Extension of a grammar of French determiners

Éric Laporte[1]
Université Paris-Est

Abstract

Assessments of the quality of parts of syntactic grammars of natural languages are useful for the validation of their construction. We extended a grammar of French determiners that takes the form of a recursive transition network and evaluated its quality. The result of the application of this local grammar gives deeper syntactic information than chunking or information available in treebanks. We performed the evaluation by comparison with a corpus independently annotated with information on determiners. We obtained 85% precision and 93% recall on text not tagged for parts of speech.

Keywords: determiner, syntax, grammar, local grammar, evaluation, annotated corpus.

1. Introduction[2]

Syntactic-semantic grammars of natural languages are complex objects and their construction takes many years. Therefore, it is desirable to assess the quality of parts of such grammars and to control their evolution before it is complete. In this paper, we report the extension and evaluation of a partial syntactic-semantic grammar of French: a grammar of determiners, including complex determiners and combinations of determiners. This grammar neglects dependencies between the determiner and the noun (N). It takes the form of a recursive transition network (RTN). As compared to chunking, the syntactic information obtained by the application of the grammar is deeper, since the grammar describes complex determiners which may contain several chunks. The output of the parser was compared to a corpus independently annotated with information on determiners.

This article is organised as follows. The next section surveys related work. In section 3, we describe the grammar of determiners. Section 4 reports how the grammar was evaluated and analyses the results.

2. Related work

In recent campaigns of evaluation of syntactic grammars (Paroubek *et al.* 2006), each grammar was assessed globally. Evaluation consisted in comparing the output of the parser to a treebank, and no evaluation of separate parts of grammars was organised. However, parts of a manually constructed grammar have not necessarily the same author or the same quality,

[1] Institut Gaspard-Monge (IGM), Université Paris-Est, eric.laporte@univ-mlv.fr

[2] This work has been supported by CNRS and by Senior Planet Co.

and are not necessarily built at the same time. Therefore, it is also desirable to assess the quality of parts of a grammar and to control their evolution during their construction.

Most partial grammars[3] are grammars for NE recognition or for chunkers. Others are components of deep syntactic grammars. Local grammars are partial grammars taking the form of RTNs and available in graphical form (cf. Fig. 1). The objective of a local grammar designed as a component of a deep syntactic grammar is (i) to represent some set of syntactic constructs with maximal recall, and (ii) to resolve syntactic ambiguity, but, in general, only when this is possible without exploring the context of these constructs[4]. Thus, precision is less relevant than recall in the assessment of a component of a syntactic grammar.

Very few quantitative data about the coverage of such local grammars are presently available[5]. We provide such data referring to a grammar of French determiners.

An alternative approach to local grammars is the use of features. For instance, (Hockey and Mateyak 2000) propose a set of features in a tree-adjoining grammar (TAG) for the recognition of complex determiners in English[6]. Such descriptions are less readable: to check whether a sequence is recognised by the grammar, you have to simulate the behaviour of a TAG parser. With a local grammar, you have to read while you follow arrows.

Another alternative is the training of a probabilistic model on an annotated corpus, as has been done for shallow parsing (Sha and Pereira 2003) and named entity recognition (Li and McCallum 2003). However, these techniques are less compatible with the introduction of syntactic-semantic information. In addition, the annotation of determiners is simplified in available treebanks (cf. section 4), and the delimitation and properties of determiners would be too difficult to infer from raw corpus.

3. The grammar

The grammar is a description of French determiners, including complex determiners and combinations of determiners. We developed it manually from three existing RTNs: two grammars of French determiners (Gross 2001; Silberztein 2003) and a grammar of numerical expressions (Constant 2000). It is freely available in the GraalWeb library[7]. In this section, we delimit the scope of the grammar and report how it was constructed.

3.1. Scope

In language engineering and traditional grammar, determiners are usually viewed as a part of speech, i.e. a morpho-syntactic category of words, rather than as a syntactic notion. This view

[3] See (Laporte 2007) for a survey.

