Spanish Adverbial Frozen Expressions

Dolors Català
Autonomous University of Barcelona
Campus Sabadell, 08202, Spain
*f*LEXSEM
dolors.catala@uab.cat

Jorge Baptista
Univ.Algarve, Campus de Gambelas
P-8005-139 Faro, Portugal
L²F – INESC-ID Lisboa, Portugal
jbaptis@ualg.pt

Abstract

This paper presents an electronic dictionary of Spanish adverbial frozen expressions. It focuses on their formal description in view of natural language processing and presents an experiment on the automatic application of this data to real texts using finite-state techniques. The paper makes an assessment of the advantages and limitations of this method for the identification of these multiword units in texts.

1 Introduction

We have undertaken the construction of an electronic dictionary of compound adverbs, or adverbial frozen expressions (Català 2003). This dictionary completes the DELACs, i.e., the dictionary of compound words of Spanish (Blanco and Català (1998)).

These adverbial frozen expressions (*a tontas y a locas = by fits and starts, como anillo al dedo = like a glove; a ojo de buen cubero = at a guess*) [1] have often been considered as exceptions but they constitute an important part of the lexicon.

Their formal description highlights many problems for NLP applications. On the one hand, they are multiword expressions functioning as meaning units, so they have to be recognized as a block and are not to be analyzed as a free sequence of simple words. On the other hand, they present, sometimes, some lexical variation that can take complex lexical syntactical patterns.

[1] Approximate translations of examples do not intend to be fully acceptable, but to illustrate syntactic phenomena.

For example, some adverbs show combinatorial constraints between discontinuous elements:

*día sí, día no / año si, año no,*día si, año no*
'on even days/years'.

Others yet present long distance dependencies:

[***Yo** estudio*] *con todas **mis/*sus** fuerzas*
'(I study) with all my/his strength');

Lexical variation of the compound elements is often constraint in an unpredictable way:

[*Juan aprobó*] *por los/*todos los/*sus/*unos pelos*

'(John passed the exam) with difficulties'

Some allow for a theoretically infinite paradigm as in the expression <Card> *veces seguidas* '<number> of times in a row', where *Card* stands for a numeral, whose meaning is compositional but whose form is fixed:

[*Eso sucedió*] *Card veces seguidas*
'(It happened) <number> of times in a row'

since the adjective does not allow for any variation:

*[*Eso sucedió*] *Card veces continuas*
'(It happened) <number> of times in a row'

In some cases, the adjective can not be reduced:

[*Juan dijo esto*] *en voz baja* / **en voz*
'(John said this) in low voice/in voice'

nor can it be placed before the noun:

[*Juan dijo esto*] *en voz baja* / **en baja voz*
'(John said this) in voice low /in low voice'

2 The Dictionary

The theoretical and methodological framework adopted is the lexicon-grammar based on the principles of the transformational grammar of Harris (1976, 1997) developed by Maurice Gross

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 33–40,
Prague, June 2007. ©2007 Association for Computational Linguistics

(1986). In this perspective, the adverbial frozen expressions are formalized in the frame of simple sentences and their network of paraphrastic relations. Adverbs are predicates that necessarily apply on other predicates and have a basic influence in their selection. For example, some adverbs are only associated with a limited number of verbs[2]:

[*Juan duerme*/*pernocta*/*pasa la noche*] *al raso*
'(John sleeps) in the open air'

While some others are only used in a negative sentence:

[*Juan* ***no*** *aceptará*] *por nada del mundo*
'(John will not accept) by no means'

*[*Juan aceptará*] *por nada del mundo*
'(John will accept) by no means'

Others impose a specific tense:

[*Juan* ***llegará***] *en breve* '(John *will* come shortly'

*[*John llegó*] *en breve* '(John has come) shortly'

2.1 Classification

We apply the notion of adverbs to syntactically different structures of traditional terminology such as underived (primary) adverbs (*bien*, 'well') or derived forms (*profundamente* 'deeply'), circumstantial complements (*al amanecer* 'at dawn'), and circumstantial clauses (*hasta que la muerte nos separe* 'until death do us part').

We considered the sequence *Prep Det C Modif*[3] as the basic structure that formally define and classify compound adverbs, adopting the concept of *generalized adverb* proposed by M. Gross (1986) for French adverbs.

Based on this, we defined 15 formal classes for Spanish compound adverbs. Table 1 (below) shows the current state of the dictionary, the internal structure of each class, an illustrative example and the number of compound adverbs collected so far.

