From Linguistic Resources to Medical Entity Recognition: a Supervised Morpho-syntactic Approach

Maria Pia di Buono, Alessandro Maisto
and Serena Pelosi
Department of Political, Social and Communication Sciences
University of Salerno
84084 Fisciano, Italy
{mdibuono, amaisto, spelosi}@unisa.it

***Abstract*—Due to the importance of the information it conveys, Medical Entity Recognition is one of the most investigated tasks in Natural Language Processing. Many researches have been aiming at solving the issue of Text Extraction, also in order to develop Decision Support Systems in the field of Health Care. In this paper, we propose a Lexicon-grammar method for the automatic extraction from raw texts of the semantic information referring to medical entities and, furthermore, for the identification of the semantic categories that describe the located entities. Our work is grounded on an electronic dictionary of neoclassical formative elements of the medical domain, an electronic dictionary of nouns indicating drugs, body parts and internal body parts and a grammar network composed of morphological and syntactical rules in the form of Finite-State Automata. The outcome of our research is an Extensible Markup Language (XML) annotated corpus of medical reports with information pertaining to the medical Diseases, Treatments, Tests, Symptoms and Medical Branches, which can be reused by any kind of machine learning tool in the medical domain.**

***Index Terms*—Medical Entity Recognition, Lexicon-Grammar, Morphosemantics, Semi-automatically Generated Lexical Resources.**

I. INTRODUCTION

The necessity to access and integrate in real time health related data that come from multiple sources, opens plenty of opportunities for the research studies in Natural Language Processing (NLP). This comes together with the need of support in the extraction and in the management of useful information and in the development of systems, which must be able to give a structure to the semantic dimension of real words and sentences.

In this paper, we present a Lexicon-grammar (LG) method, that takes advantages from word combination rules and from the lexical and syntactic structures of the natural language. Our purpose is to locate and describe the meaning of phrases, sentences and even entire documents belonging to the medical domain.

In order to overcome the poor flexibility of the existing medical databases with respect to neologisms, we exploit many Morpho-semantic strategies, which can be crucial in the automatic definition of technical-scientific lexicons, in which the global meaning of the words presents strong connections with the meaning of the morphemes that compose them. In other words, we reorganize the information derived from the semantics of the word formation elements, by making the medical words derive the meaning of the morphemes with which they are formed. In this way, starting from a small number of indicators and without any dependence to limited knowledge bases, we can any time automatically build a technical-scientific dictionary of the medical text we process.

In this work, thanks to the opportunities offered by the productive morphology, we automatically locate and define the medical entities contained in a corpus of 989 medical reports. Moreover, using the theoretical insights of the LG framework, we use syntactic rules to semantically describe the categories (e.g., Disease, Treatment, Test, Symptom and Medical Branch) of the located entities. Therefore, if our starting point is a corpus of medical records in electronic format, the output of our research is a structured version of the same corpus, which can be easily reused and queried in every kind of machine learning tool, Clinical Decision Support System (CDSS) or NLP tool in the medical domain.

The paper is structured in the following way: Section II introduces the most important works on the identification and categorization of medical entities in free texts; Section III briefly describes the Lexicon-grammar framework, the Morpho-semantic approach and the tools we used to perform our tasks; in the Section IV, we introduce the automatically generated IMED dictionary and the set of syntactic rules applied to extract entity classes from Medical Records; the Section V describes the structure of the Medical Records Corpus used to test our approach and the results of the application of our method; in the end, Section VI presents the conclusions and the further developments of our research.

 ISBN: 978-1-61208-445-9

II. Related Works

The Medical Entity Recognition (MER) can be decomposed in two main tasks: the extraction of semantic information referring to medical entities from raw texts and the identification of the semantic categories that describe the located entities [1].

As regards the first task, many medical lexical databases (e.g., Medical Subject Headings (MeSH), RxNorm, Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine (SNOMED), and Unified Medical Language System (UMLS), which includes all the other sources) can be used as knowledge base for the location of the medical entities.

Anyway, the quick evolution of entity naming and the slowness of the manual development and updating of the resources often make it necessary to exploit some word-formation strategies, that can be truly helpful in the automatic population of technical-scientific databases. Such strategies concern the Morpho-semantic approach and have been successfully applied to the medical domain by [2] on terminal morphemes into an English medical dictionary; by [3] on medical formative elements of Latin and Greek origin; by [4] on the suffix *-itis*; by [5] on suffixes *-ectomy* or *-stomy* and by [6] on the suffix *-osis*.

