isis ingénieurs
informatique et systèmes
d'information pour la santé

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

TIC Salut Social

FIB

ENCODER

Automatic encoding of natural language into ICD-10-CM / PCS

Bachelor thesis: 5th year internship rapport

February 26, 2018 – July 27, 2018

Fundació TIC Salut Social

Parc TecnoCampus Mataró Maresme - Torre TCM3

Av. Ernest Lluch, 32, 6a planta | 08302 Mataró

Student	Internship supervisor	University supervisor
Pau Garcia Gozàlvez	Ariadna Rius Soler	Georges Soto-Romero
Promotion 2018	Head of the Standards and Interoperability Office	Director of École d'ingenieurs ISIS

Company informations

Company name	Fundació TIC Salut Social
Web site	www.ticsalutsocial.cat
Contact mail	info@ticsalutsocial.cat
Contact phone	+34 93 553 26 42

Organism of	Departament de Salut, Generalitat de Catalunya
Address	Parc TecnoCampus Mataró Maresme - Torre TCM3 Av. Ernest Lluch, 32, 6a planta | 08302 Mataró

Activity area	Innovation and development of new technologies in the health sector in the Catalan hospital network
Foundation date	September 19, 2006
Number of employees	About 30

Logo:

Professional contacts

Ariadna Rius Soler	**Jordi Martinez**	**Francesc Garcia Cuyas**
Head of the Standards and Interoperability Office.	Director of Innovation in TIC Salut Social	Director of TIC Salut Social
arius@ticsalutsocial.cat	jmartinez@ticsalutsocial.cat	fgarciacuyas@ticsalutsocial.cat

Abstract

This project has been realized in the Fundació TIC Salut Social, an organization of the Health Department of Catalonia. The main objective is to study and test several natural language processing algorithms to find the best way to encode sentences into ICD-10-CM/PCS, a clinical classification of diagnostics and procedures.

The need for this project is due to the recent change from ICD-9 to ICD-10 done by the Spanish Government. For this reason, from now medical staff needs to encode into ICD-10 what can be a difficulty due to its big difference to the last version.

To study the best way to solve this problem, in this project will be developed a prototype called ENCODER. This user-friendly software corrects spelling mistakes and encodes, in several languages, an input text to get the ICD-10 code by using different natural language processing techniques. Apart from that, it uses the SNOMED CT descriptions because of its closeness to the natural way of speak of the medical staff.

ENCODER has been tested using SNOMED CT descriptions and the results have been satisfactory. It has encoded correctly the 80% of Spanish SNOMED CT descriptions and the 65% of the Catalan ones. Apart from that, the 60% of natural language diagnostics have been well-encoded.

So that, we conclude that the software developed can be a good solution to solve the problem. Another conclusion is that the difference between the Spanish and Catalan results may be due to a bad translation since ENCODER applies the same algorithm in both. Seeing the natural language results, it can be appreciated that they only differ from 5% to the Catalan ones, so we can induce that it is due to the same reason.

Keywords: Semiautomatic clinical codification, ICD-10-CM/PCS, Natural language processing, Artificial intelligence, SNOMED CT, Search engine, Python, MongoDB.

Resum

Aquest projecte ha sigut realitzat a la Fundació TIC Salut Social, una organització del Departament de Salut de Catalunya. L'objectiu principal és estudiar i provar diferents algoritmes de processament de llenguatge natural per trobar la millor manera de codificar diagnòstics en CIM-10-CM/SCP, la classificació clínica de diagnòstics i procediments.

Recentment el Govern espanyol, per real decret, ha actualitzat la versió de la CIM-9 a la CIM-10. Per aquesta raó, a partir d'ara el personal mèdic ha de codificar en CIM-10, el que pot significar una dificultat degut a les seves grans diferències amb l'antiga versió.

Per estudiar la millor manera d'abordar el problema, en aquest projecte s'ha desenvolupat un prototip anomenat ENCODER. Aquest programari proposa una codificació en CIM-10 donat un text d'entrada en possibles diferents llengües mitjançant diferents tècniques de processament de llenguatge natural. Per realitzar-ho, s'utilitzen les descripcions de SNOMED CT ja que són més pròximes a la manera natural de parlar del personal mèdic, pel que milloren la codificació.

ENCODER ha sigut avaluat fent servir les descripcions de SNOMED CT amb uns resultats satisfactoris. Ha codificat correctament el 80% de les descripcions de SNOMED CT en castellà i el 65% de les escrites en català. D'altra banda, ha codificat correctament el 60% dels diagnòstics escrits en llenguatge natural per un metge.

Com a conclusió, el programari desenvolupat pot ser una solució pel problema proposat. D'altra banda, es pot deduir que la diferència en els resultats en funció de la llengua d'entrada es pot deure a una mala traducció, ja que ENCODER aplica el mateix algoritme per totes les llengües. Per respecte als resultats del llenguatge natural, només difereixen en un 5% dels del català, així que es pot induir la mateixa raó.

Paraules clau: Codificació clínica semiautomàtica, CIM-10-MC/SCP, Processament del llenguatge natural, Intel·ligència artificial, SNOMED CT, Cercador, Python, MongoDB.

Acknowledgments

To begin with, I have to thank all the team of the Fundació TIC Salut Social for these two agreeable internships. Thanks to them, it has been a great experience with which I have grown professionally and personally. Especially, I have to thank Ariadna Rius for her support and also for all her teachings during these internships. This project has been possible thanks to her advices, ideas, and recommendations that have served me as a reference point in this trajectory.

I want to thank Dr. Francesc Garcia Cuyàs and Dr. Jordi Martinez for opening the doors of this company and for having created a pleasant work environment. On the other hand, I wish to thank all the people that have helped me with some tasks of the project: again Dr. J. Martinez for being offered to test the tool developed in this project, Miquel Martí for helping me with the interface designing of this tool and finally Anna Ceresuela for her help in all related clinical terminologies staff.

This project has been possible thanks to the knowledge acquired in the École Ingenieurs ISIS and all its teachers. Especially, I have to thank Monsieur Soto-Romero for the support and monitoring given during this project and Madame Lhôte for its internship management.

Finally, I want to thank Carolina Martín and María Teresa Abad from the FIB, UPC. They have made possible this Erasmus that has permitted me to take this way and go for two years to study in France.

Table of contents

Glossary

- **Generalitat de Catalunya** is the institutional system in which the government of Catalonia is organized politically. The Generalitat holds exclusive and wide jurisdiction in various matters of culture, environment, communications, transports, commerce, public safety and local governments. However, in aspects relating to education, health, and justice, the region shares jurisdiction with the Spanish government.

- **Departament de Salut** is the main administrative body of the Generalitat de Catalunya in healthcare decision-making. It has the exclusive competence of the organization, internal functioning, assessment, inspection and control of health centers, services, and establishments. On the other hand, it participates in the planning and coordination of health affairs.

- **SISCAT**. *Sistema sanitary integral d'utilització pública de Catalunya* (Healthcare Public System of Catalonia). It is the healthcare network that groups together all the public hospital centers, primary care centers, mental health centers, transport resources and others.

- **CatSalut** is the public insurer of Catalonia. Its objective is to guarantee full, public and quality health coverage to all the citizens of the territory.

- **OFTSI**. *Oficina d'Estàndards i Interoperabilitat.* It is the Standards and Interoperability Office of the Fundació TIC Salut Social.

- **ICD-10 (*CIM-10 in Catalan and French*)** is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs, and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases

- **ICD-10-CM/PCS (*CIM-10-MC/SCP in Catalan and French*)** is the International Classification of Diseases, 10th revision, Clinical Modification / Procedure Coding System. It is a subset of ICD-10 that contains the codes for clinical diagnostics and procedures.

- **SNOMED CT**. *Systematized Nomenclature of Medicine Clinical Terms*. It is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms, and definitions used in clinical documentation and reporting.

- **Snomed International (formerly IHTSDO).** It is an international agency that controls, manages and distributes the International version of SNOMED CT.

  http://www.snomed.org/

- **Catalan Extension of SNOMED CT.** The Catalan Extension is maintained and distributed by the OFTSI and used by different healthcare centers of SISCAT. It is a non-international subset of concepts of SNOMED CT in Catalan and Spanish managed by TIC Salut Social.

- **Artificial intelligence** is the set of theories and techniques used to create machines capable of simulating different areas of human intelligence

- **Natural Language** is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation

- **Natural Language Processing (NPL)** is a domain of computer science, artificial intelligence, and linguistics that studies the interactions between computers and human (natural) languages.

- **MongoDB** is a document-oriented NoSQL database system, developed under the concept of open source.

- A **Map** in computer science is a file that links a set of elements to another one. For each element of the source set, the map gives an element of the target set. In this project, we are using a map that links SNOMED CT with ICD-10-CM.

- **Stopwords** are meaningless words, that can be deleted from a sentence and it still has the same meaning (e.g. articles, pronouns, prepositions).

1. Introduction

My internship has been realized in the Fundació TIC Salut Social in Mataró, Spain with a duration of 6 months. This foundation is part of the Departament de Salut of the Generalitat de Catalunya and works to improve technologically the health sector of Catalonia.

In 2017 I also did my 4th year ISIS internship in TIC Salut Social. It lasted 4 months but I worked another 2 months as an employee after my internship. For this reason, I already knew the working way of the company, so my reincorporation was pretty fast and simple.

In my last internship, I developed a web application to manage a termbase of Catalan extension of SNOMED CT. This application serves, for example, to create/modify/delete SNOMED CT concepts in different languages, to manage subsets and its most important function is the version management module that allows files transformation from RF1 format to RF2.

In both of my internship periods, I worked in the OFTSI, the standards and interoperability office under Ariadna Rius supervision. About 60% of my time was spent on ENCODER project, my final thesis which will be presented in this document.

This project has been directed by Ariadna Rius and implemented by Pau Garcia counting on the collaboration of different TIC Salut members. The main objective of this project is to design and implement a search engine. This will serve to study and test several natural language processing algorithms to find the best way to encode sentences into ICD-10-CM/PCS, a clinical classification of diagnostics and procedures.

1.1. Fundació TIC Salut Social

Fundació TIC Salut Social is an agency of the Departament de Salut de Catalunya (Catalan Ministry of Health), which works to promote the development and use of ICTs[1] (Information and communication technologies) in the fields of health. Other functions of the company are to keep abreast of new trends and emerging

[1] *TIC* in Catalan and French

initiatives, to innovate and promote new projects and to offer an approval and accreditation to health products. It is also responsible for providing standardization of Catalan health products.

The Foundation mission is to be a facilitator of the transformation of the health and social care model through ICTs. Its vision is to be a benchmark in the health sector boosting innovation with the use of ICT as a tool for the transformation of the healthcare model. TIC Salut Social is guided by values of transparency, sustainability, commitment to the sector, global and local innovation and management autonomy among others.

1.1.1. OFTSI: Standards and Interoperability Office

I worked in the OFTSI, which is a department of Fundació TIC Salut Social. The department is responsible for managing and distributing some of the controlled vocabularies, terminologies and classifications that are used in the Catalan health system.

The OFTSI works to guarantee the different levels of interoperability between all the healthcare information systems. On the other hand, it promotes standards and defines documents that serve as references to carrying out interoperability.

