Genre-based Reference Chains Identification for French

Laurence Longo* and Amalia Todirascu

Université de Strasbourg, laboratoire LiLPa
22 avenue René Descartes, 67084 Strasbourg cedex, France
{longo,todiras}@unistra.fr

Abstract. In this paper we present *RefGen*, a reference chains identification module for French. *RefGen* algorithm uses genre specific properties of reference chains and an accessibility measure to find the mentions of the referred entity. The module applies strong and weak constraints (lexical, morpho-syntactic, and semantic) to automatically identify coreference relations between referential expressions. We evaluate the results obtained by *RefGen* from a public reports corpus and we discuss the importance of the genre-dependent parameters to improve the reference chains identification.

Keywords: reference chains identification, topic detection, coreference resolution

1 Introduction

In this paper, we present a new reference chains identification module *RefGen*, developed for French. Reference chains are linguistic markers indicating a topic shift or a topic continuation in the discourse [1]. *RefGen* is the main module of a topic detection system, integrated into a topic search engine. The search engine uses topic indexing to help users to retrieve relevant documents from the archives. The topic detection system takes into account the genres of the documents.

Reference chains identification is a key process for many NLP applications as topic detection or text summarization. To solve the reference, the systems identify various referential expressions (e.g. pronouns, definite noun phrase, possessives) referring the same discourse entity. A reference chain includes at least three referential expressions (e.g. *Barack Obama ... il ... lui 'Barack Obama ... he ... his'/ Le chat qui a mangé l'escalope ... le chat ... il' the cat that ate the escalope ... the cat ... it'*) which denote the same referent [2]. If this referent is common to several sentences of the same paragraph, it represents a potential topic candidate. Several paragraphs sharing the same referent indicate also a topic candidate.

Coreference resolution methods either apply manually defined heuristic rules (which select the best antecedent for a given anaphora) or rules learned from annotated corpora. While supervised learning methods [3], [4] are effective in the processing of coreference relations, they require large, manually annotated training corpora. However, there is no large reference corpus annotated with reference chains in French [5] to

* This work was supported by a French company, RBS (Ready Business system), Strasbourg (http://www.rbs.fr)

apply machine learning techniques for reference chain detection. Coreference task of the SEMEVAL2010 conference provides annotated corpora for several languages, but no French data is yet available.

To identify referential relations, we propose a new knowledge poor method as adopted for pronoun [6] and coreference resolution [7], [8], [9]. We select the reference chain elements (mentions of the same entity) using criteria about accessibility and information content of various categories of referential expressions (Accessibility theory [10]), their syntactic function, but also some genre-dependent properties of the reference chains. The *RefGen* algorithm proceeds in two steps: it first selects the starting element of a reference chain and then it selects the next elements of the reference chain applying strong and weak constraints (lexical, morpho-syntactic, and semantic) [11] between antecedent-anaphora potential pairs. The paper is organized as follows. In section 2 we describe the referential expressions, the coreference relations processed by our system, and the genre-specific parameters of the reference chains identified by a corpus analysis. In section 3, we present the *RefGen* module: the genre-dependent parameters used to identify chains, the annotation scheme adopted and the algorithm. We then discuss the first *RefGen* results obtained from a comparison with manually annotated corpora and we stress the importance of the genre-dependent parameters to improve the results of the algorithm (section 4). In section 5 we conclude and we present future developments.

2 The Reference Chains

2.1 Referential expressions

Following [2], we consider a reference chain as a relation between at least three mentions (three referential expressions). The reference chains include three types of constituents with a referential function: the proper nouns, the NPs (definite, indefinite, possessive or demonstrative) and the pronouns. The proper nouns have an important role in the discourse structure as they often open a reference chain, due to their capacity to point a unique, well-identified referent. Indeed, a study of reference chains in the journalistic portraits [12] shows their importance in organizing the discourse. Apart from cases where there is a referential competition (the repetition of the proper noun eliminates ambiguity between two referents, e.g. *"Paul and Pierre... Paul..."*), the repetition of a proper noun signals a break in the reference chain. When a referring expression is used, it triggers a "particular recruitment process" of a referent. Thus, the demonstrative (e.g. *"ce président" 'this president'*) points directly to the nearest referent while the anaphoric pronoun "il" recruits a referent that is already in mind and there is no concurrent referent [13]. The use of a particular mention (referential expression) is an indication for the reader to remember a specific referent. This referent becomes a local theme of the discourse. However, the use of a complete noun phrase instead of a pronoun is an indication of a referent change. These informations are useful to detect the end of the reference chain or the beginning of a new one.

We decide to first process single referential relations (excluding plural anaphora) between co-referent expressions within a paragraph. We treat direct coreference

cases [14] for the coreferential NPs having the same head (eg *"le changement climatique" 'climate change'/ "ce changement" 'this change'*) and some indirect coreferences between a person name and a function (e.g. *"Barack Obama... le président américain"*). Other indirect coreference cases (hyponym/ hyperonym) will be treated in the future extensions of the system.

2.2 Coreference Relations

The elements of the same reference chain are related by coreference relations. This means that they are referring to the same entity. These coreference relations are expressed by various linguistic means: agreement in gender and number between antecedent and anaphor, similar syntactic functions or semantic relations (hyponym/hyperonym, or ontological relations). These properties might be simultaneously or partially satisfied and they are usually exploited by automatic coreference resolution systems.

