Morphological Rule Induction for Terminology Acquisition

Béatrice Daille
IRIN, 2, rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3 France
daille@irin.univ-nantes.fr

Abstract

We present the identification in corpora of French relational adjectives (RAdj) such as *gazeux (gaseous)* which is derived from the noun *gaz (gas)*. RAdj appearing in nominal phrases are interesting for terminology acquisition because they hold a naming function. The derivational rules employed to compute the noun from which has been derived the RAdj are acquired semi-automatically from a tagged and a lemmatized corpora. These rules are then integrated into a termer which identifies RAdj thanks to their property of being paraphrasable by a prepositional phrase. RAdj and compound nouns which include a RAdj are then quantified, their linguistic precision is measured and their informative status is evaluated thanks to a thesaurus of the domain.

1 Introduction

Identifying relational adjectives (RAdj) such as *malarial*, and noun phrases in which they appear such as *malarial mosquitoes*, could be interesting in several fields of NLP, such as terminology acquisition, topic detection, updating of thesauri, because they hold a naming function acknowledged by linguists: (Levi, 1978), (Mélis-Puchulu, 1991), etc. The use of RAdj is particularly frequent in scientific fields (Monceaux, 1993). Paradoxically, terminology acquisition systems such as TERMINO (David and Plante, 1990), LEXTER (Bourigault, 1992), TERMS (Justeson and Katz, 1995), have not been concerned with RAdj. Even (Ibekwe-Sanjua, 1998) in her study of term variations for identifying research topics from texts does not take into account derivational variants. Our concern is:

1. To identify noun phrases in which relational adjectives appear, as well as the prepositional phrases by which they could be paraphrased. We will see through another source presented in section 2 that this property of paraphrase can be used to identify these adjectives.
2. To check the naming character of these adjectives and to evaluate the naming character of the noun phrases in which they appear.

Moreover, identifying both the adjective and the prepositional phrase is useful in the field of terminology acquisition for performing accurate term normalization by grouping synonym forms referring to an unique concept such as *produit laitier (dairy product)* and, *produit au lait (product with milk)*, *produit de lait (product of milk)*, *produit issu du lait (product made of milk)*, etc. To carry out this identification, we use shallow parsing (Abney, 1991), and then, for morphological processing, a dynamic method which takes as input a corpus labeled with part-of-speech and lemma tags. The morphological rules are built semi-automatically from the corpus.

In this study, we first define, and give some linguistic properties of RAdj. We then present the method to build morphological rules and how to integrate then into a term extractor. We quantify the results obtained from a technical corpus in the field of agriculture [AGRIC] and evaluate their linguistic and informative precision.

2 Linguistic properties of relational adjectives

According to linguistic and grammatical tradition, there are two main categories among adjectives: epithetic such as *important (significant)* and relational adjectives such as *laitier (dairy)*. The first ones cannot have an agentive interpretation in contrast to the second: the adjective *laitier (dairy)* within the noun phrase *production laitière (dairy production)* is an argument to the predicative noun *production (production)* and this is not the case for the adjective *important (significant)* within the phrase production *importante (significant production)*. Relational adjectives (RAdj) possess the following well-known linguistic properties:

- they are either denominal adjectives — morphologically derived from a noun thanks to suffix—, or adjectives having a noun usage such as *mathématique (mathematical/mathematics)*. For the former, not all the adjective-forming suffixes lead to relational adjectives. The following suffixes are considered by (Dubois, 1962) as appropriate:*-ain*, *-aire*, *-al*, *-el*, *-estre*, *-ien*, *-ier*, *-il(e)*, *-in*,*-ique*. However, (Guyon, 1993) remarks that a suffix, even the most appropriate, is never necessary nor sufficient. Several adjectives carrying a favorable suffix are not relational: this is the case with the adjectives ending with *-ique (-ic)*, which characterize chemistry and which are not derived from a noun, such as *désoxyribonucléique (deoryribonucleic)*, *dodecanoique (dodecanoic)*, etc. Other suffixes inappropriate are sometimes used such as the suffixes *-é* and *-eux*: *carbone (carbon)* → *carboné (carbonaceous)*, *cancer (cancer)* → *cancéreux (cancerous)*, etc.
- they own the possibility, in special conditions, of replacing the attributive use of a corresponding prepositional phrase. The preposition employed, as well as the presence or not of a determiner, depends on the head noun of the noun phrase:
  acidité sanguine (blood acidity) ≃ *acidité du sang (acidity of the blood)*
  conquête spatiale (space conquest) ≃ *conquête de l'espace (conquest of space)*
  débit horaire (hourly rate) ≃ *débit par heure (rate per hour)*
  expérimentations animales (animal experimentation) ≃ *expérimentations sur les animaux (experimentation on animals)*
- and several other properties such the impossibility of a predicative position, the incompatibility with a degree modification, etc.

