Syntagmatic and Paradigmatic Representations of Term Variation

Christian Jacquemin
LIMSI-CNRS
BP 133
91403 ORSAY Cedex
FRANCE
jacquemin@limsi.fr

Abstract

A two-tier model for the description of morphological, syntactic and semantic variations of multi-word terms is presented. It is applied to term normalization of French and English corpora in the medical and agricultural domains. Five different sources of morphological and semantic knowledge are exploited (MULTEXT, CELEX, AGROVOC, WordNet1.6, and Microsoft Word97 thesaurus).

1 Introduction

In the classical approach to text retrieval, terms are assigned to queries and documents. The terms are generated by a process called automatic indexing. Then, given a query, the similarity between the query and the documents is computed and a ranked list of documents is produced as output of the system for information access (Salton and McGill, 1983).

The similarity between queries and documents depends on the terms they have in common. The same concept can be formulated in many different ways, known as *variants*, which should be conflated in order to avoid missing relevant documents. For this purpose, this paper proposes a novel model of term variation that integrates linguistic knowledge and performs accurate *term normalization*. It relies on previous or ongoing linguistic studies on this topic (Sparck Jones and Tait, 1984; Jacquemin et al., 1997; Hamon et al., 1998). Terms are described in a two-tier framework composed of a *paradigmatic level* and a *syntagmatic level* that account for the three linguistic dimensions of term variability (morphology, syntax, and semantics). Term variants are extracted from tagged corpora through *FASTR*[1], a unification-based transformational parser described in (Jacquemin et al., 1997).

Four experiments are performed on the French and the English languages and a measure of precision is provided for each of them. Two experiments are made on a French corpus [AGRIC] composed of 1.2×10^6 words of scientific abstracts in the agricultural domain and two on an English corpus [MEDIC] composed of 1.3×10^6 words of scientific abstracts in the medical domain. The two experiments in the French language are [AGRIC] + Word97 and [AGRIC] + AGROVOC. In the former, synonymy links are extracted from the Microsoft Word97 thesaurus; in the latter, semantic classes are extracted from the AGROVOC thesaurus, a thesaurus specialized in the agricultural domain (AGROVOC, 1995). In both experiments, morphological data are produced by a stemming algorithm applied to the MULTEXT lexical database (MULTEXT, 1998). The two experiments on the English language are [MEDIC] + WordNet 1.6 or [MEDIC] + Word97; they correspond to two different sources of semantic knowledge. In both cases, the morphological data are extracted from CELEX (CELEX, 1998).

2 Term Variation: Representation and Exploitation

Terms and variations are represented into two parallel frameworks illustrated by Figure 1. While terms are described by a unique pair composed of a structure—at the syntagmatic level—and a set of lexical items—at the paradigmatic level—, a variation is represented by a pair of such pairs: one of them is the source term (or normalized term) and the other one is the target term (or variant).

The syntagmatic description of a term is a context free rule; it is complemented with lexical information embedded in a feature structure denoted by constraints between paths and values. For instance, the term *speed measurement* is represented by:

$$\left\{ \begin{array}{l} \text{Syntagm: } \{\mathrm{N}_0 \rightarrow \mathrm{N}_2\ \mathrm{N}_1\} \\ \text{Paradigm: } \left\{ \begin{array}{l} \langle \mathrm{N}_1\ \textit{lemma} \rangle = \textit{measurement} \\ \langle \mathrm{N}_2\ \textit{lemma} \rangle = \textit{speed} \end{array} \right\} \end{array} \right. \quad (1)$$

This term is a noun phrase composed of a head noun N_1 and a modifier N_2; the lemmas are given by the constraints at the paradigmatic level. This framework is similar to the unification-based representation of context-free grammars of (Shieber, 1992).

[1] *FASTR* can be downloaded from www.limsi.fr/Individu/jacquemi/FASTR.