[4] Recall that RTNs are equivalent to *context-free* grammars. In a syntactic grammar, the resolution of syntactic and part-of-speech (POS) ambiguity is ultimately obtained by the combination of all components, and is not the problem addressed by a single component.

[5] (Danlos 2005) claims 97% accuracy on a corpus of about 240,000 words in the discrimination between expletive and anaphoric occurrences of the French pronoun *il* "it". (Silberztein 2003) reports 100% recall on a sample of about 4200 words, in the recognition of determiners in French, but as regards precision, mentions only that it is ‘very low’. (Gross 1998-1999) claims 99.8% precision for the recognition of verb sequences in French, but does not give the size of the evaluation corpus, nor an assessment of recall.

[6] However, constraints between determinative nouns and the morpho-syntactic and syntactic-semantic features of subsequent nouns are not addressed.

[7] http://igm.univ-mlv.fr/~mconstan/library/index_graalweb.html (Constant 2004).

is only a simplification. Determiners behave according to a complex syntax, well described in (Gross 1986 [1977]). Some determiners are employed with prepositions, e.g. in *beaucoup de facteurs* ‘plenty of factors’. Some combine together, as in *les sept pays* ‘the seven countries’. In French, the interaction between the frequent preposition *de* ‘of’ and determiners involves complex rules (Gross 1967).

Some sequences containing a noun phrase behave as determiners of other nouns, as in *restituer une partie des prêts* ‘give back part of the loans’. The sequence *une partie des* ‘part of the’ behaves semantically as a determiner. In addition, the semantic head of the complement of *restituer* ‘give back’ is *prêts* ‘loans’ rather than *partie* ‘part’. With this analysis, the syntactic structure is closer to the semantic structure, which is desirable since most applications of syntactic parsing involve an interpretation of the text. Since most of such sequences comprise a determiner in turn, sequences that behave as determiners are embedded in others. We do consider such sequences as (generalised) determiners, like (Gross 1986 [1977]; Hockey and Mateyak 2000; Silberztein 2003). We refer to nouns such as *partie* ‘part’ by the term ‘determinative nouns’ (*Ndet*).

The scope of the grammar is to describe generalised determiners, defined by (Silberztein 2003) as follows: if each noun phrase is assigned a head noun on syntactic and semantic grounds, the (generalised) determiner of the noun is the sequence from the beginning of the noun phrase to the head noun, excluding the head noun itself and possible adjectives directly attached to the head noun. Thus, in *restituer une partie des prêts* ‘give back part of the loans’, selectional restrictions point to *prêts* ‘loans’, rather than to *partie* ‘part’, as the object of *restituer* ‘give back’; therefore, the determiner of the noun phrase *une partie des prêts* is the sequence *une partie des* ‘part of the’. The scope of our grammar also includes the prepositions *à* and *de* when they introduce the noun group. The sequences described in the grammar are surface forms such as *au*, and not normalized forms such as *à le*. Predeterminers are considered as parts of the corresponding determiners, as *même* ‘even’ in *même les grandes avenues* ‘even the large avenues’, except if they are separated from the determiner by a preposition, as in *même dans les grandes avenues* ‘even in the large avenues’.

However, the grammar does not specify morpho-syntactic agreement in gender and number, either between the determiner and the noun, or between the determiner and other elements of the sentence (e.g. the subject-verb agreement). This exclusion is motivated by the fact that the parser that we used, the Unitex parser (Paumier 2006), does not support unification in its present version. We plan to introduce agreement constraints with the Outilex parser[8], or when the Unitex parser is compatible with unification. Determiners occurring without a head noun are also outside the scope of the grammar. For instance, *plusieurs* ‘several’ can be a syntactic variant of *plusieurs objets* ‘several objects’. In that case, the deletion of the head noun is not accompanied by formal modifications of the determiner, but it is in other cases, e.g. in *beaucoup* ‘many’ for *beaucoup d’objets* ‘many objects’.

3.2. Method of construction of the grammar

The grammar has been developed manually from three existing RTNs (Gross 2001; Silberztein 2003; Constant 2000). We removed from Silberztein’s grammar two elements:

- the constraints involving the countable vs. uncountable feature of nouns, since this feature is absent from available lexicons of French;

[8] Outilex (Blanc and Constant 2006) allows for encoding unification constraints without blowing up the transducer created with these into a large finite-state network.

