Extracting and Visualizing Quotations from News Wires

Éric Villemonte de la Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, Victor Mignot

Extracting and Visualizing Quotations from News Wires

Éric de La Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis,
Gaëlle Recourcé, and Victor Mignot

ALPAGE, INRIA Paris-Rocquencourt & Université Paris 7
Domaine de Voluceau – Rocquencourt B.P. 105
78153 Le Chesnay Cedex – France
{eric.de_la_clergerie,benoit.sagot,rosa-devi.stern,pascal.denis}@inria.fr

Abstract. We introduce SAPIENS, a platform for extracting quotations from news wires, associated with their author and context. The originality of SAPIENS is that it relies on a deep linguistic processing chain, which allows for extracting quotations with a wide coverage and an extended definition, including quotations which are only partially quotes-delimited verbatim transcripts. We describe the architecture of SAPIENS and how it was applied to process a corpus of French news wires from the AFP news agency.

1 Introduction

The SAPIENS platform was designed to give a preview of what NLP resources and techniques would bring to a larger domain application, namely a service for press agencies, offering production, analysis, search and synthesis tools to journalists handling on a daily basis a large amount of information in the form of news wires.[1]

SAPIENS focuses on quotations detection, as information sourcing constitutes a major part of press agencies' work; quotations thus represent a significant part of news wires content. The user's need in this context lies in the retrieval of those quotations in order to synthesize information on a particular topic with regards to the persons or entities — such as organisations — who expressed any kind of verbal utterances about it. Automatization appears to be very desirable for such a task; the main requirements of the system are an exhaustive coverage of various forms of quotations found in texts, as well as their correct attribution to entities who are to be recognised.

SAPIENS has been applied on a corpus of news wires from the Agence France-Presse (France-Press Agency, AFP). The AFP produces every day 6000 news in six languages (French, English, German, Spanish, Arabic, Portuguese). The average size of a news item is 250 words. Currently, SAPIENS focuses only on French language.

2 Related Work

Automatic extraction of quotations is not among the most explored fields in NLP applications, despite the interest it raises for domain related tools such as news exploring

[1] SAPIENS has been developed within the Scribo project, funded by the French "pôle de compétitivité" System@tic (web page: http://www.scribo.ws).

Z. Vetulani (Ed.): LTC 2009, LNAI 6562, pp. 522–532, 2011.

or sentiment detection.Nevertheless, we can mention several applications developed in this perspective: Google InQuotes[2], News Explorer[3], Excom[4]. Like SAPIENS, News Explorer aims at the detection of quotations in news — although SAPIENS is specialized for press agency news wires — and their author; it is based on word lists and pattern matching. Excom implements NLP techniques and takes into account quotations surrounded by typographic quotes, which is also the case with News Explorer and InQuotes. The latter indexes quotations based on word-spotting but is limited to a predefined list of authors. By proposing an approach using advanced NLP techniques for the detection of complex types of quotations and wide entities recognition, SAPIENS addresses these limitations and implements a broader notion of quotation, i.e., not only text segments surrounded by quotes.

3 Overall Architecture

Quotations and the related issue of reported speech (hereafter RS), currently treated in the context of journalistic corpora, can be described at different levels: a surface level, which gives several indications for detection by automatic tools, and a deeper, linguistic and discursive level. The surface level points out some markers such as typographic quotes and the presence of certain verbs close to them; these verbs can be grouped in a list which can be drawn up by observation of corpora or produced by various linguistic studies on RS. Example 1 illustrates a simple verbatim quotation, that is the indirect object of the verb "s'engager" (*to promise*).

(1) Ségolène Royal s'est engagée mardi, au stade Charléty à Paris, à "rassembler toutes les énergies d'où qu'elles viennent". (AFP-May $1^{s}t$ 2007)
Ségolène Royal promised on Thursday, at the Charléty Stadium in Paris, to "put together all the help we can get, wherever it comes from".

However, the description of the deeper level shows several features of quotations usage and structure which are not captured by this surface approach. This is especially true for quotations in journalistic texts, which have the particularity to mix very often direct and indirect RS. This mixed type correspond to a very precise preoccupation of the journalist, who has to distinguish, at the word level, which parts of a speech were actually uttered and which parts he is indirectly reporting. This is illustrated in Example 2, where the quotation is between brackets.