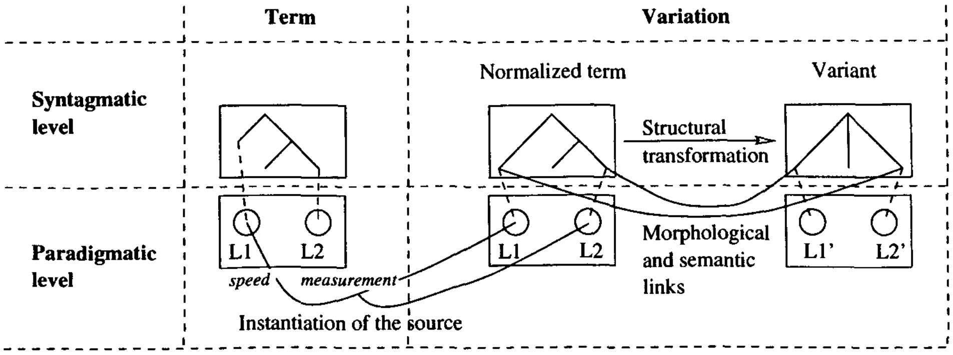

Figure 1: Two level description of terms and variations

At the syntagmatic level, variations are represented by a source and a target structure. At the paradigmatic level, the lexical elements of variations are not instantiated in order to ensure higher generality. Instead, links between lexical elements are provided. They denote morphological and/or semantic relations between lexical items in the source and target structures of the variation. For example, the variation that associates a Noun-Noun term such as the preceding term $speed_{\mathrm{N}_2}$ $measurement_{\mathrm{N}_1}$ with a verbal form of the head word and a synonym of the argument such as $measuring_{\mathrm{V}_1}$ $maximal_{\mathrm{A}}$ $shortening_{\mathrm{N}}$ $velocity_{\mathrm{N}'_2}$ is given by:

$$\text{Syntagm:}\\ \left\{\begin{array}{l}(\mathrm{N}_0 \rightarrow \mathrm{N}_2\ \mathrm{N}_1) \Rightarrow \\ (\mathrm{V}_0 \rightarrow \mathrm{V}_1\ (\mathrm{Prep}^?\ \mathrm{Det}^?\ (\mathrm{A}|\mathrm{N}|\mathrm{Part})^\star)\ \mathrm{N}'_2)\end{array}\right\} \tag{2}$$

$$\text{Paradigm:}\left\{\begin{array}{l}\langle \mathrm{N}_1\ root\rangle = \langle \mathrm{V}_1\ root\rangle \\ \langle \mathrm{N}_2\ sem\rangle = \langle \mathrm{N}'_2\ sem\rangle\end{array}\right\}$$

If this variation is instantiated with the term given in (1), it recognizes the lexico-syntactic structure

$$\mathrm{V}_1\ (\mathrm{Prep}^?\ \mathrm{Det}^?\ (\mathrm{A}|\mathrm{N}|\mathrm{Part})^\star)\ \mathrm{N}'_2 \tag{3}$$

in which V_1 and *measurement* are morphologically related, and N'_2 and *speed* are semantically related. The target structure is under-specified in order to describe several possible instantiations with a single expression and is therefore called a *candidate variation*. In this example, a regular expression is used to under-specify the structure[2]; another solution would be to use quasi-trees with extended dependencies (Vijay-Shanker, 1992).

3 Paradigmatic relations

As illustrated by Figure 2 and Formula (2), there are two types of paradigmatic relations between lemmas involved in the definition of term variations: morphological and semantic relations. The morphological family of a lemma l is denoted by the set $F_M(l)$ and its semantic family by the set $F_{S_L}(l)$ or $F_{S_C}(l)$.

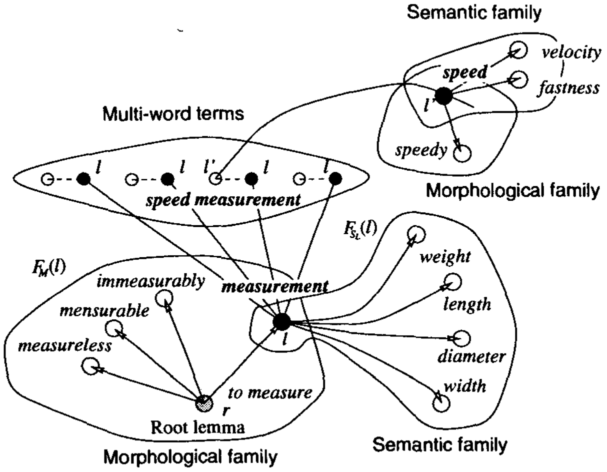

Figure 2: Paradigmatic links between lemmas

Roughly speaking, two words are **morphologically related** if and only if they share the same root. In the preceding example, *to measure* and *measurement* are in the same morphological family because their common root is *to measure*. Let $\mathcal{L}$ be the set of lemmas, morphological roots define a binary relation M from $\mathcal{L}$ to $\mathcal{L}$ that associates each lemma with its root(s): $M \in \mathcal{L} \nleftrightarrow \mathcal{L}$. M is not a function because compound lemmas have more than one root.

The morphological family $F_M(l)$ of a lemma l is the set of lemmas (including l) which share a common root with l:

$$\left\{\begin{array}{l}F_M \in \mathbb{P}(\mathcal{L}) \\ \forall l \in \mathcal{L}, F_M(l) = \{l' \in \mathcal{L} \bullet \exists r \in \mathcal{L}, (l,r) \in M \\ \quad \wedge (l',r) \in M\} = M^{-1}(M(\{l\}))\end{array}\right. \tag{4}$$

[2] A stands for adjective, N for noun, Prep for preposition, V for verb, Det for determiner, Part for participle, and Adv for adverb.

($\mathbb{P}(\mathcal{L})$ is the power-set of $\mathcal{L}$, the set of its subsets.)