Among the most used tools for the MER, we mention MetaMap [7], a reference tool which recognizes and categorizes medical terms by matching noun phrases in free texts to the corresponding UMLS Metathesaurus and Semantic Network, and MEDSYNDIKATE [8], a natural language processor able to automatically acquire data from medical findings reports.

Examples of approaches based on the MetaMap knowledge base are the one of [9], which extracts medical entities from pathologist reports, and the one of [10], which focuses on the extraction of medical problems with an approach based on the MetaMap Transfer and the NegEx negation detection algorithm.

With reference to the second task, we can find in literature rule-based, statistical and hybrid approaches.

As regards the contributions that exploit statistical methods for the identification and classification of medical entities, we mention [11], that uses decision trees or SVMs; [12], that uses Hidden Markov Models or CRFs; [13], that presents a machine learning system which makes use of both local and syntactic features of the texts and external resources (gazetteers, web-querying, etc.); and [14], that obtains the nouns of disease, medical condition, treatment and symptom types, by using MQL queries and the Medlineplus Health Topics ontology (www.nlm.nih.gov/medlineplus/xml.html).

Rule-based methods are the ones proposed by [15], who identifies, with a set of graphical patterns, cause-effect information from medical abstracts in the Medline database, and [16], that manages to extract clinical entities disorders, symptoms and body structures from unstructured text in health records, using a rule-based algorithm.

Hybrid approaches have been proposed by [17] for the extraction of gene symbols and names; by [18] for protein-name recognition and by [19], which combines terminology resources and statistical methods with sensible improvements in terms of Precision.

III. Methodology

Our methodology is based on the LG framework, set up by the French linguist Maurice Gross during the ‘60s and subsequently applied to Italian by [20].

The LG theoretical and practical framework is one of the most consistent methods for natural language formalization, automatic textual analysis and parsing. Its main goal is to describe all mechanisms of word combinations closely related to concrete lexical units and sentence creation, and to give an exhaustive description of lexical and syntactic structures of natural language.

LG theoretical approach is prevalently based on [21], which assumes that each human language is a self-organizing system, and that the syntactic and semantic properties of a given word may be calculated on the basis of the relationships that this word has with all other co-occurring words inside given sentence contexts. The study of simple or nuclear sentences is achieved analyzing the rules of co-occurrence and selection restriction, i.e., distributional and transformational rules based on predicate syntactic-semantic properties.

As described in Section IV-B, in this work, following LG methodology, we anchored the recognition of terminological ALUs (Atomic Linguistic Units) to the sentence structures that recursively occur in medical reports. This way, on the base of co-occurrence rules, which can be characterized by different levels of variability, we could correctly annotate and classify a great part of the medical entities contained in our corpus.

As it is commonly done in literature, in our work we divided the Medical Entity Recognition into two subtasks, every one of which takes advantages from different resources.

- Semi-automatically generated lexical resources, for the extraction of semantic information from raw texts (see Section IV-A);
- Syntactic rules, for the extraction of semantic and domain information. The assumption for this step is that domain terminology is strictly interlinked with syntactic combination and co-occurrence behaviors (see Section IV-B).

Table 1 shows entity types recognized in our experiment; we also provide a description for these one.

A. NLP Tool

For our TE task we use NooJ, a software developed by Max Silberztein [22]. This system allows to formalize natural language descriptions and to apply them to corpora. NooJ is used by a large community, which developed linguistic modules,

 ISBN: 978-1-61208-445-9

TABLE I
ENTITY RECOGNITION CLASSES

Entity Type	Details
Disease	disorders and medical conditions
Treatment	therapies following diagnosis
Drug	information about prescribed drugs
Test	analysis and exams
Symptom	subjective evidences of diseases or of patient's conditions
Medical Branch	specific medical subdomains

including Finite State Automata/Transducers and Electronic Dictionaries, for more than twenty languages. The Italian Linguistic Resources have been built by the Computational Linguistic group of University of Salerno, which started its study of language formalization from 1981 [20]. Our analysis is based on the Italian module for NooJ [23], which is enriched with IMED and with grammars for Text Extraction (TE).