(2) Elle rappelle que [les "premiers jalons" de la scolarisation des handicapés "furent posés par le plan Handiscol en 1999"]. (AFP-May $3^{r}d$ 2009)
She recalls that [the "first steps" towards the schooling of disabled people "were achieved by the Handiscol plan in 1999"].

The direct RS parts are surrounded by quotes and constitute *verbatims*. Quotations are consequently not only those verbatims, which can be embedded in a larger textual area

[2] http://labs.google.com/inquotes/
[3] http://press.jrc.it/NewsExplorer/ — A description of the system can be found in [5].
[4] http://www.excom.fr/

containing RS. The verbatims should always appear as such, even outside a detected segment of reported speech, in the final application, as they are for the user, i.e. the journalist, the core element of sourced utterances. By enlarging the definition of quotations beyond the verbatims, SAPIENS allows for quoted words to make sense in a broader context, corresponding to the original circumstances of utterance and thus giving consistency to the information retrieved.

More generally, quotations are relevant only when associated with their *author*, i.e., the person — or in some cases the organization — who originally uttered the reported speech. Other relevant information can be extracted in order to make the detection tool as relevant and useful as possible (e.g., the location, date and audience of the utterance, such as "mardi, au stade Charléty à Paris" in Example 1). The extraction of such information is best achieved when full syntactic structures are available: this enables the extraction of arguments and modifiers of quotation verbs such as "déclarer" (*to declare*), and lead to better anaphora resolution results (which is required when, e.g., the author of a quotation verb is referred to with a pronoun, as illustrated in Example 2 with the pronoun *elle*).

SAPIENS' processing chain thus relies on a full-featured linguistic processing chain, *NewsProcess*, which includes a series of modules for handling news wires data both at the surface level through a pre-processing phase and at the deeper linguistic level during the parsing and post-processing phase. This allows *NewsProcess* to extract the relevant information and to store it in a database, which is in turn accessed by a visualization environment (see Section 6). The linguistic processing chain in itself can be split in three main phases:

1. pre-processing with SXPipe (Section 4)
   - tokenization, segmentation in sentences, detection of a first set of "named entities" such as URLs, addresses, numbers and other numeric units, sequences in foreign languages, etc.;
   - (standard) named entities recognition (such as persons, locations, organizations, etc.);
   - verbatim quotations extraction;
2. deep parsing with FRMG (Section 5)
3. post-processing, and in particular
   - anaphora resolution (Section 5.1)
   - quotation extraction (based both on verbatim quotations and on parsing results) (Section 5.2)

4 Pre-processing with SXPipe

SXPipe [6,8] is a set of tools which performs (1) "named entities" recognition: pre-token named entities (URLs, emails, dates, addresses, numbers...), (2) tokenization and segmentation in sentences, (3) token-based named entities (phrases in foreign languages...), (4) non-deterministic multi-word units detection and spelling error correction, and (5) lexicon-based patterns detection, including but not limited to named entities.

The most relevant parts of this pre-processing chain in the context of SAPIENS are classical named entities recognition (persons, locations, organizations, companies, products and brands, artworks) and verbatims extraction. They both belong to the 5th step, and are detailed in the two next sections.

4.1 Named Entities Recognition

While developing SAPIENS, a new module for standard named entities detection has been developed in SXPipe's `dag2dag` framework [8]. This framework allows for context-free patterns definitions and for the use of dedicated gazetteers, while remaining very efficient in terms of processing time.

The first step of this module is the generation of gazetteers. For location names, we us the `Geonames` database[5] and filter it using criteria defined according to the nature of the corpus.

For each retained location name, we store the `Geonames` id and normalized name, as well as the lcoation latitude and longitude. Moreover, we compute a reasonable scale level to be used in the final interface when showing the location in *Google Maps*.

