Experiments in identifying frozen sentences in a large corpus

Graça Fernandes[1], Jorge Baptista[1,2]
[1] Universidade do Algarve, Portugal
[2] *L²F* – Spoken Language Laboratory, INESC-ID Lisboa, Portugal

Abstract

This paper describes an experiment on the identification of frozen sentences (or verbal idioms) from European Portuguese on large corpus of journalistic text. It aims at identifying the main difficulties (or shortcomings) resulting from the intersection of linguistic information encoded in the lexicon-grammar with finite-state transducers that are then applied to texts. The paper shows that, for a selection of frozen sentences, this method is capable of identifying most instances of those idioms in the corpus. In most cases, only insertions of free elements (especially, adverbs), which are external to the frozen construction, hinder better results than those here obtained, namely of precision P=0.94 and recall R=0.88.

Keywords: frozen sentences, idioms, European Portuguese, Syntax, Lexicon, Lexicon-Grammar, Natural Language Processing, Computational Linguistics.

1. Introduction

This research adopts the theoretical and methodological framework of Lexicon-Grammar (M. Gross 1982, 1989, 1996), based on the harrissian transformational operator-grammar (Z. S. Harris 1991). It is part of a larger program to build a lexicon-grammar of Portuguese frozen sentences (Baptista 2004; Baptista *et al.* 2004, 2005, Fernandes 2007, Fernandes and Baptista, *in print*) and other idiomatic expressions. In Fernandes (2007), the main syntactic (distributional, structural and transformational) properties of intransitive, frozen, verbal constructions were described and formalized into lexical-syntactic, binary matrices, in view of several applications, namely on natural language processing. About 900 frozen sentences were collected from several lexicographic sources, both general and specialized dictionaries of idioms, as well as from *corpora* and from introspection. They were then classified in three main formal classes: **CP1** (*O Pedro jogava pelo seguro* 'Peter plays safe'), **CPN** (*O Pedro não chega aos calcanhares do João* 'Peter is not a match to John') and **CPP** (*O Pedro puxou pelos cordões à bolsa* 'Peter was reluctant to make some expenses').

* Research for this paper was partially supported by MCTES-FCT (project grant POSI/PLP/14319/2001); the second author received a Fundação Calouste Gulbenkian conference grant; the authors wish to acknowledge Ana Soares (U. Algarve) for her help on the final English version of this paper.

1 Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Campus de Gambelas, P-8005-139 FARO, Portugal {jbaptis, a9110}@ualg.pt.

2 L2F – Spoken Language Laboratory, INESC-ID Lisboa, R. Alves Redol, 9, P-1000-029 LISBOA, Portugal.

Based on these structures, this paper reports an experiment conducted in order to apply the linguistic information encoded in the matrices to CETEMPÚBLICO, a freely-distributed, large-size corpus of journalistic text (approx. 180 M words) and its subsequent evaluation. This experiment on corpus is based on finite-state techniques, namely the intersection of linguistic data on matrices with reference finite-state transducers, using the UNITEX software (Paumier 2002). Due to the number of frozen sentences collected so far, observations were made on a selection of expressions, taken from those entries of the class CP1 that were intuitively deemed as of common use. Our aim is to identify the advantages and/or limitations of this method of formal representation on the application of the grammar to the *corpora* adopted here.

2. Methods

For the application of linguistic information to corpora, codified in the lexical-syntactic matrices, the following instruments are needed: (1) a reference graph, that describes the sequence in which the constituent elements of the frozen expressions can be ordered in the sentences; (2) the lexical-syntactic matrices of each formal class; (3) a corpus of text; and (4) linguistic resources (electronic dictionaries) for text processing in Portuguese. In the next sections, we briefly describe these instruments.

2.1. Reference Graphs

A reference graph is a finite-state transducer that refers to the elements and the properties presented in the matrices by the variables @X, in which X means the content of the corresponding column of the matrix, where such elements or properties are formalized. Fig.1 illustrates, in a simplified way, the construction of the reference graph for class CP1.

