A NATURAL LANGUAGE APPROACH TO MULTI-WORD TERM CONFLATION

Judith Klavans
Center for Research on Information Access
Columbia University
535 W. Broadway, 114th Street, MC 1101
NEW YORK, NY 10027, USA
klavans@cs.columbia.edu

Christian Jacquemin
IRIN, IUT de Nantes
BP 34103
3, rue du Maréchal Joffre
44041 NANTES Cedex 1
jacquemin@irin.univ-nantes.fr

Evelyne Tzoukermann
Bell Laboratories, Lucent Technology,
700 Mountain Avenue, Room 2d-448,
P.O. Box 636, Murray Hill, NJ 07974, USA
evelyne@research.bell-labs.com

Abstract

This paper presents a corpus-based system to expand multi-word index terms using a part-of-speech tagger and a full-fledged derivational morphological system, combined with a shallow parser. The unique contribution of the research is in using these linguistically based tools with filters in order to avoid the problems of semantic degradation typically associated with derivational analysis. The expansion and subsequent conflation of terms increases indexing coverage up to 30%, with precision of nearly 90% for correct identification of related terms. The system core is language independent and provides a uniform platform on which to build multilingual applications. Language specific modules have been developed for English and French. The fully implemented system is described with particular attention to the role of derivational morphology and phrasal relations. Results and evaluation will be presented in terms of precision and recall, with an analysis of errors. This paper illustrates how the use of natural language processing tools for tasks to which they are especially suited such as indexing, has the potential to improve performance in IR.

System Function and Architecture

Three NLP modules are key to the system: morphology, part of speech tagging, and surface syntactic analysis (see Figure 1). The emphasis in our research is on the computational linguistic features of the system with particular attention to the role of the morphological component, and on the utilization of a toolset to solve the multi-word indexing coverage problem in information retrieval. The system consists of the following procedures:

1. Start with a multi-word term list and a large corpus;
2. Disambiguate and part-of-speech tag the words in the multi-word term list and corpus;
3. Generate term variant patterns from the application of morphosyntactic transformations to multi-word terms;
4. Generate all morphologically derived forms of words dynamically;
5. Run a shallow parser identifying morphosyntactic variants within the target corpus;
6. Link term and term variant occurrences to the initial multi-word list for expanded indexing;
7. Build an inverted index file with pointers for term expansions.

The system permits the expansion and subsequent conflation of morphosyntactically related terms, such as *abattage d'arbre* (tree cutting) and *les arbres on été abattus* (trees have been cut down) or *valeur d'estimation* (value of estimation) and *estimer la valeur* (estimate the value), and syntactically related terms such as *plantes aromatiques* (aromatic plants) and *plantes et extraits aromatiques* (aromatic plants and extracts) or *séchage par le vide* (vacuum drying) and *séchage sous vide* (drying under vacuum).

