TreeLex Meets Adjectival Tables

Anna Kupsc

Treelex Meets Adjectival Tables

Anna Kupść
Université de Bordeaux / CLLE-ERSS, Signes, IPIPAN

Abstract

The paper presents Treelex, a valence lexicon of French adjectives automatically extracted from a treebank. The corpus contains morphological and syntactic annotations but no subcategorisation information is present for adjectives. Due to rich corpus annotations, our extraction method is guided by linguistic knowledge. The obtained lexicon (about 2000 adjectives and 40 frames) has been evaluated against hand-crafted adjectival tables described in [13] and achieved 0.46 F-measure.

Keywords

adjectives, valence, treebank, syntactic lexicon, lexicon-grammar tables, French

1 Introduction

The importance of subcategorisation information is unquestionable with respect to performance of various NLP applications [4, 15, 5, 9]. So far, creating valence lexicons has been mostly devoted to obtaining such resources for verbs whereas valence lexicons for other predicates, e.g., nouns, adjectives or adverbs, are scarce. For French, the only available resources for adjectives (lexicon-grammar tables in [8] and [13]) exist only on paper and have not been adapted yet to automatic processing. In this paper, we present a method for obtaining an electronic valence lexicon of French adjectives which can be used in various NLP applications.

The adopted technique consists in automatically extracting the lexicon from a treebank. We exploit a journalistic corpus, richly annotated with both morphological and syntactic information (constituents and functions), cf. [1]. The corpus is relatively small as it contains about 1 million words. The treebank has been automatically pre-tagged and then manually verified by human experts following annotation guidelines in [2]. We rely on linguistic knowledge and corpus annotations in order to obtain subcategorisation patterns for adjectives.

Syntactic annotations in the corpus provide information about major constituents (including adjectival phrases) but grammatical functions are indicated only for phrases related to verbs. Therefore, no distinction between argumental and non-argumental dependents of adjectives is made in the treebank. Specifying valence (i.e., arguments) of adjectives is more difficult than for verbs. First, in most cases, the syntactic realization of adjective's arguments is optional. For instance, [12] mentions just a few adjectives, such as *enclin* 'inclined', *exempt* 'exempted' or *désireux* 'desirous', among those for which a complement is obligatory. This makes the strongest 'obligatoriness' criterion, often used to identify complements of verbs, practically inapplicable to adjectives. Also results of other linguistic tests, e.g., topicalisation or pronominalisation, are in general less reliable than for verbs, cf. [13]. Second, syntactic realization of adjective's arguments is more variable than with verbs: several equivalent syntactic realizations of one semantic argument are possible. For example, *affable* 'affable' may appear with two semantically equivalent PP complements: *affable envers/avec les clients* 'affable towards/with customers'. Such variability makes the specification of required components even more challenging. Finally, adjectives may appear in many syntactic constructions, e.g., comparative or impersonal phrases. Hence, arguments specific to individual adjectives should be distinguished from elements regularly appearing in a particular construction.

Despite the difficulties in defining complements of adjectives, the subject of an adjective can be quite easily identified. In the paper, we focus on specifying these two types of arguments: both the subject and complements are incorporated into valence frames of adjectives.

2 Method

As mentioned above, we exploit corpus annotations and use linguistic knowledge in order to obtain subcategorisation information for adjectives. In particular, we refer to constituent types (to identify APs and their components) as well as to functions associated with direct dependents of verbs, especially with respect to predicative adjectives.

2.1 Arguments of Adjectives

In French, complements of adjectives can be mainly realised by three syntactic categories: prepositional phrases, subordinate clauses or infinitive clauses, tagged in the corpus, respectively, as PP, Ssub and VPinf.

(1) sûr [PP de son retour] / [Ssub qu'il reviendra] / [VPinf de revenir]
sure of his return that-he will-be-back to be-back
'sure of his return/ that he'll be back / to be back'

Nominal phrases (NP), on the other hand, can serve only as the subject of an adjective.[1] We adopt the notion of the subject also for attributive adjectives: the modified noun is a semantic argument of the adjective and can be considered its semantic subject. Therefore, we consider that the subject is present in both predicative (2) and attributive (3) uses of an adjective:

(2) predicative use:

[NP La maison] est **grande**.
the house is big

'The house is big.'

(3) attributive use:

Je vois une **grande** [N maison].
I see a big house

'I see a big house.'