Further than this classification based on their internal structure, we have proposed different types of semantic-functional groups presented in terms of Finite State Transducers (FSTs), as in Fig.1. In this graph, all adverbial expressions have the same general meaning ('quickly'). Similar graphs can be used, for example, to compare the distribution of semantically 'equivalent' expressions and to structure the co-occurrence of those adverbs with their argument predicates.

Class	Structure	Example	Size
PC	*Prep C*	*sin ambajes*	869
PDETC	*Prep Det C*	*al contado*	585
PAC	*Prep Adj C*	*sin previo aviso*	157
PCA	*Prep C Adj*	*a brazo partido*	291
PCDC	*Prep C de C*	*a cuerpo de rey*	168
PCPC	*Prep C Prep C*	*de cabo a rabo*	149
PCONJ	*Prep C Conj C*	*en cuerpo y alma*	131
PCDN	*Prep C de N*	*a condición de*	233
PCPN	*Prep C Prep N*	*de espaldas a*	51
PV	*Prep V W*	*sin querer*	127
PF	frozen sentence	*que yo sepa*	169
PECO	(*como*) *Adj que C*	*sordo como una tapia*	797
PVCO	(*V*) *como C*	*(beber) como una esponja*	532
PPCO	(*V*) *como Prep C*	*(desaparecer) como por ensalmo*	46
PJC	*Conj C*	*y no se hable más*	91
		TOTAL	4396

Table 1. Classification of Spanish compound adverbs.

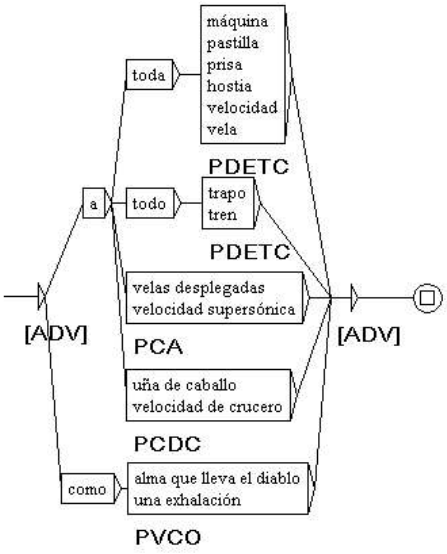

Fig.1 Finite-State graph (simplified) for semantic clustering of adverbs

2.2 Microstructure of Dictionary

The description takes the shape of binary matrices (see Table 2, for an example), in which each line corresponds to a lexical entry, and the columns represent different information. The set of matrices constitute the lexicon-grammar of adverbial frozen expressions. Next, we present a brief description of the microstructure of the dictionary.

[2] In the examples, (argument) simple sentences are given in brackets.

[3] *Prep* = preposition; *Det* = determiner; *C* = lexical constant, usually a noun; *Modif* = modifier, such as an adjective (*Adj*) or a prepositional phrase.

N_0	V	Prep	Det	C	PreMod	Mod	Prép Dét C	Prép Dét Adj C	Conj	DiaSys	English equivalent
hum	Vact	-	-	acto	-	seguido	-	-	+	-	immediately afterwards
hum	llegar	a	la	hora	-	horada	+	-	-	familiar	on the nose
hum	Vact	por	-	voluntad	-	propia	-	+	-	-	with one's own will
hum	comprar	a	el	por	-	mayor	-	-	-	commerce	wholesale
hum	dormir	con	los	ojos	medio	abiertos	-	-	-	-	with one's eyes half open

Table 2. Class PCA (extract)

The first column concerns the syntactic-semantic nature of the subject. We adopted G. Gross (1995) and Le Pesant and Mathieu-Colas (1989) basic typology, distinguishing the following semantic classes: *human, animal, vegetal, concrete,* and *abstrac*t.

The second column refers to the verb most commonly used with the adverb, for example:

[*salir*] *a cuerpo gentil*
'(to go out) without cloak';

[*cerrar Nconc*] *a cal y canto*
'(to close something) under lock and key'.

The following columns contain the elements of the structure: *Prep, Det, C,* and *Modif*, e.g.:

[*Esta gente llegó en este país*] *con las manos vacías*
'These people arrived in this country with empty hands'

Naturally, in Spanish the modifier can be placed before *C*:

[Se peleaban] *a la menor ocasión*
'(they were fighting each others) at the least occasion/opportunity'.