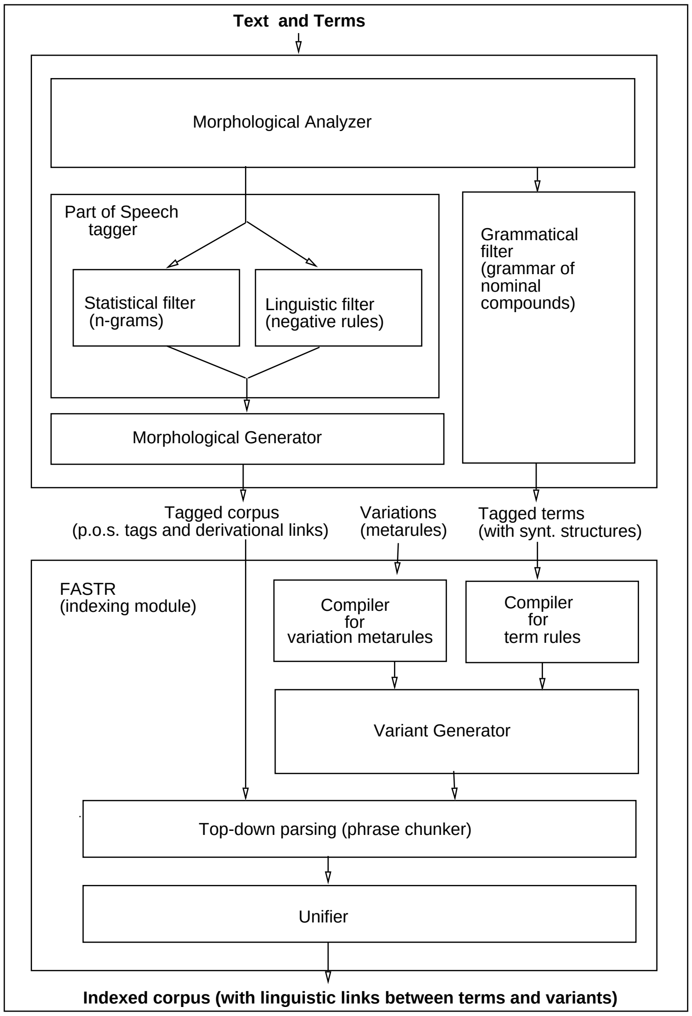
Text and Terms
Morphological Analyzer
Part of Speech tagger
Statistical filter (n-grams)
Linguistic filter (negative rules)
Grammatical filter (grammar of nominal compounds)
Morphological Generator
Tagged corpus (p.o.s. tags and derivational links)
Variations (metarules)
Tagged terms (with synt. structures)
FASTR (indexing module)
Compiler for variation metarules
Compiler for term rules
Variant Generator
Top-down parsing (phrase chunker)
Unifier
Indexed corpus (with linguistic links between terms and variants)

Figure 1: System Architecture.

Representing and Parsing Term Variations

Linguistic variations for terms have been well-studied in information access (Sparck Jones and Tait1984). When categorizing variants, three main linguistic distinctions can be drawn, as discussed more fully in (Jacquemin, Klavans, and Tzoukermann1997 to appear). The first type involves syntactic variants, where the syntactic structure of the term is modified but there is no explicit morphological change; secondly, morphosyntactic variants refer to cases where both syntactic and morphological changes occur over related terms; finally, semantic variants concern terms that are conceptually related, often using synonyms or hypernyms. Each of these types is used freely to enhance the style of writing and to provide variety in phrasing, with term truncation used as a way to avoid repetitiveness. Our research deals specifically with the application of morphological derivation for controlled multi-word term lists with syntactic, morphological, and morphosyntactic variations. Semantic variants are not considered, although they are an important potential source of additional variants (Klavans1994). Consideration of syntactic structure in the analysis of multi-word terms avoids the problems created by non-syntactic word based approaches where each item is considered independently, regardless of syntactic relation. This is often called the "bag of words" approach, since word order and hierarchical relations are ignored, e.g. *house representative* vs. *representative house* are considered related in the bag of words approach, whereas when word order and syntactic structure are taken into account, they are correctly not conflated. This problem is particularly rampant in English with poor morphology and multiple part of speech membership for many lexical forms. In (Jacquemin1996), it is shown that only 65% of the cooccurrences of content words within a 10-word window are genuine term variants. The remaining 35% are fortuitous cooccurrences without any conceptual relation with the original terms. This high ratio of spurious variants among word cooccurrences reveals that a simple window-based extractor is not likely to retrieve variants with accuracy. It is therefore necessary to use linguistically motivated procedures for correct controlled indexing.

Previous research on term variant recognition, such as (Metzler and Haas1989; Schwarz1990; Sheridan and Smeaton1992; Strzalkowski1996), have used NL parsers for recognizing noun phrase structures within corpora. These approaches are well-suited for free indexing. Similarly, for controlled indexing, it is necessary to account for the transformations of terms by morphological and syntactic operations. To this end, we have integrated a morphological processor (Tzoukermann and Jacquemin1997) and a transformation-based parser (Jacquemin and Royauté1994) into a platform for multilingual controlled indexing. Our system permits recognition of syntactic variants through coordination, substitution, and compounding.