- gender and number agreement; in Silberztein's grammar, agreement is represented by the existence of 4 versions of the grammar for the 4 combinations of the two genders and the two numbers; this redundancy makes the grammar difficult to maintain.

We introduced into the grammar various elements of Gross' and Constant's grammars[9]. From Gross' grammar, we extracted lists of modifying adverbs, of negative adverbial determiners (e.g. *jamais de* 'never any'), of adjectives that can modify determinative nouns, and of adjectives with properties of determiners (e.g. *premier* 'first'). From Constant's grammar, we extracted the description of physical magnitudes and of approximate numerical expressions. Then we enhanced the grammar with more constructions and more constraints, using the same two approaches as Gross, Silberztein and Constant to construct their grammars: the corpus-based bootstrapping method (Gross 2000) and introspection. For example, we introduced combinations of adverbial determiners such as *un peu de* with adjectival determiners such as *chaque*. We also described constraints between successive determinative nouns, as in *trois sortes de parties de* 'three kinds of parts of'.

We mentioned above that the sequences described in the grammar are surface forms such as *au*, and not normalized forms such as *à le*. However, during the construction of the grammar, we managed all the graphs in the normalized form, and we changed them to the surface form at the end of the construction, because this operation obfuscates considerably the grammar and makes it difficult to maintain. We saved the normalized version so that maintenance operations can be performed on it.

3.3. Classification of determinative nouns

The grammar of (Gross 2001) contains a selection of about 20 frequent determinative nouns: dose entièreté fraction groupe majorité maximum minimum minorité moitié morceau nombre paire part partie portion quantité restant reste sorte total totalité. In addition to this list, it contains numbers: dix, and nouns derived from numbers: dizaine, dixième. The determinative nouns in (Silberztein 2003) are a subset of those present in (Gross 2001). We included all, and a few others such as abondance catégorie classe couple ensemble espèce flopée floraison foison foisonnement foule kyrielle multitude myriade parcelle pléthore pourcentage profusion proportion ribambelle tas type volume. The resulting list is far from exhaustive: (Buvet 1994) mentions a list of about 3000 determinative nouns with a quantifying value, and (Buvet and Lim 1996) deals with other determinative nouns with an aspectual value. However, our list was large enough to demonstrate that it was necessary to classify the determinative nouns before including them into the grammar. For example, these nouns impose different syntactic restrictions on the number of the head noun:

*Voici un (kilo + *groupe) de ce papier,*

on its lexical value or countable feature:

*Voici un grand (nombre + *morceau) de bocaux*
*Voici un grand (*nombre + morceau) de métal*

and on their own determiner:

*Voici trois (groupes + *quantités) de bocaux*

The typology of determinative nouns of (Buvet 1994) was partly adequate for our purposes, but not entirely, because it does not take into account the immediate morpho-syntactic context

[9] We thank Anastasia Yannacopoulou for her valuable contribution to this work.

of the nouns. This typology is based on paraphrases showing how the interpretation of the determinative noun is related to that of the same noun used in other contexts, as in *Ces bocaux forment trois groupes* or *Ce papier pèse un kilo*.

Thus, we devised another classification based on 5 syntactic properties directly exploitable in the grammar. Each property is the acceptability of a sentence type:

voici un Ndet:s (E + Modif) de ce N:s	*Voici une partie de ce papier*
voici Dnum:p Ndet:p de Det N	*Voici trois parties de ce bocal*
voici un Ndet:s de Ncpt:s	*Voici une sorte de bocal*
N_0 *contenir un Ndet:s de Ncpt:p*	*Ce mortier contient une partie de gravillons*
N_0 *être un Ndet:s de combien de Ncpt:p ?*	*C'est un groupe de combien de personnes ?*

In these formulae, *:s* denotes the singular, *:p* the plural, *Modif* nominal modifiers, *Dnum* numeral determiners, and *Ncpt* countable nouns. The following decision tree distributes *Ndet* into 9 classes (the size of each class is given in parentheses):

