Linguistic clues for corpus-based acquisition of lexical dependencies

Cécile Fabre, Didier Bourigault
ERSS - UMR 5610
Université de Toulouse Le Mirail, 5 allées A. Machado, 31058 Cedex, France.
{cfabre,bourigault}@univ-tlse2.fr

1 SYNTEX: a tool for the extraction of lexical dependencies

Disambiguation of prepositional phrase attachment is a crucial issue for all NLP applications that need textual resources enriched with syntactic knowledge. We are confronted to this problem in the process of designing SYNTEX, a shallow parser specialised in the extraction of lexical dependencies (such as adjective/noun, or verb/noun associations) from French technical corpora. These word-to-word associations will be used as material for the construction of semantic classes on a distributional basis. In this context, the first step towards the automatic discovery of such dependencies is to determine to which word a preposition must be attached. For example, in the phrase *disséquer le plateau rocheux en chevron*, taken from a corpus in the domain of geomorphology[1], the preposition *en* may potentially be attached to any of the three words *disséquer*, *plateau*, *rocheux*, as verbs, nouns or adjectives (and also adverbs) may govern a prepositional phrase.

As lexico-syntactic information is part of what we want to extract from the text, we cannot rely on prior lexical resources: it is our belief, based on in-depth studies of corpora from technical domains, that words exhibit idiosyncratic uses from one domain to the other, not only at the semantic level, but also regarding their syntactic properties. As a consequence, our parser relies as much as possible on corpus-based information to solve the ambiguities of syntactic attachment. Our second claim is that such disambiguation process can be performed on the basis of linguistic clues, with only limited use of statistical measures.

In this paper, we describe a method which relies on the search for linguistic indications to perform syntactic disambiguation. Our parser is based on two main ideas: first, the detection of unambiguous contexts is used as a starting point for the processing of ambiguous contexts. Second, the notion of productivity (more reliable than the simpler notion of frequency) is proposed to assess the strength of a word/preposition association: we define the productivity of a (word,preposition) pair as the ability of the word to appear with this preposition in various contexts. These two basic ideas are refined with the help of further linguistic clues. In particular, we use morphological information and we take into account semantic similarity between the words to evaluate the likelihood of the association between a word and a preposition.

We first describe the modules of SYNTEX which perform prepositional phrase attachment. We present the various information that the analyser exploits to resolve ambiguity, focusing on the notion of productivity. We then analyse the linguistic relevance of this empirical notion of productivity: is it feasible to use productivity measure as a means to differentiate between various levels of lexical relations, and particularly to draw a frontier between arguments and non-arguments? The paper will present the first results in favour of this hypothesis.

2 Description of the strategy for PPs attachment

Prepositional attachment resolution is usually considered as the first procedure that permits the delimitation of phrases for automatic parsing (Brent 1993, Basili et al. 1999). This is crucial for NLP applications such as text retrieval or information extraction (Grefenstette 1994). Our objective is also to provide data about lexico-syntactic relations between words that will help to construct semantic resources. Classes of words can be defined on the basis of their distributional properties, following the propositions of Harris (Harris et al., 1999).

For example, in the geomorphology corpus, we are able to identify a set of four nouns - *alluvions*, *sable*, *dépôts*, *cendres* - all sharing at least two lexico-syntactic contexts with each other:
- same *V PREP N* contexts:

[1] All the examples in the paper are extracted from this geomorphology corpus. We are grateful to Danièle Candel (INALF) who has made it available.

disparaître sous (*des alluvions, du sable, des dépôts*)
enfouir sous (des alluvions, les dépôts)
creuser dans (les alluvions, le sable, les cendres)
tailler dans (des alluvions, des cendres)

- same *N PREP N* contexts:

manteau de (alluvions, dépôts)
banc de (alluvions, sable)

These four nouns indeed appear to be semantically close, all denoting some sort of sediment. The extraction of syntactically and lexically related pairs of words is therefore a means to discover corpus-specific classes with semi-automatic techniques. Research initiated in French towards this goal has suffered from the lack of automatic parsers for this language: (Habert et al. 1996), (Assadi and Bourigault 1995) were both taking as input the results of a NP parser, LEXTER (Bourigault, 1994), and were not able to exploit data concerning verbs. SYNTEX provides an extension to LEXTER by identifying all types of lexical dependencies involving verbs, nouns and adjectives.