For other kinds of named entities (persons, organizations, companies, products and brands, artworks), we extract information from the French Wikipedia.[6] We manually define a mapping from a set of Wikipedia "categories" to one of the above-mentioned named entities types. This allowed to type the title of each relevant Wikipedia article and to select entity pages. For each entity, variants are extracted from redirection pages variants (e.g., *CIA* in addition to *Central Intelligence Agency*, or *Marie-Ségolène Royal* in addition to *Ségolène Royal*). A "short description" is extracted from the heading of the article (in the case of Ségolène Royal, *femme politique française (22 septembre 1953, Dakar –)*). Variants allow us to segment person names into the first name, a possible middle name, the last name, and a gender if possible. New variants are then computed, in particular the omission or abbreviation of first and middle names, as in *M.-S. Royal* or *Royal*.[7]

Both lexicons are corrected and enriched by a blacklist and a whitelist of named entities, both manually drawn up. The result is a large lexicon that contains over 1 million entries (variants), associated with a normalized form, a reference URI and a "description" (e.g. date of birth, position...).

A context-free grammar consisting of 117 rules has been developed for defining patterns based on these gazetteers, as well as on specific lists for identifying relevant contexts (e.g., *ville*, *village*, *localité*, i.e., *city*, *village*, *locality*); other lists used by the grammar are a large list of first names, a list of possible titles such as *Dr.*, *Mme*, and others. Disambiguation heuristics have been activated, so that the amount of ambiguities added by this named entities module is as low as possible, although not null. The

[5] Freely available at `http://www.geonames.org`

[6] A full dump can be downloaded freely at `http://download.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2`

[7] A candidate such as *Royal*, i.e. an isolated last name, is discarded during the disambiguation step unless it refers to an entity mentioned earlier in the same news item in a more extended form, e.g. *Ségolène Royal*.

following verbatims extraction step then allows for some more disambiguation. For example, if an entity is ambiguous between a location and a person name, but is later found in the position of a verbatim author, it is disambiguated by the verbatims extraction component as a person name.

The result is a fast and high-quality named entities detection tool, integrated within SXPipe. Based on an evaluation conducted over a manually annotated corpus of 1544 entities occurrences[8], this tool achieves a 0.77 f-score, which is rather satisfying for French although giving good scope for improvement.

4.2 Verbatims Extraction

The SXPipe pre-processing chain achieves the extraction of verbatim quotations at a surface level with a dedicated module which uses symbolic patterns. The extraction is first done by identifying all parts of text surrounded by typographic quotes, i.e. verbatims. The patterns of this module allow us to detect more interesting elements usually related to a quotation when they are present, mainly the predicate supporting the quotation[9] and the author of the quotation. The latter can be of two kinds: a named entity, which we propose as a candidate for the role of quotation author according to its distance to the predicate–verbatim set, or a clitic pronoun, when found in an incident clause. This particular position of the pronoun indeed guaranties the correct detection of the quotation author, which can thus be identified later in the processing chain by the anaphora resolution module. Similarly, the candidate for the role of author can later on be confirmed or rejected by the syntactic analysis. As mentioned in 4.1, in case of a type ambiguity on a named entity, its identification as a quotation author is used to assign a type to it. Some patterns are applied to link several chunks of verbatims scattered in a sentence but which belong to the same quotation; this link can later be used after the parsing step as an indication about the extension of the current RS area.

5 Parsing and Post-processing

The *NewsProcess* processing chain is organized as a sequence of processing modules called either offline or through a webservice. Strictly speaking, the pre-processing by SXPipe is embedded within the first of these modules, which takes SXPipe's output as an input for the FRMG parser; the parsing result is then enriched by a series of post-processing modules which organize all the information retrieved along those steps. In particular, anaphora resolution is achieved, named entities are stored in a database with their normalized form and description ; quotations detected at the syntactic level are then aligned with verbatims.

Deep parsing is performed by the FRMG parser [10], a symbolic parser based on a compact TAG for French that is automatically generated from a meta-grammar. FRMG relies on the morphological and syntactic lexicon Le*fff* [7].The output of FRMG is a shared dependency parse forest that represents all derivation structures that the grammar

[8] This evaluation corpus is however restricted to person, location and organization names.

