Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction

Denys Proux [1] proux@xrce.xerox.com
François Rechenmann [2] Francois.Rechenmann@inria.fr
Laurent Julliard [1] julliard@xrce.xerox.com
Violaine Pillet [3] violaine@crrm.univ-mrs.fr
Bernard Jacq [4] jacq@lgpd.univ-mrs.fr

[1] Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France
[2] INRIA Rhône-Alpes, Institut National de Recherche en Informatique et en Automatique, 655 avenue de l'Europe, 38330 Montbonnot Saint Martin, France
[3] CRRM, Centre de Recherche Rétrospective de Marseille, Faculté des Sciences et Techniques de St-Jérôme, Université Aix-Marseille, Marseille, France
[4] LGPD, Laboratoire de Génétique et Physiologie du Développement, Parc Scientifique de Luminy, CNRS case 907, 13288 Marseille cedex 9, France

Abstract

Gathering data on molecular interactions to be fed into a specialized database has motivated the development of a computer system to help extracting pertinent information from texts, relying on advanced linguistic tools, completed with object-oriented knowledge modeling capabilities. As a first step toward this challenging objective, a program for the identification of gene symbols and names inside sentences has been devised. The main difficulty is that these names and symbols do not appear to follow construction rules. The program is thus made up of a series of sieves of different natures, lexical, morphological and semantic, to distinguish among the words of a sentence those which can only be potential gene symbols or names. Its performance has been evaluated, in terms of coverage and precision ratios, on a corpus of texts concerning *D. melanogaster* for which the list of names of known genes is available for checking.

1 Introduction

A large part of the data requested by emerging research activities in genomics are not stored in classical databases and have therefore to be searched for directly in the scientific literature. A problem arises when using this primary source of knowledge : papers written to present specific results may contain information on subjects which are secondary, or even marginal, compared to the main topic, but which may be extremely useful for researchers operating on these new genomics areas. This problem, added to the increasing volume of scientific publications, asks for the development of computer assisted methods for extracting data from texts.

Several research projects are working in that direction. For instance, Ohta *et al.* [9] describe the IFBP (Information Finding from Biological Papers) system and its application for the construction of the Transcription Factor DataBase (TFDB). M. Andrade and A. Valencia [1] extract biologically significant words related to protein functions directly from MEDLINE abstracts. The sentences in which these significant words occur might be retained as pertinent textual annotations for the proteins. In these two examples, relative frequencies of word occurrences are the basis for information retrieval and further information extraction. In another field of biology, systematics, A. Taylor [12] parses natural language taxonomic descriptions to support specimen identification. This parsing approach is possible because of the existence of a distinctive sublanguage used in these taxonomic descriptions.

The need for data on molecular interactions to be fed in specialized databases such as FlyNets and GIF-DB for *Drosophila melanogaster* [8] or GeneExpress [6] provides another strong motivation for information extraction from texts. In the context of a research project which involves biology and computer science laboratories, we are developing such a computer system using a linguistic approach based on a grammatical and semantical analysis of sentences. The ultimate goal of the project is to feed an object-oriented knowledge base on molecular interactions [3] with data on several organisms, to organize these data into networks of interactions which can be visualized, edited, analyzed and eventually simulated over time. As a first step in this direction, it is essential to correctly identify gene names in texts. This is the purpose of the present study which has been performed on a text corpus restricted to the fly *D. melanogaster*.

2 The reference corpus

Previous works have been performed by the CRRM to establish a methodology for selecting sentences dealing with genetic interactions from texts [10]. The texts have been extracted from FlyBase, the molecular and genetic database devoted to Drosophila [5]. More specifically, the "Phenotypic Information" data field was used. This field contains short abstracts of biological papers in connection with specific genes of Drosophila.

A dictionary of Drosophila gene symbols has also been built by collecting information from the "Gene Symbol" field. Using this dictionary, we have filtered the corpus and only retained sentences containing at least two gene symbols. Such sentences are indeed believed to have a higher probability to deal with interactions between two partners. The two following sentences are examples extracted from the corpus:

> "The salm gene acts independently of abd-A."
>
> "Recessive mutations of the Hab group of abd-A alleles have been isolated as revertants of a dominant gain-of-function abd-A mutation."