2.1 An inductive approach

The identification of lexical dependencies in texts may be performed by different techniques. The first consists in the projection of a-priori lexical resources. There are major well-known problems with this approach: firstly, such resources, which should not only contain information related to argument structure but also deal with complementation and modification phenomena, are simply not available. The process of constructing such lexical bases is long and costly, but most of all, constructing a-priori resources is an endless task since idiosyncratic uses of words are found in corpora. Work on technical corpora demonstrates that such texts exhibit a great variety of lexico-syntactic properties regarding how words associate with each others. Of course, corpora contain also highly predictible data: for example, we find that in the geomorphology domain, the verb *débiter* associates with the preposition *en* without a determiner, from occurrences such as *débiter en blocs, en chicots, en granules, en lamelles*, etc. This construction is described in any French dictionary. But corpus observations also show many configurations that are totally idiosyncratic. This is especially true concerning noun modification. To illustrate this point, we can point among numerous examples to the existence of a lexical pattern of the form *N pour det N* with the non argument-taking, non relational, headnoun *salle* (*room*), (in: *salle pour l'étude des minéraux, salle pour la détermination des argiles*). In this corpus, nouns very often appear with the prepositions *en* (*vallée en canyon, section en cluse*) and *à* (*section à méandre, cirque à source*), which are much more frequently used as postverbal prepositions in other types of texts. Since these properties do not concern argument selection and are very unstable from one corpus to another, such data cannot be listed to be used on any type of texts. Another argument against the use of prior resources is the variety of prepositional attachment patterns that exist for a single lexeme. If subcategorization, complementation and modification phenomena (Grimshaw 1990) are all taken into account, it appears that many verbs and nouns can potentially associate with a great range of prepositions. This is the case even in a limited domain. For example, the verb *accumuler*, besides the transitive construction, gets constructed with six different prepositions, corresponding to locative and instrumental interpretations (e.g. *dans le creux, derrière l'obstacle, sur le névé*). As a consequence, the listing of all complementation alternatives – if possible at all – will not in itself be suffichient to reduce ambiguity.

Alternatively, other approaches are proposed to learn this information from corpora in order to avoid the preconstruction of lexical resources. Our strategy is in line with research adopting a distributional methods to solve the ambiguities of syntactic analysis. (Brent 1993) was the first to take into account lexical association measures in texts to identify verbs' arguments. A similar approach has been adopted by (Hindle and Rooth 1993) or (Manning 1993), who defined methods to discover verbs' subcategorization frames in texts. More recently, (Federici et al. 1999) combine the *tabula rasa* approach and inductive technics for the parsing of Italian texts. Similarly, our parser uses information extracted from the whole corpus to infer local decisions, exploiting lexical redundancy in technical corpora to acquire patterns of prepositional attachment. But the novelty of our approach can be characterized as the conjunction of three options:

- SYNTEX deals not only with subcategorization but also with any type of lexical dependency involving prepositional attachment,
- it maximizes the use of linguistic clues, limiting the development of statistical methods to perform disambiguation,

- it is mainly based on the productivity measure, a key notion to reach a decision in ambiguous texts.

2.2 Disambiguation rules

SYNTEX is based on inductive learning of complementation properties. The parser looks for triplets (*governor*, *preposition*, *governee*) linked by a dependency relation[2]. This relation may correspond to subcategorization patterns - as in the triplets (*pénétrer*, *dans*, *pore*) or (*accessible*, *à*, *bâteau*) - or it may illustrate non argumental associations, such as verb + circumstant – (*déplacer*, *à*, *vitesse*) -, noun + expansion (*côte*, *à*, *fjord*), etc. Learning is performed in two steps: first, properties of lexical associations are acquired from unambiguous contexts; second, they are used as indications to solve ambiguous cases throughout the corpus. This strategy is very close to the approach described in (Federici et al. 1999) for shallow parsing of italian. More generally, it is not uncommon to see parsers exploiting unambiguous contexts to limit the complexity of the analysis. The novelty of SYNTEX is in the criteria used to identify the triplets likely to correspond to genuine dependency relations.