IV. LINGUISTIC RESOURCES

A. Italian Medical Electronic Dictionary

In order to automatically create the Italian Medical Electronic Dictionary (IMED) of the disease ALUs occurring in the corpus, we exploited morphosemantic strategies, which uses the semantics of a special kind of morphemes to identify and describe disease nouns or adjectives. Such kind of morphems are called neoclassical formative elements [24]. They come into being from Latin and Greek words and are generally used to form both technical-scientific words and ordinary words in a very productive way. They can combine themselves with other formative elements or with independent words.
In this paper we will talk about them using the word "confixes", which has been predominantly employed in literature [25]–[29].

The Medical Morphosemantic Module (M^3) we implemented is composed of the following resources:

- An Electronic Dictionary of Italian Morphemes belonging to the Medical Domain called *M3.dic*.
- Seven Morphological Grammars, denominated *M3#.nom*
- A syntactic Grammar, named *M3.nog*

The Dictionary *M3.dic* contains morphemes of the Italian medical domain which have been extracted from the electronic version of the GRADIT [30]. The morphemes has been divided into three classes: prefixes, suffixes and confixes, on the base of the positions of the morphemes in the words. Table II shows the morphemes extracted from the GRADIT. Each morpheme is described by a tag that specifies the meaning of the morpheme (i.e., *-oma* corresponds to the descriptions *tumori*, "tumours") and a tag that gives its medical subcategory (assigned with the support of a domain expert by dividing the macro class of the medicine into 25 subcategories, i.e., CARDIO, "cardiology"; ENDOCRIN, "endocrinology"; PSIC, "psychiatry" GASTRO, "gastroenterology"; PNEUMO, "pneumology"; NEURO, "neurology"; etc). We made use of a class UNKNOWN that has been used as residual category, in order to collect the words particularly difficult to classify.

TABLE II
MORPHEMES EXTRACTED FROM THE GRADIT

Manner of Use	Category	Number
Medicine	Confixes	485
Medicine	Suffixes	5
Medicine	Prefixes	7
Anatomy	Confixes	104
Anatomy	Prefixes	3

The morphemes that were not contained in the GRADIT's medical category have been manually added to our list, i.e., morphemes that are used in the formation of adjectives. The electronic dictionary of medical morphemes is classified in the following way:

- Confixes (CPX): neoclassical formative elements that appear in the initial part of the word (i.e., *pupillo-*, *mammo-*, *cefalia-*);
- Prefixes (PFX): morphemes that appear in the first part of the word and are able to connote it with a specific meaning (i.e., *-ipo*, *-iper*);
- Suffixes (SFX): morphemes that appear in the final part of the word and are able to connote it with a specific meaning (i.e., *-oma*, *-ite*);
- Suffixes for the adjectives formation: derivational morphemes that make it possible to derive and distinguish in the medical domain the adjectives (i.e., *polmonare*, "pulmonary") from the nouns that have a morpho-phonological relation with them (i.e., *polmone*, "lung").

The IMED has been completed with the addition of an electronic dictionary composed of more than 700 Concrete nouns of body/organism parts ("+Npc", i.e., *braccio*, "arm"; "+Npcorg", i.e., *cervello*, "brain") and of more than 400 Concrete nouns of drugs and medicines ("+Nfarm", i.e., *morfina*, "morphine ") developed by the Maurice Gross Laboratory of the University of Salerno.

The dictionaries works in combination with seven Morphological Grammars built with *Nooj*, which are able to find occurrences and co-occurrences of medical morphemes or nouns in medical documents' words. The seven grammars include the following combination of morphemes:

1) *confixes-confixes* or *prefixes-confixes* or *prefixes-confixes-confixes*;
2) *confixes-suffixes* or *prefixes-confixes-suffixes*;
3) *confixes-confixes-suffixes* or *prefixes-confixes-confixes-suffixes*;

 ISBN: 978-1-61208-445-9

4) *nouns-confixes*;
5) *prefixes-nouns-confixes*;
6) *confixes-nouns-confixes*;
7) *nouns-suffixes*;

In order to complete the IMED dictionary with medical multiword expressions, a syntactic grammar, built with NooJ, has been created with the goal of finding the following combination of Nouns (N), Adjectives (A) and Prepositions (P): N, NN, AN, NA, NNA, NAA, NPN.