As a result of this filtering process, a corpus a 1200 sentences has been obtained. It is important to notice that this corpus contains only gene symbols, so the early version of the system described in this paper targets gene names designated by a single lexical entity. The term "gene name" will be used hereafter to refer both to gene symbols and gene names based on a single lexeme. A new version of the system is planned to handle multiple word expressions which designate a single entity.

3 Problem specificity

The use of dictionaries of gene names for identification of key semantical entities is a straightforward, but unfortunately limited, solution. When such dictionaries exist, as for example in Flybase for Drosophila, they are not always up to date, as new genes are being discovered and quoted in the literature. Moreover, the authors too often do not respect the recommended gene nomenclature, so that in scientific papers different names might reference a same gene.

In this context, the development of a method which identifies gene names in sentences without relying on pre-defined gene name dictionaries appears highly pertinent.

Fukuda *et al.* [4] describe a method to identify protein names in texts. Their method mainly relies on the assumption that protein names can be identified according to lexical considerations, such as the presence of upper cases and of special characters. Our analysis of the list of gene names for Drosophila shows unfortunately that such an assumption does not seem to hold for genes.

These names can indeed be partitioned into three categories :

- names including special characters, such as upper cases, hyphen, digit, slash, or brackets, e.g. Hrp54 , Lam-B1 , Laer\mt , Lmac\bb , M(2)201 , ...

- names using lower case letters only and belonging to the (English) natural language, e.g. vamp, ogre, zip, zen, ...
- names using lower case letters only, but not belonging to the language, e.g. ynd, zhr, wp, unr, sth, ...

The last two categories gather more than 50% of the 550 gene names appearing in the previously described corpus of 1200 sentences. The existence of gene names which belong to the language is a problem when parsing a sentence. Using a classical dictionary for the lexical lookup, these words (such as "vamp" or "zip") will be recognized as ordinary words since they appear in the dictionary. So they will not be identified as proper nouns, but rather according to their classical grammatical category as it appears in the dictionary. A large part of them will be recognized as nouns which do not disrupt the grammatical structure of the sentence. For instance, in the sentence "boss activates protein synthesis in [...]", "boss" will not be identified as a proper noun, but as a noun. This occurrence does not disrupt the sentence grammatical correctness, but obviously modifies the overall meaning of the sentence.

Fortunately, gene names belonging to the language represent only a very small percentage of all existing gene names (at least for Drosophila). The exact percentage depends on the coverage of the dictionary used in the lexical lookup process. Our tests using the Xerox lexical lookup for English show that 5.6% of gene names of our reference corpus belong to the language : 4.3% are recognized as a noun and 1.3% are recognized as something else (adjective, verb, ...).

In scope	**Out of scope**	**In Conflict (verb, article, ...)**
heat	vamp	a
blood	eve	her
cell	disco	by
double	boss	for
arm	gypsy	if
lab	ocr	is
per	zip	red
...	ogre	can
	...	...

Figure 1: Examples of gene names belonging to the English language. In the reference corpus, the percentages of these names belonging to the "In scope", "Out of scope" and "In Conflict" categories are 32%, 32% and 36% respectively. "Out of scope" words have a primary meaning outside the biological domain.

Furthermore, within the category of potential gene names identified as nouns, some of them could be quite reliably recognized as gene names because they are related to topics outside the domain of molecular genetics. For example, words such as "vamp", zen", or "ogre", when appearing in Medline abstracts, are likely to denote genes instead of their original meaning. These words correspond roughly to one third of the gene names belonging to the language ("Out of scope" category in Fig. 1). Applying such a rule further reduces the percentage of ambiguous gene names to 4%.

4 Linguistic tools

Our system relies on linguistic tools developed at XRCE Grenoble and based on the Finite State Technology. The backbone of the system is built around a specific device called a "tagger" [11], a non-deterministic finite-state automaton which works in three steps : tokenization, lexical lookup, and disambiguation.

The tokenizer [2] splits the sentence into logical lexical entities, the “tokens”. A token can be a simple word, but also a multiword entity, such as “rather than” or “such as”. The lexical lookup uses an ordered stack of final state transducers to process the morphological analysis. If a token does not match with any of the classical word transducers, a last transducer is then applied (the “guesser”) to provide at least one grammatical tag. In our system, each time this “last chance” transducer is called, it sends to the system a signal to identify the corresponding token with a special “guessed” flag. The disambiguator is a program based on Hidden Markov Models [7]. Its purpose is to select the “right” category for a token according to the categories of the disambiguated words that surround it.