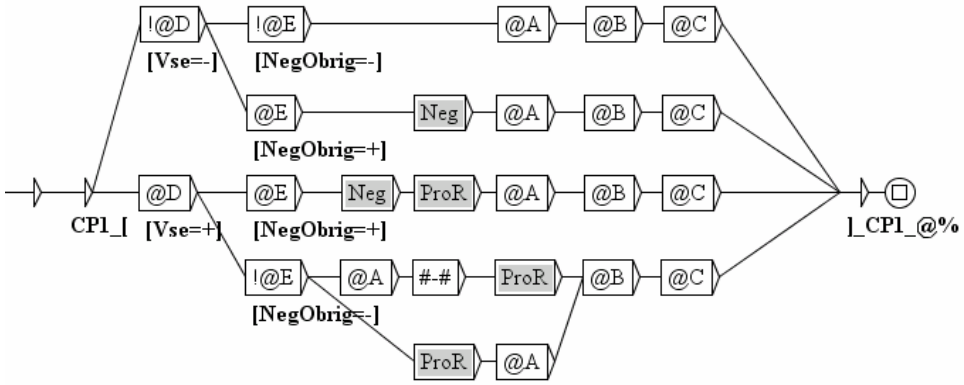

Fig.1. Reference Graph (simplified) for the class CP1

Variables @A, @B and @C represent, respectively, the verb, the (eventual contraction of) the preposition and the determinant and the lexical constant; variables @D and @E represent the properties of the intrinsically pronominal construction (*Vse*) and of the obligatory negation (*NegObrig*); variable @% indicates the subgraph number that identifies the correspondent lexical entry. Properties are given explicitly in the output: 'Vse' refers to an intrinsically pronominal construction while 'NegObrig' stands for obligatory negation.

The UNITEX programme, for each line of the matrix, goes through the reference graph replacing each variable with the content of the correspondent column in that matrix. If it finds lexical elements in that column, the variables are replaced with those words or, eventually, their lemma. If the column contains lexicon-syntactic properties, the variables work as switches that determine the construction of the graph: if the property is marked positively, the

rest of the line of that graph will be produced; if that property is shown as negative, the construction of the graph collapses at this point. It is also possible to deny a binary property (!@X), which has the reverse effect.

The system builds a subgraph for each line of the matrix. The set of subgraphs created this way is gathered in a general graph. This graph can be used to search for expressions in texts. Being a transducer, it can be used to insert linguistic information in the texts. In this case, tags like class code, entry number, and the values of the entries' syntactic properties may be merged with the text.

2.2. Working *corpus*

A working *corpus* was constituted from the CETEMPÚBLICO (v. 1.7, annotated) corpus using the concordances that result form the research of rational expressions. These expressions consist, generically, of the lemma of the verb and its frozen prepositional complement, allowing for a five-word window between these elements. So, for example, to the idiom <*dar*> *à sola* 'to take to heels' we used the rational expression:

```
[lema="dar"][]{0,5}"à""sola"
```

We used this five-word window to allow for the insertion of elements between the components of the frozen expression. This window size follows Manning & Schütze's (2003: 157-162) remark that the elements of a frozen combinatory usually do not occur at a greater distance than five words. Bellow we present the result of the research referred to previously (conventional codes of the corpus extracts were kept):

Ext 679965 (soc, 94a): [...] na eventualidade de se ter que **dar à sola** nu [...]
Ext 740390 (opi, 96a): [...] resolveu abandonar a família e «**dar à sola**» .
Ext 825518 (nd, 91b): R. -- Mesmo que a guerra fosse justa, **dava à sola** .
Ext 841573 (opi, 96a): [...] elogiando o corajoso ministro que **deu à sola** por ter fugido ao fisco .
Ext 882590 (soc, 94a): [...] e então pirei-me para Leiria, **dei à sola**», explicou .
Ext 991750 (nd, 95b): [...] o pessoal via aquilo e **dava à sola**, também tenho a noção do ridículo, ah, ah, ah... "
Ext 1466240 (soc, 92b): Entretanto, a culturista sérvia [...] assim que pôde, **deu à sola** .
Ext 1532008 (pol, 96b): Quero lá saber dos militantes que [...] são os primeiros a **dar à sola**... [...]

Fig.2. Results of the research in CETEMPÚBLICO of the expression <dar> à sola

As we can see in this concordance, from the eight occurrences resulting from the research of the rational expression in CETEMPÚBLICO, all of them correspond to the correct use of the frozen expression. Notice that no insertion has been found, in spite of the window defined between *V* and *Prep C*. This method allows to build a *corpus* of work in which we can be sure of finding all the expressions that are object of this study. It is in this sense that *recall* can be calculated. Naturally, and contrary to the case above, not all of the concordances obtained thus are free of incorrect expressions (noise).

