Automatic recognition of French expletive pronoun occurrences

Abstract. We present a tool, ILIMP, which takes as input a French raw text and which produces as output the input text in which every occurrence of the word il is tagged either with the tag [ANA] for anaphoric or [IMP] for expletive. This tool is therefore designed to distinguish the anaphoric occurrences of il, for which an anaphora resolution system has to look for an antecedent, from the expletive occurrences of this pronoun, for which it does not make sense to look for an antecedent. The precision rate for ILIMP is 97,5%. The few errors are analyzed in detail. Other tasks using the method for ILIMP are described briefly, as well as the use of ILIMP in a modular syntactic analysis system.

1 Introduction

In Natural Language Processing, a lot of research is dedicated to anaphora resolution since it is a crucial issue, for example, for Information Retrieval or Text Summarization. Among anaphora, pronouns are quite frequent and therefore widely studied. Among pronouns, there is one, *il* in French, *it* in English, with an "impersonal" ("expletive") use (*il pleut, it rains*) which should be distinguished from the anaphoric use (*il est violet, it is purple*). Therefore, the authors who have developed a pronoun resolution system acknowledge that the impersonal pronoun occurrences must be recognized first, before dealing with anaphoric pronouns.

There exists a number of works on the English pronoun *it*, among them (Lapin, Leass, 1994), (Kennedy, Bogurev, 1996) and (Evans 2001). However, no work has been done on the French pronoun *il*[1]. This paper presents a tool, ILIMP, which is designed to recognize all the occurrences of the impersonal pronoun *il* in French texts: it marks any occurrence of *il* with either tag [IMP] or tag [ANA] (for impersonal or anaphoric use, respectively). This tool is rule based (as it is the case for Lapin and Leass' system); it works on raw texts (contrarily to Lapin and Leass' system which relies on a syntactic analysis).

If ILIMP is imperative for an anaphora resolution system, it is also a tool which can integrate a processing chain within a modular approach to syntactic analysis.

[1] While the English pronoun *it* can be found in subject and object positions, the French pronoun *il* can be found only in a subject position. Moreover, when *it* is an anaphoric subject pronoun, it can have a clausal or nominal antecedent, while an anaphoric *il* can have only a nominal antecedent. The anaphoric subject pronoun *il* translates in English as *he* or *it* according to the human nature of the nominal antecedent. The impersonal (expletive) subject pronoun *il* translates in English as *it* when it appears in a French impersonal clause which translates in English as an impersonal clause.

First, it has to be underlined that the tags [IMP] and [ANA] on the pronoun *il* can be viewed as an enhancement of the part-of-speech tag set generally used in taggers: the tag "pronoun" would be replaced by two tags, "anaphoric pronoun" versus "impersonal (expletive) pronoun". It is known that the richer the tag set is, the better would be the syntactic analysis based on this tagging (Nasr, 2004). Moreover, it will be shown that tools derived from ILIMP can be used for other linguistic annotations.

Section 2 presents the method which is based, on linguistic grounds, on a French linguistic resource, the Lexicon-Grammar, and on computational grounds, on a tool, Unitex. Section 3 presents the realization of ILIMP, the difficulties which have been encountered and the choices made to solve them. Finally, Section 4 presents an evaluation of ILIMP and discusses its positioning within a modular syntactic analysis.

2 Method

2.1 Lexicon-grammar

As for most linguistic phenomena, the impersonal use of *il* depends on both lexical and syntactic conditions. For example, the adjective *violet (purple)* can never be the lexical head of an impersonal clause - see(1a); the adjective *probable (likely)* followed by a clausal complement anchors an impersonal clause - see(1b); and the adjective *difficile (difficult)* when followed by an infinitival complement introduced by the preposition *de* (resp. *á*) anchors an impersonal (resp. personal) clause - see(1c) and (1d).