1.2. Terminologies

In order to contextualize the project functionalities, in this section, the main clinical terminologies used in ENCODER will be presented. They are ICD-10, also named CIM-10 in Catalan and French and SNOMED CT which is the clinical terminology of greater breadth, precision, and importance developed until now.

1.2.1. ICD-10 (CIM-10)

According to the World Health Organization (WHO), the ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems. It is the international standard for reporting diseases and health conditions and it also serves as a diagnostic tool for epidemiology. Uses include interoperability of systems, monitoring of the incidence and study of diseases evolution and distribution.

ICD is managed by the WHO that provides all the necessary documentation and tools to use it. All of them are available online[2] and they are free access for all users. It contains codes for diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases.

1.2.2. ICD-10-CM / PCS (CIM-10-MC / SCP)

ICD-10-CM/PCS (International Classification of Diseases, 10th revision, Clinical modification/Procedures coding system) are two subsets of ICD-10 that contain the codes for clinical diagnostics (CM) and procedures (PCS).

CM refers to "Clinical Modification" and it contains all the diagnostics codified in ICD-10. On the other hand, PCS refers to "Procedure Code System" and contains all ICD-10 codes corresponding to medical procedures.

The objective of this project implies the use of this both codifications, not all the ICD-10. The expected result is a search engine that takes a natural language fragment and codifies it into either CM or PCS.

1.2.3. SNOMED CT

SNOMED CT or SNOMED Clinical Terms is a systematically organized electronic collection of medical terms providing codes, terms, synonyms, and definitions used in clinical documentation and reports.

SNOMED CT is considered one of the most comprehensive multilingual clinical terminology. The main objective of SNOMED CT is to encode most of the clinical terms used in health information systems and to support effective clinical data reporting to improve the patient care.

This terminology includes clinical outcomes, symptoms, diagnostics, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices, and specimens. SNOMED CT is maintained and distributed by SNOMED International, an international non-profit organization. SNOMED

[2] http://www.who.int/classifications/icd/en/

International is the trade name of the International Health Terminology Standards Development Organization (IHTSDO), established in 2007.

1.2.4. Natural Language Processing (NLP)

Natural Language Processing is a field of computer science, artificial intelligence and linguistics that studies the interactions between computers and human language. The PLN deals with the formulation and research of computationally effective mechanisms for communication between people and machines through natural languages.

Until the 1980s, most NLP systems were based on a complex set of rules hand-designed. However, at the end of 1980s there was a revolution in NLP with the introduction of machine learning algorithms for language processing.

The objective of NPL is to ease the human's work with the computers and allow communications between humans and computers in natural language. Natural Language Processing can be categorized in two groups: Natural Language Understanding and Natural Language Generation which evolves the task to understand and generate the text.

1.2.5. NLP categories

There are different ways to analyze a sentence through NLP in order to find its intended meaning. Therefore, there are different levels of analysis in which the system can reach. These are listed below:

The lowest level analysis is the **morphological** which objective is to split words into its roots, inflectional features, lexical units and other relevant elements. **Syntactic analysis** decomposes the syntactic structure of sentences through the grammar of the language in question.

The **semantic analysis** objective is to extract the full meaning of the sentences, by resolving lexical and structural ambiguities. On a higher level, there is p**ragmatic analysis** that identifies the text beyond the limits of the sentences, for example, to determine the referential antecedents of the pronouns. Pragmatics also includes the analysis of extralinguistic aspects such as communicative situation, knowledge shared by speakers, interpersonal relations, etc. This includes , all those factors that are not referenced in a formal communication form.

2. Framework and objectives

I have expended the most part of my time in TIC Salut Social developing the project ENCODER. Apart from this commission, I gave support to the Standards and Interoperability Office and I did some minor tasks that TIC Salut ask me to do.

In order to understand the need of ENCODER, this section will start contextualizing the law changes in Catalonia that concern to ICD-10. Then, the objectives of the project will be exposed and the last subsection will explain some of the other missions I had in the company.

2.1. Justification of ENCODER

In January 2016, by royal decree (Real Decreto 69/2015), the Ministry of Health, Consumption and Social Welfare[3] of the Spanish Government implemented and launched ICD-10 throughout the Spanish territory except for Catalonia. Due to an independent functioning of the invoice report, in Catalonia, it was applied two years later.

From January 2018, it is mandatory that all the healthcare centers report its economic activity in ICD-10 to CatSalut thus leaving ICD-9. For this reason, all the institutions of SISCAT have to change its information system in order to respond to the new requirements.

ICD-10 incorporates many new codes and important structural changes for both classifications: diagnostics (CM) and procedures (PCS). ICD-10-CM has grown from 14,025 concept codes to 71,486. This increase is due to the incorporation of

[3] *Ministerio de Sanidad, Consumo y Bienestar Social*

laterality[4] concepts, new codes for medical and surgical complications, the combination codes[5] and a greater specificity.

On the other hand, ICD-10-PCS has grown from 3,838 to 71,924 concept codes. Some of the most important changes have been the elimination of eponyms and combination codes, the clinical terminology update, the character meaningful position, a better flexibility for new extensions and a greater specificity.

Most doctors and documentarists encode quickly because they know most common codes thanks to their experience. Due to this report procedure changes, all the medical staff has to change their way of coding and learn the new encoding method.

Apart from the exposed changes, often the concept descriptions are written using too technical language which makes difficult to find a code and can provoke coding errors. The ENCODER project is born to satisfy this encoding difficulty in order to find the best way to solve the problem.

2.2. Objective of ENCODER

ENCODER aims to find an artificial intelligence algorithm that serves as the basis of an ICD-10-CM / PCS code search engine. In this way, the task of finding a code in ICD-10-CM / PCS will be simplified since the search will be done from more familiar words by doctors.

The artificial intelligence of the search engine will be able to "understand" the concept although the words introduced do not coincide directly with the description. This tool will propose a variable number of codes sorted by similarity with the user input.

This user, normally medical staff, can use those suggestions to find the best code to encoding the correspondent diagnostic or procedure. The user will be able to write the description on its preferred language because ENCODER will translate from any language to English with a language auto-detection. Furthermore, even if

[4] Indicates the part of the body side (e.g. right/left arm).

[5] It corresponds to those pathologies that need more than one code to be identified.

the user makes spelling mistakes, ENCODER incorporates an auto-corrector that will correct all sentences before encoding.

On the other hand, ENCODER matches descriptions with all the synonyms used in SNOMED CT which are most familiar for doctors, what will improve its matching rating.

2.3. Other missions in the company

Apart from ENCODER, I had other missions in the Fundació TIC Salut Social. For example, I did other tasks such as give support to some events, to improve the last year project called SCATManager, to write some articles and so on.

For example, on June 29, TIC Salut Social organized a workshop called "Bussejant entre dades" [*diving among data*] whose main topic was how artificial intelligence can help us discover and process information. In this workshop, there were several conferences about the use of natural language processing (NLP) to encode unstructured text into ICD-10/CM-PCS.

The invited companies presented some of the innovation tools they developed for semiautomatic encoding and information discovering. Finally, there was an open discussion where assistants could make questions to the speakers.

As it was much related to my final degree work, I applied to be organizer together with Ariadna Rius and communication team. I realized tasks like the workshop schedule, establishing some contacts and giving support on the day of the workshop. Apart from this workshop, I have given support to my coworkers in other ones such as "In Deep", "Hello World", "CIOs Workshop for the Trends Map" [6].

Apart from workshops organized by TIC Salut Social, I assisted in some workshops created by other companies (e.g. Interoperability in the social sector). In addition, I realized some online courses like IBM Blockchain Foundation for Developers in order to become a contact reference for the Blockchain topic in the Fundació TIC Salut Social.

[6] For more information, visit https://ticsalutsocial.cat/

The company also commissioned me to write some articles for its webpage and to give support for writing some charters for new projects taking about Blockchain. On the other hand, I have participated in the writing of a guide titled "Application development guide".

It is a good practices guide for the development of health mobile applications which gives recommendations and steps to follow from the beginning of an application project until its publication. Several companies in the sector and universities have participated and it will be published on the TIC Salut Social web in September 2018.

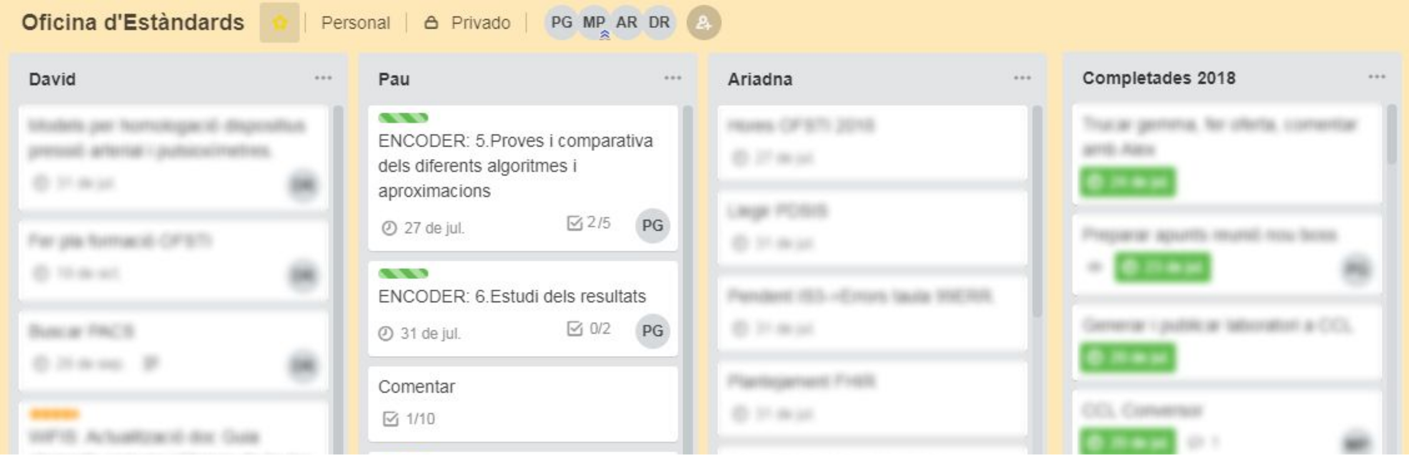
Finally, I worked to improve some functionalities of my last project in the company. Last year, in my first internship in TIC Salut Social, I developed a web application called SCATManager to manage the Catalan termbase of SNOMED CT. This year, OFTSI needed to transform all the database data into version 2 of SNOMED CT (RF2) so they ask me to fix some minor bugs of visualization and displayed information.

3. Realizations: ENCODER

3.1. Project management

The OFTSI works using Trello which is a to-do online method. On this tool, there are all tasks of the team in form of post-its. In this way, we know what our partners are working with.

For each person, Trello contains a list of to-do tasks and another with done tasks. Each one is responsible for his list and it is important to have it update to keep the team's organization.