2.2.1 Detection of unambiguous contexts

Learning is first performed by detecting attachment zones in the corpus. Such zones are delimited to the right by a preposition and to the left by a frontier, which may not, or may rarely be crossed over by prepositional attachment. Such frontiers are punctuations, verbs, prepositions other than *de* and *à* or typographical items such as parentheses. In the following examples, these zones are enclosed within brackets and underlined.

i. On tend de plus en plus à]insérer cette science dans] une géographie physique globale

ii. transformations du relief des versants], conséquences des actions de l' homme sur] le sol

Obsviously, most frontiers are not entirely reliable. The prepositional link can be established over interpolated verbal phrases, over prepositions, etc. as in the two following examples, where the preposition and its governor are underlined:

iii. La puissance est utilisée , on s'en souvient , en partie par le transport de la charge

iv. On peut mesurer la vitesse d'infiltration d'une goutte d'eau déposée sur la roche par sa disparition de la surface

The segmentation module is therefore, for the moment, very rudimentary, but it enables us to considerably limit the complexity of the attachment procedure. As a first approximation, we have measured the silence due to "preposition jump" (a preposition finds its governor over a preposition other than *de* or *à*) as around 5% on one of our technical corpora, which is a relatively small proportion.

Within these zones, all lexical units are viewed as potential governors of the preposition. Example *ii* contains three nominal candidates (*conséquences*, *actions*, *homme*), corresponding to three potential triplets: (*conséquence*, *sur*, *sol*), (*action*, *sur*, *sol*), (*homme*, *sur*, *sol*). One condition must be met for these triplets to be used in the learning process: they have to be found in unambiguous contexts - containing no other potential governor for the preposition. For example, the triplet (*glisser*,*sur*,*bord*) is extracted from the following unambiguous context:

v. des nappes de gravité]glissent sur] le bord des surélévations

2.2.2 Acquisition of reliable dependency relations

Unambiguous cases are thus the starting point of the parser. The purpose of the acquisition module is then to use unambiguous contexts as clues for the resolution of ambiguous cases. Yet, we cannot consider all information found in unambiguous contexts as reliable. Several criteria are taken into account. Given a triplet (Gvr, Prep, Gvee) found in an ambiguous context, it is considered as a possibly genuine lexical relation if:

---

[2] The *governor* is the word governing the prepositional phrase. It belongs to the categories verb, adjective or noun. The *governee* is the word governed by the preposition. It may be a noun (as in *insérer dans une géographie*) or an infinitive (as in *tendance à former*). Both appear in lemmatized form in our results.

- Rule 1[3]: The same triplet has been found in an unambiguous context.

Example: in the ambiguous context *indique une légère tendance à l'enfoncement*, where the preposition may be governed by the verb *indique* or the noun *tendance*, the latter is considered as the more probable governor, because the triplet (*tendance, à, enfoncement*) has been found in an unambiguous context.

- Rule 2: The pair (Gvr, Prep) is productive.

Productivity of a (Gvr, Prep) pair equals the number of different governees with which the pair (Gvr, Prep) occurs in the corpus. A word is considered as productive with a given preposition, if it combines with at least two different governees. For example, given the following unambiguous contexts found in the corpus, it appears that the verb *disséquer* is productive with the preposition *en*, with a productivity of 5.

disséqué parfois en récif *disséquer en récif* *disséqué en dents de scie* *disséquer en terrasses* *disséqués en bois de renne* *disséquée en chevron*	→ *disséquer en*	*récif* *dent* *terrasse* *bois* *chevron*	*prod=5*

Figure 1: productivity of a (governor, preposition) pair

The productivity measure allows us to assess the likelihood of a word being used as a governor of a preposition on a more reliable basis than the simpler frequency measure. High productivity indicates that the governor regularly associates with this preposition, in various occurrences. To illustrate this point, we can oppose three cases:
- unfrequent and unproductive association of a word and a preposition: the string *annuler en général* is found only one time in the corpus. In this case, the occurrence corresponds to the association of a verb and an adverbial phrase (meaning *generally*), and it does not indicate that the verb is constructed with the preposition *en*.
- frequent but unproductive association: the strings *an en moyenne, croûtes en Afrique, allonger dans la direction* are each found three times in the corpus. These findings may correspond to genuine lexical dependencies but they cannot be used to infer a regular association between the head and the preposition.
- productive (and necessarily frequent) association: the pair (*disséquer, en*) seems to indicate a regular relationship between the verb and the preposition because they appear together in various contexts.

