Boosting Variant Recognition with Light Semantics

Cécile Fabre
ERSS / Dépt de Sciences du Langage
Univ. Toulouse-Le Mirail
5 allées A. Machado
31058 Toulouse Cedex, France
cfabre@univ-tlse2.fr

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

Abstract

A reasonably simple, domain-independent, large-scale approach of lexical semantics to paraphrase recognition is presented in this paper. It relies on the enrichment of morpho-syntactic rules and the addition of four boolean syntactico-semantic features to a set of 1,023 words. It results in a significant enhancement of precision of 30% with a slight decrease in recall of 10%.

1 Overview

The recognition of paraphrases and variants is an important issue in several areas of information retrieval and text understanding. Merging paraphrastic sentences improves summarization by avoiding redundancy (Barzilay et al., 1999). Term variant conflation enhances recall in information retrieval by pointing at documents that contain linguistic variants of query terms (Arampatzis et al., 1998).

In (Jacquemin and Tzoukermann, 1999), a technique is proposed for the conflation of morpho-syntactic variants that relies solely on morphological and low-level syntactic features (part-of-speech category, number agreement, morphological relationships, and phrase structure). An analysis of these results shows the limitation of this approach: correct and incorrect variants cannot be separated satisfactorily on a purely morpho-syntactic basis. Some additional lexical semantics must be taken into consideration.

In this study we propose a reasonably simple, domain-independent, large-scale approach of lexical semantics to noun-to-verb variant recognition. It relies on the mere addition of two boolean syntactic features to 449 verbs and two boolean morpho-semantic features to 574 nouns. It results in a significant enhancement of precision of 30% with a slight decrease in recall of 10%. This new approach to semantics—human-based, efficient, involving simple linguistic features —convincingly illustrates the positive role of linguistic knowledge in information processing. It confirms that verbs and their semantics play a significant role in document analysis (Klavans and Kan, 1998).

2 Morpho-syntactic Approach to Nomino-verbal Variation

In order to illustrate the contribution of semantics to the detection of paraphrastic structures, we focus on a specific type of variation: the verbal variants of Noun-Preposition-Noun terms or compounds in French. For example, *les contraintes résiduelles dans les coques sont analysées* (the residual constraints in the shells are analyzed) is such a verbal variant of *analyse de contrainte* (constraint analysis).

As a baseline for the extraction of these variants, we use a set of five morpho-syntactic transformations for Noun-Preposition-Noun terms reported in (Jacquemin and Tzoukermann, 1999) (see Table 1).[1] We use the notation $\mathcal{M}(\mathrm{N}_i)_\mathrm{V}$ for the morphological link between the initial term and the transformed structure. It represents any verb in the same morphological family as N_i. For instance, in English, and according to the CELEX database, $\mathcal{M}(\mathit{analysis})_\mathrm{V} = \{\mathit{to\ analyze}, \mathit{to\ psychoanalyze}\}$.

Given a $\mathrm{N}_1\,\mathrm{P}_2\,\mathrm{N}_3$ structure, these transformations are obtained through corpus-based tuning

[1] The following symbols are used for syntactic categories: N (noun), A (adjective), Av (adverb), V (verb), C (coordinating conjunction), P (preposition), and D (determiner). In the regular expressions, ? denotes optionality and | disjunction. Morphologically related words are underlined.

Table 1: Morpho-syntactic (MS) Variants of $\mathrm{N}_1\,\mathrm{P}_2\,\mathrm{N}_3$ Terms