B. Syntactic Rules

In the corpus, we noticed the presence of recursive sentence structure, in which we recognized specific and terminological ALUs. In this way, we could identify open series compounds, that are ALUs in which one or more fixed elements co-occur with one or more variable ones.
On the basis of this evidence, we developed different Finite State Automata for the TE task and for annotating treatments, tests, symptoms.

As for semantics, we observed the presence of compounds in which the head did not occur in the first position; for instance, the open series to recognize treatments *terapia di* N , "therapy of N", places the heads at the end of the compounds, being *terapia* used to explicit the notion N0 is a part of N1.

In Figure 1(a) and Figure 1(b), we recognized Treatment and Test classes by selecting a series of nouns, as fixed part, and a variable part, as head, formed by a Noun Group and/or an Adjective. We also applied a node with the option 'unknown word' (UNK); this feature allowed us to retrieve words which have not been inserted in IMED, also in order to update our dictionary.

To extract the Symptom class we developed a Finite State Automaton (Figure 1(c)) in which semantic features can be identified using grammars that are built on specific verb classes (semantic predicate sets) - i.e., *presentare, riferire, esporre, etc...*, "to present, to report, to express, etc..."; in such cases, co-occurrence restrictions can be described in terms of lexical forms and syntactic structures.

We used the grammatical information with which dictionary entries are tagged and syntactic rules as a weighting preference for the co-occurrence selection. So, we developed matrix tables in which semantic role sets, established on the basis of those constrains (properties), are matched with grammatical and syntactic rules. Matrices list a certain number of verbal entries and a specific number of distributional and syntactic properties.

During the recognition process, labeled IMED entries and FSA are the inputs. After the phase of text processing, the result is as follow:

⟨cardiology⟩ *Il Paziente* ⟨symptom⟩ *affetto da ipertensione arteriosa e BPCO* ⟨/symptom⟩. *Nel 1998 é stato sottoposto ad* ⟨treatment⟩ *intervento di sostituzione valvolare aortica mediante protesi meccanica Carbomedics* ⟨/treatment⟩ *e in quell'occasione le coronarie erano risultate prive di lesioni significative. Dopo l'intervento il Paziente ha eseguito periodici* ⟨test⟩ *controlli cardiologici* ⟨/test⟩ *presso l'Ospedale di Montichiari per* ⟨symptom⟩ *fibrillazione atriale ad elevata risposta ventricolare e scompenso cardiaco* ⟨/symptom⟩ ⟨/cardiology⟩.

"⟨cardiology⟩ The patient is ⟨symptom⟩ suffering from hypertension and BPCO ⟨/symptom⟩. In 1998 he had ⟨treatment⟩ surgery for aortic valve replacement using Carbomedics mechanical prosthesis ⟨/treatment⟩ and coronary arteries did not have significant injuries. After surgery, the patient performed periodic ⟨test⟩ cardiology checks ⟨/test⟩ at the Hospital of Montichiari for ⟨symptom⟩ atrial fibrillation with high ventricular response and heart failure⟨/symptom⟩ ⟨/cardiology⟩ ".

V. Testing and Results

The annotation process is performed on Italian clinical texts. The corpus has been built from a collection of 989 real medical records, opportunely anonymized with regards to every kind of sensitive data they contained.

Our corpus provides information about the Family History, the Physiological Anamnesis, the Past Illnesses, the Anamnesis, the Medical Diary and the Diagnosis Review for every patient. For our analysis we kept out the Family History and the Physiological Anamnesis sections, since they did not contain concepts, assertions or relation information.

The corpus, pre-processed with Nooj exploiting the traditional NLP pipeline, includes 470591 text units, 41409 different tokens and 1529774 word forms.

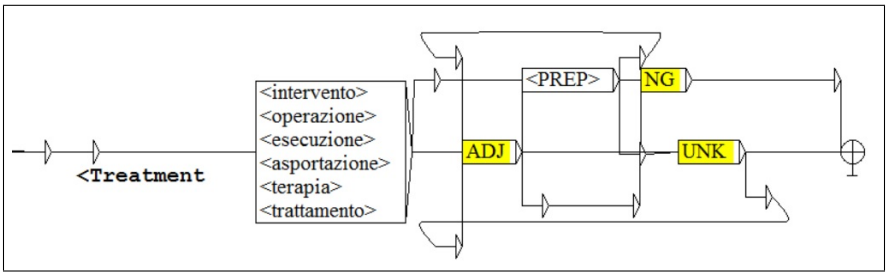
An evaluation of the results produced by our MER tool is given in Table III. We gave a measure of the validity of our method by calculating the Precision, the Recall and the F-score in the extraction of every entity class. In this phase we merged together the classes Drug/Treatment and Disease/Medical Branch because the syntactic grammars used to locate them are the same and our tool presents their results in the same list of concordances. Anyway, the tags used to annotate them are always different, so a distinction between these categories is performed any time.