- Rule 3: A word morphologically linked to Gvr is productive with the preposition.

The exploitation of such morphological links is useful because of the richness of the morphological system in French. Relations between subcategorization properties of verbs and argument-taking nominals are not systematic. Yet, they are sufficiently frequent to provide a clue for disambiguation. We use a lexical resource, *Verbaction*[4], which provides an inventory of process nominals that are constructed on a verbal base. This information is used to solve ambiguities such as the following:

vi. la formation à blocs ayant soliflué plus rapidement,]*avec glissement rapide sur*] *l' arène lente*

The only information that the corpus provides on the two potential pairs (*glissement, sur*) and (*rapide, sur*), is that the verb *glisser*, morphologically related to the noun *glissement*, has been found as governor of the preposition *sur* in eight unambiguous contexts (example *v* is one of them).
This indication is used to make the hypothesis that *glissement* may be a governor of the preposition *sur*.

- Rule 4: A word semantically linked to *Gvee* has been found as governee of the pair *(Gvr, prep).*

A word is considered as semantically related to another, if both have at least two governors in common in the corpus. They share some distributional properties. They are also said to be semantic neighbours.

---

[3] The numbering of the rules does not indicate priority of application.

[4] This morphological resource has been compiled by Nabil Hathout at the National Institute of French Language (INaLF).

vii. nous] terminerons par quelques notes sur] la morphologie de la lune et de Mars .

The noun *notes* has been found as governor in an unambiguous context with a semantic neighbour of the noun *morphologie*, namely *forme*. The proximity of the two nouns *morphologie* and *forme* has been established on account of their sharing two unambiguous contexts:

viii. compréhension de + det + (forme, morphologie)

ix. s'intéresser à + det (forme,morphologie)

The construction of semantic classes, which is the objective of our work, is thus also sketched out during the syntactic analysis to provide indications for prepositional attachment.

2.2.3 Resolution of ambiguous contexts

Ambiguous cases are solved by using this combination of clues. In the following example, rules 1 (same triplet) and 2 (productivity) are used to choose between the three potential governors.

x. L'érosion a disséqué le plateau rocheux en chevrons.

The verb is productive (prod=5) and it has been found with the same governee in an unambiguous context. The noun *plateau* has been found only one time with the preposition *en* in an unambiguous context, with another governee (*des plateaux en interfluve*); the adjective *rocheux* is not found with this preposition.

This resolution module is currently under development. At the moment, precision is 86%, which is very satisfactory, but the recall measure is only 60%. These results have been obtained by comparing the results of SYNTEX to prepositional attachment manually performed on several thousands occurrences in three different corpora. This relatively low recall corresponds to two situations: in the first case, no potential governor has been found in the attachment zone. Recall must therefore be increased by improving the segmentation module, through the definition of more flexible frontiers for the attachment zones. In the second case, no indication could be used to choose between several potential governors. It is due to lack of corpus evidence, so we must consider developping a default strategy, or using some amount of prior knowledge when all corpus-specific information has been exploited.

3 Productivity: a measure that helps to detect different levels of lexical dependency?

A further question is at issue in this experiment: what levels of linguistic information are we able to point out from the observation of lexical associations in corpora? More precisely, we want to know if the strength of the association between a governor and the preposition, that we have measured in terms of productivity, can be used to describe the type of relation - subcategorization or adjunction - that holds between the two words. (Brent 1993) claims that there is a connection between frequency of occurrence and type of complementation: two words are more frequently associated if they are associated by a grammatical relation. According to this view, heuristics based on frequency of cooccurrence should therefore enable us to make this fundamental distinction between arguments and adjuncts. On the contrary, (Basili et al. 1999) think that this distinction cannot et should not be made by automatic means, because adjuncts equally contribute to the verb semantics and are as regularly associated with the verb as arguments. Observations made on corpora indeed show that the frontier between the two types of complementation is very difficult to draw, even when we want to manually determine prepositional attachment. Yet, we wanted to know if it was feasible to go beyond the simple diagnosis of prepositional attachment, and to rely on the productivity measure to try and detect different types of prepositional phrases. The last part of this paper is devoted to the presentation of the first results regarding this issue.