Our claims are the following:

1. A principled morphological processor is capable of returning more information on the family of related forms and words than basic stemming algorithms, and thus for morphologically rich (as well as poor) languages, could provide exponential benefit.
2. Syntactic transformations are better adapted to the recognition of term variations than general grammar rules because they provide a framework for articulating syntactic modification with morphological changes.

However, our assumption is that improved results will only emerge when natural language processing tools are optimally combined for tasks to which they are especially suited.

Morphological Analyzer

The morphological system is based on generative morphology and modeled on Corbin(1987) for French; the implementation uses finite-state transducer (FST) tools (Pereira, Riley, and Sproat1994; Mohri and Sproat1996), and handles French morphology (see (Klavans and Tzoukermann1992) for an overview of the major morphological constructions and their computational processing). However, the algorithms and compilers are language independent, and have been applied to several other languages. The system is based on a dictionary of French of over 85,000 entries. The output form of the derivation, which is input to the parser, is a feature/value description. Before morphological processing is applied, part of speech tagging is performed in order to disambiguate the corpus (Tzoukermann, Radev, and Gale1997 to appear). The part of speech tagger is based on the combination of linguistic and statistical approaches. Word classes are used to regroup words which share the same part of speech, in the absence of sufficient training data on which to compute statistics. The system makes use of an additional feature available in the set of FST tools to apply weights to transitions, thereby giving preference to certain analyses over others. The tagger disambiguates about 96% of unrestricted text, which is more than acceptable for the IR task we are interested in. Derivational generation is then applied to expand variants so that the original database of complex terms can be augmented and enriched. The FST of Figure 2 describes some inflections of the French verb *recevoir* (to receive). The string *reçois* ([I or you] receive) is analyzed as a first or a second singular person of indicative present by traversing the automaton from the START state to one of the final states marked by a double circle. Similarly, the FST of Figure 3 is used for describing the various derivative forms that can be built on a set of verbal stems (*chauff* from the verb *chauffer* (to heat), *céd* and *ced* from the verb *céder* (to give), etc.). Traversing the transducer from the START state to one of the END states produces lexical forms of the verb, such as *chauff*-AGE-IST (ε represents an empty transition). A set of rewriting rules completes the transducer of Figure 3 by transforming meta-suffixes such as IST into strings, here *iste* or *istes*. Hence, application of the rules to *chauff*-AGE-IST gives *chauffagiste* (heating expert) a nominal form of the verb *chauffer* (to heat).

The morphological system presented in this paper is capable of generation and analysis like all finite-state transducer systems. It handles prefixation and suffixation. At the inflectional level, stem alternations are implemented so that the