NheadToV:	$\mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ (\mathrm{Av}^?\ (\mathrm{P}^?\ \mathrm{D} \mid \mathrm{P}\ \mathrm{D}^?)\ \mathrm{A}^?)\ \mathrm{N}_3$	*stabilisation de prix* (price stabilization) → *stabiliser leurs prix* (stabilize their prices)
NheadToVRev:	$\mathrm{N}_3\ (\mathrm{A}^?\ (\mathrm{P}\ \mathrm{A}^?\ \mathrm{N}\ (\mathrm{A}\ (\mathrm{C}\ \mathrm{A})^?)^?)^?\ (\mathrm{C}\ \mathrm{D}^?\ \mathrm{Av}^?\ \mathrm{A}^?\ \mathrm{N}\ \mathrm{A}^?)^?\ \mathrm{V}^?\ \mathrm{V}^?\ \mathrm{Av}^?)\ \mathcal{M}(\mathrm{N}_1)_\mathrm{V}$	*abattage d'arbre* (tree cutting) → *arbres ont été abattus* (trees have been cut down)
NmodifToV1:	$\mathrm{N}_1\ ((\mathrm{Av}^?\ \mathrm{A}\ (\mathrm{C}\ \mathrm{Av}^?\ \mathrm{A})^?)^?\ \mathrm{V}^?\ \mathrm{P})\ \mathcal{M}(\mathrm{N}_3)_\mathrm{V}$	*méthode d'évaluation* (method of evaluation) → *méthode pour évaluer* (method for evaluating)
NmodifToV2:	$\mathrm{N}_1\ (\mathrm{A}^?\ (\mathrm{V} \mid (\mathrm{P}\ \mathrm{D}^?\ (\mathrm{Av}^?\ \mathrm{A})^?\ \mathrm{N})^?)\ (\mathrm{Av}^?\ \mathrm{A})^?\ \mathrm{Av}^?)\ \mathcal{M}(\mathrm{N}_3)_\mathrm{V}$	*zone de déstabilisation* (region of destabilization) → *zone déstabilisée* (destabilized region)
NmodifToVRev:	$\mathcal{M}(\mathrm{N}_3)_\mathrm{V}\ (\mathrm{Av}^?\ (\mathrm{P}^?\ \mathrm{D}) \mid (\mathrm{P}\ \mathrm{D}^?)\mathrm{A}^?)\ \mathrm{N}_1$	*température de chauffage* (temperature for heating) → *chauffés à haute température* (heated at high temperatures)

and correspond basically to four configurations:

1. either N_1 or N_3 (respectively head and modifier of the initial term) is transformed into a morphologically related verb V,
2. the order of the two content words is retained or reversed,
3. the dependency relation between the two initial nouns is preserved.

For instance, rule NheadToVRev corresponds to transformations in which the head noun is morphologically related to the verb and the order of the two words is reversed; rule NmodifToV (modifier transformed, order retained) has been divided into two subrules: the first one - NmodifToV1 - requires the insertion of a preposition just before the verbal form.

3 The Limits of the Morpho-syntactic Approach

In the first step of this work, we expected the precision of variant recognition to be controlled in two ways: firstly, by searching for multi-term variants in which the two content words of the initial term are found, directly or via morphological transformation. Secondly, by defining morpho-syntactic patterns of variation in terms of part-of-speech strings that are allowed to come in between these two content words.

Yet, the sequences found on such a morpho-syntactic basis prove to be of varying quality regarding their ability to provide paraphrases of the initial term. Consider for instance some of the variants detected for the term *comparaison de résultat* (comparison of results), in which only the first two sequences are good variants:

compare les résultats (compare the results) (rule NheadToV, pattern $\mathcal{M}(\mathrm{N}_1)_\mathrm{V}\mathrm{D}\mathrm{N}_3$)

résultats expérimentaux sont comparés (experimental results are compared) (rule NheadtoVRev, pattern $\mathrm{N}_3\mathrm{A}\mathrm{V}\mathcal{M}(\mathrm{N}_1)_\mathrm{V}$)

comparés aux résultats (compared to the results) (rule NheadToV, pattern $\mathcal{M}(\mathrm{N}_1)_\mathrm{V}\mathrm{P}\mathrm{N}_3$)

résulté d'une comparaison (resulted from a comparison) (rule NModifToVRev, pattern $\mathcal{M}(\mathrm{N}_3)_\mathrm{V}\mathrm{P}\mathrm{D}\mathrm{N}_1$)

Such examples show that morpho-syntactic patterns are too coarse-grained to ensure that the dependency relation between the two pivots (*results* is the object of the predicate *comparison*) is maintained. When trying to define linguistic criteria to evaluate such variants, it appears that the frontier between good and bad variants lies between those that preserve the argument relation between the two content words and those that disrupt it. This means that, in the verbal variant, the argument relation between the verb and the noun must be the same as the relation between the deverbal noun and the other noun in the nominal term.