There are principally two types of **semantic relations**: direct links through a binary relation $S_L \in \mathcal{L} \leftrightarrow \mathcal{L}$ or classes $\mathcal{C} \in \mathbb{P}(\mathbb{P}(\mathcal{L}))$.

In the case of semantic links, the semantic family $F_{S_L}(l)$ of a lemma l is the set of lemmas (including l) which are linked to l:

$$\begin{cases} F_{S_L} \in \mathbb{P}(\mathcal{L}) \\ \forall l \in \mathcal{L}, F_{S_L}(l) = \{l' \in \mathcal{L} \bullet (l,l') \in S_L\} \cup \{l\} \\ \quad = S_L(\{l\}) \cup \{l\} \end{cases} \quad (5)$$

In the case of semantic classes, the semantic family $F_{S_C}(l)$ of a lemma l is the union of all the classes to which it belongs:

$$\begin{cases} F_{S_C} \in \mathbb{P}(\mathcal{L}) \\ \forall l \in \mathcal{L}, F_{S_C}(l) = \left(\bigcup_{(c \in \mathcal{C}) \wedge (l \in c)} c \right) \cup \{l\} \end{cases} \quad (6)$$

Links and classes are equivalent, the choice of either model depends on the type of available semantic data. In the experiments reported here, direct links are used to represent data extracted from the word processor Microsoft Word97 because they are provided as lists of synonyms associated with each lemma. Conversely, the synsets extracted from WordNet 1.6 (Fellbaum, 1998) are classes of disambiguated lemmas and, therefore, correspond to the second technique.

With respect to the definitions of semantic and morphological families given in this section, the candidate variant (3) is such that $\text{V}_1 \in F_M(\textit{measurement})$ and $\text{N}'_2 \in F_{S_L}(\textit{speed})$ or $\text{N}'_2 \in F_{S_C}(\textit{speed})$.

4 Morphological and Semantic Families

In the experiments on the English corpora, the CELEX database is used to calculate morphological families. As for semantic families, either WordNet 1.6 or the thesaurus of Microsoft Word97 are used.

Morphological Links from CELEX

In the CELEX morphological database (CELEX, 1998), each lemma is associated with a morphological structure that contains one or more root lemmas. These roots are used to calculate morphological families according to Formula (4). For example, the morphological family $F_M(\textit{measurement}_\text{N})$ of the lemmas with $\textit{measure}_\text{V}$ as root word is {*commensurable*$_\text{A}$, *commensurably*$_\text{Adv}$, *countermeasure*$_\text{N}$, *immeasurable*$_\text{A}$, *immeasurably*$_\text{Adv}$, *incommensurable*$_\text{A}$, *measurable*$_\text{A}$, *measurably*$_\text{Adv}$, *measure*$_\text{N}$, *measureless*$_\text{A}$, *measurement*$_\text{N}$, *mensurable*$_\text{A}$, *tape-measure*$_\text{N}$, *yard-measure*$_\text{N}$, *measure*$_\text{V}$}.

Semantic Classes from WordNet

Two sources of semantic knowledge are used for the English language: the WordNet 1.6 thesaurus and the thesaurus of the word processor Microsoft Word97. In the WordNet thesaurus, disambiguated words are grouped into sets of synonyms—called *synsets*—that can be used for a class-based approach to semantic relations. For example, each of the five disambiguated meanings of the polysemous noun *speed* belongs to a different synset. In our approach, words are not disambiguated and, therefore, the semantic family of *speed* is calculated as the union of the synsets in which one of its senses is included. Through Formula (6), the semantic family of *speed* based on WordNet is: $F_{S_C}(\textit{speed}_\text{N}) =$ {*speed*$_\text{N}$, *speeding*$_\text{N}$, *hurrying*$_\text{N}$, *hastening*$_\text{N}$, *swiftness*$_\text{N}$, *fastness*$_\text{N}$, *velocity*$_\text{N}$, *amphetamine*$_\text{N}$}.

Semantic Links from Microsoft Word97

For assisting document edition, the word processor Microsoft Word97 has a command that returns the synonyms of a selected word. We have used this facility to build lists of synonyms. For example, $F_{S_L}(\textit{speed}_\text{N}) =$ {*speed*$_\text{N}$, *swiftness*$_\text{N}$, *velocity*$_\text{N}$, *quickness*$_\text{N}$, *rapidity*$_\text{N}$, *acceleration*$_\text{N}$, *alacrity*$_\text{N}$, *celerity*$_\text{N}$} (Formula (5)). Eight other synonyms of the word *speed* are provided by Word97, but they are not included in this semantic family because they are not categorized as nouns in CELEX.

5 Variations

The linguistic transformations for the English language presented in this section are somehow simplified for the sake of conciseness. First, we focus on binary terms that represent 91.3% of the occurrences of multi-word terms in the English corpus [MEDIC]. Then, simplifications in the combinations of types of variations are motivated by corpus explorations in order to focus on the most productive families of variations.