The next columns correspond to their syntactic (distributional and transformational) properties: '+' indicates that the expression admits this property, and '- ' that it does not. Relevant properties depend on the class: some have to do with permutation of elements of the compound or their reduction to zero (zeroing); see §2.3, below.

Diasystem information (Hausmann 1989) is provided in next field (DiaSys) such as these categories (marked in bold, in the examples below):

– diatopy:
[*Juan trabaja*] *al cohete* (Uruguay/Argentina)
'(John works) in vain';

– diachrony :
[*Juan convoca a los estudiantes*] *a voz de apellido* (**out of use**)
'(John summons the students) by their family name';

– diafrequency :
[*Juan se sirvió*] *a barba regada* (**unusual**)
'(John served himself) abundantly'

– diastratic:
[*Juan recita*] *de carretilla* (**familiar**/ **colloquial**)
'(John recites) by heart'*;*

– diatechnical :
[*El torero clavó la banderilla*] *de sobaquillo* (**bullfighting**) '(the bull fighter has pinched the bull) on its side;

– diaintegrative :
[*Juan vino*] *motu propio* (**latinism**)
'(John came) voluntarily'.

Finally, we have included French translation equivalents. These equivalence relations are also currently being extended to other languages, such as Portuguese (Palma, *in prep.*).

2.3 Syntactic properties

We will only consider here the most prominent properties, considering all classes of adverbs under study.

One of the properties indicates the possibility to transform the initial structure in to a more analytical phrase like *de (modo + manera) C-a* 'in a *C-a* way/manner', where *C-a* is an adjective, morphologically related to the constant (frozen element) *C*; naturally the meaning of the two structures is the same:

[*La candidatura se aprobó*] *por unanimidad*
= [*La candidatura se aprobó*] *de manera unánime*

'(His application was approved) by unanimity/in an unanimous way'

[*Juan lo ha dicho*] *con todos los respetos*
= [*Juan lo ha dicho*] *de manera respetuosa*
'(John has said so) with all due respect/ in a respectful manner'.

Another, similar, property shows the possibility to transform the initial structure in an adverb based on the same type of *C-a* adjective and the suffix *-mente.* This property concerns classes PC and PDETC :

[*La candidatura se aprobó*] *por unanimidad*
= [*La candidatura se aprobó*] *unánimemente*
'(His application was approved) unanimously'

[*Juan lo ha dicho*] *con todos los respetos*
= [*Juan lo ha dicho*] *respetuosamente*
'(John has said so) respectfully'.

Property *Conj* concerns classes PC, PDETC, PAC and PCA. It highlights the eventual anaphoric effect of the adverb. We consider it as a conjunction-adverb, since in sentences like:

[*Juan estudia*] *en consecuencia*
'(John studies) in consequence'

[*Juan se marchó] por lo tanto*
'(John went away) for that much'

we need a (trans-)phrastic context such as :

[*Juan quiere aprobar*], *en consecuencia,* [*estudia*].
'(John wants to succeed in school), in consequence (he studies)'

[*Ana se enfadó con Juan*], *por lo tanto,* [*éste se marchó*]
'(Ana get bored with John), for that much (he went away)'

The next property concerns classes PCA and PAC. It describes the possible omission of the modifier:

[*Los niños andan*] *en fila india*
'(The kids walk) in Indian line'

= [*los niños andan] en fila*
'(The kids walk) in line'

Other property indicates the possibility of moving modifier from its basic position to the left of *C*; it only concerns class PCA:

[*Juan encontró a Ana*] *en hora buena*
= [*Juan encontró a Ana*] *en buena hora*
'(John met Ana) in good time/in time'

We have also noted the possibility of zeroing the second element of the compound, i.e., the free or frozen prepositional phrase. It concerns classes PCDC, PCPC, PCONJ, PCPN, and PCDN:

[*Juan estudia*] *con la mejor voluntad del mundo*
= [*Juan estudia*] *con la mejor voluntad*
'(John studies) with the best will (of the world)'

[*Juan vive*] *al margen de la sociedad*
= [*Juan vive*] *al margen*
'(John lives) at the margin (of society)'

[*Juan vive*] *de espaldas a la calle*
= [*Juan vive*] *de espaldas*
'(John lives) with his back (turned to the street)'

Certain permutations have been noted, but not dealt with in a transformational way:

[*Juan se enamoró de Ana*] *por decirlo así*
= [*Juan se enamoró de Ana*] *por así decirlo*
'(John fall in love with Ana) as it were'