2.3. Experiment

The objective of these experiment is to demonstrate a method of application to texts of the information encoded in the lexicon-grammar, and to identify its limitations and/or advantages; considering the total number of frozen sentences of this study, we have selected 100 entries (a little more than 10%) from the class CP1 (Fernandes 2007), namely, those expressions that intuitively seemed to be more of more current use.

We then collected from CETEMPÚBLICO the corresponding concordances, using the rational expressions illustrated above. Many entries of the matrices simply do not appear in this corpus in spite of its size. This is probably due to the colloquial use of the expressions and to the journalistic nature of the text. Even though it would be interesting to verify in

which way the occurrence of these expressions can vary depending on the nature of different texts (and not only journalistic text), this is beyond the aim of this paper.

From the concordance of each rational expression, we have registered the number of global occurrences (**N**) and the total number of the frozen expression (**T**[arget]), which was manually verified. Next, we have applied over that concordance the grammar built according to the methods described in §2.2.

We registered the number of cases in which the grammar correctly recognized the frozen expression (**tp**= true positives) and the number of cases in which the grammar matched a sequence that did not constitute the targeted frozen expression (**fp**= false positives). In this way we calculated the measures of precision (**P**), recall (**R**) and the F-measure (**F**) for each case, as well as the global figures of the experiment. Naturally, recall measure refers only to the targeted CP1 expressions. For lack of space, only a fragment of this table is presented here (Table 1).

Table 1. Results.

Frozen sentences (sample)	N	T	tp	fp	P	R	F
1. <dar> nas vistas	771	771	652	0	1.00	0.85	0.92
2. <usar> da palavra	402	401	380	0	1.00	0.95	0.97
3. <fugir> à regra	393	393	303	0	1.00	0.77	0.87
4. <passar> à história	274	264	254	7	0.97	0.96	0.97
5. <ir> para a rua	249	201	194	48	0.80	0.97	0.88
6. <vir> ao mundo	226	105	104	9	0.92	0.99	0.95
7. <olhar> a meios	202	201	196	0	1.00	0.98	0.99
8. <rebentar> pelas costuras	196	195	195	1	0.99	1.00	1.00
9. <renascer> das cinzas	178	171	150	3	0.98	0.88	0.93
10. <fazer> à vida	165	3	3	0	1.00	1.00	1.00
11. <jogar> pelo seguro	136	136	123	0	1.00	0.90	0.95
12. <desaparecer> do mapa	134	125	116	5	0.96	0.93	0.94
13. <ir> ao mar	133	72	72	3	0.96	1.00	0.98
14. <ganhar> para o susto	131	131	104	0	1.00	0.79	0.89
15. <dar> para o torto	109	108	100	0	1.00	0.93	0.96
16. <fazer> pela vida	109	96	89	2	0.98	0.93	0.95
...							
32. <cortar> nas despesas	55	55	49	0	1.00	0.89	0.94
33. <tratar> da vida	55	16	16	14	0.53	1.00	0.70
34. <gritar> por socorro	54	53	52	0	1.00	0.98	0.99
...							
93. <chegar> aos bons	1	1	1	0	1.00	1.00	1.00
94. <conversar> com o travesseiro	1	1	1	0	1.00	1.00	1.00
95. <dar> ao slide	1	1	1	0	1.00	1.00	1.00
96. <dar> com a coisa	1	1	1	0	1.00	1.00	1.00
97. <dar> com a coisa	1	1	1	0	1.00	1.00	1.00
98. <entender> da poda	1	1	1	0	1.00	1.00	1.00
99. <pedir> por socorro	1	1	1	0	1.00	1.00	1.00
100. <regular> da mona	1	1	1	0	1.00	1.00	1.00
Total	6,039	5,183	4,605	187	0.94	0.88	0.91

Legend: N – global number of occurrences obtained by a query to the entire corpus of CETEMPÚBLICO (180 M words) using a rational expression with the lemma of the verb and the frozen prepositional complement, and allowing for a five word window between both,; T (target) – total number of real occurrences of the frozen expression, verified manually; tp – true positives; fp - false positives; P – precision; R –recall; F – measure-F, combination of P and R, with equal reflection: 2PR/P+R. This table shows results for expressions where N > 100, N » 50 and N = 1 in the corpus.