Several linguistic theories propose valid interpretations of these properties and rules to detect topic transitions and/or pronoun antecedents. [10] proposes a hierarchy of accessible entities in the discourse. The accessibility is defined in terms of information content and rigidity. In the top of this hierarchy, there are the entities with a low accessibility (new discourse entities): proper nouns, definite descriptions; while the high accessibility entities (referring to existing entities) are demonstratives or zero pronouns. A proper noun or a definite description is self explanatory so, such expressions are preferred to introduce a new element in the discourse. If the entity is accessible in mind and it was already mentioned in the discourse, then it could be expressed by a high accessibility expression as pronoun or possessive. Its antecedent should be a previous entity with lower accessibility. When the text author introduces a new entity, by using an expression with a low accessibility, this might be a signal of a topic change. Indeed, the new entity might be an actor involved in action, or a new event described in the next paragraphs.

From another point of view, [15] treats the problem of the identification of the main discourse entity in terms of focus, or center. The Centering theory uses an order of the possible centers of the discourse, following the syntactic function (the subject of the sentence is the most probable preferred center of the next sentence). This theory predicts topic changes, by defining four categories of focus change or maintenance, but it treats only consecutive sentences and it predicts the use of pronouns as centers.

Optimality theory reformulates the rules proposed by the Centering theory [15] in terms of constraints. [16] defines the specific notion of topic sentence as

> *"the entity referred to in both the current and the previous sentence, such as the relevant referring expression in the previous sentence was minimally oblique"*.

To define the topic sentence as a sign of discourse coherence, [16] proposes a set of constraints:

- AGREE: Anaphoric expressions agree with their antecedents in gender and number;

- DISJOINT: Co-arguments of the same predicates should be disjoint;
- PRO-TOP: The topic is pronominalized;
- FAM-DEF: Each definite NP is familiar, so refers to an entity already mentioned;
- COHERE: The topic of the current sentence is the topic of the previous one;
- ALIGN: The topic is in the subject position.

These constraints, which are reformulated in terms of relations between an antecedent and an anaphor, are applied in a hierarchically manner. Moreover, the Optimality theory proposes criteria to select an antecedent from several candidates if it satisfies a maximum number of constraints. As [16] proposes, we also adopt an algorithm selecting antecedent-anaphor pairs by checking various categories of constraints.

2.3 Genre-dependent Properties

Other parameters used by our algorithm are genre dependent properties. Several studies in textual linguistics [17] aim to characterize genres, text types or registers, by a set of linguistic parameters (such as the frequency of lexical categories, the preference for some tenses, the frequency of the complex syntactic phrases). These linguistic parameters have a specific communicative purpose and they are used in a particular communicative situation. One category of these linguistic parameters is represented by the reference chains. Genres or types might influence the type of referential expressions used in the text and the choice of the various mentions of the same referent. Cohesion markers such as reference chains are dependent on the genre as [12] identifies in newspapers portraits. We assume that reference chains have their specific properties, depending on the textual genre or on the type.

To identify the genre specific properties of the reference chains and to check this hypothesis on several genres, we study the reference chains in a French corpus (about 50,000 tokens) composed of five various genres [18]: newspapers from *Le Monde* (2004), editorials from *Le Monde Diplomatique* (1980-1988), a novel *Les trois Mousquetaires* (Dumas, 1884), some European legal standards from the *Acquis Communautaire* [19] and public reports from *La Documentation Française* (2001). We manually annotate the chains to determine which reference chain properties were relevant for a particular genre.

The reference chain study is based on [12]. For each genre, we examine the chains following five criteria:

- the average length of chains (the number of referential expressions referring the same discourse entity);
- the average distance between the elements of a chain (the number of sentences separating the elements);
- the frequency of the mentions depending on their grammatical class;
- the grammatical class of the starting element of a chain;
- the identity between the sentence theme and the first element of a chain.

The study reveals several differences across genres. For example, the average length of reference chains from text laws (Acquis Communautaire) is three mentions while the length is nine mentions for the novel. The difference between the average length

Table 1. The five genres and their properties

Corpus / Criteria	*Newspapers*	*Editorials*	*Laws*	*Novel*	*Public reports*
Length of chain	4	3,7	3	9	3,4
Distance between mentions	0,8	0,9	0,6	0,4	1,1
Grammatical class of the 1st mention	proper noun	complete NP	indefinite NP	indefinite NP	definite NP
Frequence of mentions	30% proper noun	50% definite NP	40% indefinite NP	36% pronoun	- 33% pronoun - 33% definite NP
Identity theme - first mention	80%	100%	60%	60%	40%

of the two genres may be explained in that text laws involve many referents (referential competition between the referents, so many reference chains are opened) while the novel counts lots of descriptions about the main character (which maintains the current chain). Concerning the frequency of the referential categories, we notice that the newspapers contain mostly proper nouns (30.8%) while editorials contain 50% of definite noun phrases. Proper nouns are very frequent starting elements for newspapers reference chains, but indefinite noun phrases are preferred as first mention for text laws and for public reports. Indeed, the measures adopted by the European Commission have a generic scope extended to any state member of the community, which explains the massive presence of indefinites (eg. *"un Etat Membre" 'Member State'* / *"une décision" 'a decision'* / *"une mesure" 'a measure'*). In addition, the first element of the chain is identical to the sentence theme for 80% of the occurrences for the newspapers and only for 40% for the public reports. For this last criterion, we checked if it can be possible to gather the reference chains containing the same sentence topic (coreferent reference chains [2]) to identify the document topics.

Thus, the corpus analysis of the reference chains highlights their genre-specific properties (cf. table 1). We use these parameters to configure *RefGen* according to the genre of the documents.

3 The RefGen Module

In the section above, we present the study of reference chain properties on a corpus composed of several genres. Indeed, the study validates the hypothesis that reference chains have specific linguistic properties depending on the text genre and type (explanatory, narrative etc.). We exploit these properties for reference chain identification. Now, we present the *RefGen* module architecture and we present the linguistic annotations required (tagging, chunking and Named Entities Recognition). We explain the reference chains identification algorithm (*CalcRef*) before presenting the results of the evaluation.