None of the five rules ensures that the subcategorization frame is preserved. For instance, if we consider the rule NModifToVRev, we find sequences that obey this constraint and sequences that violate it[2]:

critère d'évaluation (evaluation criterion) → *évalué selon les critères* (evaluated according to the criteria)

système d'évaluation (evaluation system) *→ *évalué le système* (evaluated the system)

In the second case, the transformation is unacceptable because the instrumental relation expressed in the nominal term becomes an object relation in the verbal sequence. Even when word order is preserved, the relation between the pivots can be totally different in the term and its transformation, as in: *contrôle d'installation* (installation control) and *contrôle centralisé installé* (installed centralized control) (rule NModifToV2).

Our aim was to formulate additional constraints in order to control argument structure preservation. We thus had to cope with problem of handling nominal phrases (NP) in which one of the elements is morphologically linked to a verb. In French, as in English, the semantics of these nominal phrases is an issue for linguistic description: the two nouns can be linked by the whole range of argument-predicate relations, and very few linguistic elements can be used to decide what relation is expressed. Here is a brief list of the configurations that are likely to appear in such NPs:

- the second noun is the object of the first one: *comparaison de résultat* (comparison of result)
- the second noun is the subject of the first one: *augmentation de l'intensité* (increase in intensity)
- the second noun is an adjunct: *traitement à la chaleur* (treating with heat)
- the first noun is an adjunct: *taux d'augmentation* (increase rate)

Our aim was to find a way to use surface linguistic knowledge, as required in such an area of NLP, to deal with the interpretation of these phrases.

4 Light Semantics for Nomino-verbal Variations

Our approach consisted of two steps: firstly, defining semantic clues for accepting or discarding variants and, secondly, defining new variation patterns based on these features.

4.1 Filtering Criteria

First, using linguistic results on the semantics of French NPs (Fabre, 1996; Bartning, 1990), we identified predicate-argument configurations that cannot be matched by a given pattern ('reject' heuristics in the sense of (Lapata, 1999)). For example, when rule NmodifToVRev applies, N_1 *de* N_3 terms cannot be paraphrased by verbal sequences in which N_1 is the object of the verb, as in: *expérience d'utilisation* (experiment of use) *→ *utilisait une expérience* (used an experiment). In such a configuration, only nonthematic arguments (adjuncts) of the deverbal noun may be found inside the NP.

Similarly, when rule NheadToVRev applies, N_1 *de* N_3 terms cannot be paraphrased by verbal sequences in which N_1 is the subject of a transitive verb, as in: *utilisation de l'expérience* (use of experiment) *→ *expérience utilisant* (experiment using).

This configuration provides variants only when the verb is intransitive or ergative: ergative verbs allow for alternations of the form: NP V (*la densité augmente*) / one V NP (*on augmente la densité*).

In this case, the following transformation is correct: *augmentation de densité* (density increase) / *densité augmente* (density increases).

4.2 Enriched Metarules

Once it has been established which transformations should be rejected, we searched for surface linguistic clues that could help us to filter out these undesirable variants. It led us to the redefinition of the metarules, in two ways: putting additional constraints on the part-of-speech strings that can intervene between the two pivots, and defining new features to add linguistic control upon the application of the rules. These features are: the prepositional form, the morphological type of the noun, the transitivity of the verb, and the voice (active versus passive).