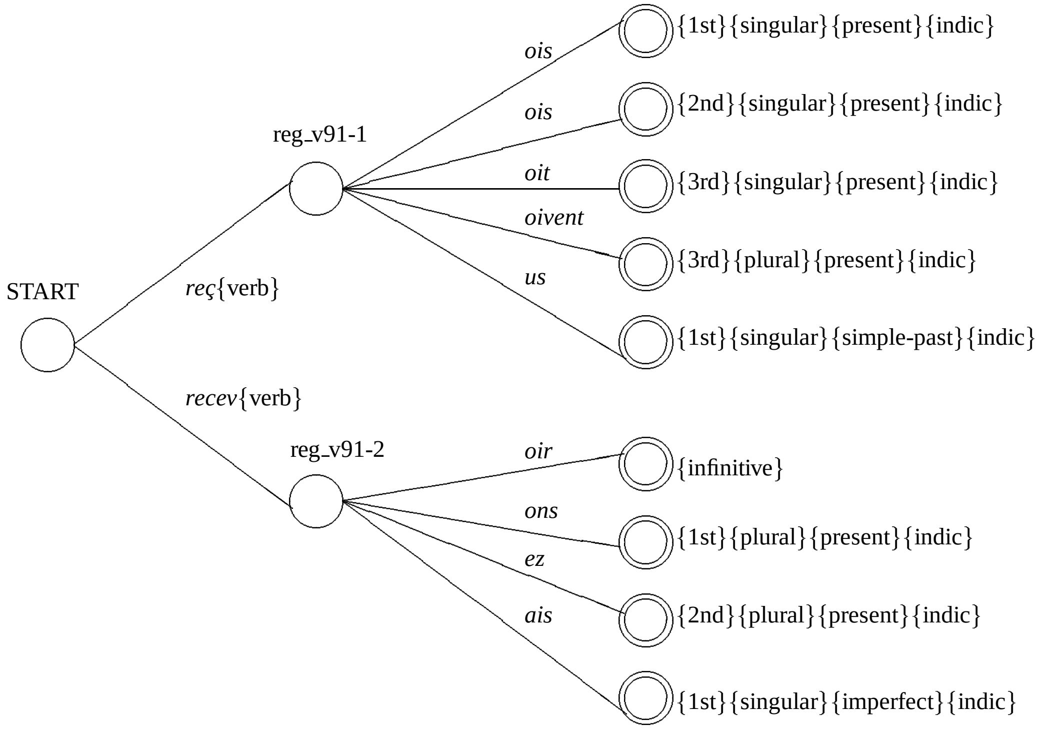
START
reg_v91-1
reg_v91-2
reç{verb}
recev{verb}
ois
ois
oit
oivent
us
oir
ons
ez
ais
{1st}{singular}{present}{indic}
{2nd}{singular}{present}{indic}
{3rd}{singular}{present}{indic}
{3rd}{plural}{present}{indic}
{1st}{singular}{simple-past}{indic}
{infinitive}
{1st}{plural}{present}{indic}
{2nd}{plural}{present}{indic}
{1st}{singular}{imperfect}{indic}

Figure 2: Sample of paradigm for the flexions of verb “recevoir”.

inflection produces only well-formed forms. At the derivational level, affixes can attach to any stem of a verb. This is designed so that coinages occuring in texts can be identified. For some applications, word overgeneration could be seen as a drawback. However, in the context of our system for the expansion of multi-word term lists, controlling word productivity is not a problem. On the contrary, it is essential to generate as many forms as possible so that the variants created from the input term expressions can be properly enriched to provide all morphological alternates. Since our focus is on multi-word terms, the overgeneration problem typically resulting from the application of derivational morphology to single words, does not present the commonly observed degradation problem. Further discussion of the types of overgeneration that are avoided is given in (Tzoukermann, Klavans, and Jacquemin1997 to appear). In our system, possible variants are matched against two types of data: (1) words occurring in the target corpus, and (2) words of the multi-word expressions.

Transformational Parser

Once all possible morphological forms are generated from index terms, and once the corpus is tagged, the parser identifies syntactic structures in the corpus in order to extract occurrences of terms and their variants. The challenge is to devise correct syntactic operators to yield valid patterns for extracting all and only correct variants. The grammar consists of a set of rules describing multi-word terms, and metarules applying over these rules in order to transform them into term variant rules.

Figure 4 gives an overview of term variant recognition. The terminological knowledge consists mainly of term rules and the linguistic knowledge is embodied in metarules representing local morphosyntactic variants. The parser proceeds as follows: first, a term rule is unified with the left part of a metarule representing a term variation. The unification propagates to the right-hand side (the target of the transformation). Secondly, the right-hand side of the metarule is copied yielding a term variant rule. Thirdly, the term variant rule is used for extracting term variants from the corpus. The parser is equipped with optimization devices (B-Tree, filtering procedures, lexicalization, etc.) which ensure an efficient processing of large corpora whatever the size of the terminological base.

The conception of a metagrammar of morphosyntactic variations for a given language is approximately a one person-week task. Variations are developed by combining linguistic introspection, observation of cooccurrences within corpora, and empirical tuning. Figure 5 shows the type of syntactic structures for describing the coordination of two terms. [1] The merging of these three structures produces the following syntactic transformation:

(1) $\text{Coord}(\text{N}_2\ \text{P}_3\ \text{N}_4) = \text{N}_2\ ((\text{C A? N A? P})\ |(\text{A C P})\ |\ (\text{P D?A? N A? C P?}))\ \text{N}_4$

Syntactic transformations like these are used to retrieve term variants from corpora and the results are manually observed. The following remarks lead to a revision of the original proposition:

[1] In the following text, N is a Noun, P a Preposition, A an Adjective, C a Coordination, Pu a Punctuation, Av an adverb. “?” and “|” are to be read like Unix regular expressions. The subscripts represent the linear position of the item in the structure.