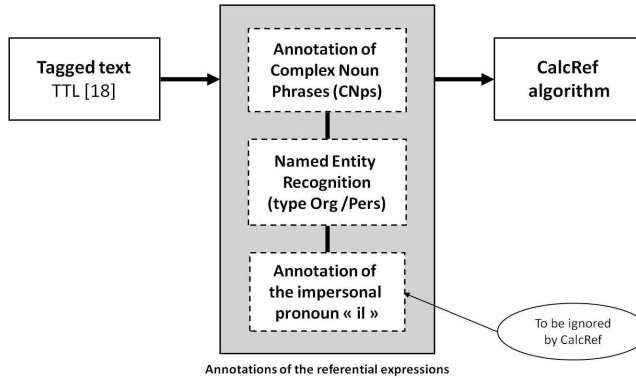

Fig. 1. The architecture of the *RefGen* module

3.1 The Architecture

RefGen is composed of several modules (fig. 1). The first processing module tags, lemmatizes and annotates the raw input text at chunk level. Then, we apply an annotation module, before proceeding to the reference chain identification step. Among the annotated phrases, we identify complex noun phrases and Named Entities, which represent potential candidates as the first mention of a reference chain. We identify the complex noun phrases among noun phrases modified by several prepositional phrases and/or modified by a relative clause. These phrases are very informative and they precisely identify the referred entity. In addition, in order to avoid wrong anaphora candidates, we annotate impersonal occurrences of the pronoun *il*. These occurrences are not taken into account by the reference computation module.

Then, after the annotation step, the reference computation module (*CalcRef*) associates a global accessibility score to each referential expression. Then, the module identifies the anaphora and their possible antecedents. To obtain only valid antecedent–anaphora pairs, the module checks several lexical, syntactic and semantic constraints.

3.2 Annotations

To identify the referential expressions, we tag, lemmatize and chunk the documents using TTL tagger [20]. This tagger identifies chunks (simple noun phrases (NP), prepositional phrases (PP), adjectival phrases (AP), verbal phrases (VP)) and it identifies some morpho-syntactic properties (tense, mode, person, gender and number). TTL uses the MULTEXT tagset [21] and the annotations provided are in XML format. The `chunk` attribute contains the chunk type, the `ana` attribute contains the tags and the `lemma` attribute contains word lemma. Then, we apply three categories of symbolic patterns to identify expressions which may be mentions of a reference chain or to identify non-anaphoric elements (as impersonal pronouns) to be ignored by our system. Thus, we identify complex NPs, named entities (person names, organization names and functions) and the French impersonal pronoun "*il*".

Identification of complex NPs – Thus, we create a set of rules (122 symbolic patterns) to identify complex NPs. A complex NP is a NP modified by at most two PP or a NP modified by a relative clause, as "*l'élévation du niveau global de la mer*"/'*the high of the global sea level*', "*le rapport qui présente les mesures contre le réchauffement climatique*"/'*the report which presents the measures against climate changes*'. Complex NPs are more informative than simple NPs. These complex NPs are self-explanatory and they often introduce a new discourse entity. For example, with the following pattern (pattern 71):

```
<pattern id="71">
 <all chunk="Np"/>
 <all chunk="Pp"/>
 <all chunk="Pp"/>
</pattern>

<action kind="addleft" tag="chunk" value="CNp" grouped="true"/>
</action>
```

the three chunks "*une élévation*" /'*a rise*', "*du niveau moyen global*"/'*the overall level*' and "*de la mer*" /'*of the sea*' are gathered in a single complex NP (CNp#3):

```
<s id="ttlfr.58">
  <w lemma="un" chunk="CNp#3,Np#8" ana="Da-fs">une</w>
  <w lemma="elevation" chunk="CNp#3,Np#8" ana="Ncfs">elevation</w>
  <w lemma="de+le" chunk="CNp#3,Pp#4,Np#9" ana="Dg-ms">du</w>
  <w lemma="niveau" chunk="CNp#3,Pp#4,Np#9" ana="Ncms">niveau</w>
  <w lemma="moyen" chunk="CNp#3,Pp#4,Np#9" ana="Ncms">moyen</w>
  <w lemma="global" chunk="CNp#3,Pp#4,Np#9,Ap#4" ana="Af-ms">
     global</w>
  <w lemma="de" chunk="CNp#3,Pp#5" ana="Spd">de</w>
  <w lemma="le" chunk="CNp#3,Pp#5,Np#10" ana="Da-fs">la</w>
  <w lemma="mer" chunk="CNp#3,Pp#5,Np#10" ana="Ncfs">mer</w>
```

Identification of named entities – While we search topic entities, we apply some heuristic rules to identify person and organization names. Persons and organizations are the main actors of the text and they are often related to the main theme of the paragraph. As well, persons and organizations refer to a unique entity and they precisely identify this entity, so they represent good candidates to be the first mention of a reference chain. Thus, we established symbolic rules and some lists of common names to identify person names and organization names. We also identified functions (even if they are not referential expressions) as proof to find the category of the named entity or to find several mentions of the same referent (a person name and its function). We use lists of nouns representing public places or group of people (school, institution, laboratory) and lists of function names (military rank "*général*, *lieutenant*, *colonel*", profession "*professeur*, *écrivain*, *banquier*"/'*professor*, *writer*, *banker*', titles "*roi*, *duc*"/'*king*, *duke*', religious function "*pape*, *prêtre*"/'*pope*, *priest*'). Our named entity extractor, which uses a set of 140 rules, proceeds in two phases. It first defines the boundaries of the entity (some of the proper names contains particles that belong to the current lexicon, as "*M. Chirac*"/'*Mr. Chirac*', "*lycée Couffignal*"/'*Couffignal college*' or numbers as "*Benoît XVI*") by adding the attribute "NER ="Ner"". Then, the extractor assigns a

value "org", "pers" or "func" to categorize the person name, the organization name and the functions.