Here are two examples for the redefinition of the metarules (further details and examples are given in table 3):

rule NmodifToVRev In this case, the metarule is transformed into a single

[2] In what follows, the symbols → and *→ respectively indicate correct and incorrect transformations

refined rule, in which the combination of parts of speech is more restricted: a preposition is required to eliminate object relations from the verbal phrase. In addition, the morphologically complex noun must be a processive deverbal. Transformations such as *expérience d'utilisation* *→ *utilisait une expérience* are filtered out.

rule NheadToVRev Here, the initial metarule is refined into three enriched rules, mainly by means of lexical constraints on the verb form. Only N_1 P_2 N_3 terms where P_2 = *de* are treated. If the verb is transitive, then the verb form must be a past participle (rule NheadtoVRev-Pass), so that the object relation still holds in the variant. If the verb is intransitive or ergative, then the verb form must be active so that the subject relation holds (rule NheadtoVRev-ActSimp (resp. NheadtoVRev-ActComp) for simple (resp. complex) verb forms). Transformations such as *utilisation de l'expérience* *→ *expérience utilisant* are filtered out.

The refinement of the metarules introduced four linguistic features which had to be encoded in the lexicon (see Table 2), namely:

- the morphological nature of the noun: the noun is either non deverbal or deverbal. In the latter case, it may correspond to an agent deverbal, which refers to the agent of the verb, e.g. *utilisateur* (user), or to a processive deverbal, which refers to the action denoted by the verb, e.g. *utilisation* (use).
- the transitivity of the verb: intransitive and ergative verbs are marked in the lexicon.

This annotation task is not time-consuming (about 3 hours for 1,023 words) and could be partly automated: characteristic endings could help to detect processive and agent deverbals. In addition, intransitive and ergative verbs form a small set of the verbal lexicon (8% of the verbs) which is likely to be partly domain-independent.

5 Experiments and Evaluations

In this section, we evaluate the variations produced from the two preceding sets of metarules: initial morpho-syntactic variations (henceforth MS) and new variations enriched through light semantics (henceforth MS+S).

The variants are obtained from a 13.2 million-word corpus composed of scientific abstracts in the agricultural domain (in French) and a set of 11,452 terms.[3] The corpus is analyzed through SYLEX, a shallow parser that builds limited phrase structures and associates each word with an unambiguous syntactic category and a lemma. Terms are acquired from the output of the SYLEX parser. Only [N [P N]] structures are selected and only terms that occur at least three times in the corpus are retained.

The numbers of variants extracted through MS and MS+S are reported in Table 4. They are arranged in such a way that corresponding variations are aligned horizontally. For instance, each of the three MS+S variations NheadToV-Comp, NheadToV-Simp or NheadtoV-Prep is a refinement of the MS variation NheadToV. In other words, the set of variants extracted by these three rich metarules is included into the set of variants extracted by the poor metarule. These two sets are not equal since the rich metarules are made more selective than the original metarule from which they are derived.

In addition to the output of rich and poor metarules, Table 4 shows, in the third column, the number of co-occurrences associated with these metarules. Co-occurrences are the least selective filters associated with morpho-syntactic variants; they are expected to extract all the possible correct nomino-verbal variations (recall value 1.0). Given a N_1 P_2 N_3 term, these co-occurrences correspond to a configuration in which N_1 co-occurs with a verb that is morphologically related to N_3 or N_3 co-occurs with a verb related to N_1. Co-occurrences are extracted from a 11-word window (9 intervening words). These co-occurrences are used to evaluate the recall values of the filtering metarules.

[3] We are grateful to Xavier Polanco, Jean Royauté and Laurent Schmitt (INIST-CNRS) for providing us with this scientific corpus.

Table 2: Semantically Enriched Lexicon.