3. Results

Generally, the expressions selected for this experiment are of current use. Yet, more than a half occur less than 25 times, in a *corpus* of 180 million words; there were 33 with 10 or less occurrences, and 8 they occur only once (hapaxes; see last section of the table). The expression that occurred more frequently (N=771), *dar nas vistas* ('to be conspicuous'), is a non-ambiguous expression, hence all its occurrences correspond to instances of the frozen sentence (N=Target). The same happens, in general, with expressions with low N.

The set of idiomatic expressions studied here presents an average frequency in the corpus of about 0.0193 occurrences per million words (values are approximate). This frequency is to be considered a value by excess, since the sample consisted of commonly used expressions. Yet, the most frequent expression studied here (*dar nas vistas*) only presents an average frequency of 1.94 per million words.

Generally, the average precision is P=0.94 and recall R=0.88; F-measure is F=0.91. In the next section, we present a more detailed analysis of these results.

4. Discussion

Our goal was only to identify the main difficulties (or shortcomings) of this finite-state method for pattern-matching idioms. Most instances in the *corpus* of the frozen sentences selected to the test-sample have been be correctly matched by using the simple (but already quite sophisticated) finite-state methods applied here. The average precision and recall are both similar and high. Naturally, we benefited from the fact that the intransitive constructions studied here show very few transformations (or alternations). On one hand, this greatly simplifies the construction of the reference graph; on the other hand, the (almost total) absence of structural variation prevents the formation of *loci* where insertion of 'disruptive' material could undermine the success rate reported above. On one hand, this greatly simplifies the construction of the reference graph; on the other hand, the (almost total) absence of structural variation prevents the formation of *loci* where insertion of 'disruptive' material could undermine the success rate reported above.

The observations made here should then be considered carefully, as the extension of this study to transitive constructions may alter results significantly. One could also argue that the sample is somehow biased by the fact that it consists of currently used idiomatic expressions. These expressions might then be so frozen as to allow an easier matching that less current idioms. This is not entirely so, as the frequency distribution of the sample entries clearly demonstrates.

However, as it will be shown below, the most prominent problem identified in this paper – namely the insertion of diverse free elements inside the 'frozen' string – is of a very general nature and it is not to be directly related to the frequency of these expressions. Still, we consider the sample to be (loosely) representative of the full lexicon-grammar of intransitive frozen sentences. For example, specific phenomena such as obligatory negation and intrinsically reflexive constructions represent a percentage of the sample similar to the general percentage in the entire lexicon-grammar. The small impact of these different insertions makes it clear that frozen sentences tend to appear in a straightforward manner, without spurious elements inserted, perhaps to become more easily detected by the reader as idioms.

Next, we show in detail the matching problems we have encountered while manually perusing the 'unmatched sequences' of the *corpus*. In general, the main hindrance to the recognition of frozen sentences in this corpus is due to the insertion of the free elements, occurring between the constitutive elements of the idiomatic expression. Since these insertions are not

anticipated by the reference graph, the system cannot find them. Some of these insertions are just formal elements, such as the quotation marks, which, in written texts, frequently indicate the idiomatic character of these expressions (in the examples, the elements of the frozen sentence are in bold, while insertions are in italic):

Ext 176357 (soc, 95b): [...] bombeiros **« brincaram** *»* **com o fogo** [...] ('firemen «played» with fire')

However, other insertions are lexical elements. Thus, for example, we frequently observe the insertion of adverbs of a very diverse nature and form between the verb and the frozen prepositional complement:

Ext1435163(des,92b):[…] alguns dos [mirones] […] **bateram** *estrategicamente* **em retirada**
('they strategically retreated')

Since these insertions are not described in the reference graph, the frozen sentences are not identified. It would be easy to adapt the graph in order to permit this insertion, but our aim in this paper was exactly to measure how much this particular aspect could make the task harder.