2.2 Linguistic Cues in the Treebank

Corpus annotations indicate adjectival phrases (AP) but they do not specify arguments of adjectives: functional annotations are absent within APs. Moreover, an argument of an adjective can be outside of an AP, for example the subject of a predicative adjective (2). Therefore, we use linguistic knowledge, applied to corpus annotations, in order to identify predicate-argument structure of adjectives. In particular, we aim at providing a 'normalized' valence, i.e., to separate constituents which occur with individual adjectives from components of productive constructions (elements which can appear with almost any adjective).

If no complement and no subject have been identified for an adjective in the corpus, we assume that its valence contains only the NP subject.

2.2.1 Arguments

In the corpus, predicative adjectives are arguments of a verb and they are assigned a grammatical function: a subject complement (ATS) or an object complement (ATO), i.e., a predicate referring either to the sentential subject (2) or to the direct object (4).

(4) [NP Jacques] trouve [AP inevitable] [Ssub qu'elle chante].
SUJ Jacques finds ATO unavoidable OBJ that-she sings

'Jacques finds unavoidable that she sings.'

In such cases, the subject of the adjective can be easily identified as it is indicated by the grammatical function of another argument of the verb: SUJ for ATS, and OBJ for ATO adjectives. As (4) shows, the subject of an adjective does not have to be nominal.

Adjectives can also appear in impersonal constructions with an accompanying Ssub or VPinf, (5). The status of the propositional components in (5) is different from those in (6), as indicated also by corpus annotations. The crucial difference is that Ssub or VPinf in (5) can be preposed to become the sentential subject, whereas this is not possible in (6).

(5) Il est [AP agréable] [Ssub qu'il fasse beau] / [VPinf de sortir].
it is ATS nice OBJ that-it makes beautiful OBJ to go out

'It's nice that the weather is good / to go out.'

(6) Paul est [AP heureux [Ssub qu'il fasse beau] / [VPinf de sortir]].
Paul is ATS happy that-it makes beautiful to go out

'Paul is happy that the weather is good / to go out.'

In (6), corpus annotations indicate that Ssub or VPinf is embedded within AP, unlike in (5). We consider the propositional constituents in (5) the extraposed subject of the adjective, i.e., in impersonal constructions (the subject is *il* or *ce*), OBJ is the subject of ATS adjective. On the other hand, if no construction-specific elements are present (sec. 2.2.2), the subordinate component in (6) is treated as a complement of the adjective.

French clitics are always attached to a verb but they can replace dependents of other predicates as well. Although clitics often pronominalise arguments, they may refer to adjuncts, for instance to locative phrases. In the corpus, clitics are direct dependents of a verb and they are assigned a function. In copular predicative constructions, as the copula itself does not have a clitic, the clitic can only indicate a dependent of the predicative adjective. The clitic function specifies whether it is a complement or an adjunct.

2.2.2 Non-arguments

Constituents which regularly appear in syntactic constructions are not related to a specific adjective and do not belong to its valence list. We filter out such components (PP, VPinf or Ssub) based on linguistic cues.

In comparative constructions, an adjective is often accompanied by a PP or Ssub, annotated in the corpus as an internal component of AP. If the adjective appears with a comparative adverb, *plus* 'more', *moins* 'less', *autant* 'as much as', etc., the embedded constituent is not considered part of the adjective frame (in contrast to (6) where there is no adverb).

(7)–(8) illustrate another type of productive constructions where the embedded constituent of AP is not an argument of the adjective. Again, the presence of an intensifier adverb, such as *si* 'so', *trop* 'too', *tellement* 'so much', etc., is decisive for the status of Ssub or VPinf constituent within AP. Only if no such adverb is present, the constituent can be considered an argument of the adjective.

(7) Paul est [AP si heureux [Ssub qu'il saute de joie]].
Paul is ATS so happy that-he jumps of joy

'Paul is so happy that he jumps out of joy.'

[1] [13] mentions two apparent exceptions: *bleu roi* 'royal blue' and *rouge cerise* 'cherry red'. Such adjectives, however, can be considered multi-word units, cf. [8].

(8) Cette histoire est [AP trop belle [VPinf
this story is ATS too beautiful
pour être vraie]].
for be true
'This story is too good to be true.'