An algorithm which the sole purpose would be to detect gene names in texts may have not required the use of a tagger. But as we intend to gather more complex information, we have to achieve grammatical analysis of sentences to identify relations between semantic groups of words. This analysis requires that a part-of-speech tagger previously tags all the lexical entities with an appropriate grammatical category.

A very first test is to make the tagger parse the collection of the 550 unique gene names extracted from the Flybase corpus of 1200 sentences. Fig. 2 displays the grammatical tags which have been attached to these names.

Tags	**Proportions**
Guessed	86.6 %
CARDINAL	1.8 %
NOUN	6.2 %
Other :	5.4 %
PROPER NOUN	3.4 %
ABBREVIATION	0.7 %
Words of the language $\neq$ NOUN	1.3 %

Figure 2: Tags and flag given to gene names by the tagger. The flag “guessed” does not represent a grammatical category, but a notification that a token has not been recognized by a classical transducer in the lexical lookup.

Several useful conclusions may be drawn from this test. First, gene names are basically distributed among the following categories : words with a “guessed” flag, proper nouns, abbreviations and nouns. The first category is by far the largest, confirming that very few gene names are words of the English language. Then, numbers in a sentence are ambiguous because they do generally correspond to numbers, but might in very few cases also appear to be gene names. To avoid the capture of wrong names, numbers are therefore never retained as possible candidates. Finally, it is difficult to set apart gene names belonging to the language and identified as nouns from the natural language.

5 Method overview

The method uses a cascade of specific processes which are sequentially executed (Fig. 3). Two levels can be distinguished.

The first level performs the lexical analysis. It is organized around the tagger previously described. A series of recovery rules are applied on the output of the tagger to remove wrong candidates (e.g. “guessed” words which are not gene names) and to bring back some “not guessed” words into the candidate list. The second level applies a series of contextual rules to validate or invalidate candidates.

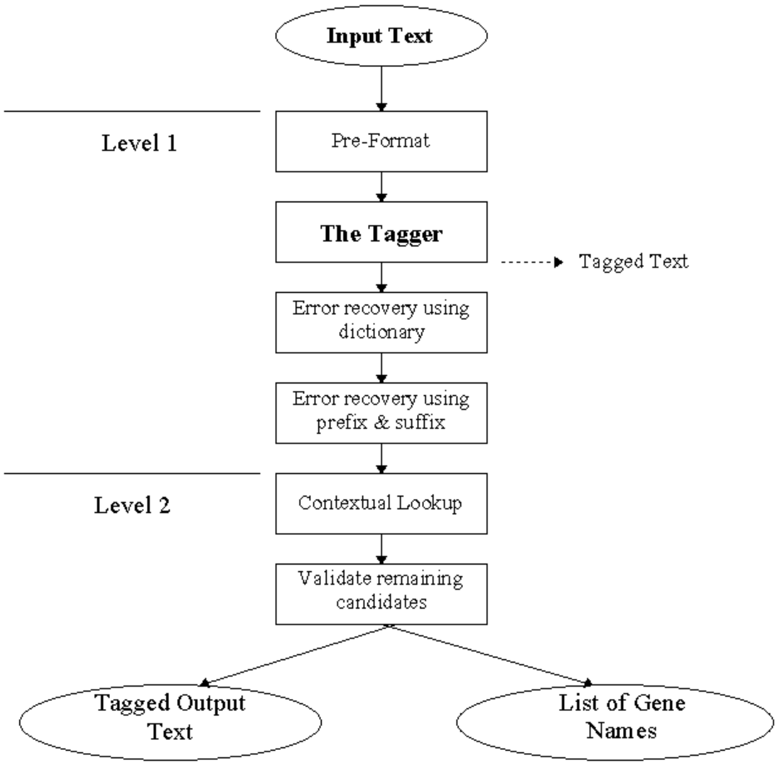

Figure 3: Cascade of finite-state transducers involved in the gene name detection system.

5.1 First level : lexical analysis

The first step is a pure pre-formatting process. Its aim is to facilitate the task of the tokenizer by modifying about fifteen specific expressions such as "23–30[o]C" which is re-written in "20 to 30 Celsius degrees". The objective is to clarify some domain specific notations to avoid bad tokenization and identification by the tagger.