[9] A list of 230 quotation verbs has been acquired mainly by corpus examination and semi-automatic extraction, detailed in [9].

can build for the input sentence.[10] This forest is then disambiguated by a heuristic-based module that outputs a unique dependency tree. A resulting dependency tree is shown in Figure 1 for the sentence *"soyons réalistes", a-t-il déclaré.* (*"be realist", he has declared.*), with *be realist* being a sentential argument and a citation governed by *to declare*. A fragment of the underlying DEP XML format used to represent these dependency trees [11] is shown in Figure 2. The DPath language (section 5.2) is used to extract information from this XML representation.

5.1 Anaphora Resolution

Next in the *NewsProcess* chain is the anaphora resolution module. This module uses the information collected by the earlier processing modules, including named entities and morphosyntactic features as extracted by SXPipe (I.e. persons gender), as well as deep grammatical analysis provided by FRMG.

Our resolution system concentrates exclusively on third person singular pronouns (e.g., *il*, *elle*, *le*, *la*, etc.), which are the most relevant in the context of quotations. This excludes possessive and demonstrative pronouns, as well as plural anaphora. The latter are more tricky in that they often take disjoint antecedents and moreover, sentiments expressed by a collective entity are not considered as quotations by the AFP guidelines, since they are not strictly sourced.

Technically, we model anaphora resolution as the task of mapping the identified referential pronouns onto one of the entities detected by the previous modules. Pleonastic *il* pronouns have been filtered out by the ILIMP system of [1] which has been integrated within SXPipe. Following [3] or [4], we model resolution as a two-step process, whereby: (1) we apply a series of *hard constraints* which has the effect of eliminating candidates semantically incompatible, and (2) we rank of the remaining candidates on the basis of salience *preferences*.

Among the hard constraints, we check compatibility in terms of gender, number, person and semantic typing by filtering all antecedent candidates whose gender (resp., number, person, and entity type) is incompatible with that of the anaphor.[11] We also filter out all entities that do not have any occurrence preceding the pronoun in the text; this in effect excludes possible cases of cataphora. For gender determination, we rely on the information provided by SXPipe, potentially updated by FRMG. Gender is assigned to an entity based on the genders of its occurrences using a simple majority vote.When gender information is not provided for an occurrence, we "back-off" to a simple gender guesser that tries to classify occurrences based on their surface form. In particular, the guesser relies on honorifics (e.g., *Mme*, *M.*, *le président*).

The ranking of the remaining antecedent candidate entities[12] is performed based on how well each candidate meets certain salience preferences. These preferences include proximity (in terms of sentence distance), grammatical functions as provided by

[10] More precisely, but this is outside the scope of this paper, the actual derivation forest is transformed into a dependency one.

[11] Currently, only PERSON entities are considered, but the constraint can easily be relaxed to include other entity types.

[12] There are a few cases in which the application of the above filters results in an empty set of candidates; the anaphora is left unresolved in those cases.