To define the boundaries of a named entity "pers" and "org" and to categorize it, we used lists of internal and external evidences [22].

- An external evidence is the context (nouns located after or before the entity) where the named entity appears. For example, the keyword "*entreprise*"/ *'company*' in the sentence "*l'entreprise RBS compte 150 employés*"/ *'the RBS company counts 150 employees*' represents a useful proof to categorize the named entity "*RBS*" as an organization name.
- An internal evidence is located within the named entity (e.g. the first name "*Jacques*" in "*Jacques Chirac*", "*Inc.*" in "*Microsoft Inc.*").

Thus, the first name is a proof that the named entity is a person, while "*Inc*" marks as an organization. For example, the following pattern

```
<pattern id="59">
  <one lemma="@list@OrgNcExt"/>
  <one ana="Af.*"/>
  <one ana="Np"/>
</pattern>

<action kind="addright" tag="Ner" value="org">
  <filter ana="Np">
  </filter>
</action>
```

annotates the named entity "*Zundapp*" as "org" in the context "*la firme allemande Zundapp*"/ *'German firm Zundapp*'. The rule uses the external proof "*firme*" (located in the list @listOrgNcExt) that denotes an organization to categorize "*Zundapp*". To annotate only the named entity "*Zundapp*" as an organization (and not the entire sequence corresponding to the pattern 59), we use a filter on the proper noun tag ('ana="Np"'). So the action is to add at the right of the existing "*Zundapp*" tag '"Ner"', the value '"org"'. We obtain the following annotation for "*Zundapp*":

```
<s id="ttlfr.17">
  <w lemma="le" chunk="Np#1" ana="Da-fs">la</w>
  <w lemma="firme" chunk="Np#1" ana="Ncfs">firme</w>
  <w lemma="allemand" chunk="Np#1,Ap#1" ana="Af-fs">allemande</w>
  <w lemma="Zundapp" chunk="Np#2" Ner="NER#1,org" ana="Np">Zundapp
     </w>
```

In addition, we create patterns to identify complex function names (more informative) as "*le ministre des affaires étrangères, le président directeur général*"/ *'foreign minister, CEO*' with help to the previous CNps annotations. We apply heuristics rule such as

> "*if the CNp countains a function name then this element is a complex function name*".

We then focus on relations between named entities, as the relation between a person name and its function: "*Marcel Klaus, directeur financier de Swiss*"/ *'Swiss Chief Financial Officer Marcel Klaus*' or "*le directeur financier de Swiss Marcel Klaus*". The coordination cases between two function names (for a person) as "*Le membre de la commission et président de l'association Marc Dupont*" are identified by our extractor. In this case, we use the non repetition of the article "*le*" as a proof to gather the two functions with the person name. This relation is useful to identify more informative referential expressions. Thus, our extractor is able to identify simple named entities ("*Chirac, Benoît XVI, président, Université de Strasbourg*"), complex function names, relations between named entities.

We have also identified some entity which are not persons or organisations. These entities are labelled "`other`" as "*l'affaire Dreyfus, la loi Falloux*"/*'the Dreyfus affair, the Falloux law*' and some location names (countries, towns, streets) to avoid confusion with "`org`" and "`pers`" named entities.

However, our extractor do not identify partial ellipsis cases ("*Michèle et Barack Obama, le couple Hollande–Royal*" /*'Michele and Barack Obama, the Hollande-Royal couple*' or nicknames ("*l'hôte de l'Elysée, le nouveau Napoléon*"/ *'the host of the Elysee, the new Napoleon*'). These phenomena might be identified with a set of equivalence rules such as:

- "*Michele and Barack Obama*" is equivalent to "*Michele Obama and Barack Obama*";
- "*l'hôte de l'Elysée*" is equivalent to "*Sarkozy*".

Identification of the French pronoun "*il*" – We create a set of morpho-syntactic patterns (382 patterns) to identify the French impersonal pronoun "*il*" (e.g. "*il pleut, il faut*"/*'it rains, must*'). We use lists of weather verbs, several impersonal past participles (with "*avoir*" /*'have*') ("*il a plu, il a neigé, il a fallu*"/*'it has rained* (*it rained*), *it has snowned* (*it snowed*), *to have to*') and adjectives (with "*être*"/*'to be*') ("*il est aisé de, il est indéniable de*"/*'it is easy, it is denied*'). We focus on reverse-subjects like "*est-il nécessaire de*"/*'is it necessary to*'. Thus, the feature "`feat ="imp"`" is added to the impersonal pronoun. For example, in "*il s'agit d'abord de*"/*'it is first*', the identification pattern is:

```
<pattern id="337">
  <one value="il"/>
  <one value="s." lemma="se"/>
  <one lemma="agir" ana="V.*"/>
</pattern>


<action kind="set" tag="feat" value="imp">
  <filter value="il"/>
  </filter>
</action>
```

The pronoun marked as "impersonal" will be ignored by our module when checking possible anaphora.