3.1 Variety of lexical dependencies

In the previous sections, we have encountered examples of prepositional attachments that illustrate the diversity of semantic relations between a governor and a PP. SYNTEX resolves prepositional attachments corresponding to different types of lexical associations, namely:

- an argument-taking element with its argument
  - verb: *s'enfoncer dans + det +* (*alluvions,eau, fond, surface*)
  - adj: *sujette à + det +* (*bouleversement, gel,variation, émiettement*)
  - noun: *déversement dans + det +* (*bassin, cuvette*)

- an argument-taking element with a complement
  verb: *disparaître dans* (*le bassin,le gouffre,le lac,le puits*)
  noun: *ruissellement en* (*nappes, rigoles, films*)
- non argument-taking element with a complement
  adj: *active dans + det + (la baie, la zone)*
  noun: *équilibre entre (forces,puissances)*
  noun: *vallée à (flancs,replats)*

All these relations prove to be useful for the construction of semantic classes: words may of course be grouped because they share arguments, or they may be grouped because they are arguments of the same words. But non-argumental relations are also useful for the construction of homogenous sets of words. To illustrate this point, we can point to the fact that nouns that appear in the same list of complements in the previous examples are closely related, such as the three nouns *nappes* (*sheet*), *rigoles* (*rills*), *films* (*films*), or the four nouns *bassin* (*basin*), *gouffre* (*chasm*), *lac* (*lake*), *puits* (*well*).

As a consequence, it would be very useful to propose some clues in order to differenciate between these types of dependencies. One track that we are currently following consists in the measurement of different types of productivity. Our objective is to find criteria to differentiate between argument relations and other levels of complementation relations. With this in mind, we have tried to compare two types of productivity: the productivity of governor-preposition pairs, exemplified so far, and the productivity of preposition-governees pairs. In the latter case, saying that the governee is productive with a preposition means that it occurs in the scope of various governors. For example, the pair (*par*, *mer*), separated by a determiner, is productive because it occurs under the government of six different verbs:

battues
coupé
déposé + *par la mer* *prod=6*
envahie
occupée
recouvert

Figure 2: productivity of a (preposition,governee) pair

We have compared the lexical information that is acquired from these two different criteria. Our first conclusions are illustrated by the observation of two prepositions: *sur* and *à*, in verbal and nominal contexts.

3.2 Verbal complementation: *V sur det N* phrases

We compare two lists: the first one (figure 3) is made up of 15 verbs that are found to be most productive as governors of the preposition *sur* (+ determiner) in our corpus. The first line of the table indicates that the verb *reposer* has been found before the preposition *sur* in unambiguous contexts with 23 different right contexts, which are all nouns (*reposer sur le banc*, *reposer sur la couche*, etc.).

governor		*governees*	*prod*
reposer		banc,couche,critère,critérium,dos,fond,galet,horizon,lit,marne,mesure,plancher,précédent,pétition de principe,reg,remarque,restitution,roche,sable,socle,substratum,surface,étude	23
renseigner		climat,condition,constitution,degré,direction,intensité,morphogenèse,mouvement,nature,provenance,rapport,relief,sens,topographie,valeur,évolution	16
situer		bord,crête,dôme,emplacement,face,front,ligne,passage,plan,trajet,versant,équateur	12
emporter		accumulation,altération,amplitude,creusement,exportation,mode,proportion,roche,élément,érosion	10
rencontrer	*sur + det*	croupe,côte,flanc,granit,niveau,paroi,pente,revers,roche	9
trouver		bord,emplacement,glacier,mer,partie,pente,planète,rivière,roche	9
établir		année,bloc,couverture,fond,marne,permien,pénéplaine,socle,surface	9
appuyer		chronologie,connaissance,couverture,interstratification,pierre,pointement,étude,îlot	9
glisser		bord,couche,flanc,fond,neige,pente,plaque,substratum	8
opérer		arbre,modèle, pente, sol, échantillon, élément	6
poser		glacier, planèze, plaque, pupitre, socle, sol	6
tomber		glacier, interfluve, planète, région, sol, versant	6
fonder		distinction, glacio-eutatisme, granulométrie, inégalité, loi, repère	6
insister		composition, fixisme, inefficacité, ouvrage, rôle, épigénie	6
localiser		bordure, contact, côte, emplacement, retombée, versant	6