In the second step, the text is sent through the cascade of transducers of the tagger as previously described in section "Linguistic tools".

–¿ 3

The goal of the third step is to recover errors generated by the tagger using a domain-specific dictionary. In order to built a more generic system and improve its reusability, this dictionary has not been integrated into the tagger and can be easily substituted when required. For our tests, this dictionary contained about 200 general expressions from biology. Typical words recovered after the usage of the domain-specific dictionary are for instance species names, units, or common protein names. Gene names belonging to the language, but out of the biological domain (see section 3), are also detected at this step. A small and very restricted dictionary referencing only gene names belonging to the language but out of the biological domain are used for this purpose. As seen in section 3 this approach enables the system to recover nearly one third of these names.

The last step of this first level uses algorithmic rules together with suffix and prefix recognition techniques to recover the remaining tagging errors. These rules are applied in a strict order : 1) algorithmic rules, 2) suffix recognition, 3) prefix recognition.

Algorithmic detection rules recognize complex expressions involving a formalized sequence of characters like nucleotide sequences (e.g. "TCAATTAAAT") or peptide notations (e.g. "VGIDLGT-TYSC"). These expressions are retagged as proper nouns, abbreviations or cardinal numbers so that

Input sentence : “Scr is required to activate fkh expression.”
At the end of Level 1 : (lexical analysis)
“#guessed#scr∧scr+PROP is∧be+VBPRES required∧require+VPAP to∧to+TO activate∧activate+VINF #guessed#fkh∧fkh+ADJPOS expression∧expression+NOUN_SG .∧.+SENT”
At the end of level 2 : (contextual analysis)
“#CANDIDATE#scr∧scr+PROP is∧be+VBPRES required∧require+VPAP to∧to+TO activate∧activate+VINF #GENE#fkh∧fkh+PROP expression∧expression+NOUN_SG .∧.+SENT”

Figure 4: Output of the system at the end of level 1 and level 2.

subsequent grammatical pattern matching will work correctly.

Suffix recognition rules are applied before prefix recognition, because a suffix carries more grammatical information than a prefix. Specific endings are checked among unknown words to remove non-gene names. Among the suffix categories which are checked we can distinguish verbs (suffixes such as “-ed” or “-ing”, e.g. “phosphotransfered”), adverbs (ending with “- ly”, ...), nouns (ending with “-ation”, ...), proteins (ending with “-ase” or “-one”) and so on. Currently the suffix list contains nearly 100 entries.

The last series of rules of the first level tries to recover errors using prefixes. Typical examples of words recovered by such rules are chemical compounds (e.g. “phosphoribosylaminoimidazole”, “ethyl-p-toluenesulphonate”, ...) and biological terms (e.g. “orthologue”). Currently the prefix list contains nearly 200 entries. As part of the system tuning process, the user can customize the parser by adding a specific dictionary of prefixes or suffixes invalidating gene name candidates. If the user includes a small dictionary, more words are passed on to later stages as gene candidates. If the dictionary is large, fewer unknown words are considered.

5.2 Second level : contextual analysis

This level makes use of a collection of validating or invalidating lexical-syntactic patterns provided by the user for the current application and applied only on candidates. Candidate words which are next to some specific expressions can be definitively accepted or rejected. For example, in the sentence “Antp and esp genes activate [...]”, the “Antp” and “esp” unknown words are located in the near context of the word “gene” and are therefore validated as gene names. Patterns aiming at the detection of bibliographic references to remove author names from the list form another example (e.g. “[Proux 97]”).

The last step is purely syntactic and involves a “cleaner” to reformat the resulting text to comply with the requirements of further semantic and grammatical analyses.

Fig. 4 describes the modifications which take place in a specific sentence during the parsing. The output sentence contains the original words, but two kinds of information have been added. The first one is the root form of each token (e.g. is → be), and the second one is the corresponding grammatical tag (e.g. VBPRES stands for the verb “to be” in the present tense and NOUN_SG stands for a singular noun. The ∧ and + symbols are separators. One noticeable thing is the presence at the end of level 1 of the “guessed” flag (#guessed#fkh∧fkh+ADJPOS) to identify words handled by the guesser module. This “guessed” word will be confirmed as a valid candidate after level 2 (#CANDIDATE#fkh∧fkh+PROP).