- *voici un Ndet:s (E + Modif) de ce N:s*
  - *voici Dnum:p Ndet:p de Det N*
    - *voici un Ndet:s de Ncpt:s*
      - N_0 *contenir un Ndet:s de Ncpt:p* *NdetPartie* (6)
      - * N_0 *contenir un Ndet:s de Ncpt:p* *NdetMorceau* (19)
    - * *voici un Ndet:s de Ncpt:s* *NdetMasse* (open)
  - * *voici Dnum:p Ndet:p de Det N* *NdetQuantité* (41)
- * *voici un Ndet:s (E + Modif) de ce N:s*
  - *voici Dnum:p Ndet:p de Det N*
    - *voici un Ndet:s de Ncpt:s* *NdetSorte* (12)
    - * *voici un Ndet:s de Ncpt:s*
      - N_0 *être un Ndet:s de combien de Ncpt:p ?* *NdetGroupe* (16)
      - * N_0 *être un Ndet:s de combien de Ncpt:p ?* *NdetDizaine* (32)
  - * *voici Dnum:p Ndet:p de Det N* *NdetNombre* (45)

Class *NdetMasse*, the largest, must be further divided into several subclasses, some of which are defined by (Buvet 1994) as C2, C3, C4, C5, C6 and C13 (measurement units); and C7a, C7b, C8 and C9 (contents); the residual subclass contains *masse*. The other classes of the decision tree above are in a complex relation with Buvet's typology.

3.4. Structure

The grammar is a network of 186 graphs. One of them id displayed in Fig. 1.

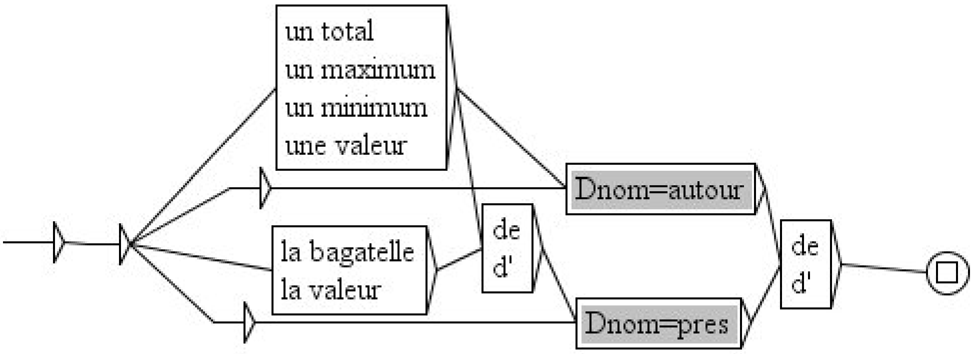

Figure 1. Graph. 'Dnom=presDe' from the local grammar

There are 3 main graphs:

- *aDet* and *deDet* for determiners preceded respectively by the prepositions *à* and *de*,
- *Det* for determiners not preceded by prepositions or preceded by other prepositions.

The compilation of these main graphs produces automata with respectively 2143, 2223 and 2044 states. The grammar is strongly lexicalised: it contains 1206 lexical tokens. The grammar recognizes embedded constructs, for instance sequences with several determinative nouns (cf. 3.1). All recursion is terminal and could be represented in a finite-state way. However, if it is done automatically, through the options of the Unitex grammar compiler, parsing with the resulting grammar is slower; and we checked that if it were done manually, the resulting grammar would be less readable.