Finally, we consider the possibility of substitution of the second element by a subordinate clause (finite or infinitive); this property concerns PCDN and PCPN:

[*Le consultará*] *en caso de duda*
= [*Le consultará*] *en caso de que haya duda*
'(He will consult him) in case of doubt/in case there is any doubt'

[*Juan se marchó*] *por miedo al fuego*
= [*Juan se marchó*] *por miedo a que haya fuego*
'(He went away) for fear of fire/there being fire'

[*Juan se sujetó*] *por miedo a una caída*
'(John hold tight) by fear of a fall'
= [*Juan se sujetó*] *por miedo a caer*
'(John hold tight) by fear of to fall'

A strictly statistically, corpus-based approach that only contemplates strings of words in view to produce lexicon entries (Manning and Schütze 2003) cannot but fail to put in relation such formal variants of equivalent expressions. On the other hand, many formal variations are very much dependent on the particular lexical combinations, and cannot be generalized, hence the need to describe their syntactic properties systematically.

While very time-consuming, our method provides a fine-grained linguistic description, and is directly exploitable by finite-state methods.

With the aim of retrieving the adverbial expressions from texts using the information encoded in the lexicon matrices, it should be noted

that most but not all properties referred to above can be directly formalized using the finite-state methods we are currently using. In the following lines, we present this methodology.

3 Formalization

In order to apply to texts the set of matrices that constitute the Lexicon-Grammar and thus to identify and tag compound adverbs, we have followed the methodology proposed by Senellart (1998) and Silberztein (2000), and adapted by Paumier (2003, 2004) for the UNITEX system [4]. This method consists of intersecting linguistic data on matrices with a finite-state graph (called a reference graph) in order to generate automatically a finite-state transducer (FST) that can be applied to a corpus[5].

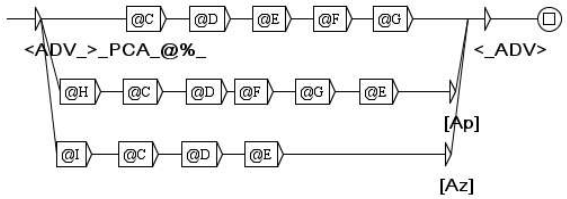

Fig.2 Reference graph (simplified) for class PCA

Fig.2 shows a (simplified) reference graph for class PCA. In the graph, variable `@X` stands for column `X` in the matrix. For each line in the matrix the system builds a sub-graph by replacing each variable for the content of the corresponding columns in the matrix. If that columns is a binary property, the corresponding variable in the graph functions as a switch, allowing for the rest of that graph's path to be build in case of a '+' or, else, collapsing the graph at that point, if a '-' is found at that property. It is also possible to deny a property (`!@X`), which has the opposite effect. Another utility of the system is the inclusion of a variable `@%` that outputs the number of each entry line in the matrix, thus enabling the user to easily put in correspondence a given result to the corresponding lexical entry. The set of sub-graphs (one per each entry in the matrix) is automatically gathered in a finite-state transducer that can be directly applied to texts.

In Fig. 2, class PCA reference graph includes: two delimiters of the compound expression, `<ADV_>` and `<_ADV>` ; the `@%` variable; the top-most path describe the full expression, while the second and third paths, below, depend on properties described by variables `@H` and `@I`; these correspond to the permutation of the adjective `[Ap]` and its reduction to zero `[Az]`, respectively.

Similar graphs have been built to other classes[6]. The set of classes thus formalized constitute an electronic dictionary of 2,930 entries (67% of all compound entries collected so far).

4 An experiment on texts

The aim of this experiment is to assess the advantages and limitations of the methodology described in §3 in the identification of multiword units, in this case, compound adverbs, in real texts in Spanish.

The FSTs were applied to a fragment of a corpus of journalistic text taken from the newspaper *El Mundo*, of about 2 Mb and 171.5 K (~24 K different) words. The system retrieved 2,276 matches, corresponding to 461 different entries.

Table 3 shows the breakdown of these matches per class and its percentage, followed by the number of different entries (types) matched by the system and the corresponding percentage of each class entries.

class	class size	matches	% matches	entries	% entries
PC	869	849	0.37	215	0,47
PCDN	233	489	0.22	12	0,03
PDETC	585	406	0.18	119	0,26
PCPN	51	238	0.10	23	0,05
PCA	291	134	0.06	19	0,04
PF	169	42	0.02	7	0,02
PAC	157	38	0.02	23	0,05
PCONJ	131	22	0.01	9	0,02
PCPC	149	21	0.01	12	0,03
PCDC	168	17	0.01	12	0,03
PV	127	16	0.01	10	0,02
	2,930	2,272		461	

Table 3. Breakdown of matches per class.