The 3 Dimensions of Linguistic Variations

There are as many types of **morphological relations** as pairs of syntactic categories of content words. Since the syntactic categories of content words are noun (N), verb (V), adjective (A), and adverb (Adv), there are potentially sixteen different pairs of morphological links. (Associations of identical categories must be taken into consideration. For example, Noun-Noun associations correspond to morphological links between substantive nouns such as agent/process: *promoter/promotion.*) Morphological relations are further divided into *simple* relations if they associate two words in the same position and *crossed* relations if they associate a head word and an argument. Combining categories and positions, there are, in all, 64 different types of morphological relations.

In (Hamon et al., 1998), three types of **semantic relations** are studied: a link between the two head words, a link between the two arguments, or two parallel links between heads and arguments. These authors report that double links are rare and that their quality is low. They only represent 5% of the semantic variations on a French corpus and they are extracted with a precision of 9% only. We will therefore focus on single semantic links. Since we are only concerned with synonyms, only two types of semantic links are studied: synonymous heads or synonymous arguments.

The last dimension of term variability is the **structural transformation** at the syntagmatic level. The source structure of the variation must match a term structure. There are basically two structures of binary terms: $X_1\,N_2$ compounds in which X_1 is a noun, an adjective or a participle, and N_1 Prep N_2 terms. According to (Jacquemin et al., 1997), there are three types of syntactic variations in French: coordinations (Coor), insertions of modifiers (Modif), and compounding/decompounding (Comp). Each of these syntactic variations is further subdivided into finer categories.

Multi-dimensional Linguistic Variations

The overall picture of term variations is obtained by combining the 64 types of morphological relations, the two types of semantic relations and the three types of syntactic variations (and their sub-types). There are different constraints on these combinations that limit the number of possible variations:

1. Morphological and semantic links must operate on different words. For example, if the head word is transformed by a morphological link, the only word available for a semantic link is the argument word.
2. The target syntactic structure must be compatible with the morphological transformations. For example, if a noun is transformed into a verb, the target structure must be a verb phrase.

These two constraints influence the way in which a variation can be defined by combining different types of elementary modifications. Firstly, lexical relations are defined at the paradigmatic level: morphological links, semantic links or identical words. Then a syntactic structure that is compatible with the categories of the target words is chosen.

The list of variations used for binary compound terms in English is given in Table 1.[3] It has been experimentally refined through a progressive corpus-based tuning. The **Synt** column gives the target syntactic structure. The **Morph** column describes the morphological link: a source and a target syntactic category and the syntactic positions of the source and target lemmas. The **Sem** column indicates whether the variation involves a semantic link and the position of the lemmas concerned by the link (both lemmas must have an identical position). The **Pattern** column gives the target syntactic structure as a function of the source structure which is either $X_1\,N_2$, $A_1\,N_2$, or $N_1\,N_2$.

For example, Variation #42 transforms an Adjective-Noun term $A_1\,N_2$ into

$$N_1\ ((\mathrm{CC\ Det}^?)^?\ \mathrm{Prep\ Det}^?\ (\mathrm{A|N|Part})^{0-3})\ N'_2$$

N_1 is a noun in the morphological family of A_1 (noted $F_M(A_1)_N$) and N'_2 is semantically related with N_2 (noted $F_S(N_2)$). This variation recognizes *malignancy in orbital tumours* as a variant of *malignant tumor* because *malignancy* and *malignant* are morphologically related, *tumour* and *tumor* are semantically related, and $\textit{malignancy}_{\mathrm{N}}$ $\textit{in}_{\mathrm{Prep}}$ $\textit{orbital}_{\mathrm{A}}$ $\textit{tumours}_{\mathrm{N}}$ matches the target pattern. Variation #56 is a more elaborated version of variation (2) given in Section 2.

Sample Syntactico-semantic Variants from [MEDIC]

The first 36 variations in Table 1 do not contain any morphological link. They are built as follows. Firstly, the different structures of noun phrases are used as target structures. Twelve structures are proposed: head coordination (#1), argument coordination (#4), enumeration with conjunction (#7), enumeration without conjunction (#10), etc.

Then each transformation is enriched with additional semantic links between the head words or between the argument words. Semantic links between argument words are found in variations $\#(3n+2)_{0\leq n\leq 11}$ and between head words in variations $\#(3n)_{1\leq n\leq 12}$. (Due to the lack of space, only variations #2 and #3 constructed on top of variation #1 are shown in Table 1.) Sample variants from [MEDIC] for the first 36 variations are given in Table 2. Some variations have not matched any variant in the whole corpus.

Sample Morpho-syntactico-semantic Variants

Morpho-syntactico-semantic variations are numbered #37 to #62 in Table 1. Only 10 of the 64 possible morphological associations are found in the list of morphological links: Noun to Adjective on arguments (#37), Adjective to Noun on arguments (#39), etc. Each of these variations is doubled by adding a semantic link between the words that are not morphologically related. For example, variation (#40) is deduced from variation (#39) by adding a semantic link between the head words. Sample variants are given in Table 3.