4. Evaluation

4.1. Method of evaluation[10]

We did not use an existing treebank for evaluation of the grammar, because the annotation derived from the grammar is richer than the information found in golden standards. For example, the French Treebank (Abeillé and Barrier 2004) analyses *J'ai appris un certain nombre d'exigences administratives* 'I got aware of a certain number of administrative requirements' with *nombre* 'number' as the head noun of the complement the verb. We annotated an 8000-word corpus with information on determiners, we ran the parser with the grammar on the raw version of the evaluation corpus, and we compared the output of a parser with the manual annotation. The annotated evaluation corpus is freely available on http://infolingu.univ-mlv.fr/corpus. It was annotated with XML tags in order to delimit the (generalised) determiners as defined in 3.1 above. The XML tag is respectively <*ad*> or <*dd*> instead of <*d*> if the preposition *à* or *de* was included. The application of the guidelines led to the annotation of 1513 occurrences of determiners: 62% with <*d*>, 27% with <*dd*> and 11% with <*ad*>. There were 248 different determiners: 66% with <*d*>, 21% with <*dd*> and 14% with <*ad*>. We ran a test transducer invoking the 3 main graphs *Det, aDet* and *deDet* on the raw, untagged version of the evaluation corpus, with the Unitex system (version 1.2). The annotation inserted by the parser was compared to the manual annotation. The annotation of a sequence in the two files was considered to agree only if both the opening tag and the closing tag occurred at the same place. Two comparisons were performed. In the first one, the three kinds of tags <*d*>*,* <*ad*> and <*dd*> were confused: for example, an annotation with <*d*> in the output of the parser was considered to agree with an annotation of the same sequence with <*dd*> in the reference corpus. In the second comparison, two annotations were considered to agree only if the value of the tag was the same.

4.2. Results

We computed the precision (proportion of sequences annotated in the reference corpus among those annotated by the parser) and the recall (proportion of sequences annotated by the parser among those annotated in the reference corpus). The results of the comparison are displayed in Table 1. The 'All' column corresponds to the comparison in which the three kinds of tags are considered equal.

[10] See (Laporte 2007) for a detailed description of the method of evaluation and the annotation guidelines.

Table 1. Comparison between parser output and manual annotation

	All	*Det*	*aDet*	*deDet*
Precision	85%	71%	97%	28%
Recall	93%	95%	95%	15%

These results show that the grammar is able to detect determiners with some accuracy, even on text which is not tagged for parts of speech. The causes of non-recognition of determiners are the following (figures refer to the first comparison protocol, with confusion of tags):

- Ambiguities (43%), e.g. in *<dd>d'un</dd> côté* 'on the one hand' analysed as *<dd>d'</dd> un côté*, because of the lexical ambiguity of *un*, a frequent determiner ('a') or a rare adjective ('united').

- Under-generation of the grammar (31%), e.g. *<d>toutes sortes d'</d>abus* 'all kinds of abuse'. These constructions will be described in the next version of the grammar[11].

- Presence of multi-word units (15%), e.g. *impôt sur <d>le</d> revenu* 'income tax', which disturbs the recognition of internal determiners.

- Other causes (11%) : (i) Words not in the lexicon, e.g. *<d>son</d> antisyndicalisme* 'their anti-trade unionism'[12]. (ii) Overlappings with other analyses, e.g. *depuis 1985 <dd>des</dd> achats* 'since 1985 of purchases', where the overlapping analysis is *<d>1985 des</d> achats* '1985 among the purchases'. (iii) Spelling errors. (iv) Unknown causes.

With the second comparison protocol, the main cause (83%) of non-recognition of *<dd>* sequences as such is their recognition as *<d>* sequences. The surface form *de* can be analysed either as a preposition, or as a determiner, or as a combination of a preposition and a determiner. When the choice depends on syntactic context, the grammar cannot discriminate between these cases. The parser's tagging of the text involves a linearization. When several analyses are exclusive of one another, the parser picks one of them. These cases of non-recognition are thus an artefact of the evaluation process and do not depend on the grammar.

5. Conclusion

We extended a grammar of French determiners and evaluated its quality by comparison with an independently annotated corpus. The application of the grammar gives deeper syntactic information than chunking or information available in treebanks: in particular, it contributes to a more accurate detection of heads of noun phrases. The grammar achieves 85% precision and 93% recall. The analysis of errors showed directions for improvement of both figures. The size of the grammar shows that delimitation of determiners is not predictable from simple rules.These facts suggest that the local grammar is worth using as a component of a deep syntactic grammar of French

[11] 23% of the cases of non-recognition consist in the inclusion vs. non-inclusion of pre-nominal adjectives in the determiner. Following the guidelines, a pre-nominal adjective is included if and only if it cannot be employed predicatively with the same meaning, e.g. in *<d>leur propre</d> caisse* 'their own retirement system', vs. *<d>de</d> mauvaises moussons* 'bad monsoons'. The grammar and the lexicon do not contain this syntactic feature and analyse the above as *<d>leur</d> propre caisse* and *<d>de mauvaises</d> moussons*.