6 Results and discussion

The reference corpus is composed of the 1200 sentences extracted from Flybase and for which all the 550 occurring distinct gene names are known. This corpus has been divided into two sets. A first set of 450 sentences has been used to tune the system which has then been experimented on a second set of 750 sentences. The performance achieved on this set can be measured through the recall and precision ratios :

Recall	=	94.4 %	
Precision	=	91.4 %	(81 % at the end of the lexical analysis level)

The recall ratio is computed as the number of pertinent items captured by the system over the total number of pertinent items in the corpus. The precision ratio is computed as the number of pertinent items captured by the system over the total number of items captured by the system.

Most of the false positive are words which are specific to biology but are not found in the dictionaries used by the system, such as "ommatidial", "placode" or "sensillum". Some of them are alpha-numeric strings, for instance chromosomal locations, such as "102Efc" or "91B". Other are due to typing errors, such as "compnents" or "phentype". Country names and foreign words also appear in the list.

Gene names not detected by the system are basically dispatched among the following categories:

- numerical expressions (e.g. 1.6.99.7, 0.3M, ...)
- names belonging to the language and with a primary meaning that can belong to the biological domain (e.g. per, red, ...)
- biological vocabulary (autoantiserum, ...)
- proper nouns (e.g. B. Mori, Duchenne, ...)
- other (7q21q22, 1::LYS2, ...)

The recognition of these expressions can be improved by acting at several levels. First, new detection rules can be added for specific expressions such as chemical compounds, numerical expressions or chromosome locations. Then, domain specific vocabulary together with prefix and suffix dictionaries can be expanded. Another important way of improvement relies on a better analysis of the semantic context, to remove more efficiently wrong candidates, as it is planned in the next step.

As for gene names belonging to the language and inside the "In Conflict"category (see section 3), we plan to use a detection of incorrect grammatical structure in sentence parsing to modify the corresponding tag of these words. Several combinations of tags (proper noun *versus* classical grammatical category) will be tried until the sentence is recognized with a correct grammatical structure.

7 Conclusion and perspectives

A new algorithm based on a cascade of transducers to automatically extract gene names from biological texts has been designed and experimented in the context of a research project aiming at gathering and analyzing data on genetic interactions. It is built around finite-state lexical tools, a probabilistic HMM part-of-speech tagger and a collection of finite-state error recovery modules. The analysis is sequentially performed at the lexical and contextual levels.

The system has been designed for the purpose of detecting gene names in Drosophila, a model organism for which a vast amount of data on genetic and molecular interactions can be found in the literature. Tests performed by our system on a selected Flybase corpus have reached a 94% level in recall and a 91% level in precision. It has to be noted that our test study has been performed on sentences extracted from FlyBase, the Drosophila database, in which all gene names are described using their symbol. Such a situation, in which all gene names have been verified and standardized by database annotators may represent an ideal case and it is likely that new problems will arise when using a corpus of journal abstracts or texts. On the one hand, the fact that full gene names are likely to be found in texts (with or without their corresponding symbol) is expected to improve the recall since there are less chances that a full gene name matches a name of the scientific or normal language than a symbol does. On the other hand, a small percentage of gene names are compound words and correct identification of this group of terms as only one gene name will require new tools presently under development. Another problem is that our corpus did not make it possible to test adequately the question of erroneous gene tagging, since all the sentences contained two gene symbols (this was so because this corpus has then been used for interaction prediction). We are presently tackling this problem by using a corpus of Medline abstracts in which neither gene names nor symbols are present.

As part of a larger evaluation of new potential problems when using literature abstracts, preliminary experiments have also been run on a corpus of 25,000 abstracts extracted from Medline. The precision ratio achieved on this corpus has been estimated by sampling. As expected, it is quite lower (70%) than on the Flybase corpus. Several reasons can however explain this performance and provide hints for improvement.

The first one is the problem of compound names already discussed above. Second, the list of identified gene names has been compared to the list of Drosophila gene names only. Although these abstracts have been selected on the basis of the "drosophila" MeSH term, many of them refer in fact to research performed on other species than Drosophila and therefore do contain gene names specific to these species. Clearly, these names have been erroneously considered as false positive (they are actual gene names, but not Drosophila ones), thus lowering the precision ratio. Another reason is that, since the scope of this Medline corpus is larger than the FlyBase one, the error recovery and pattern recognition modules are not yet scaled to such a vocabulary and specific expressions rich environment.