Word	**Processive Deverbal**	**Agent Deverbal**	**Intransitive**	**Ergative**
abaisser	−D	−A	−I	−E
abaissement	+D	−A	−I	−E
absorber	−D	−A	−I	−E
absorbeur	+D	+A	−I	−E
accorder	−D	−A	−I	−E
accord	+D	−A	−I	−E
accumuler	−D	−A	−I	−E
accumulateur	+D	+A	−I	−E
accumulation	+D	−A	−I	−E
accélérer	−D	−A	−I	+E

Table 3: Semantically Enriched Morpho-syntactic (MS+S) Variants of $N_1\,P_2\,N_3$ Terms

NheadToV-Comp: $\textit{avoir}\,\mathrm{Av}^?\,\mathcal{M}(\mathrm{N}_1)_\mathrm{V}\,\mathrm{Av}^?\,\mathrm{D}\,\mathrm{A}^?\,\mathrm{N}_3$

$\{\langle \mathrm{N}_1\ \textit{dev}\rangle = \textit{processive}\ \wedge\ \mathrm{P}_2 = \textit{de}\ \wedge\ \langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{tense}\rangle = \textit{pastparticiple}\}$

comparaison de résultats (comparison of results)
→ *a comparé les résultats* (has compared results)

NheadToV-Simp: $\mathcal{M}(\mathrm{N}_1)_\mathrm{V}\,\mathrm{Av}^?\,\mathrm{D}\,\mathrm{A}^?\,\mathrm{N}_3$

$\{\langle \mathrm{N}_1\ \textit{dev}\rangle = \textit{processive}\ \wedge\ \mathrm{P}_2 = \textit{de}\ \wedge\ \langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{tense}\rangle \neq \textit{pastparticiple}\}$

évaluation de risques (evaluation of risks) → *évaluer les risques* (to evaluate risks)

NheadtoV-Prep: $\mathcal{M}(\mathrm{N}_1)_\mathrm{V}\,\mathrm{Av}^?\,\mathrm{P}_2\,\mathrm{D}\,\mathrm{A}^?\,\mathrm{N}_3$

$\{\langle \mathrm{N}_1\ \textit{dev}\rangle = \textit{processive}\}$

exposition à la lumière (exposure to light) → *exposées à la lumière* (exposed to light)

NheadtoVRev-Pass: $\mathrm{N}_3\,(\mathrm{A}^?\,(\mathrm{P}\,\mathrm{A}^?\,\mathrm{N}\,(\mathrm{A}\,(\mathrm{C}\,\mathrm{A})^?)^?)^?\,(\mathrm{C}\,\mathrm{D}^?\,\mathrm{Av}^?\,\mathrm{A}^?\,\mathrm{N}\,\mathrm{A}^?)^?\,\mathrm{V}^?\,\textit{être}^?\,\mathrm{Av}^?)\,\mathcal{M}(\mathrm{N}_1)_\mathrm{V}$

$\{\langle \mathrm{N}_3\ \textit{agreement}\rangle = \langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{agreement}\rangle\ \wedge\ \mathrm{P}_2 = \textit{de}\ \wedge\ \langle \mathrm{N}_1\ \textit{dev}\rangle = \textit{processive}\ \wedge$
$\langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{tense}\rangle = \textit{pastparticiple}\ \wedge\ \langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{valence}\rangle = \textit{transitive}\}$

répartition de charge (weight distribution) → *charge également répartie* (equally distributed weight)

NheadtoVRev-ActSimp: $\mathrm{N}_3\,(\mathrm{A}^?\,(\mathrm{P}\,\mathrm{A}^?\,\mathrm{N}\,(\mathrm{A}\,(\mathrm{C}\,\mathrm{A})^?)^?)^?\,(\mathrm{C}\,\mathrm{D}^?\,\mathrm{Av}^?\,\mathrm{A}^?\,\mathrm{N}\,\mathrm{A}^?)^?)\,\mathcal{M}(\mathrm{N}_1)_\mathrm{V}$

$\{\mathrm{P}_2 = \textit{de}\ \wedge\ \langle \mathrm{N}_1\ \textit{dev}\rangle = \textit{processive}\ \wedge\ \langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{tense}\rangle \neq \textit{pastparticiple}\ \wedge$
$\langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{valence}\rangle = (\textit{ergative}|\textit{intransitive})\}$

chute de température (drop in temperature) → *température chute* (temperature drops)