(1) a Il est myope
(He is short-sighted)
b Il est probable que Fred viendra
(It is likely that Fred will come)
c Il est difficile de résoudre ce problème
(It is difficult to solve this problem)
d Il est difficile à résoudre, ce problème
(It is difficult to solve, this problem)

Therefore, the French lexicon-grammar developed by Maurice Gross and his group (Gross 1994, Leclre 2003) is an appropriate linguistic resource for ILIMP since it describes, for each lexical head of a clause, its syntactic arguments and the possible alternations. From the lexicon-grammar, I have (manually) extracted all the items that can be the lexical head of an impersonal clause while recording their syntactic arguments. Below a brief overview of the lexical heads that I have recorded.

First, one has to distinguish verbal phrases for which the subject can only be *il*, from those whose subject is impersonal *il* when their "deep subject" is "extraposed" in a post-verbal position.

- Among the former ones, I have compiled 45 meteorological verbal phrases *(Il neige (It snows), Il fait beau (It is a nice day))*, 21 verbs from Table 17 of (Gross 1975) (*Il faut que Fred vienne/du pain*) and 38 frozen expressions (*Il tait une fois (once upon a time), quoi qu'il en soit (whatsoever))*.
- Among the latter ones, one has to distinguish those with a clausal extraposed subject from those with a nominal extraposed subject.
  - Among the former ones, I have compiled 682 predicative adjectives *(Il est probable que Fred viendra (it is likely that Fred will come))*, 88 expressions of the form Prép X (Danlos 1980) *(Il est de règle de faire un cadeau (It is standard practice to make a present))*, and around 250 verbs from (Gross 1975) *(Il est dit que Fred viendra (It is said that Fred will come))*.
  - Among the latter ones with a nominal extraposed subject, some are quite frequent verbs such as *rester* or *manquer*, while others are verbs in the passive form only used in a refined register *(Il est venu trois peronnes (Three persons came))*.

2.2 Unitex

Unitex[2] is a tool which allows us to write linguistic patterns (regular expressions or automata) which are located in the input text, with a possible addition of information when an automaton is in fact an transducer. A raw text, when given as input to Unitex, is first pre-processed: it is segmented into sentences, some compound expressions are recognized as such, and each token is tagged with all the parts of speech and inflexion features recorded in its entry (if any) in the French full-form morphologic dictionary DELAF (Courtois 2004). There is no disambiguation at all; in other words, the pre-processing in Unitex does not amount to a tagging.

For ILIMP, the basic idea is to manually write patterns (transducers) such as the pattern presented in (2) in a simplified linear form. ⟨être.V:3s ⟩ targets the inflected forms of the verb *être* conjugated at the third person singular; ⟨Adj1:ms⟩ targets the masculine singular adjectives that belong to the class Adj1, which groups together adjectives behaving as *difficult*; ⟨V:K ⟩ targets the verbs in the infinitive form. [IMP] is a tag which is added in the input text to the occurrences of *il* that appear in clauses which follow the pattern in (2). The occurrence of *il* in (1c) is therefore marked with the tag [IMP].

(2) Il[IMP] ⟨être.V:3s ⟩ ⟨Adj1:ms⟩ de ⟨V:K⟩

Tag [ANA] is the default value: it marks the occurrences of *il* that have not been annotated with [IMP]. The occurrence of *il* in (1d) is therefore marked with the tag [ANA]. Nevertheless, the matter is a little bit more complex, since there is a third tag [AMB], which is explained in Section 3.2.

[2] Unitex is a GPL open source system, which is similar to Intex (Silberstein 1994). Documentation and download of Unitex can be found at the following URL: http://ladl.univ-mlv.fr.

The output of ILIMP is therefore the input text in which each occurrence of *il* is marked with one of the tags [IMP], [ANA] and [AMB]. After this presentation of the theoretical principles underlying ILIMP, let us examine its realization.