ENCODER: 5.Proves i comparativa dels diferents algoritmes i aproximacions
27 de jul. 2/5 PG

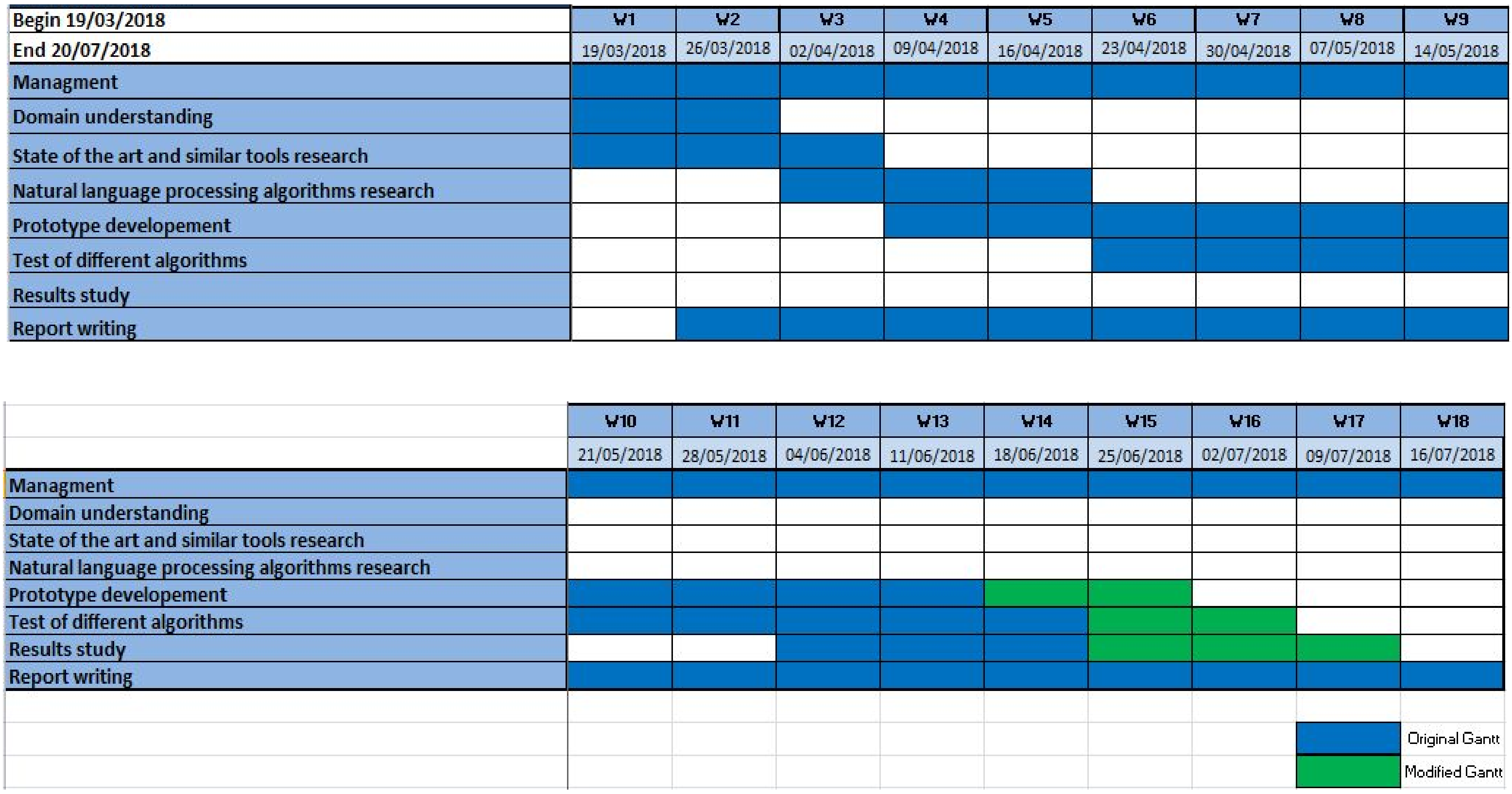
ENCODER: 6.Estudi dels resultats
31 de jul. 0/2 PG

Comentar
1/10

3.1 Screenshot of the OFTSI Trello's

On the other hand, for each project, we use a Gantt. To begin the project, it is required a Charter[7] which is the project proposal. In this document, there is defined the duration of the task of the project although it is can be modified during the project execution in function of needs.

[7] This document was presented in French by e-mail to Monsieur Soto-Romero at the beginning of the project



Begin 19/03/2018	W1	W2	W3	W4	W5	W6	W7	W8	W9
End 20/07/2018	19/03/2018	26/03/2018	02/04/2018	09/04/2018	16/04/2018	23/04/2018	30/04/2018	07/05/2018	14/05/2018
Managment	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Domain understanding	🟦	🟦							
State of the art and similar tools research	🟦	🟦	🟦						
Natural language processing algorithms research			🟦	🟦	🟦				
Prototype developement				🟦	🟦	🟦	🟦	🟦	🟦
Test of different algorithms						🟦	🟦	🟦	🟦
Results study									
Report writing		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

	W10	W11	W12	W13	W14	W15	W16	W17	W18
	21/05/2018	28/05/2018	04/06/2018	11/06/2018	18/06/2018	25/06/2018	02/07/2018	09/07/2018	16/07/2018
Managment	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Domain understanding									
State of the art and similar tools research									
Natural language processing algorithms research									
Prototype developement	🟦	🟦	🟦	🟦	🟩	🟩			
Test of different algorithms	🟦	🟦	🟦	🟦	🟦	🟩	🟩		
Results study			🟦	🟦	🟦	🟩	🟩	🟩	
Report writing	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

🟦	Original Gantt
🟩	Modified Gantt

3.2 Initial (blue) and real (green) Gantt of ENCODER

In the previous image [3.2], it is shown how initial Gantt was modified during the project ENCODER. It is divided by the main tasks of the project, and each one has assigned some period to be realized.

In total, the project has lasted 18 weeks. Green cells are modifications with respect to the initial Gantt. At week 11, we replaned tasks and we extended the time duration of the last three tasks: development, test, and study of the results. As we planed the last weeks with less workload, there were no problems when modifying the initial plan and all tasks were done finally in time.

3.2. Project organization

The team of ENCODER is essentially formed by Ariadna Rius and Pau Garcia but there were different collaborations of members of TIC Salut. Ariadna Rius is the project manager, who defines and plans different tasks and who have a global vision of the project. On the other hand, I am the main programmer analyst of the project. My task is to follow instructions from the project manager and report all the activity in order to carry out a constant feedback.

The first collaboration in the project was Anna Ceresuela that realized some tasks related to ICD-10. She is a documentarist student, with knowledge about medical terminologies, so she helped the team with the ICD-10 structure understanding, providing some of the documents used to import the terminology in the database and so on.

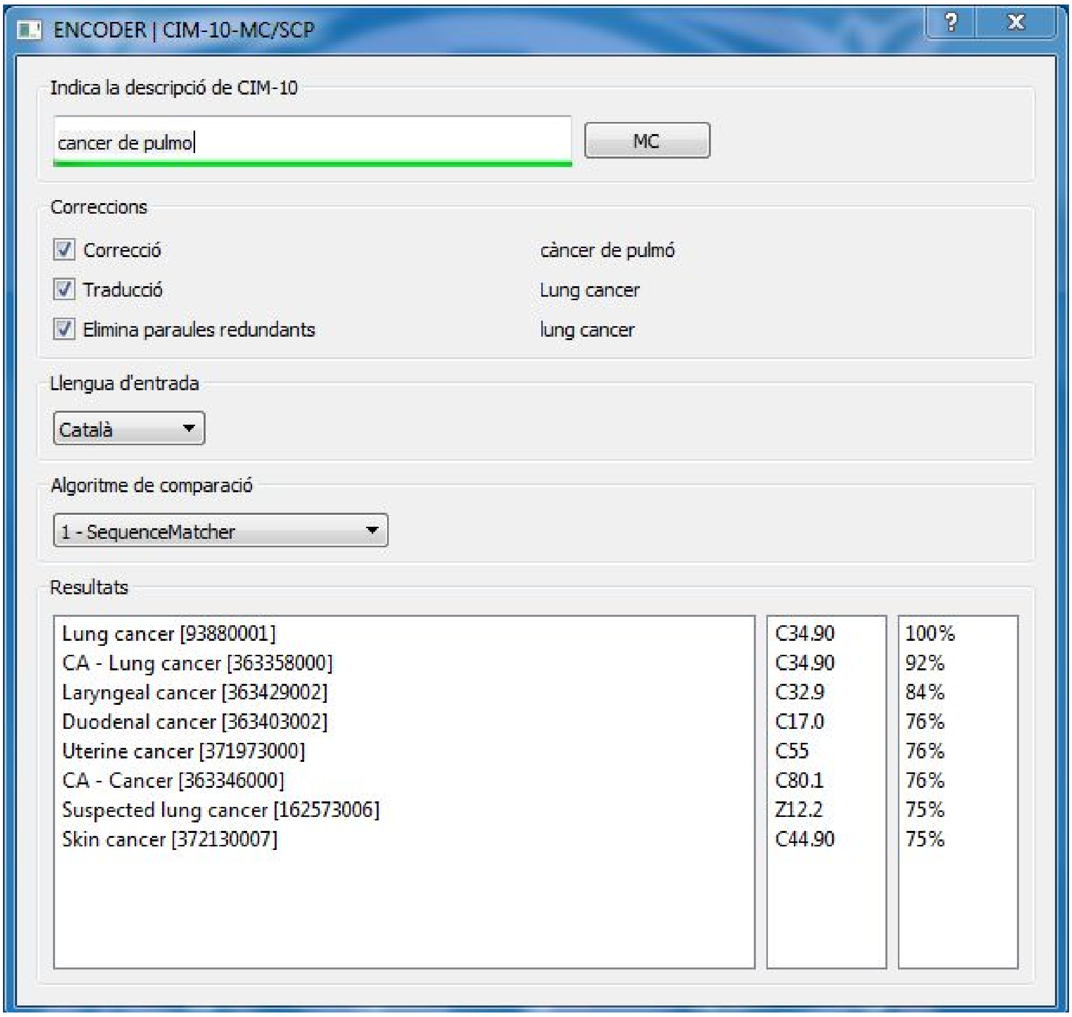
Then, Miquel Martí gave support in the user interface design. He studied videogames design, so he has experience in designing ergonomic and useful interfaces. Finally, Jordi Martinez helped the team writing some diagnostics that we used to test ENCODER. He is the innovation director of TIC Salut and family doctor, so he could write some of most typical diagnostics in medicine. These diagnostics were one of the most important points to test because they were written in real natural language by a doctor.

3.3. Project realization

As a programmer analyst, I did most of the programming part of ENCODER. The result is a python multiplatform program (Windows and Linux) that helps the doctors on the ICD-10 codification task.

Its utilization is pretty simple: the user has to input some ICD-10-CM/PCS code and press a button. After few seconds, ENCODER will search in SNOMED CT and ICD-10 database in order to find the best matches with the input description.

When the search ends, the tool displays a list of results (by default 8) ordered by a matching ratio with the original sentence. Each result is an ICD-10-CM code and for each one, it is shown its SNOMED CT description and ICD-10 description.