2.2.3 Lexicon of Prepositions

Apart from comparative phrases, PPs do not appear in adjectival constructions. Therefore, no other linguistic observations can help us specify the status of PPs in APs. In particular, there is no general rule which would permit to distinguish a PP complement of an adjective from a PP attached to an adjective in complex NP restructured constructions, [11]. Instead, we use PrepLex [7], a lexicon of argumental and non-argumental prepositions, i.e., prepositions which can or cannot introduce an argument. We adopt it to filter out PPs which cannot be complements of an adjective. We added to the list of non-argumental prepositions a few complex ones found in the treebank which are not present in PrepLex.

3 Extracted Frames

The described method results in a lexicon of 2153 adjectives[2] and 40 frames. The vast majority of adjectives (1849) appear only with a basic frame, i.e., with the nominal subject, whereas the remaining 304 adjectives were found with a different frame. Tab.1 presents 23 extracted frames which appeared more than once along with their frequency counts and the number of corresponding adjective entries.

Before proceeding to a quantitative evaluation (sec. 4), we provide a brief impressionisitic analysis of the obtained results. As far as Ssub and VPinf arguments are concerned, their identification should be quite reliable since elimination of these non-argumental phrases is targeted by the adjectival constructions. However, a few issues still remain. First, our list of intensifier adverbs is not exhaustive and 2 adjectives were mistakenly assigned a VPinf[pour] complement. Second, certain impersonal constructions have not been recognized. At the beginning of the sentence, a predicative adjective is often followed by its extraposed VPinf subject, as in a regular impersonal construction (5), but neither the impersonal pronoun nor the copula are present. In such cases, as no construction-specific adverb is present either, the embedded VPinf is misinterpreted as a complement. Even more problematic is recognition of PP-complements since it is based on purely lexical rather than contextual information. Most prepositions listed in PrepLex are ambiguous, i.e., whether they introduce a complement or not depends on the context. Another issue related to PP-arguments is a verification of their semantic content. Although adjectives can admit several different PP-realisations of a single semantic argument, the corresponding prepositions should be semantically equivalent. At present, we have no means of verifying this requirement other than manually. Finally, a few singleton frames (i.e., of frequency 1, not listed in Tab. 1) resulted from occasional annotation problems, mostly related to incorrectly assigned syntactic structure.

[2] Numerals, quantifiers and interrogative adjectival pronouns have been excluded.

FRAME	freq.	#adjs
SUJ:NP (basic)	15485	2087
SUJ:NP|P-OBJ:PP[à]	278	81
SUJ:NP|P-OBJ:PP[de]	204	94
SUJ:NP|P-OBJ:VPinf[de]	83	44
SUJ:VPinf[de]	66	29
SUJ:NP|P-OBJ:VPinf[à]	53	16
SUJ:NP|P-OBJ:PP[pour]	35	29
SUJ:NP|P-OBJ:PP[en]	30	23
SUJ:NP|P-OBJ:VPinf[pour]	24	6
SUJ:NP|P-OBJ:PP[dans]	22	14
SUJ:SsubI[que]	18	11
SUJ:NP|OBJ:Ssub[que]	18	4
SUJ:NP|P-OBJ:PP[par]	13	12
SUJ:NP|OBJ:SsubI[que]	12	3
SUJ:NP|P-OBJ:PP[sur]	11	11
SUJ:NP|P-OBJ:PP[avec]	9	6
SUJ:NP|P-OBJ:PP[loc]	8	8
SUJ:NP|P-OBJ:PP[entre]	5	3
SUJ:SsubS[que]	6	5
SUJ:NP|P-OBJ:PP[chez]	4	3
SUJ:NP|P-OBJ:PP[depuis]	3	3
SUJ:VPinf[de]|P-OBJ:PP[à]	3	3
SUJ:NP|P-OBJ:PP[après]	2	2

Table 1: *Extracted frames with their frequency and the number of adjectival entries in which they appear. Abbreviations: functions: SUJ – subject, P-OBJ – PP or VPinf object, OBJ – object without an introducing element, categories: NP – noun phrase, PP – prepositional phrase, Ssub – a subordinate clause, either in subjunctive (SsubS) or indicative (SsubI) mode, VPinf – an infinitive clause.*

4 Comparison with Adjective Tables

In order to get a more objective evaluation of Treelex, we compared it with adjectives listed in lexicon-grammar tables in [13], the only available syntactic lexicon of French adjectives we are aware of. This reference resource is not ideal for our purpose. First, the tables exist only on paper so they cannot be directly used. Second, they contain constructions rather than 'normalized' frames we aim at producing here. Finally, tables do not describe adjectives that appear only with the NP subject, which leaves the status of missing adjectives unclear. Despite these inconveniencies, we decided to use the tables for our preliminary evaluation.

