions for their verification and thus can desynchronize the steganography *vs.* steganalysis. The text should not be semantically specific, i.e. not to be a mere list of names or sequence of numbers. In this respect newswire flow or political articles are quite acceptable. Any long fragments of inappropriate type increase the total length required for steganographic use.

The steps of the algorithm are the following:

A1. Syntax Analysis. The text is tagged using the TnT tagger trained with the Spanish corpus LEXESP [17]. This is reported to be 94% accurate for Spanish [14]. Then the text is lemmatized trying morphological variants against a dictionary [12]. Syntactic structure is extracted with the aid of selectional preferences, using the disambiguation method presented in [9].

A2. Identification of word combinations that can be paraphrased. The following patterns are extracted for each sentence; subordinate clauses are treated as separated sentences, so that there is only one verb per sentence:

1. noun — verb
2. verb — noun
3. noun — preposition — noun
4. verb — preposition — noun

The symbol "—" stands for a dependency link that does not presuppose adjacency of the linked words within the sentence. All other words between those linked—adverbs, adjectives, articles, etc.—are discarded on this step.

A3. Evaluation of synonyms using selectional preferences. Synonyms are generated for each word—except for prepositions—in the pattern. Then, different combinations are tested against a previously acquired selectional preferences database—details for extracting this database were described in Section 0. This database yields a *score* for a given combination. This score is calculated by using a mutual information formula—*freq*(*w1*,*w2*) / [*freq*(*w1*) +*freq*(*w2*) + *freq*(*w1*,*w2*)]. Different formulae for calculating mutual information are presented in [13]. If the score of a combination is greater than a threshold, the combination is listed as a possible substitution. Some patterns may have more than one possible substitution. Each one of them is listed in a particular order, e.g. starting from the one with higher value in the selectional preferences database, to the one closer to the threshold. The original construction is ranked also, using the same selectional preference database.

A4. Enciphering. Each bit of the information to be encoded decides which synonym paraphrasing will be done. As for some patterns there are several options for substitution, each paraphrasing may represent more than one bit. For example, given four possible substitutions, it is possible to represent four combinations of two bits, namely 00,01,10 and 11.

A5. Reagreement. If there are any substitutions that require simple syntactic structure changes, these are done at this stage. For example, in Spanish *historia* 'story' can be substituted by *cuento* 'tale,' but *historia* is feminine and *cuento* es masculine. So it is necessary to change the article *la* 'the$_{\text{fem}}$' for *el* 'the$_{\text{masc}}$,' resulting *el cuento* and avoiding **la cuento*.

At the receiver side, it is necessary to decode the hidden information. It is the task of a specific decoder (steganalyzer). It possesses the same resources as the encoder: the selectional preference database and the tagging module. The syntactic structure of the text is obtained as in A1; the patterns are extracted as in A2. The synonym paraphrases are ranked as in A3; bits are extracted by mapping each possible combination in the same way as in A3 and A4. Reagreement does not represent a problem while decoding because articles and other words were discarded as in A1.

6 Application Example

To illustrate the algorithm presented in the previous section, we will apply our method to hide a small amount of information in a fragment of a text in Spanish extracted from a local newspaper (La Jornada)—see Figure 2. The translation of this fragment is:

> 'Sheltered in the Madison Square Garden to protect themselves against "terrorist" and complainer threats, the republicans began their celebration with self-congratulations on how they faced September 11. Indeed, when NY was selected to celebrate the convention, the idea was to return under the shadow of the Twin Towers with G. W. Bush as commander in chief of Iraq, Afghanistan, and heading the great struggle of good against "axes of evil". But the reality has obliged to change the angle of the program...'

For this example, we have listed several possible synonyms for words as listed by a dictionary [12]. Not every substitution is verifiable, since our selectional preferences

$\left\{\begin{array}{l}\text{Atrincherados}\\ \text{Resguardados}\\ \text{Guarecidos}\end{array}\right\}$ en el Madison Square Garden para $\left\{\begin{array}{l}\text{asegurarse}\\ \text{consolidarse}\end{array}\right\}$ contra amenazas

"terroristas" y de manifestantes, los republicanos $\left\{\begin{array}{l}\text{iniciaron}\\ \text{comenzaron}\\ \text{emprendieron}\\ \text{* originaron}\end{array}\right\}$ su festejo con auto

elogios de cómo $\left\{\begin{array}{l}\text{encararon}\\ \text{enfrentaron}\\ \text{* desafiaron}\\ \text{* retaron}\end{array}\right\}$ el 11 de septiembre. De hecho, cuando se $\left\{\begin{array}{l}\text{seleccionó}\\ \text{eligió}\end{array}\right\}$

a NY para $\left\{\begin{array}{l}\text{celebrar}\\ \text{hacer}\\ \text{realizar}\\ \text{festejar}\end{array}\right\}$ la convención, la $\left\{\begin{array}{l}\text{idea}\\ \text{* concepto}\\ \text{proyecto}\\ \text{* creencia}\end{array}\right\}$ era regresar bajo la $\left\{\begin{array}{l}\text{sombra}\\ \text{* silueta}\\ \text{* opacidad}\end{array}\right\}$ de