ENCODER | CIM-10-MC/SCP

Indica la descripció de CIM-10

cancer de pulmo

Correccions

☑ Correcció — càncer de pulmó

☑ Traducció — Lung cancer

☑ Elimina paraules redundants — lung cancer

Llengua d'entrada: Català

Algoritme de comparació: 1 - SequenceMatcher

Resultats

Lung cancer [93880001]	C34.90	100%
CA - Lung cancer [363358000]	C34.90	92%
Laryngeal cancer [363429002]	C32.9	84%
Duodenal cancer [363403002]	C17.0	76%
Uterine cancer [371973000]	C55	76%
CA - Cancer [363346000]	C80.1	76%
Suspected lung cancer [162573006]	Z12.2	75%
Skin cancer [372130007]	C44.90	75%

3.3 Screenshot of ENCODER

The first task was to design the NLP algorithm. Ariadna has a master of artificial intelligence, so the team had periodical meetings to define the best theoretical way of process natural language descriptions written by doctors. In the first meetings, Python was selected as a development language because of its flexibility, simplicity, and speed.

Thanks to her knowledge in artificial intelligence, Ariadna proposed the best way to treat a natural language sentence, proposing the spell checking, translation, truncating, and stopwords treatment. After some researches, we selected the sentences comparison algorithm and she designed how the flow of information should be in ENCODER.

After these meetings, we draw a schema of how the program should process the input text. Then, with a collaboration of another member of the OFTSI (Miquel Martí), we designed the user interface. He has a degree in video games, so he could help in this task giving ideas of how to make a UI as clear and simple as possible.

At the beginning of the development, I need to find the ICD-10 documents in order to fill the database. At this point, the team had the intervention of Anna Ceresuela. She is a documentarist student that knows well several clinical terminologies, so she was very helpful.

In order to fill the databases with the SNOMED CT elements, we used the data from the work of last year's internship (SCATManager). It was helpful because there are the Catalan, Spanish and international versions in MongoDB format, so it was pretty easy to import to the new project that also uses MongoDB.

In parallel, I planed the development part choosing the development tools and researching the appropriate Python libraries to implement the artificial intelligence part. Apart from this, I chose MongoDB as a database because of its flexibility, speed, and good performance.

Apart from this, I designed the database structure, the Python classes of the program ENCODER and the internal dataflow. However, I always had the help from Ariadna Rius who validated all the decisions and proposed several improvements to the design. Once all these items were defined, we started the development part, where I take charge of all the software implementation having periodical meetings to validate the work.

At the beginning of the project, we realized that there was not an implemented map of ICD-10-PCS to SNOMED CT. This fact difficult a lot the task of encoding because the map is an essential part to obtain the right results. So that, we decided to simplify the objectives and work only the CM part.

3.4. Technical choices and tools

The program has been developed using Python[8] and its main libraries such as sys, os, threading, HTML, JSON and so on. On the other hand, the database of the

[8] https://www.python.org/

program has been programmed using MongoDB[9] technologies and Studio 3T[10] as a database manager.

Another Python library used to implement the user interface part has been QT, specifically PyQt5. To implement the spell checking we have used Language Tool, with the corrector of SoftCatala for the Catalan. For the translation, ENCODER uses the Google Cloud technologies via RESTFUL APIs.

Other libraries used have been NLTK (Natural language toolkit), SKLearn, DiffLib, TextBolb and Jellyfish to make all the natural language processing. This includes the truncating the sentences matching.

3.5. Main flowchart

Once ENCODER starts, it waits until user inputs some text in the first field and he clicks to the search button.

When the button is pressed, the program launches a thread [**(1)** *see the image 3.4*] which will take charge of all the calculations to obtain the results. The main thread is responsible for controlling the user interface. This division into threads is used to prevent the program from hanging.

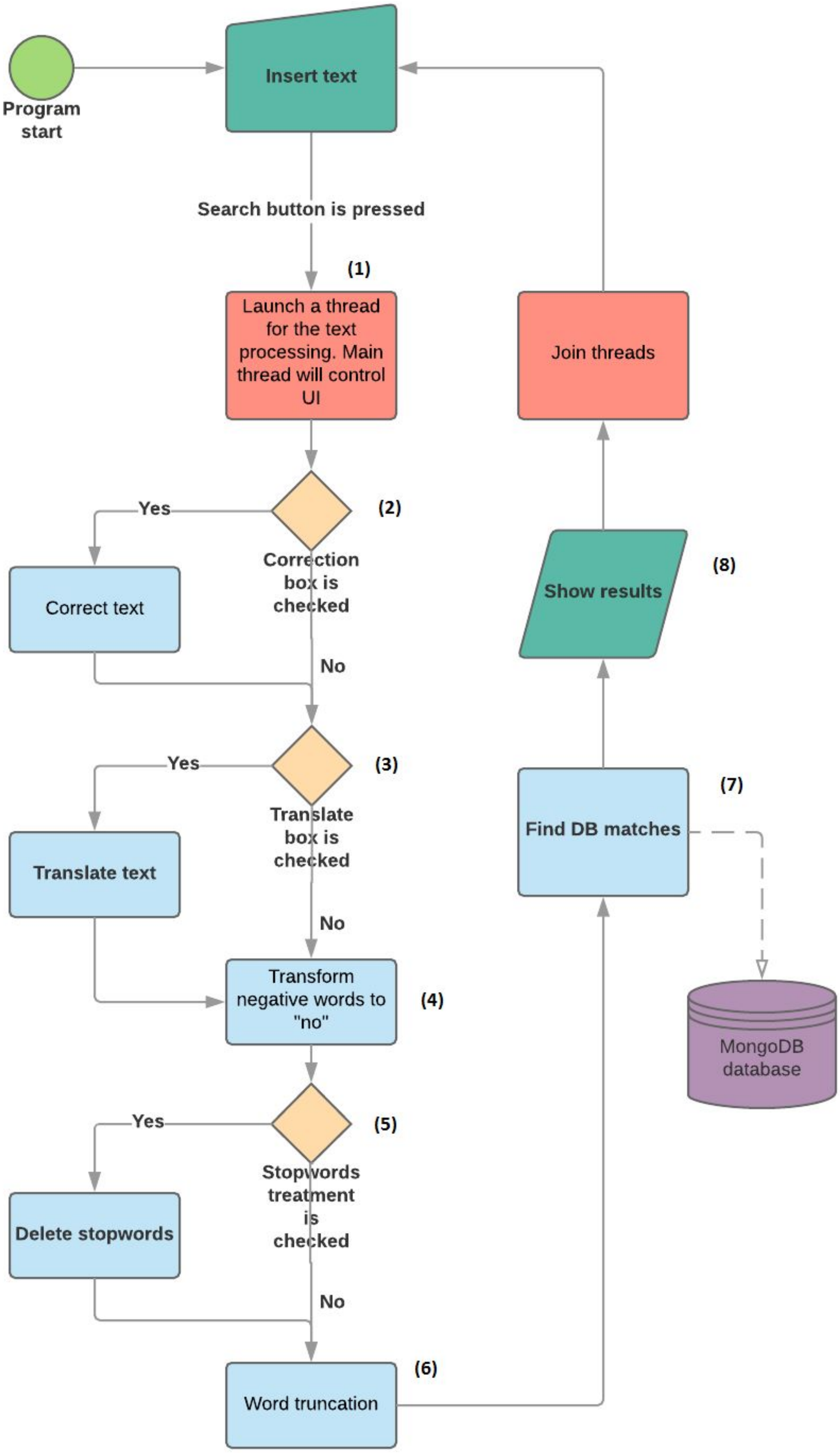
The calculations thread starts verifying if checkboxes are marked. If the correction box is checked **(2)**, ENCODER will correct spelling of the input text using the LanguageTool API provided by a web service.

Next, it checks if it has to translate the text into English **(3)**. If so, it will use the Google Cloud Translation web service to do this task. The following step is to transform all negative words into 'no' **(4)**. This modification is not configurable by the user and it has to be always carried out.

Then, the program will delete all stopwords if the user has checked the corresponding box **(5)**. The last step of text processing is word truncation whereby all sentences have to pass **(6)**.

[9] https://www.mongodb.com/

[10] https://studio3t.com/



3.4 Main flowchart

Once the text is processed, it is obtained a simplified sentence in English that strictly contains the meaning of the sentence without language complements. The following step is to find SNOMED CT descriptions matching with the processed sentence **(7)**. In this step, the user can choose which of the 5 available algorithms of sentence matching will be used.

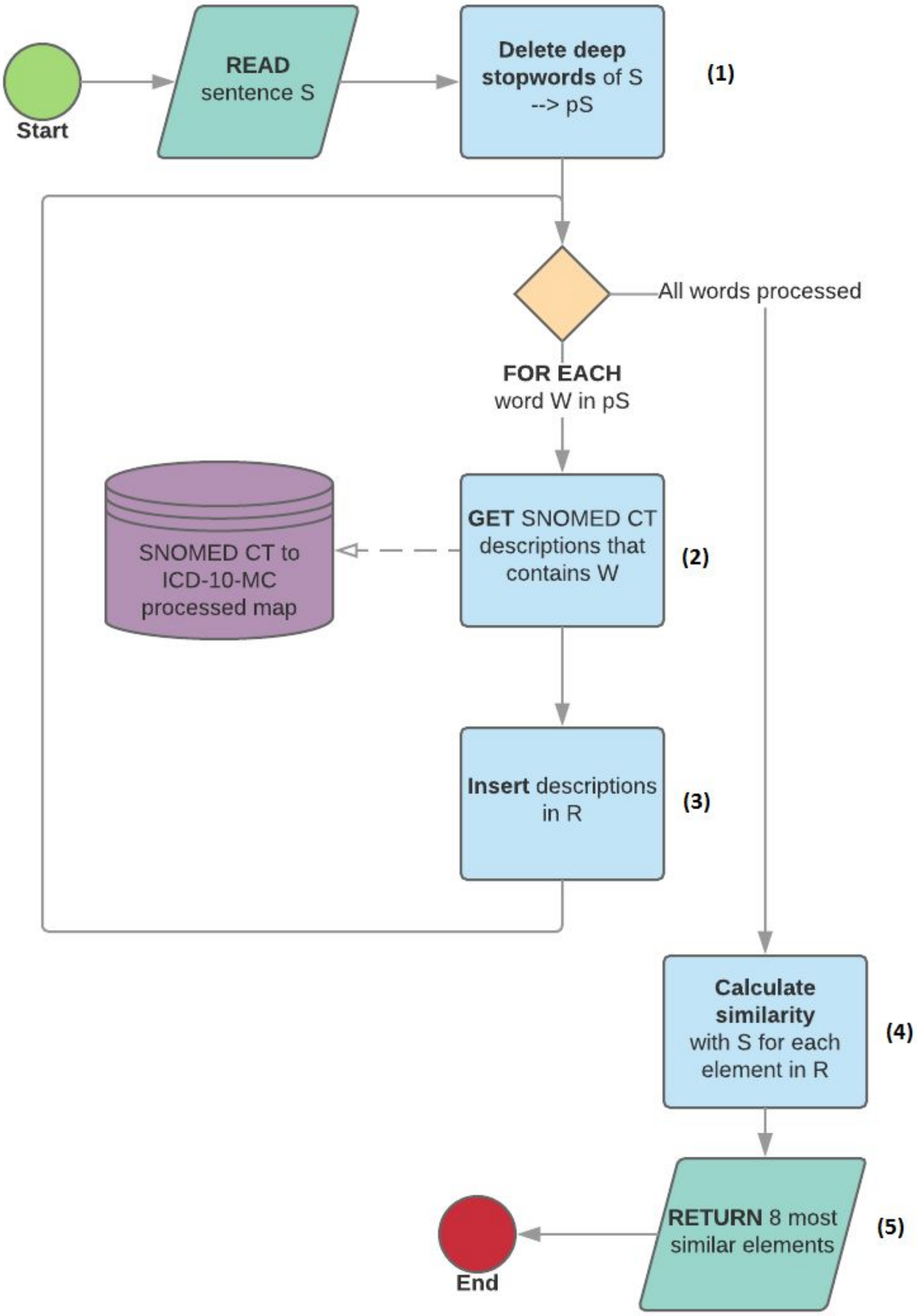
To conclude, the application shows the most similar matchings in the results table (by default the first 8 are displayed) **(8)** and the calculations thread joins the main thread. Then, the user can write another diagnostic and restart the process.