Fig. 2. is an example of annotations including lemmas (`lemma`), chunks (simple `Np`, `Pp`; complex `CNp`), morpho-syntactic properties (`ana`), named entities (`ner`) and

```
<w lemma="le" chunk="Np#1" ana="Da-fs">L'</w>
<w lemma ="union" chunk="Np#1" ana="Ncfs" ner="NER#1, org">Union</w>
<w lemma ="européen" chunk="Np#1, Ap#1" ana="Af-fs" ner="NER#1, org">européenne</w>
<w lemma ="avoir" chunk="Vp#1" ana="Vaip3s">a</w>
<w lemma ="adopter" chunk="Vp#1" ana="Vmps-s">adopté</w>
<w lemma ="il" ana="Pp3ms" feat="imp">il</w>
<w lemma ="y" ana="Pp3">y</w>
<w lemma ="avoir" ana="Vaip3s">a</w>
<w lemma ="peu" chunk="Ap#2" ana="R">peu</w>
<w lemma ="de_le" chunk="CNp#5, Pp#1, Np#2" ana="Dg-mp">des</w>
<w lemma ="acte" chunk="CNp#5, Pp#1, Np#2" ana="Ncmp">actes</w>
<w lemma ="législatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">législatifs</w>
<w lemma ="relatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">relatifs</w>
<w lemma ="à+le" chunk="CNp#5, Pp#2, Np#3" ana="Dg-ms">au</w>
<w lemma ="changement" chunk="CNp#5, Pp#2, Np#3" ana="Ncms">changement</w>
<w lemma ="climatique" chunk="CNp#5, Pp#2, Np#3, Ap#4" ana="Af-ms">climatique</w>
```

Fig. 2. Annotations used to detect reference chains

impersonal pronouns *il* (`feat="imp"`). We use these linguistic annotations to identify the reference chains and the anaphora pairs.

3.3 The CalcRef Module

CalcRef is the main module of *RefGen* and it proceeds to reference chain identification by using genre-specific parameters and the linguistic annotations presented in the previous sections. Thus, we specify the genre of the indexed documents and we configure *CalcRef* according to the properties of the reference chains (average distance between the mentions, average chain length, and the preferred category of the first element of a chain). For example, for a corpus of public reports, the length of the chains is 4, the average distance is 2 sentences and the preferred type of the first element is a complete definite NP. To select the mentions of the reference chain, we define a set of weak and strong constraints, explained later in this section.

The algorithm follows the next steps (see Algorithm 1).

For each paragraph, *CalcRef* selects candidates for the first mention of the reference chains. Then, *CalcRef* identifies the next elements of the reference chains, by selecting a set of pairs of antecedent-anaphora candidates. Most of the candidates are filtered out by the application of several constraints. Then, we apply the transitivity of the coreference relation to compose the reference chains.

Ordering the referring expressions – For each paragraph, *CalcRef* selects candidates as first mention of the reference chains among expressions with a high degree of accessibility. [10] defines an accessibility hierarchy to classify the referential expressions according to their accessibility: less the referent is accessible, the referential expression should be longer, self-explanatory and rigid. Thus, indefinite NPs, proper nouns or complex NP, occupying the thematic position are used to mention a new entity (cf. table 2), while short mentions as pronouns might be used to refer to entities already specified in the discourse. The global accessibility GA is computed by combining three elements: informativity (the amount of lexical information), rigidity (the possibility to

Algorithm 1 The *CalcRef* Algorithm

1: For each p paragraph:
2: n : first phrase of the paragraph;
3: d : the average distance between candidates (number of phrases)
4: Select the first mention candidates from the complex noun phrases and the Named Entities (their global accessibility GA is greater than or equal to 190);
5: Order the list of the first mentions (according to GA, the function and the type), open several chains;
6: For each phrase, select the list of anaphora candidates (GA less than 190). We exclude the anaphoric use of the impersonal pronouns.
    a. for each anaphoric candidate, select the candidate pairs checking the strong constraints (without reflexive pronouns) and apply weak constraints.
    b. order the pairs according to the number of checked constraints and select the pair that satisfies the maximum number of constraints.
    c. recompose the reference chain from the identified pairs
7: n=n+d and start again at A.

pick up a specific referent) and attenuation (phonological size). With respect to the initial accessibility hierarchy, we also add indefinite noun phrases in the accessibility scale, even if this category of candidates generates several erroneous candidates. Indeed, indefinite noun phrases might be considered first mentions of a reference chain if there are lexical repetitions of the lexical head : *un rapport/a report... le rapport/the report... ce rapport/this report*. Reference chains starting with an indefinite noun phrase are very frequent in descriptive or informative texts.

We use weights on a scale of 10 to 110 for each of these elements (e.g. the global weight of the complete proper noun *"Le président Barack Obama"* is 220 while it is 150 for the pronoun *"elle"*).

CalcRef computes the global weight of the candidates as a sum of the global accessibility weight and the syntactic role weight. We also define a scale for the syntactic role weights: 100 for the subject position, 50 for the direct object position, 30 for the indirect object and 20 for other syntactic functions. Then, genre dependent parameters (as the preference for the first element type) are used to increase the weight (+50) of some candidates (for example, if we treat law texts, indefinite NPs are preferred as starting elements of a chain). We order the first element candidates according to the global weight and we select the highest weight candidate as a first element of the current chain. We open a new chain for each element of this candidate list.

In addition, we use the accessibility scale to propose possible antecedent-anaphora pairs. The antecedent should have the global accessibility higher than the anaphor.

Searching valid antecedent-anaphora candidates – *CalcRef* selects the next elements of the reference chain from highly accessible expressions (pronouns, demonstratives etc.). We establish a set of possible antecedents from low accessible expressions. We combine elements from the two sets and we identify potential antecedent-anaphora pairs. The distance between the two elements of the pair should be less than the average distance defined by the genre-specific parameters. Then, we adapt the method proposed by [11]. This method checks several constraints between antecedent and anaphora to fil-

Table 2. Accessibility Table

Referential expressions	Informativity	Rigidity	Attenuation	Global Accessibility
Indefinite NP	110	110	10	230
Complete proper noun	100	100	20	220
Proper noun	90	90	30	210
Complex definite NP	80	80	40	200
Simple definite NP	70	70	50	190
Last name	60	60	60	180
First name	50	50	70	170
Demonstrative	40	40	80	160
Pronoun	30	30	90	150
Reflexive pronoun	20	20	100	140
Possessive	10	10	110	130

ter out impossible pairs. Indeed, [11] present a system implementing a constraint-based method for pronoun resolution inspired by the Optimality theory [16]. [11] adapts these constraints to several languages (English, Korean) and proposes an implementation of the algorithm. In addition, the order of the constraints might be changed to obtain better results.