las Torres Gemelas con G. W. Bush como comandante en $\left\{\begin{array}{l}\text{jefe}\\ \text{* líder}\\ \text{* patrón}\end{array}\right\}$ en Irak, Afganistán y $\left\{\begin{array}{l}\text{encabezando}\\ \text{* iniciando}\\ \text{* empezando}\\ \text{conduciendo}\end{array}\right\}$ la gran $\left\{\begin{array}{l}\text{lucha}\\ \text{* torneo}\\ \text{* riña}\\ \text{rivalidad}\end{array}\right\}$ del $\left\{\begin{array}{l}\text{bien}\\ \text{* patrimonio}\\ \text{* fortuna}\\ \text{* sí}\end{array}\right\}$ contra los "ejes del mal".

Pero la realidad ha obligado a cambiar el $\left\{\begin{array}{l}\text{tono}\\ \text{* fuerza}\\ \text{aire}\end{array}\right\}$ del programa...

Fig. 2. Text with synonyms for paraphrasing—bad substitutions are marked with *

Table 1. Verified combinations and their score (s) for example in Figure 2

word combination	s	translation
atrincherar en Madison	0	entrenched in M.
resguardar en Madison	0	protected in M.
asegurar contra amenaza	**3**	secured against threat
consolidar contra amenaza	0.2	consolidate against threat
iniciar festejo	**0.7**	begin party
comenzar festejo	**0.8**	start party
emprender festejo	0.4	undertake party
originar festejo	0	originate party
encarar 11	0	face 11
retar 11	0	threaten 11
seleccionar a NY	**1.3**	select NY
elegir a NY	**1.2**	choose NY
celebrar convención	**1.9**	celebrate convention
hacer convención	**1.8**	make convention
realizar convención	**1.6**	do a convention
festejar convención	**0.5**	celebrate convention
idea ser	**0.7**	idea is
proyecto ser	**0.6**	project is
concepto ser	0.4	concept is
creencia ser	0.31	belief is
sombra de torre	0	shadow of tower
silueta de torre	0	silhouette of tower
opacidad de torre	0	opacity of tower
comandante en jefe	**6.7**	commander in chief
comandante en líder	0.45	commander in leader
comandante en patrón	0.4	commander in employer
líder en Irak	0	leader in Iraq
patrón en Irak	0	employer in Iraq
encabezar lucha	**2.1**	head struggle
iniciar lucha	**1.75**	start struggle
conducir lucha	**0.8**	conduce struggle
encabezar rivalidad	**0.67**	head rivalry
iniciar torneo	0.47	begin tournament
empezar lucha	0.4	begin struggle
empezar torneo	0.3	begin tournament
encabezar torneo	0.28	head tournament
conducir rivalidad	0.08	lead rivalry
conducir torneo	0.03	lead tournament
lucha de bien	0.7	struggle of good
lucha de patrimonio	0	struggle of wealth
lucha de fortuna	0	struggle of fortune
lucha de sí	0	struggle of ok
aire de programa	**0.66**	tone of program
tono de programa	**0.54**	angle of program
fuerza de programa	0.23	strength of prog.

database does not contain every possible instance. That is the case of combinations including *Madison Square Garden*, for example. The next combinations: *asegurar(se) contra amenazas* 'insure (themselves) against threats' vs. *consolidar(se) contra amenazas* 'consolidate (themselves) against threats' can be verified in our selectional preferences database: the first one yields a score of 3 against the second one which has a score of 0.2. If we set our threshold around 0.5, the second option will be discarded.

Table 1 shows additional combinations of nouns verified by the selectional preferences database. Entries not contained in the selectional preferences database are marked with 0.

In Table 1, combinations which are above the threshold (0.5) are shown in bold. Alternative possibilities which allow the representation of one bit are marked in light-grey; those which can represent two bits are marked with dark grey.

This fragment can hide 8 bits (i.e. 1 byte) of information. The text has around 500 bytes; hence the ratio of the hidden information size and the size of original text (steganographic bandwidth) equals approximately .002. This means that the text should be 500 times longer than the hidden information.

7 Conclusions

As in previous works, the proposed method of linguistic steganography conserves the meaning of the carrier text, as well as its inconspicuousness. The main advantage of our method is that it does not require a manually collected database of collocations. Instead, a large database of automatically extracted selectional preferences is used. Since the method presented in this paper is based on automatically acquired resources, it is possible to extend its application to many languages. In this paper we have shown an example of application to Spanish.

On the other hand, the mean value .002 of steganographic bandwidth reached with the local synonymous paraphrasing seems rather low. An example of the maximum synonym paraphrasing that can be reached may be found in [3]. It argues that starting from the samples of synonymous paraphrasing by I. Mel'cuk, the maximum bandwidth of the paraphrasing method in steganography can reach approximately 0.016. This is done by considering synonym paraphrasing for complete phrases, such as *ayudar* 'to help' ↔ *dar ayuda* 'to give help'. The automatic compiling of immense lists of paraphrases for each sentence seems a problem to be resolved in a remote future.