Classes PC, PCDN, PDETC, PCPN and PCA are the only classes with over 100 matches; together they constitute 93% of the matches, all other classes have residual expression.

[4] www.univ-mlv.fr/~unitex.

[5] See Paumier (2004), for further details.

[6] In this paper, however, we did not deal with classes of comparative adverbs (PECO, PVCO and PPCO) or class PJC, which pose particular problems to their recognition.

On the other hand, classes PC and PDETC present the larger number of dictionary entries matched. Notice that, despite the number of entries in the matrices, only 461 entries (16%) were found in the corpus.

Class PC alone represents 47% of the total entries matched by the system (215/461), immediately followed by class PDETC, with 26% of matched entries (119/461). Matches for these two classes together constitute 55% of the total of strings matched by the system (1,255/2,272). These two figures make PC and PDETC the most prominent classes for this experiment, in view of the assessment of the finite-state methods here used to identify compound adverbs in texts. For lack of space, analysis of results will thus focus on these classes and only major phenomena, i.e., those situations with major impact on results, will be taken in consideration here.

5 Results and discussion

We went through the concordances manually, and confirmed a **precision** of 77.4% (974/1,255) [7]. We discuss these results below.

The major reason for incorrect matching has been found to correspond to cases where the matched sequence is not the target compound adverb but part of a longer, free word sequence, or part of a compound word; in the following example, the adverb *de accidente* 'accidentally' is an ambiguous string since it overlaps with the compound noun *seguros de accidente* 'accident insurances'

Antes de iniciar un rodaje, se prevé cualquier eventualidad. Se contratan **seguros *de accidente***, *enfermedad y muerte para las personas clave del proyecto* [PC_0010]

while in the next example, the string *de derecho* 'by law/right' overlaps a (free) prepositional phrase which includes a compound noun *derecho de veto* 'right of veto':

Yo creo que no se puede pretender ejercer una especie ***de derecho*** **de veto**, *porque esto querría decir que el Gobierno es rehén* [PC_0243]

In some few cases, incorrect matches were the result of an inadequate treatment of contractions of prepositions and determiners. In classes PCDN, PCPN, the second preposition often appears contracted with the determiner of the free NP. In the next example, contraction of *a + el = al* has not been correctly described:

coches serán introducidos en el mercado nipón en el mes de octubre, con ocasión del Salón de Tokio. ***Con respecto*** **al Tigra**, *que se produce en exclusiva para todo el mundo en Figuer* [PC_0686]

This problem is to be fixed on a next version of the reference FSTs.

In some cases, especially when the adverb is marked as a conjunction-adverb (*Conj*), it often appears between comas or at the beginning of sentences, followed by coma.

*se había montado su particular Guerra de los Mundos de tema ferroviario. También hay quien piensa****, por cierto,*** *que a este Gobierno se lo van a cargar no sus errores, sino las cos* [PC_0145]

privatizar el 99,9% de las empresas y entes públicos de la Comunidad y ya está trabajando en ello. ***Por cierto,*** *le ha arrebatado el control del Canal de Isabel II a Pedroche y lo* [PC_0145]

We have annotated these cases so that this information can be added to the matrices and used in disambiguation tasks.

Finally, many temporal adverbs have only partially been identified.

puede seguir así»- exigió al Gobierno de González que fije un calendario electoral **antes del 17 de *este mes***. *Tras de lo cual, el aún secretario general de CDC sostuvo que, si* [PDETC_0076]

zo de Erez, consiguió dos objetivos. En primer lugar, Israel se comprometió a iniciar, **a finales de *este mes***, *la evacuación gradual de tres ciudades palestinas: Jenin, Kalkilia* [PDETC_0076]

This occurs because matrices only included simple word combinations. As others have noted previously (Baptista and Català 2002; Baptista 2003a,b), time-related adverbs may be described by FST methods as those used here. Those local grammars could easily be integrated in the system.

[7] Since we started with a previously, manually build, electronic dictionary, we can not compute *recall*. We define *precision* as the number of correct matches on total matches.

6 Conclusion

The taxonomic approach adopted here, the systematic survey of the lexicon and its formal representation, resulted in a complex linguistic database of Spanish compound adverbs. This may have many applications, not strictly in Linguistics, but also in Didactics and in Lexicography.