3 Morphological Rule Induction

To identify RAdj trough a term extractor, we use their paraphrastic property which includes the morphological property, the morphological property being insufficient alone. We need rules to recover the lemma of the noun from which the lemma of the RAdj has been derived.
These rules follow the following schemata:
R = [-S +M]{exceptions} where:

S is the relational suffix to be deleted from the end of an adjective. The result of this deletion is the stem **R**;

M is the mutative segment to be concatenated to **R** in order to form a noun;

exceptions list the adjectives that should not be submitted to this rule.

For example, the rule [*-é* +*e*]{*agé*} says that if there is an adjective which ends with *é*, we should strip this ending from it and append the string *e* to the stem except if this adjective belongs to the list of exceptions, namely *agé*.
We extract these morphological rules from the corpora following the method presented in (Mikheev, 1997) with the difference that we don't limit the length of the mutative segment. The relational suffixes are known, only the mutative segments have to be guessed. For the lemma of an adjective ending with a relational suffix in the corpus Adj$_i$, we strip this suffix of Adj$_i$ and store the resulting stem in R. Then, we try to segment this stem R to each noun Noun$_j$ appearing in the corpus. If the subtraction result in an non-empty string, the system creates a morphological rule where the mutative segment is the result of the subtraction of R to Noun$_j$. We thus obtained couples (Adj$_i$, Noun$_j$) associated to a morphological rule. For example: (*gazeux*, *gaz*) [-*eux* +””].
This schemata doesn't take into account stem alternants such as:

e/é *alphab**e**t/aphab**é**t-ique*

è/é *hygi**è**ne/hygi**é**n-ique*

e/i *poll**e**n/poll**i**n-ique*

x/c *thora**x*** / *thora**c**-ique*

In order to handle this allomorphy, we use the Levenshtein's weighted distance (Levenshtein, 1966) which determines the minimum number of insertions or deletions of characters to transform one word into another. (Wagner and Fisher, 1974) presents a recursive algorithm to calculate this distance.

$$\begin{aligned} & dist(w_{1,i}, w'_{1,j}) = \\ & \quad \min(dist(w_{1,i-1}, w'_{1,j}) + q, \\ & \quad dist({}_{w1,i}, w'_{1,j-1}) + q), \\ & \quad dist(w_{1,i-1}, w'_{1,j-1}) + p * dist(w_{i,i}, w'_{j,j})) \end{aligned}$$

with $w_{n,m}$ being the substring beginning at the n^{th} character and finishing after the m^{th} character of the word w,

$$\begin{aligned} dist(x,y) &= 1 \quad if\ x = y \\ &= 0 \quad if\ x \neq y \end{aligned}$$

and

q cost of the insertion/deletion of one character

p cost of the substitution of one character by another.