In other cases, coordination phenomena are behind the incomplete recognition of the frozen expression. Two frozen sentences, with the same verb and structure but different *C*, v.g. *dar à lingua* ('to talk') and *dar ao dente* ('to eat'), have conflated in the next example:

Ext1271901(nd,94a): Muito **darão à língua** *e* ***ao dente*** […]
(lit: Much will [they] give to_the tongue and to_the tooth, 'they will talk and eat very much')

This type of limitation is perfectly natural with these finite-state methods, since they are based on the mere recognition of strings of words. The possibility to coordinate the frozen complements of the two distinct idiomatic expressions built with the same verb shows that in spite of its idiomatic status and distributional constraint, frozen sentences keep some syntactic structure and are not to be lightly dismissed as a simple matter of string matching. On the contrary, this and other similar cases lead one to think that only after some degree (probably even shallow) syntactic analysis (or parsing), will it be possible to rigorously and adequately identify this kind of expressions.

A similar case, but of different syntactic nature: the system does not recognize the clitic indefinite pronoun *–se* (one), which is then an insertion between the verb and its frozen prepositional phrase:

Ext38967(nd,93a): No Ocidente chamamos-lhe brindes, mas na Geórgia **não** *se* **brinca em serviço**
(In the West we call it gifts, but in Georgia, one does not play while at work; 'one is really very serious about it'; it = 'bribes')

Once more, only a correct syntactic analysis of *–se* as well as of the sentence in which this is inserted would allow for the correct identification of the frozen expression.

A general case of incomplete recognition of frozen expressions occurs when an auxiliary accompanies the verb; in the following example, the past participle of the verb *meter* (*metido*) is used with a past tense auxiliary *ter*:

Ext635725(clt, 94a): Diana retorquiu: detesta «junk food» e não apanha muitas tosgas, apesar de na noite anterior **se** *ter* **metido nos copos**, porque fazia anos de casada
(Diana answered back: she hates junk food and she does not get drunk frequently, even if on the night before she had given in to the bottle, for it was her marriage anniversary)

The system only matches the participle and the prepositional phrase. However, since auxiliaries form a verbal complex with the main verb, they should perhaps be recognized as an integrant part of the frozen expression. Obviously, we have not yet applied a syntactic module to analyse auxiliary verbs; therefore, all expressions with auxiliaries have been only partially recognized. This example also shows another shortcoming of the problem with

auxiliary verbs. This is an intrinsically reflexive construction (*meter-se nos copos*), hence the reflexive pronoun should be matched by the grammar along with the verb. Furthermore, the sentence has been set in a subordinate clause, and consequently the reflexive pronoun *–se* had to be placed before the verb. In itself, this syntactic situation can be predicted and formalized in the reference graph. However, since the verb has an auxiliary (*ter*), this is placed between the clitic and the verb. For that reason, the expression cannot be matched. A similar situation occurs in the case of the intrinsically negative expressions: we also considered that the negation adverb (*Neg*) is part of the frozen expression and that it should be matched by the system:

Ext484626(eco,95a): É que advogado [...] **não** *pode* **brincar em serviço** [...] ('a lawyer [...] cannot play while at work')

In the example above, the auxiliary verb *poder* ('can') occurs between the negation adverb and the main verb, a situation that is not described in the reference graph. This prevents the frozen expression from being matched by the system.

As a result of the decision to enforce the matching of all elements of the idiomatic expression, some variants of the same lexical entry had to be entered twice in the matrices. This is the case of *armar*(*-se*) *ao pingarelho* ('to presume, to take liberties'), an intrinsically reflexive construction that can also be used without the reflexive pronoun. Also, rare ambiguous frozen sentences (e.g. *ir para a rua* 'to be sacked', 'to be expelled (from class)', 'to demonstrate/protest') were duplicated in the matrices, which accounts for some high false positives in Table 1. These duplicated entries do not affect overall results.

Finally, in some cases, the graph correctly recognizes the target sequence but this is only part of a longer expression. This is the case of the expression *cortar nas despesas* ('to cut expenses'), where the noun *despesas* ('expenses'), while being frozen wit the verb *cortar* ('to cut'), can be freely modified. In some cases, these combinations may even constitute compound nouns, mostly from economy in general or from public accounting: *despesas de investimento* ('investment expenses'), *despesas públicas* ('public expenses'), *despesas sociais* ('social expenses'), *despesas correntes* ('current expenses'), *despesas salariais* ('salary expenses'), *despesas de funcionamento* ('general expenses'), *despesas orçamentais* ('budget expenses'), etc. We considered, then, that the matching was only partially achieved and considered those cases as false negatives (silence).