As we can notice, the values present a variability with reference to the different categories, but we consider the average results very satisfying; nevertheless we are already planning to enrich our research outcomes with many other improvements.

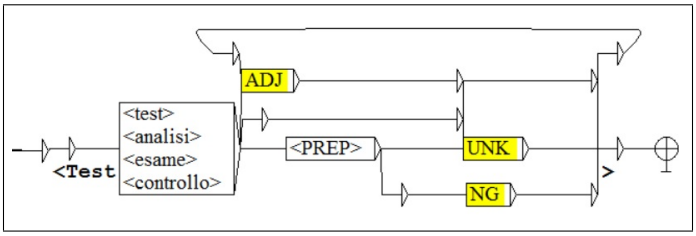
TABLE III
Evaluation

Entity Name	Precision	Recall	F-score
Symptom	0,75	0,52	0,61
Drug and Treatment	0,83	0,51	0,63
Test	0,96	0,51	0,67
Disease and Medical Branch	0,69	0,76	0,72
Average	0,80	0,58	0,66

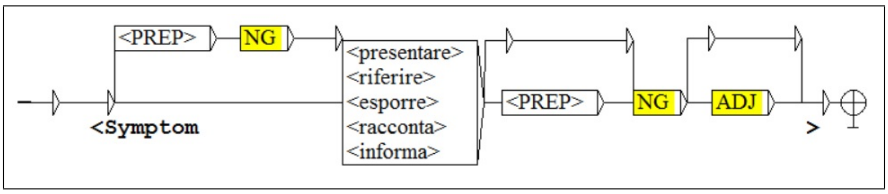
 ISBN: 978-1-61208-445-9



(a) Sample of FSA for Entity Recognition of Treatment



(b) Sample of FSA for Entity Recognition of Test



(c) Sample of FSA for Entity Recognition of Symptom

Fig. 1. Samples of FSA

All the annotations produced by the application (almost 5000 with the morpho-semantic method; more than 4000 with the syntactic strategies and about 2500 applying the preexistent dictionaries of the Italian module of Nooj) of our method and resources can be reused to enrich lexical databases or ontologies referred to the medical domain. Obviously, the size and the quality of the enrichment is strictly dependent on the largeness and on the content of the corpus on which the Nooj resources are applied. Therefore, in order to obtain widespread medical databases, it is preferable to use corpora able to cover the larger group of medical branches.

VI. Conclusion and Future Work

In this work, we presented our methodology for the extraction of entity classes from medical records, conducting different levels of linguistic analysis. Our framework is based on a robust definition language, which is used for creating and extending grammars and lexicons.

In our experiment, we considered the issue of entity boundaries carefully, but result analysis shows a request of improvement in ALUs recognizing. This comes from the specific use of medication abbreviations and word separators and from a nonstandard method to compile free-text notes. Challenging areas are the presence of ambiguous phenomena, e.g., fractions or numbers without unit or time references, and the use of brand name of drugs or a class of products, i.e., eye drops.

The combination of computational morphology and semantic distribution proposed here indicates very promising perspective: processing different corpora could instruct our system to recognize more recursive phenomena and more stop words in order to overcome partial matching issues. Future works aim at integrating our tools with the alignment of ontology constraints to syntactic relations in order to improve extraction of clinical concepts from notes, increasing the interoperability and the utility of clinical information.

References

[1] A. B. Abacha and P. Zweigenbaum, "Medical entity recognition: A comparison of semantic and statistical methods," in *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, 2011, pp. 56–64.

[2] A. W. Pratt and M. Pacak, "Identification and transformation of terminal morphemes in medical english." *Methods of information in medicine*, vol. 8, no. 2, pp. 84–90, 1969.