Generally, a substitution is considered as a deletion followed by an insertion, thus p = 2q. We apply this algorithm to each stem R, obtained after the deletion of the relational suffix, that had not been found as a stem of a noun. But, we add the constraint that R and the noun must share the same two first characters, i.e. the substring computed begin at character 3. We only retain couples composed of an adjective and a noun with a Levenshtein's weighted equal to 3 (i.e. one substitution + one insertion) . From these couples, we deduce new relational suffixes to be added to list of allowed suffixes. More precisely, we consider that such suffixes are allomorphic variants of the relation suffixes. We also add new morphological rules. For example, for the couple (*hygiène*, *hygiénique*), we add the suffix *-énique* which is considered as an allomorph of the suffix *-ique*, and create the rule: [*-énique* +*ène*]. However, this method doesn't retrieve RAdj built from non autonomous bases of noun classes such as *cœur/card (heart/card)*, nor from Latin noun bases such as *père/pater (father/pater)*, *ville/urb (town/urb)*.

We check manually the rules obtained and added to the list of exceptions the wrong derivations obtained. Table 1 presents the number of rules retained and the number of variants for each suffix.

Relational Suffix	Number of allomorphs	Number of rules
-al	3	5
-aire	4	8
-é	2	2
-el	1	2
-er	1	2
-eux	1	3
-ien	1	2
-ier	1	2
-if	2	6
-in	1	2
-ique	8	18
-iste	1	1
-oire	1	1
Total	25	54

Figure 1: Number of variants and rules by relational suffix

4 Term Extractor

First, we present the term extractor chosen then, the modifications perform to enable the application of the derivational rules.

4.1 Initial Term Extractor

ACABIT (Daille, 1996), the term extractor used for this experiment eases the task of the terminologist by proposing, for a given corpus, a list of candidate terms ranked, from the most representative of the domain to the least using a statistical score. Candidate terms which are extracted from the corpus belong to a special type of cooccurrences:

- the cooccurrence is oriented and follows the linear order of the text;
- it is composed of two lexical units which do not belong to the class of functional words such as prepositions, articles, etc.;
- it matches one of the morphosyntactic patterns of what we will call "base terms", or one of their possible variations.

The patterns for base terms are:

Noun1 Adj *emballage biodégradable (biodegradable package)*

Noun1 Noun2 *ions calcium*

Noun1 (Prep (Det)) Noun2 *ions calcium (calcium ion) protéine de poissons (fish protein)*, *chimioprophylaxie au rifampine (rifampicin chemoprophylaxis)*

Noun1 à Vinf *viandes à griller (grill meat)*

These base structures are not frozen structures and do accept several variations. Those which are taken into account are:

1. Inflexional and Internal morphosyntactic variants:
    - graphic and orthographic variants which gather together predictable inflexional variants: *conservation de produit (product preservation)*, *conservations de produit (product preservations)*, or not: *conservation de produits (products preservation)* and case differences.
    - variations of the preposition: *chromatographie en colonne (column chromatography)*, *chromatographie sur colonne (chromatography on column)*;
    - optional character of the preposition and of the article: *fixation azote (nitrogen fixation)*, *fixation d'azote (fixation of nitrogen)*, *fixation de l'azote (fixation of the nitrogen)*;
2. Internal modification variants: insertion inside the base-term structure of a modifier such as the adjective inside the Noun1 (Prep (Det)) Noun2 structure: *lait de brebis (goat's milk)*, *lait cru de brebis (milk straight from the goat)*;
3. Coordinational variants: coordination of base term structures: *alimentation humaine (human diet)*, *alimentation animale et humaine (human and animal diet)*;
4. Predicative variants: the predicative role of the adjective: *pectine méthylée (methylate pectin)*, *ces pectines sont méthylées (these pectins are metylated)*.

The corpus is tagged and lemmatized. The program scans the corpus, counts and extracts collocations whose syntax characterizes base-terms or one of their variants. This is done with shallow parsing using local grammars based on regular expressions (Basili et al., 1993). These grammars use the morphosyntactic information associated with the words of the corpus by the tagger. The different occurrences are grouped as pairs formed by lemmas of the candidate term and sorted following an association measure which takes into account the frequence of the cooccurrences.