5. Conclusions and future work.

The results presented and discussed above, while yet preliminary since drawn from a small sample of frozen sentences, may already allow us to produce some (tentative) conclusions. Obviously, these are to be confirmed on a large-scale investigation of the entire lexicon-grammar of frozen sentences. Firstly, most frozen sentences can be easily matched by using the simple (but already quite sophisticated) finite-state methods here applied. The average precision (P=0.94) and recall (R=0.88) are both similar and high. Secondly, as expected, different types of insertions prevent the correct matching of these expressions. For the most part, these are free adverbials and modifiers; auxiliary verbs also form a significant part of the unmatched strings. However, results show that for the most part, idiomatic expressions tend to appear without spurious elements inserted, perhaps to become more easily detected by the reader as idioms. In view of this, purely statistical methods (as those discussed by Manning & Schütze 2003; see also Jurafsky and Martin 2000) are likely to produce similar quality results, at least for expressions that are sufficiently frequent in corpora (the data sparseness problem). However, average frequency is so low that it may not be sufficient to render those methods effective or their results linguistically adequate. Therefore, the data gathered here may also be used to validate/calibrate those methods. The fragments of the

corpus with both the correct and incorrect matches, along with the unmatched extracts, constitute a language resource that may be used to improve the existing reference graphs and lexicon-grammar matrices, as well as an embryo of a training corpus for the validation or calibration of methods statistical methods. Finally, we expect that a significant reduction of silence (1-R=0,12) might be achieved if linguistic data here presented could be used after a preliminary, shallow text parsing. This will constitute the mater for future work.

References

BAPTISTA J. (2004), «Compositional vs. Frozen Sequences», in *Papers* presented at the *Lexicon-Grammar Workshop*, Beijing, October 2004, in LAPORTE E. and CHENG TING-AU (Eds.) *Journal of Applied Linguistics – Special Issue On Lexicon-Grammar*: 81-92, Institute of Applied Linguistics, Beijing. [in Chinese]

BAPTISTA J., CORREIA A. and FERNANDES G. (2004), «Frozen Sentences of Portuguese: Formal Descriptions for NLP», in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, EACL'2004: 72-79, July 26, 2004, ACL: Barcelona, Spain.

FERNANDES G. and BAPTISTA J. (*in print*), «Frozen sentences with obligatory negation: linguistic challenges for natural language processing». in *Cadernos de Fraseoloxía Galega*, nº especial, Proceedings of the International Conference on Phraseology and Paremiology, Santiago de Compostela, Espanha, Universidade de Santiago de Compostela, September 19‗22, 2006.

BAPTISTA J., CORREIA A. and FERNANDES G. (2005), «Léxico Gramática das Frases Fixas do Portugués Europeo», in *Cadernos de Fraseoloxía Galega*, 7: 41-53, Xunta de Galicia/Centro Ramón Piñero para a Investigación en Humanidades, Santiago de Compostela.

FERNANDES G. (2007), *Léxico-Gramática das Frases Fixas do Português Europeu – Construções Intransitivas*, Masters' thesis, Univ. Algarve, Faro.

GROSS M. (1982), «Une classification des phrases "figées" du français», in *Revue Québécoise de Linguistique*, 11 2: 151-185, UQAM, Montréal.

GROSS M. (1989), Les Expressions Figées, Une description des expressions françaises et ses conséquences théoriques, RT nº 8, PRC-IL, Université Paris 7/LADL, Paris.

GROSS M. (1996), «Lexicon-Grammar», in BROWN K. and MILLER J. (eds.), *Concise Encyclopaedia of Syntactic Theory*: 224-259, Pergamon Press, Oxford.

HARRIS Z. S. (1991), *A Theory of Language and Information, A Mathematical Approach*, Clarendon, Press Oxford.

JURAFSKY D. and MARTIN J. H. (2000), *Speech and Language Processing*, Prentice Hall, New Jersey.

MANNING CH. and SCHÜTZE H. (2003), *Foundations of Statistical Natural Language Processing*, MA: MIT Press, London/Cambridge.

PAUMIER S. (2002), *Unitex - Manuel d'utilisation*, IGM, Univ. Marne-la-Vallée, Paris.