3 Realization

3.1 Left context of the lexical head

In example (1c), the left context of the lexical head - the sequence of tokens on the left of *difficile (difficult)* - is reduced to *Il est (it is)*. However, what is frequently found in real texts are sentences such as (3a) or (3b) in which the left context of the lexical head is more complex. In (3a), it includes (from right to left) the adverb *très (very)* which modifies the adjective, the verb *paraître (seem)* in the infinitive form which is a "support verb" ("light verb") for adjectives, the pronoun *lui (to him)* and finally the modal verb *peut (may)* preceded by *il (it)*. In (3b), it includes the support verb *s'avérer* which is conjugated in a compound tense (*s'est avéré*) and negated (*ne s'est pas avéré*).

(3) a Il peut lui paraître très difficile de résoudre ce problème
(It may seem very difficult to him to solve this problem)
b Il ne s'est pas avéré difficile de résoudre ce problème
(It didn't turn out to be difficult to solve this problem)

Therefore, for each type of lexical heads (adjectival, verbal) that anchors an impersonal clause, all the elements that may occur in its left-context have to be determined and integrated in patterns. There is no real difficulty, let's say it is time consuming[3]. In contrast to this, we are faced with tough ambiguities when coming to the right context, as we are going to show it.

In the rest of the paper, patterns are presented with simplified left-contexts - as in (2) - for the sake of readability.

3.2 Right context of the lexical head

Syntactic ambiguities. There is a number of syntactic ambiguities in the right context since, as it is well known, a sequence of parts of speech may receive several syntactic analyses. As an illustration, consider the pattern in (4a), in which the symbol Ω matches any non-empty sequence of tokens. This pattern corresponds to two syntactic analyses: (4b) in which *il* is impersonal and the infinitival phrase is subcategorized by *difficile*, and (4c) in which *il* is anaphoric and the infinitival phrase is part of an NP. These two analyses are illustrated in (4d) and (4e) respectively - these sentences differ only in the adverb *ici/juste*.

[3] This work can be re-used in a tool which aims at identifying the lexical head of a clause, that is the predicative element which sub-categorizes the form and semantic features of the complement(s) (if any).

(4) a Il est difficile pour Ω de $\langle$V:K$\rangle$
b Il[IMP] est difficile pour $(\Omega)_{NP}$ de $\langle$V:K$\rangle$
c Il[ANA] est difficile pour $(\Omega$ de $\langle$V:K$\rangle)_{NP}$
d Il est difficile pour les étudiants qui viennent ici de résoudre ce problème
(It is difficult for the students who came here to solve this problem)
e Il est difficile pour les étudiants qui viennent juste de résoudre ce problème
(It is difficult for the students who have just solved this problem)

To deal with syntactic ambiguities, one solution is to state explicitly that a pattern such as (4a) is ambiguous by means of the tag [AMB] which is to be interpreted as "ILIMP cannot determine whether *il* is anaphoric or impersonal". However this tag may be of no help for later processing, especially if it is used too often. Another solution is to call upon heuristics based on frequencies. For example, sentences which follow the pattern in (4a) are more frequently analyzed as (4b) than as (4c). Therefore *il* in (4a) can be tagged as [IMP] even if this tag is wrong in some rare cases. I have adopted this latter solution. The heuristics I use are either based on my linguistic knowledge and intuition and/or on quantitative studies on corpora.

Lexical ambiguities. In a few cases (around ten cases), a lexical item may anchor both impersonal and personal clauses with the same subcategorization frame, e.g. the adjective *certain* with a clausal complement as illustrated in sentence (5a). Since both readings of (5a) seem equally frequent, *il* in the pattern (5b) is tagged as [AMB][4].

(5) a Il est certain que Fred viendra
(He/it is certain that Fred will come)
b Il[AMB] est certain que S

Other difficulties. A last type of difficulties is found with impersonal clauses with an extraposed nominal subject. Consider the pair in (6) in which the only difference is *du/de*, whereas (6a) is impersonal and (6b) personal. Along the same lines, consider the pair in (7) in which the only difference is *valise/priorit*, whereas (7a) is impersonal and (7b) personal.

