# Medication Extraction and Guessing in Swedish, French and English

**Thierry Hamon[a], Natalia Grabar[b], Dimitrios Kokkinakis[c]**

[a] *LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, France*
[b] *CNRS UMR 8163 STL, Université Lille 1&3, France*
[c] *CLT & Språkbanken, Department of Swedish Language, University of Gothenburg, Sweden*

**Abstract**

*Extraction of information related to the medication is an important task within the biomedical area. Our method is applied to different types of documents in three languages. The results indicate that our approach can efficiently update and enrich the existing drug vocabularies.*

***Keywords:*** *Natural language processing, Vocabulary, Medication names, Medical Informatics, France, Sweden*

## Introduction

Drugs occupy an important place in the biomedical area, although information on them is often not easily accessible. Moreover, only a small part of medical area information is available as structured data. This representation practically prohibits large scale data mining, data analytics and personalized care. Natural Language Processing (NLP) methods are used for accessing drug information. Drugs mentions are often detected thanks to drug lexica [1]. Several studies have addressed drug detection in clinical documents and research literature [2-3]. Two previous studies addressed detection of new drug names in narratives through internal [4] and external [5] methods. We propose to exploit both internal and contextual clues for the detection and guessing of the medication names.

## Approach and Results

We process clinical and scientific documents in three languages (English, Swedish and French), and exploit dedicated resources: existing drugs nomenclatures, terminologies [1], and typical endings of medication names, i.e., *ine, one, ase, ate, rex* (for disambiguation when detecting new drug names). The reference data have been built manually. Medication names detection is done with drug nomenclatures and contextual method designed for the guessing unknown medication names. All documents are all processed in the same way: (1) NLP pre-processing; (2) extracting known medication names; (3) guessing new medication names; (4) evaluating results.

The main clue for the detection of new medication names relies on the fact that medication names often occur in specific semantic contexts together with medication-related information, such as dosage, frequency, mode of administration, etc:

SE*: inj Furosemid 80 mg x 2, T trombyl 75 mg x 1*

EN*: methadone 20 bid, ofloxacin 200 mg p.o. q 12*

FR*: tahor 40mg 1 cp le soir, Temerit 5 mg : 1-0-0*

The corresponding semantic pattern to these contexts is *m do mo? f*, with the following variables: medication name *m*, dosage *do*, administration mode *mo*, and frequency *f*. This pattern has can be exploited for extraction of the medication-related information, establishing the relations between the elements of the pattern and guessing elements of the pattern. We exploit it for this last function. For instance, if the first entity *m* is missing but other entities (*do*, *mo* and *f*), or some of them, are instantiated, then we can infer the entity positioned before the dosage information as medication name. If necessary, internal filtering of the extracted element is applied.

Medication extraction results performance varies between 81%-97% for precision and 85%-100% for recall. This evaluation covers known and new drug names (*Oxix Turbuhaler, Felodipin, Cellvept, Advagraf, Novonom, Bitildiem, Tacro, immunosup, Pravastatin, Sertraline, Serax, Quetiapine, Restoril)*. Future work will propose a better filtering of the new drug names and exploitation of other kinds of corpora.

## References

[1] RxNorm, a standardized nomenclature for clinical drugs. Technical report, National Library of Medicine, Bethesda, Maryland, 2009.

[2] Poon E, Blumenfeld B, Hamann C, et al. Design and implementation of an application and associated services to support interdisciplinary medication reconciliation efforts at an integrated healthcare delivery network. J Am Med Inform Assoc 2006;13(6):581–92.

[3] Chen E, Hripcsak G, Xu H, Markatou M, and Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 2008;15(1):87–98.

[4] Segura-Bedmar I, Martinez P, and Segura-Bedmar M. Drug name recognition and classification in biomedical texts. Drug Safety 2008;13(17-18):816–23.

[5] Xu R, Morgan A, Das AK, and Garber A. Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. In: BioNLP, 2009:63–70.

**Address for correspondence**

Thierry Hamon, LIM&BIO (EA3969), UFR SMBH Léornard de Vinci

Université Paris 13, 74, rue Marchel Cachin 93017

Bobigny Cedex, France (Email: thierry.hamon@univ-paris13.fr)