If the antecedent and the anaphor refer to the same discourse entity, they satisfy a set of constraints defined in section II (chapter B). These constraints are syntactic (similar syntactic function between the antecedent and the anaphor), morpho-syntactic (agreement in gender or in number) or semantic (hyponyms/hyperonyms). We check the contraints for all the possible antecedents of a selected anaphor. If there is a unique antecedent-anaphor pair satisfying a maximum number of constraints, this is a valid candidate to be included in a reference chain. If there are several candidate pairs satisfying the same number of constraints, then several possible reference chains should be generated.

The Optimality theory limits the search space of the antecedent at the previous sentence. [11] propose the algorithm only for pronoun resolution. We extend the set of constraints to other anaphora categories (definite expressions, reflexive pronouns).

Following [11], we adapt the constraints for French. For each pair, we check some strong and weak constraints. If a pair fails to satisfy a strong constraint, then the pair is deleted from the candidate list. For each candidate pair satisfying all the strong constraints, we check the number of the weak constraints that are satisfied. If several pairs satisfy the same number of constraints, we keep the valid pairs into a large list.

Weak constraints mean that they might be violated, even if there is a valid antecedent-anaphor pair:

- **MORPHO** – agreement in gender or number (between the personal pronoun and the candidate);
- **SYN** – the antecedent and the anaphora should have similar syntactic function;

- **SEM** – semantic relations between the antecedent and the anaphora (for example, person names might be valid antecedents of a NP expressing a function (*B. Obama - le président des Etats-Unis*);
- **PROX** – the antecedent and the anaphor are near neighbours (for possessives and demonstratives).

Strong constraints must be satisfied:

- **IMB** – the imbrications mean that an element must not be nested in its antecedent, as *[la soeur [de Marie]]*), or co-arguments of a verb should not be coreferent;
- **TETELEX** – the identity between NP's head and the partial repetition of the same proper noun.

Moreover, for some specific anaphora, it is necessary to define the set of strong constraints to be satisfied. For example, for possessives or reflexives, the constraint **IMB** (checking the arguments of the verbs) is not useful.

For the semantic constraints, we apply the method proposed by [23]. We use a resource extracted from a 500,000 tokens corpus from computer science newspapers: we extract the occurrences of all the main verbs and their subjects. This resource is used to select a valid antecedent for the pronoun, when several possible antecedents satisfy the same number of constraints. For example, we search the antecedent of the pronoun "*il*" in:

> *"Un virus a été trouvé dans mon ordinateur. A cause de ce virus, l'ordina-teur tourne lentement.* ***Il*** *envoie des messages de publicité". / 'A virus has been found in my computer. Because of this virus, the computer is very slow. It sends spam e-mails.*'

The two candidates *"Un virus"/'a virus'* and *"l'ordinateur"/'the computer'* satisfy the same number of constraints. To decide which is the valid candidate, we consult the list of subject-verb pairs. The verb *"envoie"/'to send'* has as subject *"le virus"/'the virus'* but no occurrences of a subject *"ordinateur"/computer'* are present in this list. We deduce the preferred antecedent for *"il"* is *"virus"/'virus'*, because the function of sending message is specific to an application and not of the computer itself.

Building reference chains – Then, we start from the first element of the chain and we search the pairs having this candidate as antecedent in the list. To build the reference chain, we apply the transitivity of the reference relation: if A is antecedent of B and B is antecedent of C, then they are part of the same chain. For example if we have three pairs *"J. Chirac - il"*; *"il - il"*; *"le président - il"* we can deduce that we have a reference chain with four mentions: [*J. Chirac, il, le président, il*]. We continue the process until the length of the current reference chain is greater than the average gender-specific length. We annotate the candidate pairs identified as part of the current reference chain.

We restart the whole process after selecting the next first candidate element from the ordered list of the current paragraph. The process is launched for every paragraph of the document.

4 An Example

We present a full example processed by *RefGen*. We note the various entities mentioned in the discourse with small letters (*i, ii, j, k,k1, k2, l, m, o, n, p, q, r, t, s, v, w, z*). The example is extracted from a white paper of the European Commission about the climate change.

[La lutte contre [le changement climatique]*ii*]*i* doit se faire [à deux niveaux]*j*. Il s'agit d'abord et avant tout de réduire [les émissions [de gaz [à effet de serre]*k2*]*k1*]*k* ([au moyen [de mesures d'atténuation]*l*]*m*), puis de prendre [les mesures d'adaptation qui s'imposent]*o* pour faire [face aux conséquences inévitables de [**ce changement**]*p*]*n*. [L'Union européenne]*q* ([UE]*r*) a adopté il y a peu [des actes législatifs relatifs [au changement climatique]*t*]*s*, qui définissent [les mesures concrètes nécessaires à la réalisation de l'objectif fixé par [l'UE]*v*]*w*, à savoir réduire [les émissions de 20% par rapport aux niveaux de 1990]*z* d'ici à 2020.