(6) a Il manque du poivre (dans cette maison)
(There is pepper missing (in this house))
Il manque de poivre (ce rôti de porc)
(It is lacking pepper (this roasting pork))

(7) a Il reste la valise du chef (dans la voiture)
(There remains the boss' suitcase (in the car))
b Il reste la priorité du chef (le chômage)
It remains the boss' priority (unemployment)

[4] S is the symbol for the pattern aiming at representing a sentence. This pattern is made up of a non empty sequence of tokens which includes a finite verb.

I have tried to set up heuristics to deal with these subtle differences. However, I did not attempt (perilous) enterprises such as using the ± abstract feature for nouns.

In conclusion, ILIMP relies on a number of heuristics so as to avoid a too frequent use of the tag [AMB]. These heuristics may lead to errors, which are going to be examined.

4 Evaluation

I have worked on the French newspaper *Le Monde*. More precisely, I have worked on a corpus of 3.782.613 tokens extracted from the corpus *Le Monde'94*. Unitex segments this corpus into 71.293 sentences. It contains 13.611 occurrences of the token *il*, and 20.549 occurrences of third person subject pronouns, i.e. *il, elle, ils, elles (he, she, it, they)*. So *il* is the most frequent third person subject pronoun, with a rate of 66%.

From this corpus, 8544 sentences which include at least one occurrence of *il* have been extracted, and they add up around 10.000 occurrences of *il* (a complex sentence with embedded clauses may include several occurrences of *il*). These sentences have been given as input to ILIMP and the results - the tags [IMP], [ANA] and [AMB]- have been manually evaluated by friends, colleagues or students. These evaluators were asked to follow only their intuition: the tagging of *il* is then immediate in almost all the cases.

The result of this evaluation is the following: the precision rate is 97,5%. We are going to examine the 2,5% errors, putting aside tag [AMB].

4.1 Errors from morphological ambiguities

Errors coming form morphological ambiguities are (of course) counted as the other errors coming from the realization of ILIMP (which are examined in the next sections).

Recall (Section 2.2) that the pre-processing in Unitex does not include any desambiguation: it is not a tagger. To illustrate the consequences of this point, consider the pattern in (7a) in which ⟨V6:K⟩ targets verbs of Table 6 at the past participle, e.g. *choisi (chosen)*, and S a sequence of tokens which includes a finite verb (see note 3). This pattern aims at targeting impersonal clauses such as (7b). Nevertheless, it also targets (7c), in which the pronoun *il* is thus wrongly tagged as [IMP]. This error comes from the fact that the dictionary DELAF rightly includes two entries for the word *mètres* - finite form of the verb *métrer* and plural form of the noun *mètre* - and Unitex does not make any distinction between these two entries. Therefore, the sequence *le béton pour soutenir une toiture de 170 mètres* follows S (it includes a finite verb).

(7) a Il[IMP] ⟨avoir.V:3s⟩ été ⟨V6:K⟩ (ADV) que S
b Il a été choisi que les séances se feraient le matin vers 9h
(It has been chosen that sessions would take place around 9 am)

c Il a été choisi plutôt que le béton pour soutenir une toiture de 170 mètres
(It has been chosen rather than concrete to support a 170 meter roof)

Any tagger should tag the word *mètres* in (7c) as a noun. If ILIMP would take as input not a raw text but the output of a tagger, the error on *il* in (7c) would be avoided. One can contemplate this strategy. However ILIMP would then become dependent of the errors of a tagger. What is the best?

In a more general way, assuming that a syntactic parser relies upon a modular approach in which collaborates a set of modules - tagger, named entity recognition module, ILIMP, chunker, etc. - the question arises of determining in which order these modules should be chained. Let this question open, and come back to the errors of ILIMP taking as input a raw text.