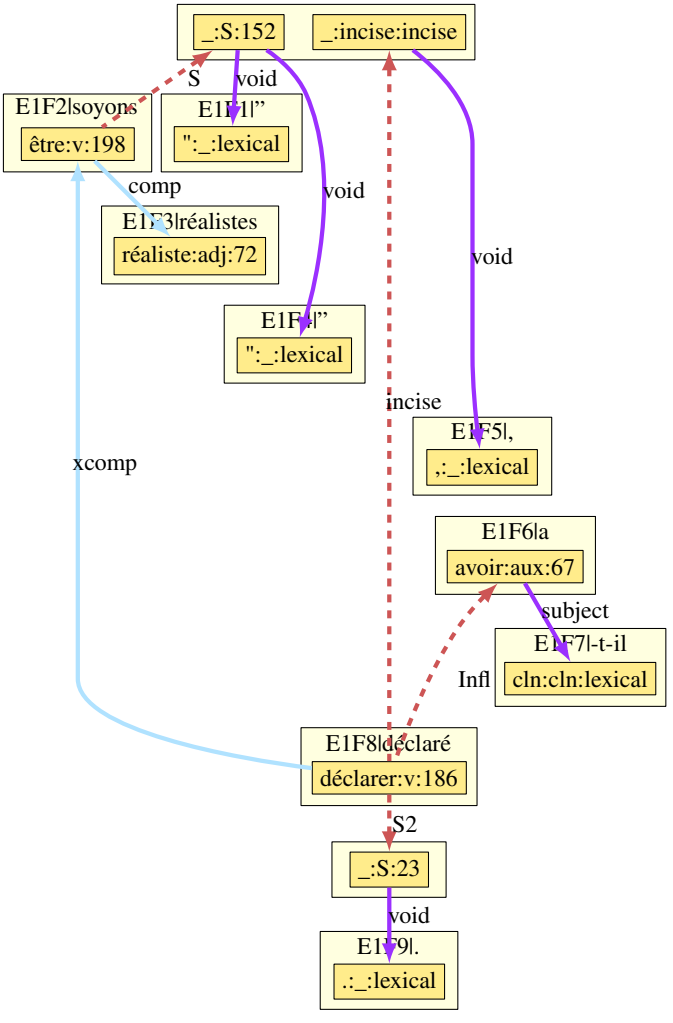

Fig. 1. FRMG dependencies for "*'be realist', he has declared*"

FRMG's output (in particular, subject, object, indirect object), the number of occurrences of the entity that precede the pronoun (i.e., how often is an entity repeated prior to the pronoun), and the appearance inside a quote. When a full parse fails to be given for a sentence, we assign grammatical functions following the order of the text: the first NP is labeled subject, the second NP object and the third NP indirect object. Each of these preferences is associated with a weight, based on a scoring inspired from [3] and determines the final resolution for each pronoun. The information is then used later to assign the detected quotations to the adequate entity (cf. section 5.2).

```
<cluster id="E1c_7_8" left="7" right="8"
         token="déclaré" lex="E1F8|déclaré"/>
<cluster id="E1c_1_2" left="1" right="2"
         token="soyons" lex="E1F2|soyons"/>
<node deriv="E1d000148" xcat="S" id="E1n023" cat="v" tree="186"
      lemma="déclarer" cluster="E1c_7_8" form="déclaré"/>
<node deriv="E1d000010" xcat="S" id="E1n005" cat="v" tree="198"
      lemma="être" cluster="E1c_1_2" form="soyons"/>
<edge id="E1e008" source="E1n023" target="E1n005"
      type="subst" label="xcomp">
```

Fig. 2. Fragment of DEP XML

5.2 Quotation Extraction

As mentioned earlier, one of the main features of journalistic quotations is their mixed characteristic. It is indeed very common to find, in journalistic content, a mixed type of quotations, with both typographic quotes and for instance a finite subordinate clause containing the RS. It is also frequent that not the whole RS is between quotes, but only one or several chunks of it, as in Example 2. Besides, those chunks do not necessarily correspond to constituents. Thus, the simple matching of text surrounded by quotes does not provide a complete retrieval of RS parts. This is why SAPIENS expands verbatims by reconstructing a coherent RS based on parsing results, hence putting forward information not directly retrievable by the user, while still distinguishing between different types of speech reporting.

Linguistic studies about RS offer different ways for us to define those configurations, and thus quotations, in the current context of journalistic corpora. They usually distinguish between various forms of RS, mainly direct and indirect, based on the presence or absence of typographic signs and particular syntactic structures, among other criteria. Without discussing here detailed aspects of RS theories, we can mention the most frequent configurations used for reporting speech, especially in journalistic content. In those constructions, the RS part is considered to be the grammatical object of the verb introducing it. As such, it can be realized as a noun phrase or a finite subordinate clause. It is also considered as the object, although at a weaker degree, when the RS verb appears in an incident clause, after or in the middle of the quoted text. For a more precise and complete study of quotations in news wires and of formal features of quotation syntactic structures, we refer to [9] and [2].