[3] Punctuations are noted Pu and coordinating conjunction CC.

Table 1: Patterns of semantic variation for terms of structure X_1 N_2.

#	Synt.	Morph.	Sem.	Pattern
1	Coor	—	—	X_1[sin] ((A$\vert$N$\vert$Part)$^{0-3}$ N Pu[',']$^?$ CC) N_2
2	Coor	—	Arg	$F_S(X_1)$[sin] ((A$\vert$N$\vert$Part)$^{0-3}$ N Pu[',']$^?$ CC) N_2
3	Coor	—	Head	X_1[sin] ((A$\vert$N$\vert$Part)$^{0-3}$ N Pu[',']$^?$ CC) $F_S(N_2)$
4	Coor	—	-	X_1[sin] (CC (A$\vert$N$\vert$Part)$^{0-3}$) N_2
7	Coor	—	-	X_1 (Pu (A$\vert$N$\vert$Part) Pu$^?$ CC (A$\vert$N$\vert$Part)) N_2
10	Coor	—	-	X_1[sin] (Pu (A$\vert$N$\vert$Part) Pu (A$\vert$N$\vert$Part) Pu$^?$ CC (A$\vert$N$\vert$Part)) N_2
13	Coor	—	-	X_1[sin] ((A$\vert$N$\vert$Part)$^{0-3}$ N Pu[','] CC) N_2
16	Modif	—	—	X_1[sin] ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
19	Modif	—	—	X_1[sin] (N Prep Det$^?$ A$^?$) N_2
22	Modif	—	—	X_1[sin] (Pu[')'] (A$\vert$N$\vert$Part)$^?$) N_2
25	Modif	—	—	X_1[sin] (Pu['('] CC$^?$ (A$\vert$N$\vert$Part)$^{1-2}$ Pu[')']) N_2
28	Modif	—	—	X_1[sin] (Pu[','] (A$\vert$N$\vert$Part)) N_2
31	Perm	—	-	N_2 (V['be']$\vert$Pu['(']) X_1
34	Perm	—	-	N_2 (V$^?$ Prep Det$^?$ (A$\vert$N$\vert$Part)$^{0-3}$ ((N) CC Det$^?$)$^?$) N_1
37	Modif	N→A (Arg)	—	$F_M(N_1)_A$ ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
38	Modif	N→A (Arg)	Head	$F_M(N_1)_A$ ((A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_2)$
39	Modif	A→N (Arg)	—	$F_M(A_1)_N$ ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
40	Modif	A→N (Arg)	Head	$F_M(A_1)_N$ ((A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_2)$
41	Perm	A→N (Arg)	—	$F_M(A_1)_N$ ((CC Det$^?$)$^?$ Prep Det$^?$ (A$\vert$N$\vert$Part)$^{0-3}$) N_2
42	Perm	A→N (Arg)	Head	$F_M(A_1)_N$ ((CC Det$^?$)$^?$ Prep Det$^?$ (A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_2)$
43	Perm	A→N (Arg)	—	N_2 ((Prep Det$^?$)$^?$ (A$\vert$N$\vert$Part)$^{0-3}$) $F_M(A_1)_N$
44	Perm	A→N (Arg)	Head	$F_S(N_2)$ ((Prep Det$^?$)$^?$ (A$\vert$N$\vert$Part)$^{0-3}$) $F_M(A_1)_N$
45	Modif	A→Adv (Arg)	—	$F_M(A_1)_{Adv}$ ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
46	Modif	A→Adv (Arg)	Head	$F_M(A_1)_{Adv}$ ((A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_2)$
47	Modif	A→A (Arg)	—	$F_M(A_1)_A$ ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
48	Modif	A→A (Arg)	Head	$F_M(A_1)_A$ ((A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_2)$
49	Modif	N→N (Head)	—	X_1 ((A$\vert$N$\vert$Part)$^{0-3}$) $F_M(N_2)_N$
50	Modif	N→N (Head)	Arg	$F_S(X_1)$ ((A$\vert$N$\vert$Part)$^{0-3}$) $F_M(N_2)_N$
51	Modif	N→N (Arg)	—	$F_M(N_1)_N$ ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
52	Modif	N→N (Arg)	Head	$F_M(N_1)_N$ ((A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_2)$
53	Perm	N→N (Head)	—	$F_M(N_2)_N$ (Prep (A$\vert$N$\vert$Part)$^{0-3}$) N_1
54	Perm	N→N (Head)	Arg	$F_M(N_2)_N$ (Prep (A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_1)$
55	VP	N→V (Head)	—	$F_M(N_2)_V$ (Adv$^?$ Prep$^?$ (Det (N)$^?$ Prep)$^?$ Det$^?$ (A$\vert$N$\vert$Part)$^{0-3}$) N_1
56	VP	N→V (Head)	Arg	$F_M(N_2)_V$ (Adv$^?$ Prep$^?$ (Det (N)$^?$ Prep)$^?$ Det$^?$ (A$\vert$N$\vert$Part)$^{0-3}$) $F_S(N_1)$
57	VP	N→V (Head)	—	N_1 ((N)$^?$ V['be'] $^?$) $F_M(N_2)_V$
58	VP	N→V (Head)	Arg	$F_S(N_1)$ ((N)$^?$ V['be'] $^?$) $F_M(N_2)_V$
59	NP	N→V (Head)	—	A_1 ((A$\vert$N$\vert$Part)$^{0-2}$ ((N) Prep)$^?$) $F_M(N_2)_V$
60	NP	N→V (Head)	Arg	$F_S(A_1)$ ((A$\vert$N$\vert$Part)$^{0-2}$ ((N) Prep)$^?$) $F_M(N_2)_V$
61	NP	V→N (Arg)	—	$F_M(V_1)_N$ ((A$\vert$N$\vert$Part)$^{0-3}$) N_2
62	NP	V→N (Arg)	Head	$F_M(V_1)_N$ ((A$\vert$N$\vert$Part)$^{0-3}$)$F_S(N_2)$