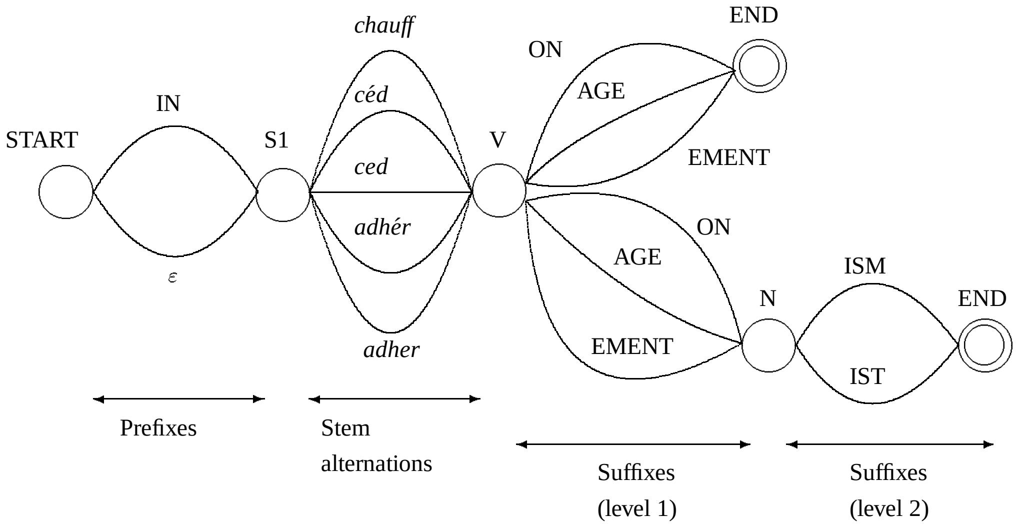

Figure 3: Sample of derivational processing.

1. A determiner is allowed in the prepositional phrase: for example, *absorption des éléments minéraux et de l'eau* "absorption of mineral elements" and of the water). In metarule (1), N_4 is transformed into D? A? N_4.
2. The coordinating conjunction can be preceded by a comma or followed by an adverb: for example, *la production, et surtout la diffusion des semences* "production, and above all the diffusion of seeds"). In metarule (1), C is transformed into Pu? C Av?, as seen in metarule (2.
3. In argument coordinations, the conjunction can be followed by a preposition and a determiner: for example, *de l'humidité et de la vitesse de l'air* "of the humidity and of the speed of the air." In metarule (1), C is transformed into C P? D?.

The final form of metarule (1) is given by the following equation:

$$(2) \quad \mathrm{Coord}(N_2\ P_3\ N_4) \;=\; \begin{array}{l} N_2\ (((\underline{Pu?}\ C\ \underline{Av?}\ \underline{P?}\ \underline{D?}\ A?\ N\ A?\ P) \mid (A\ C\ \underline{Av?}\ P) \\ \mid (P\ D?\ A?\ N\ A?\ C\ \underline{Av?}\ P?))\ \underline{D?\ A?})\ N_4 \end{array}$$

A similar tuning of metarules has been performed on French and English. A comparison of these two grammars indicates that the amount of work necessary does not depend on the language. Since these two languages belong to different linguistic families, some morphosyntactic transformations are specific to only one of the languages. It can be assumed that the transformations for Spanish and French might be much closer than those for French and English. The relative ease with which a grammar of variations can be constructed for a given language shows that the formalism of syntactic transformations is well-suited for describing term variation.