When the dependencies produced by FRMG contains such configurations, the quotation extraction module looks up the list of 230 quotation verbs at our disposal and whose making is described in [9]. The module then appoints the sentence as an RS segment if the verb head of the main clause appears in it; the verb object, whether it is an NP or a subordinate clause, is selected as the quotation segment, while the subject of the verb is selected as the author. A few patterns were defined using a Perl-based query-langage over dependencies inspired by XPath, an exemple being given below to retrieve edges linking an active-voice citation verb (as source) with a sentential object (`is_xcomp`):

```
dpath is_xcomp
  { $citation_verbs->{$_->source->lemma} }
  { $_->apply(dpath source is_active) }
```

Adverbial and prepositional clauses are finally examined in search of possible other quotation satellites, such as the date of utterance or the audience in front of which it was uttered. If previously detected verbatims are included in the current sentence, the RS segment is extended from the beginning of the clause containing the first quote to the end of the clause containing the last one. Thus the quotation can include inserted non-verbatim chunks, and stretch out on several clauses. Besides, the presence of verbatims in a sentence forces the appointment of the sentence as an RS segment even when no specific configuration or quotation verb could be found by the parser. In such a case the same extension is applied in order to put the verbatim back in a syntactical interpretable context.

RS are also frequently introduced without a specific verb, mainly in prepositional attribution phrases; in this case they are found before or after prepositional phrases such as "selon X" or "pour X" (*"according to X", "for X"*). The parser also looks for these configurations. The clause or clauses introduced by the prepositional phrase is or are selected as the quotation, and the NP following the preposition is selected as its author.

We have performed a limited evaluation of our work, mostly to guide our future efforts. We manually sampled 40 quotations from 40 different news items and evaluated both the span of the quotation and the correctness of the author. 32 quotations are found and in 19 cases, both the span and the author of the quotation are correct. Most other quotations lack an author (12 cases) or receive an incorrect one (7 cases, incl. 2 because of an erroneous anaphora resolution); 4 receive an incorrect span. More interestingly, 28 quotations exhibit patterns that would prevent a parsing-free processing chain from detecting the entire quotation span sucessfully (in most cases because not all the quotation lies between quotes, as explained above and illustrated by Example 2). Thanks to the use of the parser, SAPIENS correctly identifies the span of 21 of those 28 cases.

6 Web Interface for Visualization

The SAPIENS Web interface provides a visualization of AFP news items with a focus on detected quotations and related named entities. The access is organized by quotations authors: one can choose among entities, showed in a cloud, to whom one or more quotations have been assigned. An access *via* a search menu will soon be available, in order to enable search of quotations made by a particular person, as well as a keyword search for the retrieval of quotations related to a particular topic. On this latter possibility, we can argue that the association between a topic and a quotation will be larger and thus richer than the one offered for instance by InQuotes: InQuotes makes out this link only if the keyword itself is present within the quotation, whereas in SAPIENS' case the keyword is part of the news item metadata among other related terms which can thus be linked to the quotation in an indirect although relevant fashion. This possibility of linkage is also due to the thematic homogeneity of a news item.