[3] S. Wolff, "The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding," *Methods of Information in Medicine*, vol. 23, no. 4, pp. 195–203, 1984.

[4] M. G. Pacak, L. Norton, and G. S. Dunham, "Morphosemantic analysis of-itis forms in medical language." *Methods of Information in Medicine*, vol. 19, no. 2, pp. 99–105, 1980.

[5] L. Norton and M. G. Pacak, "Morphosemantic analysis of compound word forms denoting surgical procedures." *Methods of Information in Medicine*, vol. 22, no. 1, pp. 29–36, 1983.

[6] P. Dujols, P. Aubas, C. Baylon, and F. Grémy, "Morpho-semantic analysis and translation of medical compound terms." *Methods of Information in Medicine*, vol. 30, no. 1, p. 30, 1991.

[7] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[8] U. Hahn, M. Romacker, and S. Schulz, "Medsyndikatea natural language system for the extraction of medical information from findings reports,"

 ISBN: 978-1-61208-445-9

International journal of medical informatics, vol. 67, no. 1, pp. 63–74, 2002.

[9] G. Schadow and C. J. McDonald, "Extracting structured information from free text pathology reports," in *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 584.

[10] S. M. Meystre and P. J. Haug, "Comparing natural language processing tools to extract medical problems from narrative text," in *AMIA annual symposium proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 525.

[11] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002, pp. 1–7.

[12] Y. He and M. Kayaalp, "Biological entity recognition with conditional random fields," in *AMIA Annual Symposium Proceedings*, vol. 2008. American Medical Informatics Association, 2008, p. 293.

[13] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting context for biomedical entity recognition: from syntax to the web," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. Association for Computational Linguistics, 2004, pp. 88–91.

[14] M. de la Villa, F. Aparicio, M. J. Maña, and M. de Buenaga, "A learning support tool with clinical cases based on concept maps and medical entity recognition," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 2012, pp. 61–70.

[15] C. S. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 336–343.

[16] M. Skeppstedt, M. Kvist, and H. Dalianis, "Rule-based entity recognition and coverage of snomed ct in swedish clinical text." in *LREC*, 2012, pp. 1250–1257.

[17] D. Proux, F. Rechenmann, L. Julliard, V. Pillet, B. Jacq *et al.*, "Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction," *Genome informatics series*, pp. 72–80, 1998.

[18] T. Liang and P.-K. Shih, "Empirical textual mining to protein entities recognition from pubmed corpus," in *Natural Language Processing and Information Systems*. Springer, 2005, pp. 56–66.

[19] A. Roberts, R. J. Gaizauskas, M. Hepple, and Y. Guo, "Combining terminology resources and statistical methods for entity recognition: an evaluation." in *LREC*, 2008.

[20] A. Elia, M. Martinelli, and E. D'Agostino, *Lessico e Strutture sintattiche. Introduzione alla sintassi del verbo italiano.* Napoli: Liguori, 1981.

[21] Z. S. Harris, "Notes du cours de syntaxe, traduction française par maurice gross," *Paris: Le Seuil*, 1976.

[22] M. Silberztein, "Nooj manual," *Available for download at: www. nooj4nlp.net*, 2003.

[23] S. Vietri, "The italian module for nooj." in *In Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it 2014*. Pisa University Press, 2014.

[24] A. M. Thornton, *Morfologia*, R. Carocci Editore, Ed., 2005.

[25] A. Martinet, *Syntaxe generale*, A. Colin, Ed., 1985.

[26] A. Kirkness, "Aero-lexicography: Observations on the treatment of combinemes and neoclassical combinations in historical and scholarly european dictionaries," *Willy Martin ua (Hrsg.): Euralex*, pp. 530–535, 1994.

[27] S. C. Sgroi, "Per una ridenizione di "confisso": composti confissati, derivati confissati, parasintetici confissati vs etimi ibridi e incongrui," *Quaderni di semantica*, vol. 24, pp. 81–153, 2003.

[28] P. D'Achille, *L'italiano contemporaneo*, B. Il Mulino, Ed., 2003.

[29] T. De Mauro, *Nuove Parole Italiane dell'uso*, ser. GRADIT, T. UTET, Ed., 2003, vol. 7.

[30] ——, *Grande Dizionario Italiano dell'Uso*, T. UTET, Ed., 1999, vol. 8.

 ISBN: 978-1-61208-445-9