Evaluation

Experiments and evaluation were performed on two corpora, a training corpus based on newspaper articles extracted from "Le Monde" (1.28 million words) and a test corpus consisting of scientific papers (1.35 million words). The term list for the training corpus, consisting of 76,884 terms, was generated automatically; for the test corpus, the 4,717 terms were manually created by professional terminologists. The purpose of using a test corpus which differs greatly in nature was to test the limits of the system. We hypothesized that the system, since it is based on linguistically motivated and sublanguage independent morphological and syntactic constructions, would function well on a corpus from any domain. Our hypothesis turned out to be correct for this test; we intend to further validate the hypothesis on other corpora from different domains.

The evaluation was performed using the IR criteria precision and recall, where precision refers to the ratio of correct variants among the variants extracted by the system, and recall refers to correct variants identified from all possible variants. Determining precision is relatively clear for this task since the notion of "correct variant" is not difficult to establish. However, recall is more complex since the task of identifying possible variants within documents in order to establish ground truth is a tedious and time-consuming manual task. Nevertheless, approximately 15% of the test corpus (about 200,000 words) was checked for recall.

Syntactic variants were identified with an overall average of 91.2% precision; morphosyntactic variants achieved 86.4% precision, the combined precision thus reaching 89.4%. This means that of the term variants identified, only 10.6% were spurious under strict interpretation (and even some of the errors were of a minor nature). For recall, 75.0% of the possible syntactic variants were correctly retrieved, and 75.6% of the morphosyntactic variants, yielding a total of 75.2% recall. The