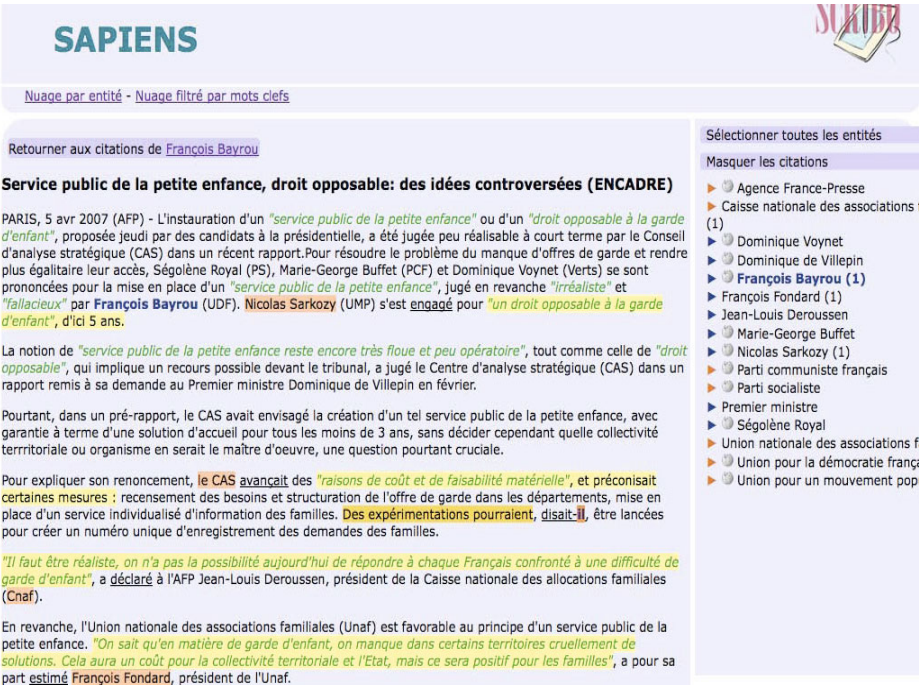

Fig. 3. Example of an enriched news item as visualized in SAPIENS

Once an entity has been chosen, the user is directed to a clickable list of news items including quotations from the selected author. It gives access to the enriched view of each of these news items, i.e. with a set of highlighted text elements: quotation(s), with distinction between verbatims and non verbatim parts, the named entity selected as quotation author - if the author is referred to with a pronoun, a tooltip indicates to which entity the anaphora has been resolved; all entities detected in the news item can be highlighted, and a link is provided for each of them, to *Google Maps* for locations and to the corresponding Wikipedia page for the other types of entities.

7 Conclusions and Perspectives

In this paper, we introduced SAPIENS, a platform for quotations extraction that relies on a deep linguistic processing chain. In particular, we have described different modules for named entities extraction, verbatims extractions, deep parsing, anaphora resolution and quotation extractions, as well as a visualization interface. We showed how we applied this chain on a corpus of news wires from the Agence France-Presse (AFP) news agency. All components of SAPIENS, including the processing chain and the resources it relies on, are free software. The information made available by SAPIENS are richer and more accurate than other systems such as Google InQuotes, in part thanks to the use of a deep parser within the chain.

In the future, SAPIENS should evolve into an operational tool used by AFP journalists. More precisely, the SAPIENS webservice will be queried by the news items editor used by journalists as soon as a news item is written, in order to automatically provide a list of descriptors to be used as metadata; these disambiguated descriptors may include named entities, topics[13], quotations associated with their author, or others; the journalist will then validate or correct these descriptors, which are meant to be used for sub-wires generation and for indexing purposes.

References

1. Danlos, L.: ILIMP: Outil pour repérer les occurrences du pronom impersonnel il. In: Proceedings of TALN 2005, Dourdan, France (2005)
2. Danlos, L., Sagot, B., Stern, R.: Analyse discursive des incises de citation. In: Actes du Deuxième Colloque Mondial de Linguistique Française, p. (à paraî tre), La Nouvelle-Orléans, Louisiane, USA (2010)
3. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics 20(4), 535–561 (1994)
4. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: Proceedings of COLING-ACL, pp. 869–875 (1998)
5. Pouliquen, B., Steinberger, R., Best, C.: Automatic detection of quotations in multilingual news (european commission - joint research centre). In: Proceedings of RANLP 2007 (2007)
6. Sagot, B., Boullier, P.: From raw corpus to word lattices: robust pre-parsing processing with SXPipe. Archives of Control Sciences, special issue on Language and Technology 15(4), 653–662 (2005)
7. Sagot, B., Clément, L., Villemonte de La Clergerie, E., Boullier, P.: The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In: Proc. of LREC 2006 (2006), http://atoll.inria.fr/~sagot/pub/LREC06b.pdf
8. Sagot, B., Boullier, P.: SXPipe 2: architecture pour le traitement présyntaxique de corpus bruts. Traitement Automatique des Langues (T.A.L.) 49(2), 155–188 (2008)
9. Sagot, B., Danlos, L., Stern, R.: A lexicon of french quotation verbs for automatic quotation extraction. In: Proceedings of LREC 2010, La Valette, Malte (2010)
10. Thomasset, F., Villemonte de la Clergerie, E.: Comment obtenir plus des méta-grammaires. In: Proceedings of TALN 2005, ATALA, Dourdan, France (June 2005), ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/mg05.pdf
11. Villemonte de la Clergerie, E.: Convertir des dérivations TAG en dépendances. In: Proc. of TALN 2010 (July 2010)

[13] Chosen within the IPTC ontology used by AFP (http://www.iptc.org/).