6 Evaluation

We provide two evaluations of term variant conflation. First, we calculate precision rates through a manual scanning of the variants. Secondly, we evaluate the numbers of variations extracted through the four experiments.

Precision

Because of the large volumes of data, only experiments on the French corpus are evaluated. [AGRIC] + AGROVOC produces 2,739 variations and 2,485 of them are selected as correct. Since the number of synonym links proposed by Word97 is higher, the number of variants produced by [AGRIC] + Word97 is higher: 3,860. 3,110 of them are accepted after human inspection.

The two experiments produce the same set of non-semantic variants (syntactic and morpho-syntactic variants). Associated values of precision are reported in Tables 4 and 5. The semantic variations are divided into two subsets: "pure" semantic variations and semantic variations involving a syntactic transformation and/or a morphological link. Their precisions are given in Tables 6 and 7.

As far as precision is concerned, these tables show that variations are divided into two levels of quality. On the one hand, syntactic, morpho-syntactic and pure semantic variations are extracted with a high level of precision (above 78%, see the "Total" values in Tables 4 to 6). On the other hand, the

Table 2: Sample variants from [MEDIC] using the variations from Table 1 (#1 to #36).

#	Term	Variant
1	*cell differentiation*	*cell growth and differentiation*
2	*primary response*	*basal secretory activity and response*
3	*pressure decline*	*pressure rise and fall*
4	*adipose tissue*	*adipose or fibroadipose tissue*
5	*extensive resection*	*wide or radical resection*
6	*clinical test*	*clinical and histologic examinations*
7	*adipic acid*	*adipic, suberic and sebacic acids*
8	*morphological change*	*morphologic, ultrastructural and immunologic changes*
9	*clinical test*	*clinical, radiographic, and arthroscopic examination*
10	*electrical property*	*electrical, mechanical, thermal and spectroscopic properties*
12	*hypothesis test*	*hypothesis, comparability, randomized and non-randomized trials*
16	*acidic protein*	*acidic epidermal protein*
17	*absorbed dose*	*ingested human doses*
18	*cylindrical shape*	*cylindrical fiberglass cast*
19	*assisted ventilation*	*assisted modes of mechanical ventilation*
20	*genetic disease*	*hereditary transmission of the disease*
21	*early pregnancy*	*early stage of gestation*
22	*intertrochanteric fracture*	*intertrochanteric) femoral fractures*
25	*arteriovenous fistula*	*arteriovenous (AV) fistulas*
27	*pressure measurement*	*pressure (SBP) measure*
28	*identification test*	*identification, sensory tests*
29	*electrical stimulus*	*electric, acoustic stimuli*
31	*combined treatment*	*treatments were combined*
32	*genetic disease*	*disease is familial*
33	*increased dose*	*dosage was increased*
34	*acrylonitrile copolymer*	*copolymer of acrylonitrile*
35	*development area*	*areas of growth*
36	*cell death*	*destruction of the virus-infected cell*

Table 3: Sample variants from [MEDIC] using the variations from Table 1 (#37 to #62).

#	Term	Variant
37	*cell component*	*cellular component*
38	*work place*	*workable space*
39	*embryonic development*	*embryo development*
40	*angular measurement*	*angles measure*
41	*deficient diet*	*deficiency in the diet*
42	*malignant tumor*	*malignancy in orbital tumours*
43	*cerebral cortex*	*cortex of the cerebrum*
44	*surgical advancement*	*advance in middle ear surgery*
45	*inappropriate secretion*	*inappropriately high TSH secretion*
46	*genetic variant*	*genetically determined variance*
47	*fatty meal*	*fat meals*
48	*optical system*	*optic Nd-YAG laser unit*
49	*drug addiction*	*drug addicts*
50	*simultaneous measurement*	*concurrent measures*
51	*saline solution*	*salt solution*
52	*flow limit*	*airflow limitation*
53	*bile reflux*	*flux of bile*
55	*measurement technique*	*measuring technique*
57	*age estimation*	*estimating gestational age*
58	*density measurement*	*measured COHb concentrations*
59	*blood coagulation*	*blood coagulated*
60	*concentration measurement*	*density was measured*
61	*combined treatment*	*combination treatment*

Table 4: Precision of syntactic variant extraction ([AGRIC] corpus).