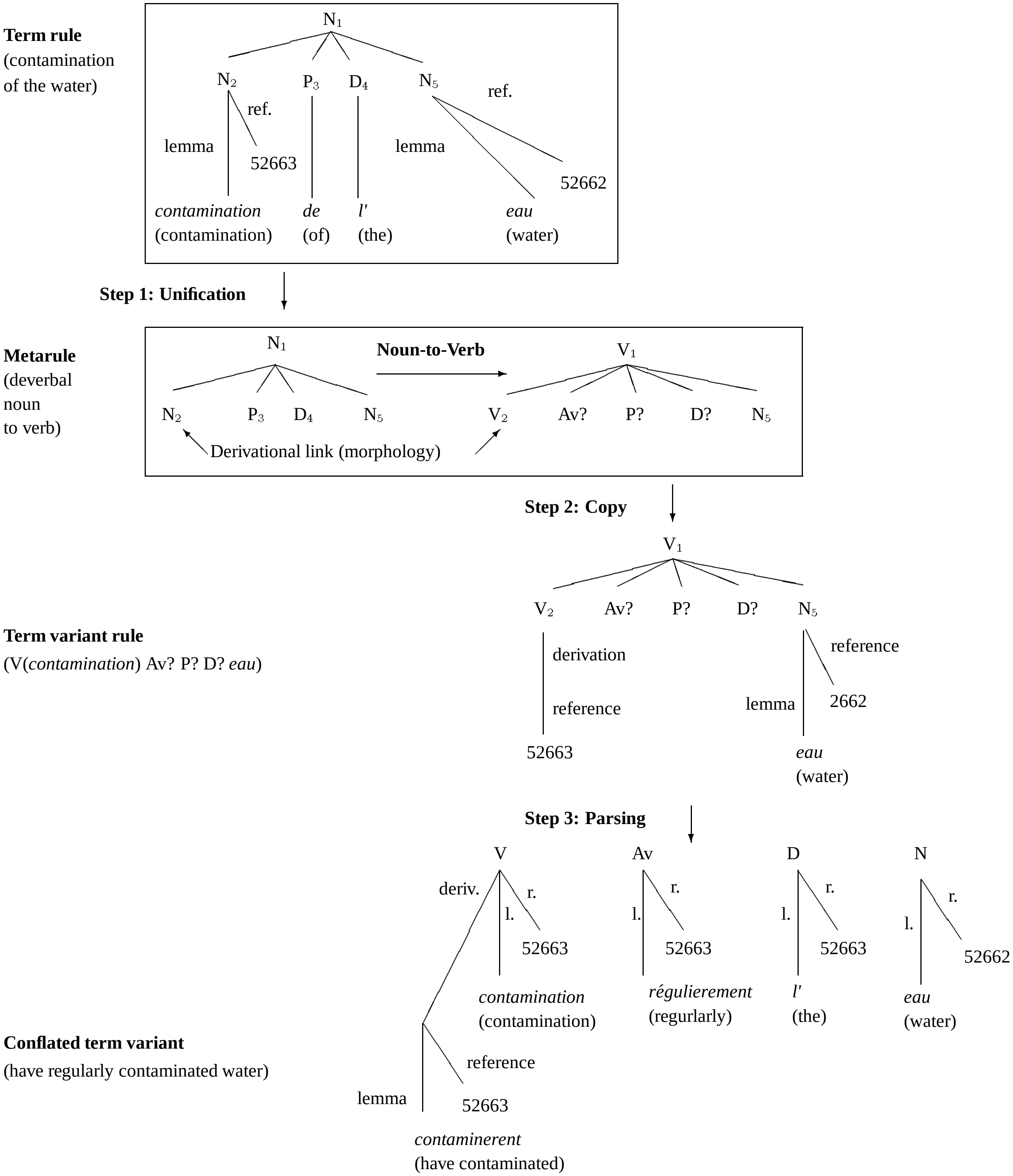

Figure 4: Variant Generation and Conflation.

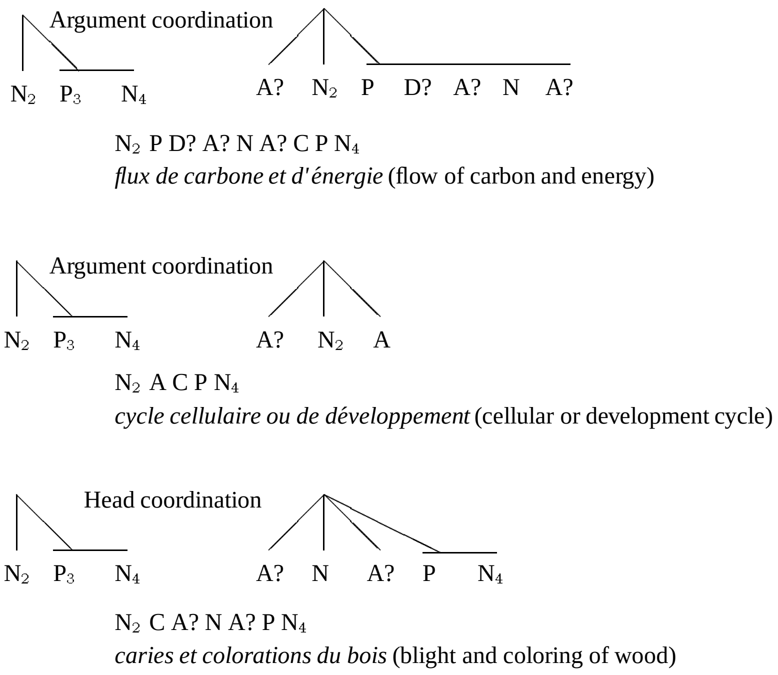
Argument coordination
N2 P3 N4
A? N2 P D? A? N A?
N2 P D? A? N A? C P N4
flux de carbone et d'énergie (flow of carbon and energy)
Argument coordination
N2 P3 N4
A? N2 A
N2 A C P N4
cycle cellulaire ou de développement (cellular or development cycle)
Head coordination
N2 P3 N4
A? N A? P N4
N2 C A? N A? P N4
caries et colorations du bois (blight and coloring of wood)

Figure 5: Conception of Coordination Transformation.

trade-off that we chose for this test was to omit certain syntactic rules which could have spuriously generated undesirable patterns. In future research, our aim is to improve recall, while retaining our very high scores on precision.

Although failure in precision occurred only about 10% of the time, an analysis of errors can be used to improve future versions of the system. Failures are caused by over- and under-generation decisions, causing relationships to be formed between unrelated terms. Examples are *organisme* “organism” in the Biology domain, and *organisation* “organization” in the Management domain. By relating these two words via morphology, relationships are established incorrectly between terms such as *organisme du sol* “soil organism” and *organisation du sol* “soil organization”. Future research will incorporate results of error analysis to filter spurious associations and to establish relationships where they actually occur.