4.2 il wrongly tagged as [IMP] instead of [ANA]: 0,3%

Very few errors: 33. This is surprising when considering the frequent appeal to "brutal" heuristics. As an illustration, *il* in the pattern *Il y ⟨avoir.V:3s⟩* is systematically tagged as [IMP]. This heuristic gives two errors, as in (8a), but around 1500 right tags, as in (8b).

(8) a Il revient de Rimini. Il y a donné la réplique à Madeleine.
(He is back from Rimini. He gave there the cue to Madeleine.)
b Il y a beaucoup de trafic à 8h
(There is a lot of traffic at 8 am)

4.3 il wrongly tagged as [ANA] instead of [IMP]: 2%

More errors. This type of errors comes from the fact that [ANA] is the default value. These errors are thus directly imputable to gaps in the patterns making up ILIMP. Among these gaps, there are first those coming from my laziness/tiredness/lack of time. For example, I have introduced quotation marks at some places in patterns but not everywhere. Hence, *il* is wrongly tagged as [ANA] in (9a). In the same lines, I wrote some automata for the cases with subject inversion, but I did not take time to write all of them, hence the error in (9b). Moreover I skipped over any co-ordination cases, hence the error in (9c).

(9) a Il[ANA] était "même souhaitable" que celui-ci soit issu ...
(It was "even desirable" that this one be from ...)
b Est-il [ANA] inconcevable que ...
(Is it inconceivable that ...)
c Il [ANA] est donc indispensable et légitime de les aider
(It is thus essential and legitimate to help them)

Secondly, there are lexical gaps in patterns. In particular, adjectives which can be the head of impersonal clauses are missing. The list of 682 adjectives I have compiled needs to be completed.

Thirdly, there are syntactic gaps in patterns. For example, I have considered any extraposed clausal subject as obligatory, whereas there exist cases where such a subject is not realized. For example, in phrases introduced by *comme (as)*, (10). I have created a pattern to take into account such phrases but it does not handle all of them.

(10) Comme il a été annoncé; comme il est bien connu
(As it has been said; as it is well known)

Finally, gaps are found for impersonal clauses with a nominal extraposed subject. On the one hand, as explained in Section 3.2.3, these clauses are hard to identify for common verbs such as *manquer* or *rester*. On the other hand, I have written no pattern at all for verbs in the passive form used in a refined register, see section 2.1.

To conclude this section of the occurrences of *il* wrongly tagged as [ANA], I would like to add that the first three types of errors can be avoided with a little effort, however that is not the case for the last type.

4.4 Other errors: 0,2%

Some errors come from the fact that the word *il* is not used as a subject pronoun (recall (note 1) that it is the only French use of this word) but as part of a named entity in a foreign language, see (11)[5]. There exist also errors coming from typing or spelling mistakes.

(10) Elle a publié cette revue appelée Il[ANA] Caffè
(She published this magazine called Il Caffè)

4.5 Evaluation on other corpora

An evaluation of ILIMP has also been realized on French literary texts written in the XIXth century. It concerns 1858 occurrences of *il*. The precision rate falls compared to the journalistic genre: it goes from 97,5% to 96,8%. This fall comes, on the one hand, from impersonal expressions which are not anymore used, (11), on the other hand, from a high number of sentences with subject inversion, as in (9b). Recall (Section 4.3) that I have not handled this case systematically at the time being.

(11) Mais peut-être était-il un peu matin pour organiser un concert, ...
(But maybe was it a little bit morning to organize a concert, ...)

[5] This kind of error would be avoided if ILIMP take as input a text in which the named entities are recognized.

The percentage of impersonal *il* in literary texts increases compared to *Le Monde* corpus: it goes from 42% to 49,8%. In a more general way, I expect important differences on the percentage of *il* with an impersonal use according to the genre of corpora[6], however I don't expect significant differences on the precision rate of ILIMP (especially, if I take the time to correct the three first types of errors decribed in Section 4.2). This is because the list of lexical heads for impersonal clauses is closed and stable.