Coor	Modif	Comp	Total
97.2%	88.7%	98.0%	**95.7%**

Table 5: Precision of morpho-syntactic variant extraction ([AGRIC] corpus).

A to N	N to A	N to N	N to V	Total
68.5%	69.6%	92.1%	75.3%	**84.6%**

Table 6: Precision of semantic variant extraction ([AGRIC] corpus).

	Word97	AGROVOC
Sem Arg	76.3%	88.9%
Sem Head	82.7%	91.3%
Total	**78.1%**	**91.0%**

Table 7: Precision of semantico-syntactic variant extraction ([AGRIC] corpus).

	Word97	AGROVOC
Coor + sem	44.8%	62.6%
Modif + sem	55.6%	87.5%
A to N + sem	44.9%	0.0%
N to A + sem	21.3%	0.0%
N to N + sem	0.0%	60.0%
N to V + sem	24.2%	44.4%
Total	**29.4%**	**55.0%**

combination of semantic links with syntax or with morphology results in poor precision (55% precision in average with the AGROVOC semantic links and 29.4% precision with the Word97 links, see line "Total" in Table 7).

The lower precision of hybrid variations is due to a cumulative effect of semantic shift through combined variations. For instance, *former un réseau continu* (build a continuous network) is incorrectly extracted as a variant of *formation permanente* (continuing education) through a Noun-to-Verb variation with a semantic link between argument words. The verb *former* and the associated deverbal noun *formation* are two polysemous words. In *formation permanente*, the meaning is related to a human activity (*to train*) while, in *former un réseau continu*, the meaning is related to a physical construction (*to build*).

Despite the relatively poor precision of hybrid variations, the average precision of term conflation is high because hybrid variations only represent a small fraction of term variations (5.4% and 0.9%, see lines "+ sem" in Table 8 below). The average precision on [AGRIC] + Word97 is 79.8% and the average precision on [AGRIC] + AGROVOC is 91.1%.

The exploitation of semantic links extracted from WordNet in term variant extraction does not suffer from the problem of ambiguity pointed out for query expansion in (Voorhees, 1998). The robustness to polysemy is due to the fact that we are dealing with multiword terms that build restricted linguistic contexts in which words are disambiguated.

Numbers of Variants

Table 8 shows the numbers of term variants extracted by the four experiments. For each experiment and for each type of variation, three values are reported: the number of variants v of this type and two percentages indicating the ratio of these variants. The first percentage is $\frac{v}{V}$ in which V is the total number of variants produced by this experiment. The second percentage is $\frac{v}{V+T}$ in which T is the number of (non-variant) term occurrences extracted by this experiment.

The last line of Table 8 shows that variants represent a significant proportion of term occurrences (from 27.3% to 37.3%). The distribution of the different types of variants depends the semantic database and on the language under study. WordNet 1.6 is a productive source of knowledge for the extraction of semantic variants: In the experiment [MEDIC] + WordNet, semantic variants represent 58.6% of the variants, while they only represent 4.9% of the variants in the [AGRIC] + AGROVOC experiment. These values are reported in the line "Tot. Sem" of Table 8. Such results confirm the relevance of non-specialized semantic links in the extraction of specialized semantic variants (Hamon et al., 1998).

7 Conclusion

The model proposed in this study offers a simple and generic framework for the expression of complex term variations. The evaluation proposed at the end of this paper shows that term variations are extracted with an excellent precision for the three types of elementary variations: syntactic, morphosyntactic and semantic variations. The best performance is obtained with WordNet as source of semantic knowledge. Ongoing work on German, Japanese and Spanish shows that such a transformational and paradigmatic description of term variability applies to other languages than French and English reported in this study.

Acknowledgement

We would like to thank Jean Royauté and Xavier Polanco (INIST-CNRS) for their helpful collaboration. We are also grateful to Béatrice Daille (IRIN) for running her termer ACABIT on the data and to Olivier Ferret (LIMSI) for the Word97 macrofunction used to extract the thesaurus.

References

AGROVOC. 1995. *Thésaurus Agricole Multilingue.* Organisation de Nations Unies pour l'Alimentation et l'Agriculture, Roma.

Table 8: Numbers of term variants.