It can further be used in several applications on natural language processing. The relatively high precision (77,4%) of the finite state methods used in this paper are very encouraging, and in some cases, discussed above, they can and will be improved in a future version both of the reference graphs and of the lexicon-grammar matrices.

However, the major difficulty to a better identification of compound adverbs in texts seems to reside in the fact that no syntactic analysis (parsing) has been performed on the text. Therefore, there is no possibility of using information regarding (sub-)phrases and other constituents of the compounds in order to preclude incorrect matching.

Another aspect that hinders better results has to do with the formal variation of compound adverbial expressions. Adverbs present more problems for their recognition as the limit between free sequence and fixed sequence is more difficult to establish than in others categories of compounds. The building of electronic dictionaries may benefit from a (more) corpus-based approach, so as to retrieve variants of a given lexical entry, but a careful and time-consuming verification is needed in order to group variants as different expressions of the same meaning unit.

Finally, the relatively small portion of the dictionary matched on the corpus imposes that it should be tested on texts of a more diverse nature and of a larger size, thus probably yielding a larger perspective of the use of these idiomatic expressions. Still, it is now possible to consider the study of the distribution of these adverbs, trying to specify the type of predicates (verbs, nouns, adjectives, mainly) on which they operate.

Acknowledgement

This research was supported by the Spanish Ministerio de Ciencia y Tecnologia in the framework of the project grant HP-2004-0098, and Conselho de Reitores das Universidades Portuguesas, project grant E-111/-05.

References

Jorge Baptista 2003a. Some Families of Compound Temporal Adverbs in Portuguese. Proceedings of Workshop on *Finite-State Methods for Natural Language Processing*: 97-104, ACL, Hungary.

Jorge Baptista 2003b. Evaluation of Finite-State Lexical Transducers of Temporal Adverbs for Lexical Analysis of Portuguese Texts. *Computational Processing of the Portuguese Language*. Proceedings of *PROPOR'2003*. Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence 2721: 235-242, Springer, Berlin.

Jorge Baptista and Dolors Català 2002. Compound Temporal Adverbs in Portuguese and in Spanish. *Advances in Natural Language Processing*, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence 2389: 133-136, Springer, Berlin.

Jorge Baptista and Dolors Català 2006. Les adverbes composés dans le domaine du travail. *Mots, Termes, et Contextes*: 249-263, AUF/LTT and Éd. Archives Contemporaines, Paris.

Xavier Blanco and Dolors Català 1998. Quelques remarques sur un dictionnaire électronique d'adverbes composés en espagnol. *Linguisticae Investigationes Supplementa* 11 (2): 213-232, John Benjamins Pub. Co., Amsterdam/Philadelphia.

Gaston Gross 1995. À propos de la notion d'humain. *Lexiques Grammaires Comparés en Français. inguisticae Investigationes Supplementa* 17: 71-80, John Benjamins Pub. Co., Amsterdam/Philadelphia.

Maurice Gross 1986. *Grammaire transformationnelle du français: syntaxe de l'adverbe,* ASSTRIL, Paris.

Zellig S. Harris 1976 *Notes du cours de syntaxe,* Le Seuil, Paris.

Zellig S. Harris. *A Theory of Language and Information. A Mathematical Approach*, Clarendon Press, Oxford.

Franz J. Haussmann 1989. Die Markierung in allgemeinen einsprachigen Wörterbuch: eine Übersicht, *Wörterbücher, Dictionaries, Dictionnaires*, vol 1: 651, Berlin/ New York, Walter de Gruyter.

Denis Le Pesant and Michel Mathieu-Colas. 1998. Introduction aux classes d'objets *Langages* 131: 6-33, Larousse, Paris.

Ch. Manning and H. Schütze 2003. *Foundations of Statistical Natural Language Processing*, MIT Press, London/Cambridge, MA

Cristina Palma (in preparation). *Estudo Contrastivo Português-Espanhol de Advérbios Compostos*, Univ. Algarve, Faro.

Sébastien Paumier 2004. *Unitex - manuel d'utilisation*, Univ. Marne-la-Vallée, Paris.

Jean Senellart 1998. Reconnaissance automatique des entrées du lexique-grammaire des phrases figées. *Le Lexique-Grammaire*. *Travaux de Linguistique* 37: 109-125, Duculot, Bruxelles.

Max Silberztein 2000. *Intex* (Manual), ASSTRIL/LADL, Paris.