3.6. Find DB matches flowchart

Below there is a zoom of the flowchart of sentence matching process. It is the step **(7)** in the previous flowchart (*3.5 main flowchart*). This process is carried out once sentence has been transformed into a simpler one.

In the database, there is a version of English SNOMED CT descriptions that has passed the same treatment. So they are descriptions without stopwords, with the negative words transformed to 'no' and truncated.

As sentences has passed the same treatment, is easier to compare and check if the meaning is the same. Below there is the flowchart passed by each sentence processed by ENCODER.


Start
READ sentence S
Delete deep stopwords of S --> pS
(1)
All words processed
FOR EACH word W in pS
SNOMED CT to ICD-10-MC processed map
GET SNOMED CT descriptions that contains W
(2)
Insert descriptions in R
(3)
Calculate similarity with S for each element in R
(4)
RETURN 8 most similar elements
(5)
End

3.5. Flowchart of getting a matching description from database

Let **S** be the input processed sentence. For each word of S, we need to take all descriptions in the database containing this word.

If this is done, we are going to obtain a lot of insignificant results although we have deleted all stopword. For example, words like 'no', 'normal', 'body', 'product' are important words for the meaning of a sentence, but we do not need to take each description of the database that contains these ones, because them are not significant enough.

So that, we need to delete what we called *deeep stopwords* defined as words that need context to have a determining meaning [**(1)** *3.5. flowchart*]. Once this process is done, we obtain the final processed sentence **pS**. Now, for each word of this sentence, we get all descriptions in the database containing this word **(2)**. We insert **(3)** all the results into a set of results named **R**.

Once all words of pS are treated, it is calculated the similarity[11] between the original sentence S with each result in R **(4)**. Finally, it orders the results and most similar elements[12] in R are returned **(5)**.

This flowchart is a simplification of the original one. In order to increase speed of ENCODER, it takes just the words of pS that produces less results. We can ignore the words that produce more than certain value because they does not delimitate the results set what means that the word is meaningless.

There are a variable called *resultsLimit*[13] which determines the maximum of results that a word produce to take it. If the number of results is greater than this value, we ignore this word (we can say it is a deep stopword). However, if this process does not produce any result with any word, we increase the *resultsLimit* by multiplying it by another value called *correctionInc* and we repeat the process until we obtain some results.

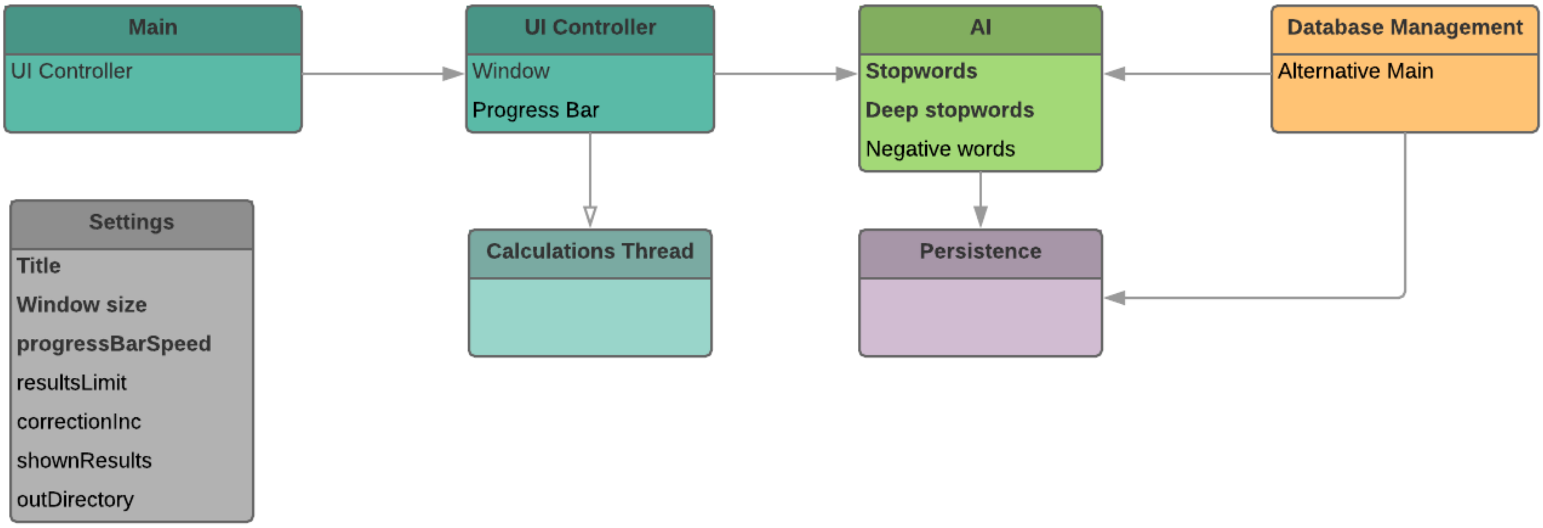
3.7. Class diagram

The source code of ENCODER is divided into seven classes related among them as shown in the class diagram below. ENCODER has two main functions: the first one is found in the Main class and the second one is in Database Management class.

[11] This algorithm is described in section *5.6*

[12] The number of results returned can be modified in *Settings.py* by modifying variable *ShownResults*

[13] This variable is in *Settings.py*



3.6. Class diagram of ENCODER

The class **Database Management** contains an alternative main and three useful functions. One of these functions produces the Processed CM map found in the database.

The other two functions of DataBase Management are used to find the "holes" in the original map. One of these functions checks which SNOMED CT codes have not a correspondence in ICD-10 and vice versa.

On the other hand, the normal execution of the program is the **Main class**. The functions of this class are to create the user interface of ENCODER. Then, the **UI Controller** class will take charge to control it.

The latter creates an instance of **Calculations Thread** that realizes all the hard work of the program (to process text and find matches). This thread calls **AI** class to process the text and this calls **Persistence** in order to obtain the similar descriptions.

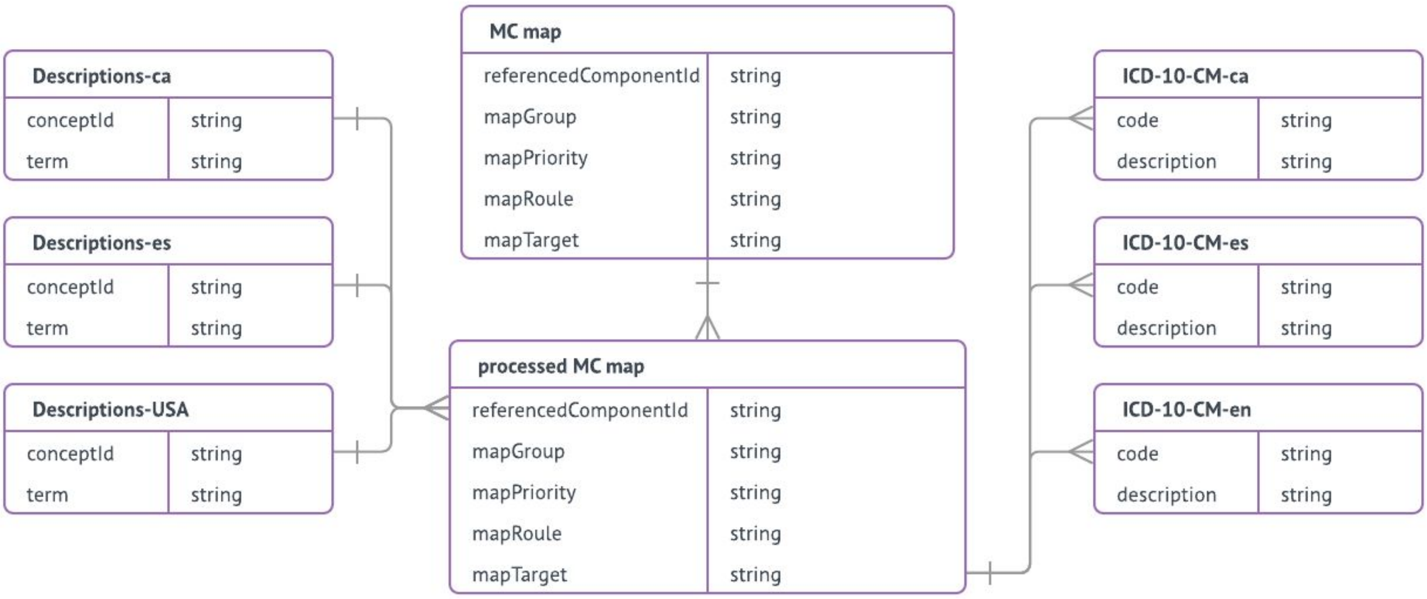
Finally, there is a **Settings** class which contains some variables that can be modified in order to change some parameters of the program such as the number of results shown.

3.8. Database

The database works under MongoDB technology, which is a document-oriented NoSQL database system, developed under the concept of open source. Documents are written codified using JSON format.

The database has 8 collections grouped in 3 categories configured as shown below:

SNOMED CT **Maps** **ICD-10**



3.7. Diagram of ENCODER database structure

The first 3 collections (the first column in the previous diagram) called Descriptions-ca, Descriptions-es, and Descriptions-USA belongs to the first group called SNOMED CT explained in section 3.7.1. The collections in the second column called CM map and processed CM map belongs to group SNOMED CT to ICD-10-CM map and it is described in section 3.7.2. Finally, the last 3 collections belong to the ICD-10-CM group and they are described in section 3.7.3.

3.7.1. SNOMED CT

The classes Descriptions-ca, Descriptions-es and Descriptions-USA contains SNOMED CT descriptions for languages Catalan, Spanish, and English.

Descriptions-ca[14] has the descriptions of Catalan SNOMED CT extension which contains descriptions in Catalan and Spanish.

Descriptions-es[15] has the descriptions of Spanish SNOMED CT extension which contains descriptions in Spanish.

Descriptions-USA[16] has the descriptions of American SNOMED CT extension written in English. As the mapping file uses codes of USA extension, we have had to use the American version instead of the international one.

3.7.2. ICD-10-CM

The classes ICD-10-CM-ca[17], ICD-10-CM-es[18], ICD-10-CM-en[19] contain all ICD-10-CM codes and its description in Catalan, Spanish and English respectively.

3.7.3. SNOMED CT to ICD-10-CM map

The original map[20] (called CM map in the database) does not directly link a SNOMED CT concept to an ICD-10-CM concept. It is constructed by using map rules, which indicates how to get the target code by asking the user at runtime.