	[AGRIC] + Word97			[AGRIC] + AGROVOC			[MEDIC] + WordNet			[MEDIC] + Word97		
	v	$\frac{v}{V}$	$\frac{v}{V+T}$	v	$\frac{v}{V}$	$\frac{v}{V+T}$	v	$\frac{v}{V}$	$\frac{v}{V+T}$	v	$\frac{v}{V}$	$\frac{v}{V+T}$
Terms (T)	5325	×	63.1%	5325	×	68.2%	25561	×	62.7%	25561	×	72.7%
Coor	173	5.6%	2.1%	173	7.0%	2.2%	531	3.5%	1.3%	531	5.5%	1.5%
Modif	346	11.1%	4.1%	346	14.0%	4.4%	1985	13.1%	4.9%	1985	20.7%	5.6%
Comp	1045	33.6%	12.4%	1045	42.1%	13.4%	×	×	×	×	×	×
Perm	×	×	×	×	×	×	1146	7.5%	2.8%	1146	11.9%	3.3%
Tot. Synt	1564	50.3%	18.5%	1564	62.9%	20.0%	3662	24.1%	9.0%	3662	38.1%	10.4%
A to A	17	0.5%	0.2%	17	0.7%	0.2%	191	1.3%	0.5%	191	2.0%	0.5%
A to Adv	×	×	×	×	×	×	35	0.2%	0.1%	35	0.3%	0.1%
A to N	89	2.9%	1.1%	89	3.6%	1.1%	640	4.2%	1.6%	640	6.7%	1.8%
N to A	78	2.5%	0.9%	78	3.1%	1.0%	102	0.7%	0.3%	102	1.1%	0.3%
N to N	545	17.5%	6.5%	545	21.9%	7.0%	416	2.7%	1.0%	416	4.3%	1.2%
N to V	70	2.2%	0.8%	70	2.8%	0.9%	1230	8.1%	3.0%	1230	12.8%	3.5%
V to N	×	×	×	×	×	×	21	0.1%	0.1%	21	0.2%	0.1%
Tot. Mor	799	25.7%	9.5%	799	32.2%	10.2%	2635	17.3%	6.5%	2635	27.4%	7.5%
Sem Arg	180	5.8%	2.1%	16	0.6%	0.2%	912	6.0%	2.2%	629	6.6%	1.8%
Sem Head	397	12.8%	4.7%	84	3.4%	1.1%	2555	16.8%	6.3%	698	7.3%	2.0%
Coor + sem	30	1.0%	0.4%	5	0.2%	0.1%	183	1.2%	0.4%	102	1.1%	0.3%
Modif + sem	100	3.1%	1.2%	7	0.3%	0.1%	3467	22.8%	8.5%	1067	11.1%	3.0%
Perm + sem	×	×	×	×	×	×	788	5.2%	1.9%	369	3.8%	1.0%
A to A + sem	0	0.0%	0.0%	0	0.0%	0.0%	82	0.5%	0.2%	42	0.4%	0.1%
A to Adv + s.	0	0.0%	0.0%	0	0.0%	0.0%	22	0.1%	0.1%	8	0.1%	0.0%
A to N + sem	22	0.7%	0.3%	0	0.0%	0.0%	256	1.7%	0.6%	118	1.2%	0.3%
N to A + sem	10	0.3%	0.1%	0	0.0%	0.0%	72	0.5%	0.2%	28	0.3%	0.1%
N to N + sem	0	0.0%	0.0%	6	0.2%	0.1%	102	0.7%	0.3%	58	0.6%	0.2%
N to V + sem	8	0.3%	0.1%	4	0.2%	0.1%	454	3.0%	1.1%	185	1.9%	0.5%
N to V + sem	×	×	×	×	×	×	11	0.1%	0.0%	2	0.0%	0.0%
Tot. Sem	747	24.0%	8.9%	122	**4.9%**	1.6%	8904	**58.6%**	21.8%	3306	34.4%	%9.4
Variants (V)	3110	×	36.9%	2485	×	31.8%	15201	×	**37.3%**	9603	×	**27.3%**

CELEX. 1998. *www.ldc.upenn.edu/readme_files/celex.readme.html*. Consortium for Lexical Resources, UPenn.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Thierry Hamon, Adeline Nazarenko, and Cécile Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In *Proceedings, COLING-ACL'98*, pages 498–504, Montreal.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multiword terms for indexing and retrieval using morphology and syntax. In *ACL - EACL'97*, pages 24–31, Madrid.

MULTEXT. 1998. *www.lpl.univ-aix.fr/projects/multext/*. Laboratoire Parole et Langage, Aix-en-Provence.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, NY.

Stuart N. Shieber. 1992. *Constraint-Based Formalisms*. A Bradford Book. MIT Press, Cambridge, MA.

Karen Sparck Jones and John I. Tait. 1984. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.

K. Vijay-Shanker. 1992. Using descriptions of trees in a Tree Adjoining Grammar. *Computational Linguistics*, 18(4):481–518, December.

Ellen M. Voorhees. 1998. Using wordnet for text retrieval. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 285–303. MIT Press, Cambridge, MA.