Figure 3: 15 most productive verbal governors with the preposition *sur*

The second list (figure 4) is made up of the 15 nouns that are found most productive (productivity>=4) as governees of the preposition *sur* in our corpus after verbs. The first line of the table indicates that the noun *fond* (*bottom*) has been found after the preposition *sur* with 15 different right contexts (*affleurer sur le fond*, *ancrer sur le fond*, etc.).

governors		*governee*	*prod*
affleurer,ancrer,arrêter,balayer,concrétionner,frotter,glisser,reposer,rouler,stratifier,transporter,traîner,triturer,élever,établir		*fond*	15
accélérer,arriver,cascader,condenser,descendre,disperser,déboucher,faire,former,glisser,influer,observer,opérer,rencontrer,trouver		*pente*	15
affleurer,descendre,exercer,faire,lire,localiser,paralyser,passer,remonter,situer,tomber,travailler,épandre		*versant*	13
déposer,développer,emporter,faire,fixer,manifester,rencontrer,reposer,trouver		*roche*	9
désintégrer,obtenir,reposer,réfléchir,voir,égaliser,épancher,établir	*sur + det*	*surface*	8
agir,basculer,opérer,poser,rouler,séjourner,tomber		*sol*	7
déferler,exister,localiser,rencontrer,retrouver,régulariser		*côte*	6
coller,exercer,plaquer,produire,rencontrer,réfléchir		*paroi*	6
compléter,effectuer,indiquer,pouvoir,repérer,étudier		*terrain*	6
déplacer,manquer,trouver,échelonner,étendre		*partie*	5
effectuer,glisser,situer,trouver		*bord*	4
affleurer,glisser,jouer,rencontrer		*flanc*	4
aligner,épancher,étaler,étendre		*kilomètre*	4
couler,peser,reposer,étaler		*lit*	4
carboniser,détruire,situer,submerger		*passage*	4

Figure 4: nouns most productive as governees with the preposition *sur*

If we compare the two tables, we see that verbs that are most productive with the preposition *sur* all instantiate one of these two cases:

- the PP headed by the preposition *sur* is subcategorized by the verb (*reposer*, *renseigner*, *situer*, *emporter*, *appuyer*, *opérer*, *poser*, *fonder*, *insister*),
- the PP denotes a localization that is expected given the semantics of the verb, which are all spatial verbs (*rencontrer*, *trouver*, *établir*, *glisser*, *tomber*, *localiser*).

In the second list, we see that all PPs (*sur le fond*, *sur la pente*, *sur le versant*) tend to behave as autonomous phrases, conveying spatial information.

3.3 Noun complementation: *N à N* phrases

The second illustration regards the study of simple nouns, with no morphological link to a verb, related by the preposition *à*. We wanted to know whether different types of *à N* expansions would emerge by the application of the two productivity measures. The parser extracted 50 nominal governors and 54 nominal governees with the preposition *à*. A subset of these results (productivity >= 3) is presented in figures 5 and 6.

governor	prepositional link	governees
carte	*à*	1.10 000, 1.200 000,1.80 000
cas	*à*	Java, Montserrat, Nantasket
cas	*à + dét*	lahar, pays; pays-bas
craie	*à*	bélemnite,micraster,silex
crête	*à*	Porolithon,cheminée,clocheton
côte	*à*	falaise,fjord,plage,ria,skjär,structure
kilomètre	*à + dét*	Nord,dizaine,pôle
méthode	*à + dét*	potassium,strontium,uranium
roche	*à*	diaclase,feldspath,feldspathoïde,grain
roche	*à + dét*	extérieur,minéral,soleil
région	*à*	cuesta,nappe,permafrost,plateaux,saison,sous-sol
zone	*à*	cristal,pergélisol,pluie

Figure 5: (noun, *à*) productive pairs, with or without a determiner

governors	prepositional link	governees
ablation; plage; terrasse	à + dét	*aval*
degré; maximum; éprouvette	à + dét	*dessous*
Ouest; actif; haut; rigole	à	*droite*
calcaire; ensemble; granit; granite; grès; leucogranit; roche	à	*grain*

gneiss; granit; micaschiste	à	*mica*
altitude; base; forme	à	*peine*
dépression; glacis; grès; zone	à + dét	*pied*
courant; glacis; plan	à + dét	*sens*
affaire; ciselure; face	à + dét	*surface*
concavité; côte; côté; pente	à + dét	*vent*