Our method reaches 12.5% of the maximum possible level, without considering adjective variants. The reachable value of synonym paraphrasing bandwidth evidently depends on the saturation of linguistic resources. Hence, they should be developed further, without any clear-cut limits of perfection. In particular the bandwidth attained with our method can be improved by considering variants of adjectives. This is part of our future work.

As to our algorithm, we can hardly consider it faultless. The following issues seem now especially acute:

- Large chains of word combinations, such as *encabezando la lucha del bien contra los "ejes del mal"* 'heading the great struggle of good against "axes of evil"' can lead to wrong selections of synonyms, since each combination is considered only by pairs ignoring every combination as a whole.
- A large database of named entities is needed to be able to recognize phrases such as *el 11 de septiembre* 'september 11' or *Madison Square Garden.* Particulary, using the selectional preferences model can help because knowing that *Madison Square Garden* is a place helps to evaluate combinations such as *Atrincherados / resguardados / guarecidos en el Madison Square Garden* ''Entrenched / protected / sheltered / in the Madison Square Garden.'
- Threshold adjustments should be made automatically.

All of these problems are to be investigated in depth in our future work.

References

1. Bogatz, H. *The Advanced Reader's Collocation Searcher* (ARCS). http://www.elda.fr/catalogue/en/text/M0013.html
2. Bolshakov, I.A. Getting One's First Million... Collocations. In: A. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing.* Proc. 5th Int. Conf. CICLing-2004, LNCS 2945, Springer, 2004, p:229–242.

3. Bolshakov, I.A. A Method of Linguistic Steganography Based on Collocation-Proven Synonymy. In: Proceedings of Int. Information Hiding Workshop IH2004, Toronto, Canada, May 2004. LNCS, Springer, 2004 (to apprear).
4. Bolshakov, I.A. Two methods of synonymous paraphrasing in linguistic steganography (in Russian, the abstract in English). Proc. Intern. Conf. Dialogue'2004, Verhnevolzhskij, Russia, June 2004, p. 62-67.
5. Bolshakov, I.A., A. Gelbukh. A Very Large Database of Collocations and Semantic Links. In: M. Bouzeghoub *et al.* (Eds.) *Natural Language Processing and Information Systems.* Proc. Int. Conf. on Applications of Natural Language to Information Systems NLDB-2000, LNCS 1959, Springer, 2001, p:103–114.
6. Bolshakov, I.A., A. Gelbukh. Heuristics-Based Replenishment of Collocation Databases. In: E. Ranchhold, N.J. Mamede (Eds.) *Advances in Natural Language Processing.* Proc. Int. Conf. PorTAL 2002, Faro, Portugal. LNAI 2389, Springer, 2002, p:25–32.
7. Bolshakov, I.A., A. Gelbukh. Synonymous Paraphrasing Using WordNet and Internet. In: F. Meziane, E. Métais (Eds.) Proc. 9th International Conference on Application of Natural Language to Information Systems NLDB-2004, LNCS 3136, Springer, 2004.
8. Calvo, H., A. Gelbukh. Acquiring Ontology-Linked Selectional Preferences from Unannotated Text. In: *Progress in Pattern Recognition, Speech and Image Analysis*, CIARP'2004. LNCS, Springer, 2004 (to appear)
9. Calvo, H., A. Gelbukh. Acquiring Selectional Preferences from Untagged Text for Prepositional Phrase Attachment Disambiguation. In: *Natural Language Processing and Information Systems*, NLDB 2004, LNCS 3136, Springer, 2004.
10. Chapman, M., G.I. Davida, M. Rennhard. A Practical and Effective Approach to Large-Scale Automated Linguistic Steganography. In: G.I. Davida, Y. Frankel (Eds.) *Information security.* Proc. of Int. Conf. on Information and Communication Security ICS 2001. LNCS 2200, Springer, 2001, p. 156–165.
11. Choueka, Y. Looking for needles in a haystack or locating interesting collocational expressions in large textual database. In Proc. *Conf. User-Oriented Content-Based Text and Image Handling* (RIAO'88), 1988, p. 609–623.
12. Lázaro Carreter, F. (Ed.) *Diccionario Anaya de la Lengua*, Vox, 1991.
13. Manning, C., H. Schutze. *Foundations of Statistical Natual Language Processing.* The MIT Press, Cambridge, MA, 1999.
14. Morales-Carrasco, R., A. Gelbukh: Evaluation of the TnT Tagger for Spanish. In Proc. *Fourth Mexican International Conference on Computer Science*, Tlaxcala, Mexico, 2003.
15. *Oxford Collocations Dictionary for Students of English.* Oxford University Press, 2003.
16. Pearce, Darren. A Comparative Evaluation of Collocation Extraction Techniques. In: Proc. *Third International Conference on Language Resources and Evaluation,* Las Palmas, Canary Islands, Spain, 2002.
17. Sebastián, N., M. A. Martí, M. F. Carreiras, and F. Cuestos. Lexesp, léxico informatizado del español, *Edicions de la Universitat de Barcelona*, 2000.