4.2 Term Extractor modifications

The identification of relational adjective takes place after extraction of the occurrences of the candidate terms and their syntactic variations. The algorithm below resumes the successive steps for identifying relational adjectives:

1. Examine each candidate of Noun Adj structure;
2. Apply a transformational rule in order to generate all the possible corresponding base nouns. We added morphosyntactic constraints for some suffixes, such as for the suffix *-er*, that the identified adjective is not a past-participle;
3. Search the set of candidate terms for a pair formed with Noun1 (identical between a Noun1 (Prep (Det)) Noun2 and a Noun1 Adj structures) and Noun2 generated from step 2.
4. If step 3 succeeds, group the two base structures under a new candidate term. Take out all the Noun Adj structures owing this adjective from the set of Noun Adj candidates and rename them as a Noun RAdj structure.

In Step 2, morphological rules generate one or several nouns for a given adjective. We generate a noun for each relational suffix class. A class of suffixes includes the allomorphic variants. This overgeneration method used in information retrieval by (Jacquemin and Tzoukermann, 1999) gives low noise because the base noun must not only be an attested for in the corpus, but must also appear as an extension of a head noun. For example, with the adjective *ionique (ionic)*, we generate both *ionie (ionia)* and *ion (ion)*, but only *ion (ion)* is an attested form; with the adjective *gazeux (gaseous)*, the noun forms *gaz (gas)* and *gaze (gauze)*; are generated and the two of them are attested; but, the adjective *gazeux (gaseous)* appears with the

Number of occurrences	1	≥ 2	Total
base structures			
Nom1 Prep (Det) Nom2	17 232	5 949	23 181
Nom Adj	12 344	4 778	17 122
Nom à Vinf	203	16	219
Total	29 912	10 895	40 807

Figure 2: Quantitative data on base structures

noun *échange (exchange)* which is paraphrased in the corpus by *échange de gaz (gas exchange)* and not by *échange de gaze (gauze exchange).* For adjectives with a noun function, as for example *problème technique (technical problem)* and *problème de technique (problem of technics)*, we have accepted that a candidate term could share several base structures: one of type Noun1 (Prep (Det)) Noun2 and another of type Noun1 Adj. No computation is needed to see that Noun2 as Noun2 and Adj share the same lemma.

5 Results and Evaluation

Our corpus, called [AGRIC], is made up of 7 272 abstracts (430 000 words) from French texts in the agriculture domain and extracted from PASCAL. We used the Brill part-of-Speech Tagger (Brill, 1992) trained for French by (Lecomte and Paroubek, 1996)) and the lemmatizer developed by F. Namer (Toussaint et al., 1998).

5.1 Quantitative results

Table 2 resumes the number of base structures extracted from [AGRIC] corpus. From these base structures, 395 groupings were identified. The linked presence of noun phrases of which the extension is fulfilled either by a relational adjective, or be a prepositional phrase the number is rare —a little bit more than 1 % of the total of occurrences—. But, these groupings allow us to extract from the numerous hapax — more than 70 % of the total of occurrences— candidates which, we presume, will be highly denominative and to increase the number of occurrences of a candidate term. The number of relational adjectives which have been identified is 129: *agronomique (agronomical)*, *alimentaire (food)*, *arachidier (groundnut)*, *aromatique (aromatic)*, etc.

5.2 Linguistic Precision

We checked the linguistic accuracy of the 395 structural variations which group a Noun1 Prep (Det) Noun2 structure and a Noun1 RAdj structure. Reported errors concern 3 incorrect groupings due to the homography, and the non homonymy, of the adjective and the noun: *fin (thin (Adj)/end (Noun))*, *courant (ordinary(Adj)/current(Noun))*, *potentiel (potential)*. This lead us to a linguistic precision of more than 99 % in the identification of relational adjectives. As a matter of comparison, (Jacquemin, 1999) obtained a precision of 69,6 % for the Noun to Adj morpho-syntactic variations calculated according to the morphological families produced by a stemming algorithm applied to the MULTEXT lexical database (MULTEXT, 1998) on the same French corpus [AGRIC].