The algorithm first identifies the entities with a high global accessibility (proper nouns, indefinite descriptions or complex definite descriptions). In this genre (public reports), complex definite descriptions are very frequent. So *RefGen* identifies as potential first mentions the following candidates:

- **[La lutte [contre le changement climatique]*ii*]***i*,
- [les émissions [de gaz [à effet de serre]*k2*]*k1*]*k*,
- [au moyen [de mesures d'atténuation]*l*]*m*,
- [les mesures d'adaptation qui s'imposent]*o*
- [face aux conséquences inévitables de [ce changement]*p*]*n*,
- [L'Union européenne]*q*,
- [UE]*r*,
- [des actes législatifs relatifs [au changement climatique]*t*]*s*,
- [les mesures concrètes nécessaires à la réalisation de l'objectif fixé par [l'UE]*v*]*w*
- [les émissions de 20% par rapport aux niveaux de 1990]*z*.

The candidates are sorted by their global weight (the sum of the global accessibility, the syntactic function and the preference for the first mention category). The most probable first mention are *i*, *q*, *s*, *w*. We open 4 reference chains and we try to find the next pairs.

Then, for each first mention candidate, we establish a list of anaphor candidates (having the accessibility less or equal than 190): definite descriptions (*ii*, *k1*, *k2*, *l*, *m*, *o*, *n*, *p*, *z*), pronouns (*il*, *il*). Both occurrences of *il* are impersonal, so we check the validity of the constraints for definite descriptions. In this case **TETELEX** is the first constraint to be checked. For example, for the entity *i*, there is no other mention explicitly referring to *"la lutte"*. But we found a reference chain starting from the entity *ii*.

We notice that several strong constraints **TETELEX** or **IMB** are violated for "*[ce changement]p*" (table 3). A constraint violated is marked as '*', a space means that the constraint is checked. We find many direct coreference cases, while pronouns are quite few and their use is impersonal.

In contrast, we note an example extracted from the newspapers *Le Monde diplomatique*.

Table 3. Validation of the constraints for the candidate *p*

	Id	MORPHO	IMB	SYN	SEM	PROX	TETELEX
p	i						
	ii			*			
	l	*					*
	m			*			*
	o						*
	k	*		*			*
	k1	*		*			*
	k2	*		*			*
	n	*	*				*

[**M. Pons**]*i* affirme en outre, dans [un entretien publié par le Figaro du 21 septembre]*j*, que "[l'immense majorité [des député RPR]*k*]*l* souhaite [le calme et la sérénité]*m* et qu'[ils] se détermineront [le moment venu]*n*". Minimisant [la fracture ouverte]p entre "balladuriens" et "chiraquiens", **il** rappelle que [Jacques Chirac]*r* **lui** apparaît comme "[le candidat légiti-me]"*q* de **son** parti.

First, the algorithm identifies the entities with a high global accessibility (proper nouns, indefinite descriptions, complex definite descriptions). So *RefGen* identifies as potential first mentions the following candidates:

– [**M. Pons**]i,
– [l'immense majorité [des députés RPR]k]l,
– [le calme et la sérénité]m,
– [le moment venu]n,
– [la fracture ouverte]p,
– [Jacques Chirac]r,
– [le candidat légitime]q.

The candidates are sorted by their global weight (global accessibility, and the syntactic function). The most probable first mention are *i*, *l*, *r*.

Then, we establish a list of anaphor candidates (having the global accessibility less than 190): definite descriptions (*l*, *m*, *n*, *p*, *q*), pronouns (*ils*, *il*, *lui*), possessives (*son*).

We notice that several strong constraints are violated between *i* and the definite descriptions. The pairs (*i*, *il*) and (*i*, *lui*) are the most probable (table 4).

5 Evaluation

We present the first results of the evaluation of *RefGen*, we compare the reference chains extracted automatically against the manually annotated corpus. We present the results obtained for the CNp annotation module, for the NER module and for the chain identification module. The evaluation corpus is a small corpus (7,230 tokens) composed of public reports of the European Commission about the measures adopted by EU to

Table 4. Validation of the constraints for the pronoun "*il*" and "*lui*"

	Id	MORPHO	IMB	SYN	SEM	PROX	TETELEX
il	i					*	-
	j			*		*	-
	l	*				*	-
	k	*		*		*	-
	m	*		*	*		
	n			*	*	*	
	p	*		*		*	-
	r				*		
	q				*	*	
lui	i			*		*	-
	j			*		*	-
	l	*		*		*	-
	k	*		*		*	-
	m	*		*	*		
	n			*	*	*	-
	p	*		*		*	-
	r				*		
	q				*		

limit the effects of the climate changes. We compute the recall, the precision and the f-measure of the intermediate modules, as well as the results for *CalcRef*. We check the results obtained for independent antecedent-anaphora pairs, as well as for reference chains (table 5).

The NER annotation errors are due to the wrong identification of some acronyms or abbreviations (e.g. *GES : gaz à effet de serre*) which were annotated as organization names. Some NER annotation errors are due to tagging errors. The CNp identification module fails to identify several CNps (an NP modified by more than three PP), which were not described by the existing set of patterns. Indeed, the test corpus is characterized by very frequent, complex, informative NP.

The evaluation corpus contains 118 anaphoric pairs, but only 24 reference chains. Several antecedent-anaphora pairs were wrongly selected, due to tagging errors or due to the lack of external knowledge sources. For example, some of the antecedent-anaphora pairs were selected because they satisfy the same number of constraints (number, gender, syntactic function). An ontology might help to select the right antecedent.

We tested the system for three various configurations of the genre parameters. First, we use three genre-specific parameters:

1. distance=2; length=4; preferred type=definite description;
   We also tested the system after changing these parameters:
2. we ignore these parameters and use only a default distance of 20 sentences;
3. we use the newspapers parameters (distance=1; length=3; preferred type = proper nouns).