For this reason, before use this map, we have had to transform this map in order to make a direct mapping. In a few words, we have processed all the map rules so as to obtain a target code for each source code.

[14] Data extracted from TIC Salut Social

[15] Data extracted from http://browser.ihtsdotools.org/

[16] Data extracted from: http://browser.ihtsdotools.org/

[17] Data extracted from: http://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/diagnostics-procediments/cim-10-mc-scp/

[18] Data extracted from: https://eciemaps.msssi.gob.es/ecieMaps/documentation/documentation.html

[19] Data extracted from: http://www.who.int/classifications/icd/en/

[20] File extracted from: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html

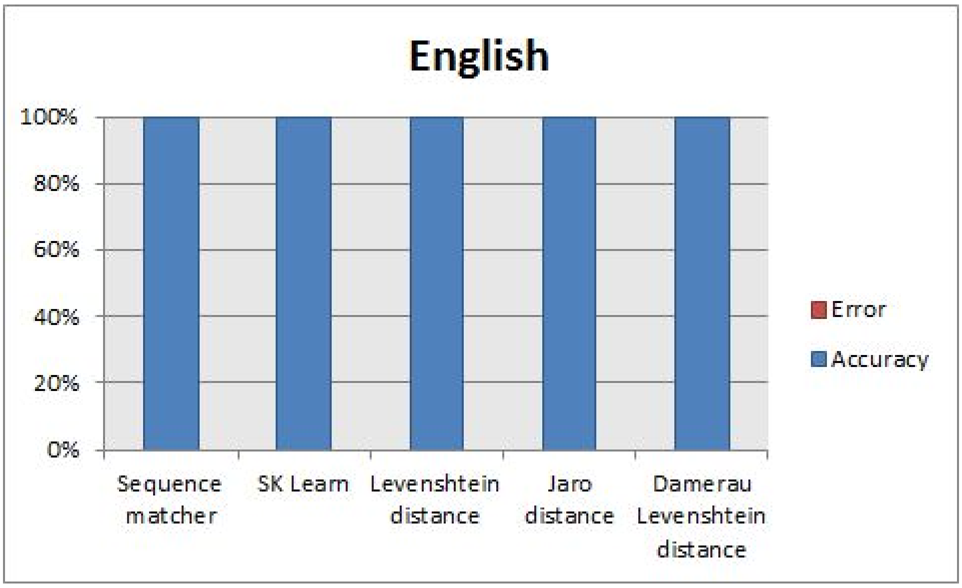
In conclusion, the processed map contains an ICD-10 code for each SNOMED CT code. Apart from this, in order to make ENCODER faster, we have processed all the SNOMED CT descriptions of the American version. We have replaced all negative words by "no", we have truncated the suffix of the words and we have deleted stopwords from the sentences which will simplify the task of comparing one sentence to the entire database.

3.9. Results

In order to test the accuracy of ENCODER, we have realized several tests in different languages using all the available algorithms. In this way, the results show the precision of encoding into ICD-10-CM.

For language English and Spanish, the test kit is a random subset of 1000 SNOMED CT descriptions. However, the SNOMED CT Catalan extension contains 130 diagnoses, the available ones in the Catalan extension.

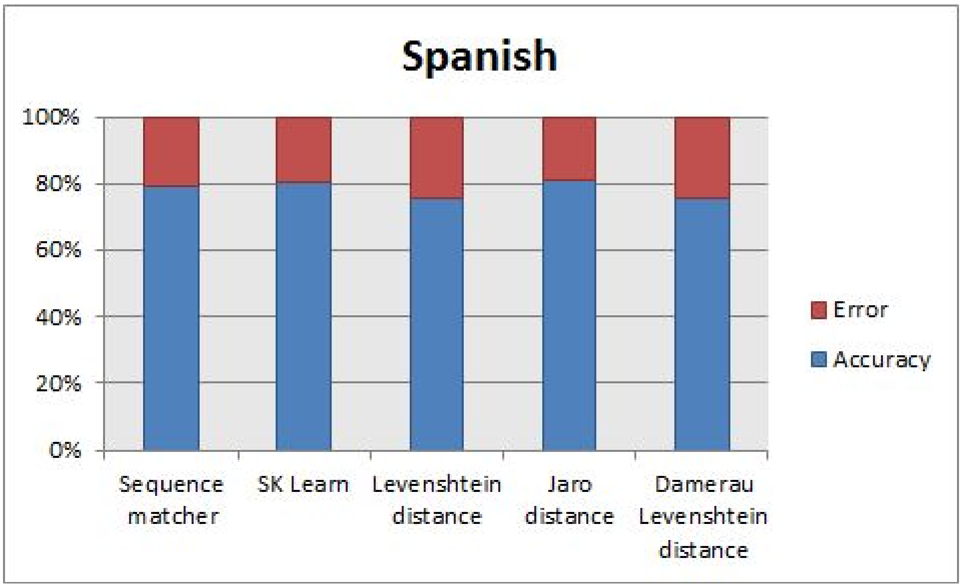
Apart from these three tests, we have asked Dr. Jordi Martinez to write 100 diagnostics. We call this the test of Natural Language because this kit was written by a doctor so it contains typical errors like typos, abbreviations, usage of Catalan and Spanish in the same sentence, etc.



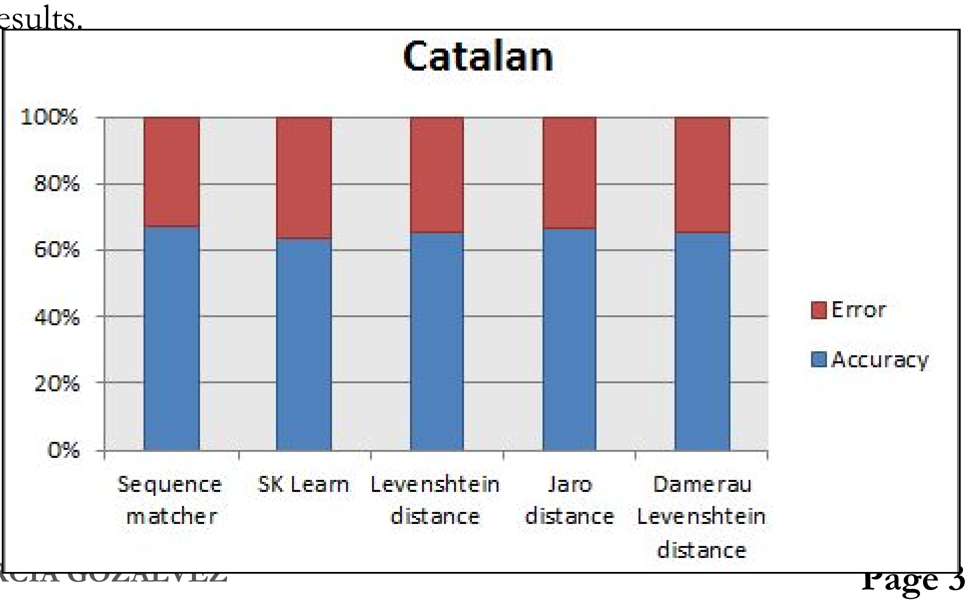
The diagram below shows results obtained using 1000 random English SNOMED CT descriptions. It shows a match for each description, so the accuracy is always

100% what makes an error of 0%. This is due to the usage of SNOMED CT descriptions. The algorithm finds always the correct description because it searches the original description in the SNOMED CT English database, so it matches always.

It is a dummy test because we knew that the result should be of 100% matching. However, we considered necessary to test the good performance of ENCODER. Seen this results, we can accept the correctness of the software.



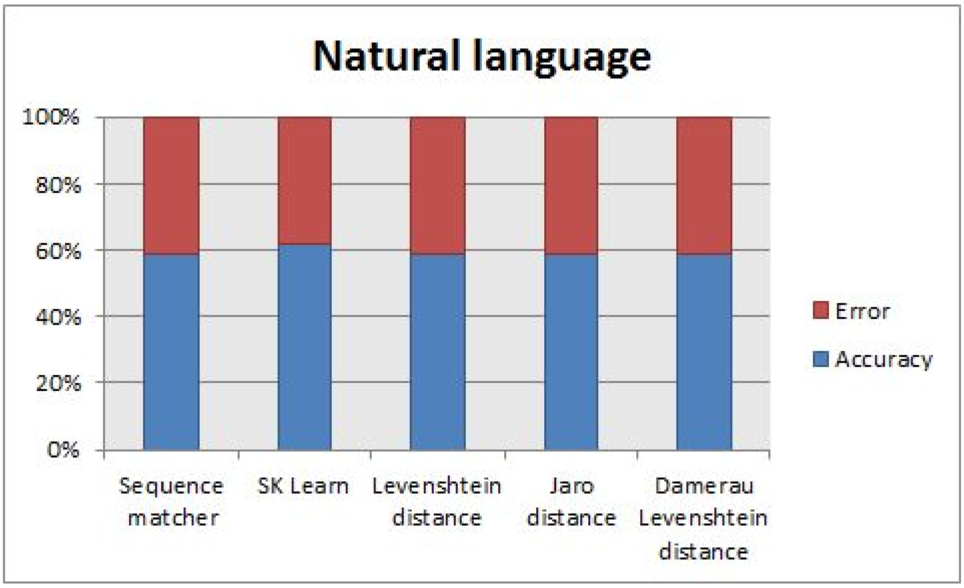
In this test, there are the results for each algorithm using 1000 random SNOMED CT Spanish descriptions as input. We can extract two ideas: the first one is that for each algorithm, the result has been always about 80%, what signifies that ENCODER well-encodes eighth of every ten descriptions. The second thing is that the difference between algorithms is almost negligible and all them gives similar results.



For the Catalan tests, we have obtained worse results than the Spanish ones with all the algorithms. ENCODER applies the same transformations for both inputs, so the only thing that changes is source language and its translation.

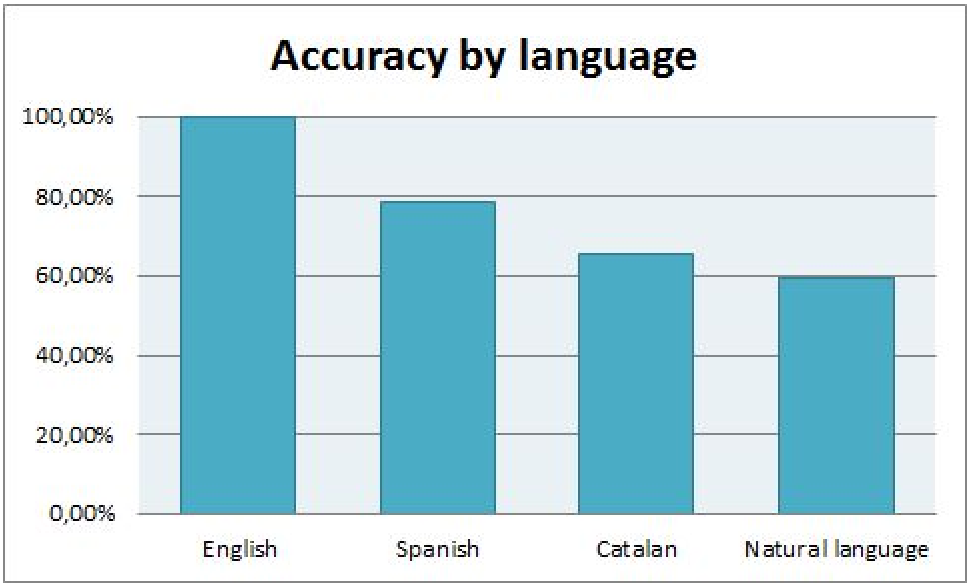
So for this reason, probably these results may be due to a worse translation provided by Google cloud translation. For experience, we know that the Catalan translation is not as accurate as the Spanish one.