5.3 Informative Precision

The thesaurus (AGROVOC, 1998) is a taxonomy of about 15 000 terms associated with synonyms in a SGML format, which leads to 25 964 different terms. AGROVOC is used for indexing with data fitting agricultural retrieval systems and indexing systems. We made two comparisons with AGROVOC: we first checked whether these RAdjR were really part of terms of it and second, we compared the candidate terms extracted with a RAdj with its terms. We consider that the presence of the RAdj in AGROVOC confirms its informative character, and that the presence of a candidate term attests its terminological value.

5.3.1 Relational adjectives alone

From the 124 correct RAdj, 68 appear inside terms of the thesaurus in epithetic position, and 15 only under their noun form in an extension position, for example *arachidier (groundnut)* does not appear but *arachide* is used in an extension position. Moreover, among the 124 adjectives, 73 appear in AGROVOC under their noun term as uniterms. The adjectives which are not present in the thesaurus in an extension position under either their adjectival or noun form are 11 in number. So 93 % of them are indeed highly informative.

5.3.2 Candidate terms with a relational adjective

Pour 9 AdjR belonging to AGROVOC, we compute the following indexes:

T$_A$ the number of terms in AGROVOC in which the relational adjective appears in an epithetic position, i.e. the terms of Noun RAdj structure. For example T$_A$=15 for the adjective *cellulaire (cellular)* because it appears in 15 terms of AGROVOC such as *différenciation cellulaire (cellular differenciation)*, *division cellulaire (cellular division).*

T$_N$ the number of terms in AGROVOC in which the noun from which has been derived the relational adjective appears inside a prepositional phrase, i.e. the terms of Noun1 Prep (Det) Noun$_{RAdj}$ structure. For example T$_N$=4 for the noun *cellule (cell)* because it appears in 4 terms of AGROVOC such as *banque de cellules (cell bank)*, *culture de cellules (culture of cells).*

C$_A$ the number of candidate terms of Noun RAdj structure. For example, C$_A$=61 for the adjective *cellulaire (cellular)* because it appears in 61candidate terms such as *acide cellulaire (cellular acid)*, *activité cellulaire (cellular activity)*, *agrégat cellulaire (cellular aggregate).*

C$_N$ the number of candidate terms of Noun1 Prep (Det) Noun$_{RAdj}$ structure. For example C$_N$=58 for the noun *cellule (cell)* because it appears in 58 candidate terms such as *ADN de cellule (cell DNA)*, *addition de cellules (cell addition).*

Then, for each candidate term of C$_A$ and C$_N$, we checked for their presence in AGROVOC. The only matches that we have accepted are exact matches. With this comparison, we obtained the following indexes:

a the number of candidate terms of Noun RAdj structure found in AGROVOC under the Noun RAdj structure.

b the number of candidate terms of Noun RAdj structure found in AGROVOC under the Noun1 Prep (Det) Noun$_{RAdj}$ structure.

	Noun RAdj	N1 Prep (Det) N$_{RAdj}$
Precision	0,34	0,04
Recall	0,46	0,14

Figure 3: Averages of precisions and recalls

c the number of candidate terms of Noun1 Prep (Det) Noun$_{RAdj}$ structure found in AGROVOC under the Noun RAdj structure.

d the number of candidate terms of Noun1 Prep (Det) Noun$_{RAdj}$ structure found in AGROVOC under the Noun1 Prep (Det) Noun$_{RAdj}$ structure.

These indexes allow us to compute precision P and recall R for each Noun RAdj structure and each Noun1 Prep (Det) Noun$_{RAdj}$ structure with the help of the following formula:

$$P_{NounRAdj} = \frac{(a+b)}{C_A} \quad (1)$$

$$P_{NounPrep(Det)Noun_{RAdj}} = \frac{(c+d)}{C_N} \quad (2)$$

$$R_{NounRAdj} = \frac{(a+b)}{T_A} \quad (3)$$

$$R_{NounPrep(Det)Noun_{RAdj}} = \frac{(c+d)}{T_N} \quad (4)$$

The averages of precision and recall for the two structures are summarized in table 3. This comparison of the average of precision computed shows that candidate terms with a Noun RAdj structure are 10 times more likely to be terms than their equivalent in Noun1 Prep (Det) Noun$_{RAdj}$. The analysis of the average of recall is also impressive: it is generally difficult to obtain a recall superior to 25 % when comparing candidate terms extracted from a corpus and a thesaurus of the same domain (Daille et al., 1998). The average of recalls obtained thanks to the identification of RAdj shows that nearly half of the terms built with the defined RAdj are identified. These good values of precision and recall have been obtained on linguistic criteria only without taking into account frequency.