NheadtoVRev-ActComp: $\mathrm{N}_3\,(\mathrm{A}^?\,(\mathrm{P}\,\mathrm{A}^?\,\mathrm{N}\,(\mathrm{A}\,(\mathrm{C}\,\mathrm{A})^?)^?)^?\,(\mathrm{C}\,\mathrm{D}^?\,\mathrm{Av}^?\,\mathrm{A}^?\,\mathrm{N}\,\mathrm{A}^?)^?\,\textit{avoir}^?\,\mathrm{Av}^?)\,\mathcal{M}(\mathrm{N}_1)_\mathrm{V}$

$\{\mathrm{P}_2 = \textit{de}\ \wedge\ \langle \mathrm{N}_1\ \textit{dev}\rangle = \textit{processive}\ \wedge\ \langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{tense}\rangle = \textit{pastparticiple}\ \wedge$
$\langle \mathcal{M}(\mathrm{N}_1)_\mathrm{V}\ \textit{valence}\rangle = (\textit{ergative}|\textit{intransitive})\}$

fermentation de jus (juice fermentation) → *jus de raisins fermentés* (fermented grape juice)

Precision and Recall

In order to calculate the precision and recall of the rich and poor metarules and to estimate the gains of semantic enrichment, a set of 1,000 co-occurrences has been randomly chosen among the 159,898 co-occurrences retrieved by the system. They have been divided into three sets: S_1 (500 co-occurrences) and S_2 and S'_2 (250 co-occurrences). S_1 has been evaluated independently by the two judges (i.e. the two authors)

Table 4: Counts of variants of $\mathrm{N}_1\,\mathrm{P}_2\,\mathrm{N}_3$ terms

MS		**MS+S**		**Co-occurrences**	
38,693	NheadToV	874	NheadToV-Comp	69,056	N1N2toV1N2
		15,583	NheadToV-Simp		
		7,644	NheadtoV-Prep		
20,453	NheadToVRev	14,248	NheadtoVRev-Pass		
		197	NheadtoVRev-ActSimp		
		26	NheadtoVRev-ActComp		
6,803	NmodifToV1	2,749	NmodifToV1-Ppr	42,882	N1N2toN2V1
2,588	NmodifToV2	1,160	NmodifToV2-Inf1	26,971	N1N2toN1V2
		0	NmodifToV2-Inf2		
		1	NmodifToV2-Inf3		
9,363	NmodifToVRev	1,892	NmodifToVRev-Prep	20,989	N1N2toV2N1
77,900		**44,374**		**159,898**	

in order to test the level of agreement and S_2 and S'_2 have been evaluated separately by only one judge each. Each cooccurrence has been marked as positive (a correct variation), negative (an incorrect variation) or inevaluable. Inevaluable cases correspond either to tagging errors or to incorrect terms such as *coque de forme* (shell of shape) which is an incomplete term structure because it should be followed by an adjective such as *coque de forme ovale* (oval-shaped shell). Only the cases of agreement between the two judges are used for the computation of recall and precision values.

The addition of semantics results in an increase of precision of 0.29: from 0.499 for MS to 0.789 for MS+S. The corresponding decrease of recall is much smaller: 0.11 from 0.696 for MS to 0.586 for MS+S. Precision and recall can be combined into a single measure such as the effectiveness measure E_α given by Formula (1) in which α is a parameter ($0 \leq \alpha \leq 1$) (van Rijsbergen, 1975):

$$E_\alpha = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)} \qquad (1)$$

E_α varies from 0 to 1.0. Low values of E_α correspond to combined high recall and high precision. If we use $\alpha = \frac{1}{2}$ in order to assign an equal importance to precision and recall, the $\mathrm{E}_{\frac{1}{2}}$ values are 0.419 for MS and 0.327 for MS+S. They indicate that the addition of semantics has significantly improved the quality of variant extraction. Detailed values of recall and precision are shown in Table 5.