Figure 6: (*à*, noun) productive pairs, with or without a determiner

First, a few remarks about these two tables. The results are not as good as those found on verbs, certainly because this second type of dependency patterns is more unstable, less recurrent than subcategorization patterns. Both tables contain erroneous triplets, illustrating some defects of the analysis. Some errors are due to the tagging procedure (*actif*, *haut* should have been tagged as adjectives). Others also come from the non recognition of verbal phrases, such as *être le cas*: the noun *cas* cannot be considered as an autonomous governor. Despite these problems, if we observe the (preposition, governee) pairs, we can see that the two tables exhibit rather different lexical relationships. Productive governees appear in three structures:

- prepositional locutions (*à peine*, *au sens*)
- circumstancial PPs (*à l'aval*, *au dessous*, *à droite*, *au pied*, *à la surface*)
- in only 2 out of 10 cases, *N à N* compounds (*à mica*, *à grain*).

As for productive governors, they are heads of N *à* N compounds in 9 out of 14 cases. In these cases, the prepositional expansion denotes a qualification (according to the classification made by (Cadiot 1997)).

These first observations indicate that the productivity of the governor is an indication that we are dealing with PPs which are cohesive with the head noun or verb. Conversely, a high productivity of (preposition, governee) pairs is rather an indication for autonomous PPs.

4 Conclusion

This paper reports the results obtained by our parser, SYNTEX, in the task of prepositional attachment disambiguation. The attachment strategy, which does not limit the focus to subcategorization patterns but processes any type of prepositional dependency involving verbs, nouns and adjectives, is based upon a combination of linguistic clues, namely: productivity of a (word, preposition) pair, evidence about morphologically related words or about words showing similar distributional behaviours. These first results indicate that the productivity measure, which finds echoes in other areas of linguistic research (see for example (Bayyen 1996) in morphology), is a very reliable criterion to assess the likelihood of a prepositional attachment and to extract lexical patterns form texts. Further work is needed to improve recall, and particularly to find further indications to make a decision between several potential governors when the system lacks corpus evidence.

In this paper, we have also reported our first observations concerning the use of the productivity measure in the differenciation of arguments and non-arguments among prepositional phrases. The opposition between cohesive and autonomous phrases, that emerges from the data we have presented concerning both verbal and nominal patterns, must be further investigated. But these first results certainly indicate a contrast between the PP phrases that are detected by these two productivity measures. Our next objective is to integrate this distinction in the disambiguation strategy.

5 References

Assadi H, Bourigault D 1995 Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation des connaissances. In *Actes des 3èmes Journées internationales d'analyse des données textuelles (JADT95)*, Rome.

Basili R, Pazienza MT, Vindigni M 1999 Adaptive Parsing and Lexical Learning, in *Proceedings of VEXTAL'99*, Venise.

Bayyen RH, Renouf A 1996 Chronicling the times: productive lexical innovations in an english newspaper. *Language*, 72(1): pp 69-96

Bourigault D 1994 *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*. Unpublished PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Brent M 1993 From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19(2): 243-262.

Cadiot P 1997 *Les prépositions abstraites en français*. Paris, Armand Colin/Masson.

Federici S, Montemagni S, Pirrelli V, Calzolari N 1998 Analogy-based extraction of lexical knowledge from corpora: the SPARKLE experience. In *Proceedings of the first international conference on Linguistic Resources and Evaluation*, Grenada, pp 75-82.

Grefenstette G 1994 *Exploration in Automatic Thesaurus Discovery*. Londres, Kluwer Academic Publishers.

Grimshaw J 1990 *Argument structure*. Cambridge, Massachussets, the MIT Press.

Habert B, Naulleau E, Nazarenko A 1996 Symbolic word-clustering for medium-size corpora. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING*, Copenhague, pp 490-495.

Harris Z, Gottfried M, Ryckman T, Mattick Jr P, Daladier A, Harris T, Harris S *1989 The Form of information in science, analysis of immunology sublanguage*. Kluwer Academic Publisher, Dordrecht.

Hindle D, Rooth M 1991 Structural Ambiguity and Lexical Relations. *In Proceedings of the 29th meeting of the association for Computational Linguistics, ACL*, Morristown, pp 229-236.

Manning D 1993 Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31th meeting of the association for Computational Linguistics, ACL*, Columbus, pp 235-242.