This may be because Google uses a machine learning algorithm to translate sentences, so it depends on the number of speakers of the source language. The minor number of Catalan speakers can explain that the algorithm has not been able to “learn” as well as it has done with Spanish.



Finally, in this plot, it is shown results obtained by matching the descriptions written by a real doctor. Using a real natural language, we have obtained a minor accuracy, with all algorithms about 60%. However, they are not so different from the Catalan ones, on average them differ in 5%.

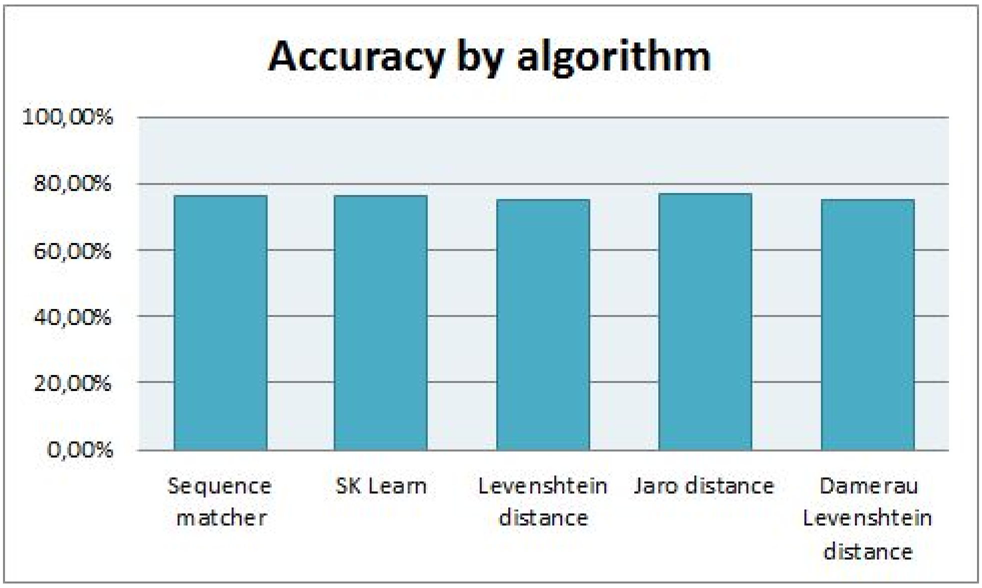
Most of these descriptions were written using informal abbreviations that cannot be matched with a SNOMED CT description and even so results are not so different from the Catalan test. On the other hand, again we see that there is not a significant difference between the results provided by the different algorithms.



In this plot, it is shown the difference between the averages of accuracy for each language. As we see, English is always 100% followed by the Spanish test kit. Finally, in the last group, probably due to a worse translation there are the Catalan and Natural Language test kits.

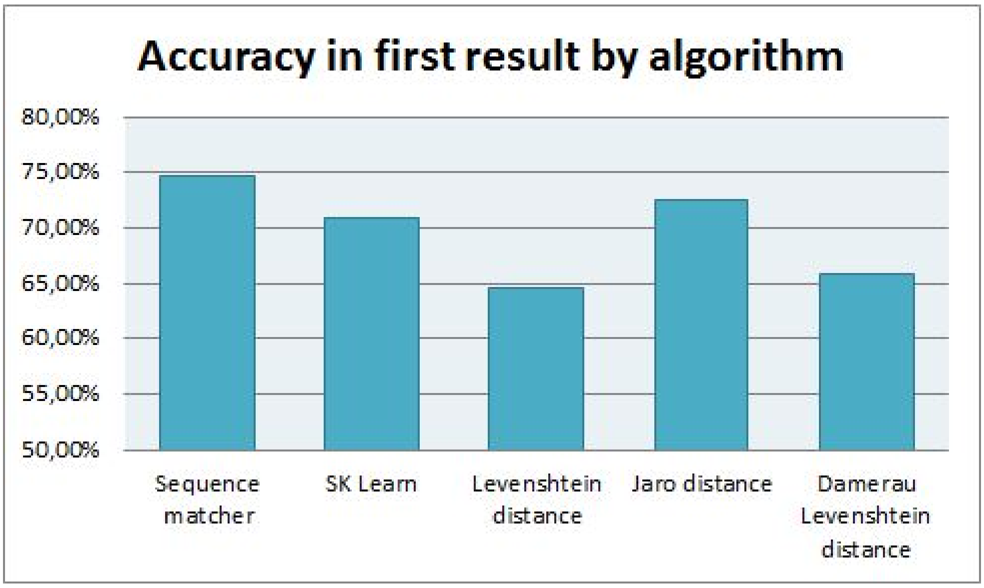
The important point here is to see that the difference between Natural Language results and the Catalan correct ones is about 5%. This means that ENOCER has a well-performance with natural language in comparison to well-written descriptions.

On the other hand, there is a difference of 15% between the Spanish and Catalan descriptions. As it was exposed, considering that the same operations were applied to treat the text, we consider that the problem is found in the translation. So maybe, with a better translation of the Catalan, we could found better results, up to 80%.



Finally, in this plot, there is the difference of accuracy between algorithms. As we see, there is not so much difference between them. Once a major set of descriptions it is provided to the algorithm, the task of this one is to select the 8 most sentences to the original one by using different methods.

From these results, it can be proved that the five algorithms are calculating the similarity pretty well and they differ just in concrete cases.



In order to test the accuracy more meticulously, we have tested which algorithms had provided the solution in the first position. All the other tests were done considering that they had a match if the solution were contained in the first 8 results.

In this plot, it can be appreciated how the algorithms provide the result in the first position. As the results show, the best results are given by Sequence matcher with more than 74% accuracy. This means that the 97% of times that that algorithm shows a correct result, it is found in the first position.

Contrarily, Levenshtein distance shows just the 64% of the right results in the first position. This means that the 86% times it shows the correct answer, it is found in the first position. Even so, we have to take into account that all these algorithms contained the right result among the 8 first positions in almost 80% times.

3.10. Problems found

In general, I have not found organizational problems in the company. I suppose the accommodation was easy for me because I already had worked in TIC Salut in my last internship. Nevertheless, our team found some technical problems during the project development. Our biggest problem probably was to not have the map between SNOMED CT and ICD-PCS.

Due to this fact, we have not been able to develop a program capable of encoding procedures. However, we reached to develop a tool that codifies, with good results, all the diagnostics. Furthermore, it can be easily adapted to codify procedures if ever somebody develops the map SNOMED CT/ICD-10-PCS.

Apart from these problems, we have found some "holes" in the ICD-10-CM map used in ENCODER. We have studied this map, and there are some elements that do not have a link to the other terminology. For instance, taking all the diagnostics codified in SNOMED CT, we found that only the 76% have an entry in the map that codifies this diagnostic into ICD-10-CM. That means that if we take some SNOMED CT diagnostic outside of this 76%, the software will not be able to encode it into the target terminology.

Otherwise, checking the ICD-10-CM elements, we found that 50% of the elements in this terminology have not an entry in the map. However, we have to consider that the map has a direction: it allows to transform a SNOMED CT code to an

ICD-10-CM one. For this reason, this lack of elements is not as important as the first one because the map is just used in the other direction.

Moreover, this is probably due to the laterality and similar attributes of the codes in the ICD-10. This means that for each diagnostic, there exist different codes that specify the location, the gravity of the diagnostic, the part of the body and so on. For example, there exist two different codes for "*Burn of third degree of* ***right*** *palm*" and "*Burn of third degree of* ***left*** *palm*". In SNOMED CT, this will be encoded to the same code, using another code that codifies just *right* or *left.*

However, we have considered that these difficulties were alien to us, so we do not have the way to solve it because it depends on the terminology form. Furthermore, it does not affect the ENCODER performance because it will be proposed, for example, both codes.

Apart from these problems, we have found some technical problems that have been solved by modifying a fragment of the algorithm or by changing some part of the source code. All these problems were not of great magnitude and it can be easily solved in the periodical meetings with the team.

4. Conclusions and perspectives

4.1. Future work

This tool is now a Python program that needs a specific compilation to be executed. A perspective to ENCODER could be to transform this software into a web-application. In this way, it would be executed on any platform independently of the operating system, facilitating its use.

Moreover, the natural language processing could be improved by applying some different techniques. For instance, an improvement to the algorithm could be a more advanced detection of negations, detecting what negations are denying. Besides, it could be done a treatment of abbreviations and synonyms in order to approach the results to the reality since doctors usually make use of them.

Another improvement could be to study a way to ameliorate the translations of Catalan to get better results. However, this could be difficult as it depends on an external service. In addition, if ever ICD-10-PCS is created, the ENCODER project can be extended to the procedures encode.

Finally, maybe the most important future work to this project can be to study the way of adapting ENCODER to a workstation of doctors and documentarists. It requires some work, as for example to test its correctness more carefully to make sure that the codes provided are strictly always correct. It is important to make this work because it is a health-related tool, so it has to pass a lot of tests after being launched in the health centers.

4.2. Conclusion

As a personal conclusion, this has been a great internship because I have learned a lot of new things about health and computing. On the one hand, I have gained experience on the Python programming and the use its libraries. Moreover, I have learned a lot about artificial intelligence: how to treat the natural language to be able to compare the meaning of two sentences, different algorithms of comparison and so on. Apart from this, I have improved my knowledge in MongoDB which I think that will be very useful for future projects due to its No-SQL paradigm.

On the other hand, I have learned how the two most important health terminologies work internally. I had some knowledge about SNOMED CT thanks to my previous internship, but these 6 months have been very useful to go deeper into this field. Apart from this, I have learned the encoding way of ICD-10 and the manner to pass from one terminology to another using a map.

Apart from the knowledge that I have acquired from doing this project, I have been able to do some tasks such as organizing events, write several articles and know people in the health sector of Catalonia. I have also learned about Blockchain, FHIR, Gantt, and Trello thanks to different minor projects done in the company.

With regard to ENCODER, we have reached the initial objective: to be a tool of support to encode natural language into ICD-10-CM. The objective of encode ICD-10-PCS could not be done due to the inexistence of a map from SNOMED CT to this terminology as it was exposed previously. However, it is easy to implement this new functionality once some institution develops the convenient map.

As the tests have asserted, ENCODER has codified correctly the 80% of the Spanish SNOMED CT diagnostics. Regarding the Catalan ones, it has codified about 65% using the same algorithm. So for this reason, we conclude that this fact is due to a wrong translation provided by Google Cloud translating services.

As it was exposed previously, Catalan is a language with fewer speakers than Spanish. So that, the machine learning algorithms used by Google to translate Catalan is less effective because it has fewer cases to learn from. As a result, the translation is worst coming from Catalan than from Spanish, as it is well-known.