5 Conclusion and Future work

The method used in ILIMP to locate the occurrences of *il* in an impersonal use, which gives good results, can be used for other languages (obviously for English) and for other tasks. For Engish, ILIMP can be straightforwardly adapted to disambiguate the impersonal versus anaphoric use of *it* as a subject pronoun.

It has already been said (Section 3.1) that a tool derived from ILIMP can be designed to identify the lexical head of a clause. Another tool can be designed to enhance a module in charge of the computation of syntactic functions, thanks to the notion of "deep extraposed subject", which is relevant for impersonal clauses.

Finally, the method I have proposed to disambiguate an ambiguous and very frequent word as *il* (impersonal versus anaphoric uses), can be used for other ambiguous frequent functional words such as the (French) word *que* (which can be a complementizer, a relative pronoun, or an adverb in discontinuous restrictive or comparative expressions, (Jacques 2005)) - very roughly, the English translation of *que* is *that.*

The goal or ILIMP or related "little" tools is obviously modest and restricted when compared to the goal of a robust parser which would give for any sentence THE correct and complete analysis, with a precision rate closed to 98%. However, it has to be acknowledged that such an ideal parser does not exist, neither for French nor for English, in spite of so many years of effort. So it could be a wise strategy to follow the saying which goes *Les petits ruisseaux font les grandes rivires (Little streams make big rivers).* If this strategy is followed, there is left the problem of canalizing the little streams. This means organizing a research effort first to develop such "little" tools, second to determine how to order them in a sequential processing chain.

This article was processed using the LaTeX macro package with LLNCS style

[6] *Le Monde* is a standard high level newspaper, although it contains a number of long papers which describe in detail the life and work of famous persons. These papers, when they describe the life of a man, link up numerous occurrences of anaphoric *il (he)* referring to this man. One may expect that the percentage of impersonal *il* increases in newspaper handling only news or economy.

REFERENCES

COURTOIS B. (2004), Dictionnaires électroniques DELAF anglais et français, *Syntax, Lexis and Lexicon-Grammar. Papers in honour of Maurice Gross* , Lingvisticæ Investigationes Supplementa 24, Amsterdam/Philadelphia : Benjamins, pp. 113–133.

DANLOS L. (1980), *Représentation d'informations linguistiques: les constructions N être Prép X.* Thèse de troisième cycle, Paris: Université Paris 7.

EVANS R. (2001), Applying Machine Learning toward an Automatic Classification of *it*, *Literary and Linguistic Computing*, Vol. 16, n°1, pp. 45-57.

GROSS M. (1975), *Méthode en syntaxe*, Paris, Hermann.

GROSS M. (1994), Constructing Lexicon-Grammars, *Computational Approaches to the Lexicon*, Oxford, Oxford University Press, p. 213-263.

JACQUES M.P. (2005), *Que* : la valse des étiquettes, *Actes de TALN 05*, Dourdan.

KENNEDY C., BOGURAEV B. (1996), Anaphora for Eveyone; Pronominal Anaphora Resolution without a Parser, in *COLING'96*, Copenhagen.

LAPIN S., LEASS H.J. (1994), An algorithm for pronominal anaphora resolution, *Computational Linguistics*, 20(4), p. 535-561.

LECLERE . C. (2003), The lexicon-grammar of French verbs: a syntactic database, In *Proceedings of the First International Conference on Linguistic Informatics,* Kawaguchi Y. et alii (eds.), UBLI, Tokyo University of Foreign Studies.

NASR A. (2004), *Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement,* Habilitation à diriger des recherches, Université Paris 7

SILBERZTEIN M. (1994), INTEX: a corpus processing system, in *COLING'94*, Kyoto, Japon, vol. 1, pp. 579-583.