From 419 adjectives in [13], 266 are also present in our lexicon and we used them for evaluation. This list contains 177 adjectives found only with a basic frame in the corpus and there are 127 adjectives occurring with different frames in text. Out of all 40 frames discovered in Treelex (sec. 3), 30 are present in the evalua-

Baseline Results	
Precision	0.69
Recall	0.19
F-measure	0.30
Results for evaTreelex	
Precision	0.74
Recall	0.33
F-measure	0.46

Table 2: *The baseline results and the overall evaluation obtained for Treelex frames*

tion sublexicon (evaTreelex). We manually translated the corresponding entries in Picabia into our format, which produced 75 frames, and then compared each evaTreelex entry with frames obtained for the adjective in Picabia (evaPicabia). If a Treelex frame was equivalent to the corresponding construction present in the evaPicabia entry, the format difference was not taken into account and the frame was marked as appearing in both lexicons.

To set a baseline for our evaluation, we assumed that all adjectives have only a basic frame. We adopted standard evaluation metrics: Precision (P), Recall (R) and F-measure (F), following their definitions in [10]:

(9) $$P = \frac{\text{evaTreelex} \cap \text{evaPicabia}}{\text{evaTreelex}}$$

(How many entries in evaTreelex are correct?)

(10) $$R = \frac{\text{evaTreelex} \cap \text{evaPicabia}}{\text{evaPicabia}}$$

(How many evaPicabia entries found in evaTreelex?)

(11) $$F = \frac{2PR}{P+R}$$ (harmonic mean)

The results obtained for all frames in evaTreelex and the baseline figures are given in Tab. 2. The overall results do not seem very impressive: [14] obtain F-measure of 0.719 for English adjectives. However, our evaluation sample is much bigger (266 vs. 30 test adjectives used for English) and so is the number of frames in the reference lexicon (75 vs. 30). On the other hand, our extraction precision is quite high (0.75) whereas the low recall (0.33) is probably due to the corpus size and the choice of the reference resource. Note that 177 out of 266 evaluated adjectives were not found with a complement in the corpus (being listed in Picabia's tables, they should have a non-subject argument). Since the adjectival tables do not come from a corpus investigation, another explanation of the low recall is a possible rarity of adjective-frame uses (constructions) presented in Picabia. For example, many contructions containing a Ssub introduced by a complex complementizer *de ce que* 'of that' or *à ce que* 'to that' were not found in the corpus.

It is clear nevertheless that Treelex does much better than the baseline, especially in identifying non-basic frames. The difference in precision is smaller due to the fact that many of evaTreelex entries only have the frame used as the baseline.

20 out of 30 frames present in evaTreelex are found also in Picabia's tables. Evaluation of each individual frame present in the common part is shown in Tab. 3. The numbers confirm the observation made for the overall performance: the precision of each frame is higher than its recall. Again, this discrepancy is directly related to the amount of data available. There is no clear correlation between extraction accuracy for propositional (Ssub and VPinf) and prepositional (PP) arguments as could have been expected from the adopted technique. Note however that, in addition to the problems mentioned in sec. 3, the frame frequencies are counted with respect to adjective entries (rather than to their frequency in text). Hence, the numbers in Tab. 3 are quite low and not fully reliable.

5 Related Work

As mentioned in sec. 4, a method for automatically extracting various syntactic lexica for English, including adjectives, has been proposed in [14]. This approach is also corpus-based but it uses a set of pre-defined frame patterns to classify adjectives in the corpus rather than discovers frames themselves. Although the authors use a reference standard for evaluation, it has been extracted from a corpus and, for adjectives, it has been specifically created for this purpose. This allows them to provide an evaluation with respect to the same-origin resource rather than use a completely independent lexicon as a reference standard.

In a recent study, [3] provides 15 classes of French adjectives based on their combinatorial properties, i.e., roughly corresponding to valence frames we presented here. His classes, however, are more general than our frames. For example, most prepositions in PP complements are not explicitly indicated, nor is the type of a propositional argument (VPinf or Ssub). More importantly, his work does not aim at creating a lexicon and he uses only a few adjectives to illustrate specific classes.