Table 5. First evaluation results

	NER	CNp	CalcRef (pairs)	CalcRef (chains)
recall	0,85	0,87	0,69	0,58
precision	0,91	0,91	0,78	0,70
f-measure (1)	0,88	0,89	0,73	0,63
f-measure (2)	0,88	0,89	0,71	0,51
f-measure (3)	0,88	0,89	0,70	0,54

If we ignore the parameters (case 2)), we obtain more antecedent-anaphora pairs than in the first case, due to the bigger distance between the mentions. Meanwhile, we obtain less reference chains because several smaller reference chains are grouped together. For all the cases, we obtain quite similar results for the pairs, but for the reference chain identification we obtain significant performance decreasing (2) f-measure: 0,51; (3) f-measure: 0,54 (table 5).

6 Conclusion

We presented *RefGen*, a reference chain identification method, developed for French. This new knowledge poor method uses a set of detailed linguistic annotations (complex noun phrases, named entities) and the accessibility hierarchy of the referring expressions to select possible antecedent-anaphora pairs. Then, a set of lexical, syntactic and semantic constraints are used to filter some invalid pairs. In addition, *RefGen* uses some genre-dependent properties of the reference chains (average length, preferred type of the first element, average distance separating several mentions of the same referent). These genre-dependent properties were identified from a corpus-based analysis. We describe a new algorithm identifying reference chains and we present a first evaluation of the module. The evaluation is done with several genre-specific parameters. The evaluation results show an improvement of the results when we use the public reports parameters.

The system is flexible; it is possible to add extra constraints to improve the quality of the output. For future evaluations, we are currently checking the reference corpus, composed of several genres, by a second human annotator, in order to improve the quality of the annotations. The annotation platform is GLOZZ [24], designed for discourse annotations. To compare the results of *RefGen* with the reference corpus, we work on a module transforming GLOZZ output into the SEMEVAL format [25]. Also, *RefGen* output will be similar to SEMEVAL format, to apply the existing SEMEVAL measures.

In the future, the module will be integrated into the topic detection system to be tested in real-life applications. The module will be extended to treat other cases of coreference: plural anaphora, hyponym/hyperonym equivalents, by adding knowledge sources as ontologies or synonym databases.

RefGen will be used to annotate large French corpora with coreference relations. It will then contribute to the development of a reference corpus for French, comparable with those provided by SEMEVAL for other languages.

In addition, future work concerns the adaptation of the system for other languages.

References

1. F. Cornish, Références anaphoriques, références déictiques, et contexte prédicatif et énonciatif. *Sémiotiques*, 8, pp. 31–57, 1995.
2. C. Schnedecker, Nom propre et chaînes de référence. Recherches Linguistiques 21. Paris : Klincksieck, 1997.
3. V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution", in *Proceedings of the ACL (Association For Computational Linguistics)*, Morristown, pp. 104–111, 2002.
4. V. Hoste, Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, 246 p, 2005.
5. S. Salmon-Alt, Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel. PhD thesis, Université H. Poincaré, Nancy, 2001.
6. R. Mitkov, "Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems," *Applied Artificial Intelligence: An International Journal*, 15, pp. 253–276, 2001.
7. S. Hartrumpf, "Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics," in *Proceedings of CoNLL (Computational Natural Language Learning Workshop)*, 2001.
8. A. Popescu-Belis, Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation. PhD thesis, Université Paris-XI, 1999.
9. K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham, "Shallow methods for named entity coreference resolution," in *Proceedings of TALN 2002*, 2002.
10. M. Ariel, *Accessing Noun-Phrase Antecedents*, London: Routledge, 1990.
11. W. Gegg-Harrison and D. Byron, "PYCOT: An Optimality Theory-based Pronoun Resolution Toolkit," in *Proceedings of LREC 2004*, Lisbonne, 2004.
12. C. Schnedecker, "Les chaînes de référence dans les portraits journalistiques : éléments de description," *Travaux de Linguistique* 51, pp. 85–133. Duculot, 2005.
13. G. Kleiber, *Anaphores et Pronoms*. Louvain-la-Neuve : Duculot, 1994.
14. H. Manuélian, Description Définies et Démonstratives : Analyses de Corpus pour la Génération de Textes. PhD thesis, Nancy 2, 2003.
15. B. J. Grosz, S. Weinstein, and A. K. Joshi, "Centering: a framework for modeling the local coherence of discourse," *Computational Linguistics* 21(2), pp. 203–225, 1995.
16. D. Beaver, "The optimization of discourse anaphora," *Linguistics and Philosophy*, 27(1): pp. 3–56, 2004.
17. D. Biber, "Representativeness in corpus design," *Linguistica Computazionale*, IX-X, Current Issues in Computational Linguistics: in honor of Don Walker, 1994.
18. L. Longo and A. Todirascu, "Une étude de corpus pour la détection automatique de thèmes," in *Proceedings of the 6th journées de linguistique de corpus* (JLC 09), Lorient, 2010.
19. R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages," in *Proceedings of the 5th LREC Conference*, pp. 2142–2147, 2006.
20. R. Ion, TTL: A portable framework for tokenization, tagging and lemmatization of large corpora. Bucharest: Romanian Academy, 2007.
21. N. Ide and J. Véronis, "MULTEXT (Multilingual Tools and Corpora)," in *Proceedings of the 14th International Conference on Computational Linguistics*, Kyoto, 1994.
22. D. Mcdonald, "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names," in *proceedings of Corpus Processing for Lexical Acquisition*, MIT press, pp. 21–39, 1996.

23. I. Dagan, A. Itai. "A statistical filter for resolving pronoun references". In Y. A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*, pp. 125–135. Elsevier Science Publishers B.V, 1991.
24. Y. Mathet, A.Widlöcher, "La plate-forme d'annotation Glozz : environnement d'annotation et d'exploration de corpus," in *Proceedings of theTALN 2009*, Senlis, France, 2009.
25. G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation," in *Proceedings of LREC 2004*, pp. 837–840, 2004.