6 Conclusion

The method proposed in this study to acquire morphological rules from corpora in order to recover derivational term variations trough a term extractor and identify relational adjectives shows an excellent precision. We have also proved that noun phrases including a RAdj are far more informative than their equivalent in Noun1 Prep (Det) Noun$_{RAdj}$ structure. We still have to write the program whose task will be to merge new morphological rules acquired from another corpus with the existing ones.

References

S. Abney. 1991. Parsing with chunks. In R. Berwick and C. Tenny, editors, *Principle-Base Parsing*, pages 257–278. Kluwer Academoc Publishers.

AGROVOC, 1998. *AGROVOC - Multilingual Agricultural Thesaurus*. Food and Agricultural Organization of the United Nations. http://www.fao.org.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. 1993. Acquisition of Selectional Patterns in Sublanguages. *Machine Tranlation*, 8:175–201.

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING'92*, pages 977–981, Nantes, France.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *ANLP'92*, pages 152–155, Trento, march.

Béatrice Daille, Eric Gaussier, and Jean-Marc Langé. 1998. An evaluation of statistical scores for word association. In Jonathan Ginzburg, Zurab Khasidashvili, Carl Vogel, Jean-Jacques Lévy, and Enric Vallduvi, editors, *The Tblisi Symposium on Logic, Language and Computation: Selected Papers*, pages 177–188. CSLI Publications.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act - Combining Symbolic and Statistical Approaches to Language*, chapter 3, pages 28–49. MIT Press.

Sophie David and P. Plante. 1990. Le progiciel termino : De la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes. In *ICO*, volume 2.

J. Dubois. 1962. *Etude sur la dérivation suffixale en Français moderne et contemporain.* Larousse, Paris.

Anne Guyon. 1993. *Les adjectifs relationnels arguments de noms prédicatifs.* Ph.D. thesis, Université Paris 7.

Fidelia Ibekwe-Sanjua. 1998. Terminological variation, a mean of identifying research topics from texts. In *COLING-ACL'98*, volume 1, pages 564–570, Montral, Canada.

Christian Jacquemin and Evelyne Tzoukermann. 1999. Npl for term variant extraction: Synergy between morphology, lexicon and syntax. In T. Strzalkowski, editor, *Natural Language Processing and Information Retrieval.* Kluwer, Boston, MA.

Christian Jacquemin. 1999. Syntagmatic and Paradigmatic Representation of Term Variation. In *ACL'99*, University of Maryland.

J. Justeson and S. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. In *Journal of Linguistic Engineering*, volume 1.

Josette Lecomte and Patrick Paroubek. 1996. Le catégoriseur d'eric brill. mise en œuvre de la version entranée à l'inalf. Technical report, CNRS-INALF.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys.-Dokl.*, 10(8):707–710.

Judith Levi. 1978. *The syntax and the semantics of complex nominals.* Academic Press, London.

A. Mélis-Puchulu. 1991. Les adjectifs dénominaux : des adjectifs de "relation". *Lexique*, 10:33–60.

Andrei Mikheev. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Anne Monceaux. 1993. *La formation des noms composés de structure NOM ADJECTIF.* Thèse de doctorat en linguistique théorique et formelle, Université de Marne la Vallée.

MULTEXT, 1998. Laboratoire Parole et Langage. http://www.lpl.univ-aix.fr.

Yannick Toussaint, Fiametta Namer, Béatrice Daille, Christian Jacquemin, Jean Royauté, and Nabil Hathout. 1998. Une approche linguistique et statistique pour l'analyse de l'information en corpus. In *TALN'98*, pages 182–191. Paris.

R.A. Wagner and M.J. Fisher. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.