6 Conclusion

The lexicon presented in the paper has been automatically extracted from a treebank, providing a resource of over 2000 adjective entries and discovering 40 frames. The quality of the lexicon has been evaluated with respect to adjectival tables listed in [13]. Although the quantitative results do not achieve the state-of-the-art performance yet, they are well above the baseline we set, which clearly indicates that adjective valence cannot be ignored in the text.

In order to obtain a better coverage and improve the quality, the lexicon should be extended. We plan to complement it by adopting statistical techniques on a much larger corpus, e.g., [6]. This will also allow us to validate the remaining Treelex frames and verify performance for individual adjectives.

The lexicon is freely available from the site: `http://erssab.u-bordeaux3.fr/spip.php?article150`.

Frame	Freq.	P	R	F
SUJ:NP	183	0.77	0.94	0.85
SUJ:NP|P-OBJ:PP[à]	40	0.89	0.66	0.75
SUJ:NP|P-OBJ:PP[de]	36	0.80	0.42	0.55
SUJ:NP|P-OBJ:VPinf[à]	11	1.00	0.15	0.26
SUJ:NP|P-OBJ:VPinf[de]	10	0.77	0.30	0.43
SUJ:VPinf[de]	8	0.62	0.80	0.69
SUJ:SsubI[que]	7	1.00	1.00	1.00
SUJ:NP|P-OBJ:PP[pour]	5	0.50	0.36	0.42
SUJ:VPinf[de]|P-OBJ:PP[à]	3	1.00	1.00	1.00
SUJ:NP|P-OBJ:PP[en]	3	0.25	0.23	0.24
SUJ:NP|P-OBJ:PP[loc]	2	1.00	0.67	0.80
SUJ:NP|P-OBJ:PP[avec]	2	0.50	0.25	0.33
SUJ:NP|OBJ:SsubI[que]	2	1.00	1.00	1.00
SUJ:NP|P-OBJ:PP[de]|P-OBJ:PP[à]	1	1.00	0.25	0.40
SUJ:Ssub[que]	1	1.00	0.04	0.07
SUJ:NP|P-OBJ:PP[sur]	1	0.33	0.50	0.40
SUJ:SsubS[que]	1	1.00	1.00	1.00
SUJ:VPinf[de]|P-OBJ:PP[pour]	1	1.00	0.33	0.50
SUJ:NP|P-OBJ:PP[dans]	1	0.25	0.50	0.33
SUJ:NP|P-OBJ:VPinf[pour]	3	0.00	0.00	NONE

Table 3: *Evaluation measures for 20 frames present both in Treelex and in Picabia's tables*

References

[1] A. Abeillé, L. Clément, and F. Toussenel. Building a treebank for French. In *Treebanks*. Kluwer, 2003.

[2] A. Abeillé, F. Toussenel, and M. Chéradame. Corpus le Monde. annotations en constituants. guide pour les correcteurs. LLF, UFRL, Paris7, 2004.

[3] N. Barrier. *Une méta-grammaire des adjectifs du français. Une appalication aux TAG*. PhD thesis, Université Paris 7, 2009.

[4] J. Carroll and A. Fang. The automatique acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Conference on Natural Language Processing*, Sanya City, China, 2004.

[5] L. Danlos. *La génération automatique de textes*. Masson, 1985.

[6] C. Fabre and D. Bourigault. Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(1):87–102, 2008.

[7] K. Fort and B. Guillaume. Preplex: a lexicon of French prepositions for parsing. In *ACL SIGSEM07*, 2007.

[8] M. Gross. *Méthodes en syntaxe*. Hermann, 1975.

[9] C. Han, J. Yoon, N. Kim, and M. Palmer. A Feature-Based Lexicalized Tree Adjoining Grammar for Korean. Technical report, IRCS, 2000.

[10] F. I., F. G., and G. C. Evaluating synlex. In *Proceedings of TALN 2007*, Toulouse, 2007.

[11] M. Meydan. La restructuration du GN sujet dans les phrases adjectivales à substantif approprié. *Langages*, (133):59–80, 1999.

[12] M. Noailly. *L'adjectif en français*. Ophrys, 1999.

[13] L. Picabia. *Les construction adjectivales en français*. Droz, Genève-Paris, 1978.

[14] J. Preiss, T. Briscoe, and A. Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *ACL 2007*, 2007.

[15] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. Using predicate-argument structures for information extraction, 2003.