Agreement on Judgment

Agreement on a classification task can be measured through the kappa coefficient (K). It evaluates the pairwise agreement among a set of coders making category judgment, correcting for expected chance agreement (Carletta, 1996). In our case the results of the ternary classification task are given by Table 6. The simple kappa coefficient is

$$K = \frac{P_0 - P_e}{1 - P_e} \qquad (2)$$

in which $P_0 = \Sigma_i \frac{n_{ii}}{n}$ and $P_e = \Sigma_i (\frac{n_{i.}}{n} \frac{n_{.i}}{n})$ (Cohen, 1960). P_0 is the proportion of times the coders agree and P_e is the proportion of times we would expect them to agree by chance. The value of the kappa coefficient is 0.91 indicating a good reliability of the evaluation performed by the two independent judges.

6 Conclusion

On a linguistic point of view, this experiment demonstrates that NLP applications can provide new issues for the description of linguis-

Table 5: Precision and recall in variant extraction for MS and MS+S variations

$\mathbf{P}_{\mathrm{MS}}$		$\mathbf{P}_{\mathrm{MS+S}}$		$\mathbf{R}_{\mathrm{MS}}$	$\mathbf{R}_{\mathrm{MS+S}}$
0.438	NheadToV	0.875	NheadToV-Comp	0.806	0.664
		0.938	NheadToV-Simp		
		0.565	NheadtoV-Prep		
0.735	NheadToVRev	0.902	NheadtoVRev-Pass		
		1.000	NheadtoVRev-ActSimp		
		—	NheadtoVRev-ActComp		
0.111	NmodifToV1	0.308	NmodifToV1-Ppr	0.674	0.578
0.769	NmodifToV2	1.000	NmodifToV2-Inf1	0.357	0.214
		—	NmodifToV2-Inf2		
		—	NmodifToV2-Inf3		
0.448	NmodifToVRev	0.000	NmodifToVRev-Prep	0.765	0.765
0.499		**0.789**		**0.696**	**0.586**

Table 6: Frequencies of pairwise judgments for the ternary classification of nomino-verbal variations ($\star$ = inevaluable, $+$ = correct, $-$ = incorrect).

n_{ij}	$\star$	$+$	$-$	$n_{i.}$
$\star$	120	9	1	130
$+$	1	184	6	191
$-$	4	10	165	179
$n_{.j}$	125	203	172	500

tic phenomena. The problem of linguistic variation in information processing forces the linguist to reconsider paraphrase and transformation mechanisms in a new perspective, based on real linguistic data and on systematic corpus exploration. The paraphrase judgment is evaluated in a new way, from a practical point of view: two sequences are said to be a paraphrase of each other if the user of an information system considers that they bring identical or similar information content. Regarding linguistic methodology, this work led us to find ”light” solutions in terms of lexical encoding to describe complex semantic phenomena. This approach is promising because it demonstrates that linguistic knowledge can really enhance the results of term recognition beyond the morphology level, and that semantics can be taken into account to some extent.

References

A. T. Arampatzis, T. Tsoris, C. H. A. Koster, and Th. P. van der Weide. 1998. Phrase-based information retrieval. *Information Processing & Management*, 34(6):693–707.

Inge Bartning. 1990. Les syntagmes binominaux en *de* - les types interprétatifs subjectifs et agentifs. In *Proceedings, dixième congrès des romanistes scandinaves.*

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1999. Informational fusion in the context of multi-document summarization. In *Proceedings of ACL'99*, pages 550–557, University of Maryland.

Jean Carletta. 1996. Asessing agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2):249–254.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cécile Fabre. 1996. *Interprétation automatique des séquences binominales en français et en anglais.* Ph.D. thesis, Université Rennes I.

Christian Jacquemin and Evelyne Tzoukermann. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25–74. Kluwer, Boston, MA.

Judith Klavans and Min-Yen Kan. 1998. Role of verbs in document analysis. In *Proceedings of COLING-ACL'98*, pages 680–686, Université de Montréal, Montreal, Canada.

Maria Lapata. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of ACL'99*, pages 397–404, University of Maryland.

C. J. van Rijsbergen. 1975. *Information Retrieval.* Butterworth, London.