[12] The test transducer (Laporte 2007) recognises determiners only if they precede nouns marked as such in the lexicon.

References

ABEILLÉ A. and BARRIER N. (2004), « Enriching a French Treebank », in *Proceedings of the International Conference on Language Resources and Evaluation* (LREC), Lisbon.

BLANC O. and CONSTANT M. (2006), « Outilex, a platform for Text Processing », in *Proceedings of Coling-ACL on Interactive Presentation Sessions*, Sydney: 73-76.

BUVET P.-A. (1994), « Détermination : les noms », in *Lingvisticae Investigationes,* n° 18, Benjamins, Amsterdam/Philadelphia: 121-150.

BUVET P.-A. and LIM J. (1996), « Les déterminants nominaux aspectuels », in *Lingvisticae Investigationes,* n° 20, Benjamins, Amsterdam/Philadelphia: 271-285.

CONSTANT M. (2000), « Description d'expressions numériques en français », in *Revue Informatique et Statistique dans les Sciences Humaines,* n° 36: 119-135.

CONSTANT M. (2004), « Vers la construction d'une bibliothèque en-ligne de grammaires Linguistiques », in *Lexicometrica*. Numéro spécial, Actes du colloque *L'analyse de données textuelles : De l'enquête aux corpus littéraires*, Québec, 2002.

DANLOS L. (2005), « Automatic Recognition of French Expletive Pronoun Occurrences », in *Proceedings of the International Joint Conference on Natural Language Processing* (IJCNLP), Companion Volume, Jeju (Korea): 73-78.

GROSS M. (1967), « Sur une règle de cacophonie », in *Langue française,* n° 7.

GROSS M. (1986 [1977]), *Grammaire transformationnelle du français*, vol. 2, *Syntaxe du nom*. Cantilène, Paris.

GROSS M. (1998-1999), « Lemmatization of Compound Tenses in English », in *Lingvisticae Investigationes,* n° 22, Benjamins, Amsterdam/Philadelphia: 71-122.

GROSS M. (2000), « A Bootstrap Method for Constructing Local Grammars », in Bokan, N. (Ed.), *Proceedings of the Symposium on Contemporary Mathematics*, University of Belgrad (Serbia): 229-250.

GROSS M. (2001), « Grammaires locales de déterminants nominaux », in *Détermination et formalisation*, LIS series (23), Benjamins, Amsterdam/Philadelphia: 177-193.

HOCKEY B.A. and MATEYAK H. (2000), « Determining determiner sequencing: A syntactic analysis for English », in Abeillé and Rambow (Eds.), *Tree Adjoining Grammars: Formalisms, Linguistic Analyses and Processing*, CSLI: 221-249.

LAPORTE É. (2007), « Evaluation of a grammar of French determiners », in *Workshop in Technology of Information and Human Language* (TIL): 1625-1634.

LI W. and MCCALLUM A. (2003), « Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction », in *TALIP*, vol. 2(3): 290-294.

PAROUBEK P., ROBBA I., VILNAT A. and AYACHE CH. (2006), « Data, Annotations and Measures in EASY, the Evaluation Campaign for Parsers of French », in *Proceedings of the International Conference on Language Resources and Evaluation* (LREC)*,* Genoa.

PAUMIER S. (2006), *The Unitex Manual*. http://igm.univ-mlv.fr/~unitex/.

SHA F. and PEREIRA F. (2003), « Shallow parsing with conditional random fields », in *HLTNAACL*, Edmonton (Canada).

SILBERZTEIN M. (2003), « Finite-State Description of the French Determiner System », in *Journal of French Language Studies,* n° 13 (2), Cambridge University Press: 221-246.