Our results show that identifying morphological variants from a multi-word controlled list of terms expands the indexing coverage by at least 30%. This number was obtained by manually estimating the types of expansions possible for terms. Syntactic variants account for an average of 18% expansion, and morphosyntactic derivations yield an average of 12% expansion. In fact, given the expansion factor, recall should be viewed as satisfactory given the potential for a 30% expansion to degrade recall even further. In light of these facts, recall results are very successful.

Conclusion

Although IR is a relatively mature area of research in computing, the successful incorporation of multilingual NLP techniques has been relatively recent. The identification of solvable IR problems with realistic NLP tools has taken time due to the sheer complexity of NL tasks, and to the time required to develop effective and time-efficient NL tools. The tools described are fully implemented, and function as separate modules of other NLP applications. The NLP module proposed in this paper consists of components for large-scale morphological analysis and generation, syntactic analysis using a full-fledged unification-based formalism capable of high performance on large terminological and textual data, and the construction of constrained morphosyntactic variants. Such results reflect the potential for the application of NLP techniques in indexing and information retrieval.

References

Corbin, Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique*. Niemeyer Verlag, Tübingen.

Jacquemin, Christian. 1996. What is the tree that we see through the window: A linguistic approach to windowing and term variation. *Information Processing & Management*, 32(4):445–458.

Jacquemin, Christian, Judith L. Klavans, and Evelyne Tzoukermann. 1997, to appear. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the Thirty-fifth Annual Meeting of the Association for Computational Linguistics*, San Mateo, California. Morgan Kauffman.

Jacquemin, Christian and Jean Royauté. 1994. Retrieving terms and their variants in a lexicalized unification-based framework. In *Proceedings, 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 132–141, Dublin.

Klavans, Judith L., 1994. *Visions of the Digital Library:Views on using Computational Linguistics and Semantic Nets in Informational Retrieval*. Festschrift for Donald E Walker, Kluwer Academic Press, Dordrecht.

Klavans, Judith L. and Evelyne Tzoukermann. 1992. Morphology. In Stuart C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 2. John Wiley & Sons, New York, second edition, pages 963–970.

Metzler, Douglas P. and Stephanie W. Haas. 1989. The Constituent Object Parser: Syntactic structure matching for information retrieval. *ACM Transactions on Information Systems*, 7(3):292–316.

Mohri, Mehryar and Richard Sproat. 1996. An efficient compiler for weighted rewrite rules. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 231–238, Santa Cruz, Ca. Association for Computational Linguistics.

Pereira, Fernando, Michael Riley, and Richard Sproat. 1994. Weighted rational transductions and their application to human language processing. In *ARPA Workshop on Human Language Technology*, pages 249–254. Advanced Research Projects Agency.

Schwarz, Christoph. 1990. Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, 41(6):408–417.

Sheridan, Paraic and Alan F. Smeaton. 1992. The application of morpho-syntactic language processing to effective phrase matching. *Information Processing & Management*, 28(3):349–369.

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

Strzalkowski, Tomek. 1996. Natural language information retrieval. *Information Processing & Management*, 31(3):397–417.

Tzoukermann, Evelyne and Christian Jacquemin. 1997. Analyse automatique de la morphologie dérivationnelle. In *Colloque Mots possibles et mots existants*, SILEX, Lille, France. CNRS INaLF. *Forthcoming*.

Tzoukermann, Evelyne, Judith L. Klavans, and Christian Jacquemin. 1997, to appear. Effective use of natural language processing techniques for automatic conflation of multi-word terms: the role of derivational morphology, part of speech tagging, and shallow parsing. In *Procedings of SIGIR '97*, New York. Association for Computing Machinery.

Tzoukermann, Evelyne, Dragomir R. Radev, and William A. Gale, 1997, to appear. *Tagging French Without Lexical Probabilities*. Natural Language Processing using Very Large Corpora, Kluwer Academic Press, Dordrecht.