To improve the performance of the system on large corpora, new rules for the identification of specific expressions are to be implemented and the coverage of the specific user dictionaries is to be expanded. The planned connection with a semantic analyzer will enable a finest validation of candidates. The efficiency of such a system would of course be much greater if a formal nomenclature was strictly applied for gene naming.

The last point to recall is that this system is a tool to help identifying gene names in a sentence. It is not designed to be used as a stand-alone application, but to be part of a larger knowledge-based system aiming at the extraction of data on genetic and molecular interactions from texts.

Acknowledgements

A copy of the text corpus used for the present work is available on request. This work has been supported in part by a grant from CNRS (Centre National de la Recherche Scientifique) in the context of its "Génome" research program, and by ANRT (Association Nationale de la Recherche Scientifique) through a CIFRE grant for Denys Proux.

We are grateful to Luc Quoniam and Ambroise Ingold from the CRRM for helpful discussions and for providing the dictionary of FlyBase gene names and for formatting the Medline corpus.

References

[1] Andrade, M. and Valencia, A., Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, *Bioinformatics*, 14(7):600-607, 1998.

[2] Chanod, J. P., Gilman, and Tapanainen, P., A Non-deterministic Tokenizer for Finite-State Parsing, *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI-96)*, Workshop on Extended Finite-State Models of Language, Budapest, Hungary, August 12–16, 1996.

[3] Euzenat, J., Chemla, C., and Jacq, B., A knowledge base for *D. melanogaster* gene interactions involved in pattern formation, *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology (ISMB97)*, Halkidiki, Greece, 108–119, June 21–25, 1997.

[4] Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T., Toward Information Extraction: Identifying Protein Names from Biological Papers, *Proceedings of the Pacific Symposium on Biocomputing (PSB98)*, Hawaii, January 4-9, 707–718, 1998. (available at : http://www-smi.stanford.edu/projects/helix/psb98/)

[5] Gelbart , W.M., Crosby, M., Matthews, B., Rindone, W.P., Chillemi, J., Russo Twombly, S., Emmert, D., Ashburner, M., Drysdale, R.A., Whitfield, E., Millburn, G.H., De Grey, A., Kaufman, T., Matthews, K., Gilbert, D., Strelets, V., and Tolstoshev, C., FlyBase: a Drosophila database. The FlyBase consortium, *Nucleic Acids Research*, 25(1):63–66, January 1, 1997.

[6] Kolchanov, N.A. *et al.*, GeneExpress : A Computer System for Description, Analysis, and Recognition of Regulatory Sequences in Eucaryotic Genome, *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB98)*, AAAI Press, 95–104, Montréal, Canada, 1998.

[7] Kupiec, J., Robust Part-of-speech Tagging Using a Hidden Markov Model, *Journal of Computer Speech and Language*, 6(3):225–242, 1992.

[8] Mohr, E., Horn,F., Janody, F., Sanchez, C., Pillet, V., Bellon, B., Röder L., and Jacq, B., FlyNets and GIF-DB, two Internet databases for molecular interactions in *Drosophila melanogaster*, *Nucleic Acids Research*, 26(1):89–93, 1998.

[9] Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I., and Takagi, T., Automatic Constructing of Knowledge Base from Biological Papers, *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB97)*, AAAI Press, 218–225, Halkidiki, Greece, June 1997.

[10] Pillet, V., Roudani, B., Quoniam, L., and Jacq, B., Extraction automatique et représentation graphique de données biologiques : les interactions génétiques et moléculaires. Application à un organisme modèle et au génome humain, *Colloque sur la Veille Stratégique Scientifique et Technologique (VSST98)*, October 19–23, 1998.

[11] Schiller, A., Multilingual Part-of-Speech Tagging and Noun Phrase Mark-up, *Proceedings of the 15th European Conference on Grammar and Lexicon of Romance Languages (ECGLRL96)*, University of Munich, Germany, September 1996.

[12] Taylor, A., Extracting Knowledge from Biological Descriptions, In *Toward Very Large Knowledge Bases – Knowledge Building and Knowledge Sharing 1995*, N.J.I. Mars, Ed., IOS Press, 114–120, 1995.