On the other hand, the codifications of real natural language in Catalan (and some in Spanish) have reached the 60% of accuracy. It just differs from 5% of the Catalan results, so it is a good result because it means that it can be useful for doctors to encode diagnostics written in their usual way.

If we regard the results of the algorithms, we have not found a big difference of accuracy. All the five have provided the same average of correct descriptions. Nevertheless, watching it in more detail, it can be appreciated that Sequence matching, SK Learn, and Jaro distance provides more times the correct solution in the first place than Levenshtein distance and Damerau L. distance.

This makes not an important difference because it means that if one of the algorithms provides the good result, probably the other four, will also. However, for a doctor can be useful to select, e.g. Sequence matching, because it will display the correct codification in a higher position than the other ones, so it will be easy to find it.

As a global conclusion, I am satisfied with the work done and I think that it can be a useful tool that the Fundació TIC Salut Social will be able to use in future projects. Moreover, in the future, with some more correctness tests, this tool can be used by medical staff that needs to encode each day a lot of diagnostics into ICD-10.

As it was exposed in the top of this document, the change between ICD-9 and 10 is currently being made. So that, doctors and health personal does not know the

ICD-10 codes, so this software can be helpful as a daily workplace tool. Even more, if some institution develops the map between the whole ICD-10, this tool may be a complete program to encode any diagnostic, procedure and so on into this terminology.

5. Bibliography

[1] Suzanne PEREIRA, Aurélie NÉVÉOL, Philippe MASSARI, Michel JOUBERT and Stefan DARMONI. Construction of a semi-automated ICD10 coding help system to optimize medical and economic coding. Technologies for Better Health in Aging Societies A. Hasman et al. (Eds.) IOS Press, 2006.

[2] P. Franz, A. Zaiss, S. Schulz, U. Hahn, and R. Klar. Automated coding of diagnoses--three methods compared. Proc AMIA Symp. 2000: 250–254.

[3] Sue Bowman. Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems. Perspectives in Health Information Management Spring 2005 (May 25, 2005).

[4] Patrick RuchEmail, Julien Gobeill, Christian Lovis, and Antoine Geissbühler. Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Decision Making20088(Suppl 1): S6.

[5] Juan Antonio Goicoechea Salazar, María-Adoración Nieto-García, Antonio Laguna Téllez, Vicente David Canto Casasola, Juliana Rodríguez Herrera, Francisco Murillo Cabezas. Desarrollo de un sistema de codificación automática para recuperar y analizar textos diagnósticos de los registros de servicios de urgencias hospitalarios. Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias, ISSN 1137-6821, Vol. 25, Nº. 6 (Diciembre), 2013, págs. 430-436.

[6] S. Nitsuwat, W. Paoin. Development of ICD-10-TM Ontology for a Semi-automated Morbidity Coding System in Thailand. Methods Inf Med 2012; 51(06): 519-528.

[7] Pestana Delgado Roberto, Llanos Zavalaga Luis Fernando, Cabello Morales Emilio Andrés, Lecca García Leonid. Concordance between medical diagnosis and informatics coding, considering ICD 10, at the Hospital Nacional Cayetano Heredia, Lima, Peru. Rev Med Hered v.16 n.4 Lima oct./dic. 2005.

[8] Diccionari Clínic project. Funsació TIC Salut Social.
http://www.ticsalut.cat/estandards/terminologia/diccionari-clinic/

[9] SNOMED International webpage. http://www.snomed.org/

[10] SNOMED CT Browser. http://browser.ihtsdotools.org/

[11] SNOMED CT documentation.
.https://confluence.ihtsdotools.org/display/DOC

[12] MongoDB official webpage where there are information and courses.
https://www.mongodb.com/webinars

[13] World Health Organization webpage http://www.who.int/es

[14] WHO ICD-10 release and documentation
http://www.who.int/classifications/icd/icdonlineversions/en/

[15] ICD-10 in Catalonia (CIM-10)
http://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/diagnostics-procediments/cim-10/

[16] ICD-10-MC/PCS in Catalonia (CIM-10-CM/SCP)
http://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/diagnostics-procediments/cim-10-mc-scp/

ENCODER

Automatic encoding of natural language into ICD-10-CM / PCS

5th-year internship rapport

February 26, 2018 – July 27, 2018

Fundació TIC Salut Social

Parc TecnoCampus Mataró Maresme - Torre TCM3

Av. Ernest Lluch, 32, 6a planta | 08302 Mataró

Student	Internship supervisor	University supervisor
Pau Garcia Gozàlvez	Ariadna Rius Soler	Georges Soto-Romero
Promotion 2018	Head of the Standards and Interoperability Office	Director of École d'ingenieurs ISIS

Company informations

Company name	Fundació TIC Salut Social
Web site	www.ticsalutsocial.cat
Contact mail	info@ticsalutsocial.cat
Contact phone	+34 93 553 26 42

Organism of	Departament de Salut, Generalitat de Catalunya
Address	Parc TecnoCampus Mataró Maresme - Torre TCM3 Av. Ernest Lluch, 32, 6a planta | 08302 Mataró

Activity area	Innovation and development of new technologies in the health sector in the Catalan hospital network
Foundation date	September 19, 2006
Number of employees	About 30

Logo:

Professional contacts

Ariadna Rius Soler	**Jordi Martinez**	**Francesc Garcia Cuyas**
Head of the Standards and Interoperability Office.	Director of Innovation in TIC Salut Social	Director of TIC Salut Social
arius@ticsalutsocial.cat	jmartinez@ticsalutsocial.cat	fgarciacuyas@ticsalutsocial.cat

Abstract

This project has been realized in the Fundació TIC Salut Social, an organization of the Health Department of Catalonia. The main objective is to study and test several natural language processing algorithms to find the best way to encode sentences into ICD-10-CM/PCS, a clinical classification of diagnostics and procedures.

The need for this project is due to the recent change from ICD-9 to ICD-10 done by the Spanish Government. For this reason, from now medical staff needs to encode into ICD-10 what can be a difficulty due to its big difference to the last version.

To study the best way to solve this problem, in this project will be developed a prototype called ENCODER. This user-friendly software corrects spelling mistakes and encodes, in several languages, an input text to get the ICD-10 code by using different natural language processing techniques. Apart from that, it uses the SNOMED CT descriptions because of its closeness to the natural way of speak of the medical staff.

ENCODER has been tested using SNOMED CT descriptions and the results have been satisfactory. It has encoded correctly the 80% of Spanish SNOMED CT descriptions and the 65% of the Catalan ones. Apart from that, the 60% of natural language diagnostics have been well-encoded.

So that, we conclude that the software developed can be a good solution to solve the problem. Another conclusion is that the difference between the Spanish and Catalan results may be due to a bad translation since ENCODER applies the same algorithm in both. Seeing the natural language results, it can be appreciated that they only differ from 5% to the Catalan ones, so we can induce that it is due to the same reason.

Keywords: Semiautomatic clinical codification, ICD-10-CM/PCS, Natural language processing, Artificial intelligence, SNOMED CT, Search engine, Python, MongoDB.

Acknowledgments

To begin with, I have to thank all the team of the Fundació TIC Salut Social for these two agreeable internships. Thanks to them, it has been a great experience with which I have grown professionally and personally. Especially, I have to thank Ariadna Rius for her support and also for all her teachings during these internships. This project has been possible thanks to her advices, ideas, and recommendations that have served me as a reference point in this trajectory.

I want to thank Dr. Francesc Garcia Cuyàs and Dr. Jordi Martinez for opening the doors of this company and for having created a pleasant work environment. On the other hand, I wish to thank all the people that have helped me with some tasks of the project: again Dr. J. Martinez for being offered to test the tool developed in this project, Miquel Martí for helping me with the interface designing of this tool and finally Anna Ceresuela for her help in all related clinical terminologies staff.

This project has been possible thanks to the knowledge acquired in the École Ingenieurs ISIS and all its teachers. Especially, I have to thank Monsieur Soto-Romero for the support and monitoring given during this project and Madame Lhôte for its internship management.

Finally, I want to thank Carolina Martín and María Teresa Abad from the FIB, UPC. They have made possible this Erasmus that has permitted me to take this way and go for two years to study in France.

Table of contents

Glossary

- **Generalitat de Catalunya** is the institutional system in which the government of Catalonia is organized politically. The Generalitat holds exclusive and wide jurisdiction in various matters of culture, environment, communications, transports, commerce, public safety and local governments. However, in aspects relating to education, health, and justice, the region shares jurisdiction with the Spanish government.

- **Departament de Salut** is the main administrative body of the Generalitat de Catalunya in healthcare decision-making. It has the exclusive competence of the organization, internal functioning, assessment, inspection and control of health centers, services, and establishments. On the other hand, it participates in the planning and coordination of health affairs.

- **SISCAT**. *Sistema sanitary integral d'utilització pública de Catalunya* (Healthcare Public System of Catalonia). It is the healthcare network that groups together all the public hospital centers, primary care centers, mental health centers, transport resources and others.

- **CatSalut** is the public insurer of Catalonia. Its objective is to guarantee full, public and quality health coverage to all the citizens of the territory.

- **OFTSI**. *Oficina d'Estàndards i Interoperabilitat.* It is the Standards and Interoperability Office of the Fundació TIC Salut Social.

- **ICD-10 (*CIM-10 in Catalan and French*)** is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs, and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases

- **ICD-10-CM/PCS (*CIM-10-MC/SCP in Catalan and French*)** is the International Classification of Diseases, 10th revision, Clinical Modification / Procedure Coding System. It is a subset of ICD-10 that contains the codes for clinical diagnostics and procedures.

- **SNOMED CT**. *Systematized Nomenclature of Medicine Clinical Terms*. It is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms, and definitions used in clinical documentation and reporting.

- **Snomed International (formerly IHTSDO).** It is an international agency that controls, manages and distributes the International version of SNOMED CT.

  http://www.snomed.org/

- **Catalan Extension of SNOMED CT.** The Catalan Extension is maintained and distributed by the OFTSI and used by different healthcare centers of SISCAT. It is a non-international subset of concepts of SNOMED CT in Catalan and Spanish managed by TIC Salut Social.

- **Artificial intelligence** is the set of theories and techniques used to create machines capable of simulating different areas of human intelligence

- **Natural Language** is any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation

- **Natural Language Processing (NPL)** is a domain of computer science, artificial intelligence, and linguistics that studies the interactions between computers and human (natural) languages.

- **MongoDB** is a document-oriented NoSQL database system, developed under the concept of open source.

- A **Map** in computer science is a file that links a set of elements to another one. For each element of the source set, the map gives an element of the target set. In this project, we are using a map that links SNOMED CT with ICD-10-CM.

- **Stopwords** are meaningless words, that can be deleted from a sentence and it still has the same meaning (e.g. articles, pronouns, prepositions).

1. Introduction

My internship has been realized in the Fundació TIC Salut Social in Mataró, Spain with a duration of 6 months. This foundation is part of the Departament de Salut of the Generalitat de Catalunya and works to improve technologically the health sector of Catalonia.

In 2017 I also did my 4th year ISIS internship in TIC Salut Social. It lasted 4 months but I worked another 2 months as an employee after my internship. For this reason, I already knew the working way of the company, so my reincorporation was pretty fast and simple.

In my last internship, I developed a web application to manage a termbase of Catalan extension of SNOMED CT. This application serves, for example, to create/modify/delete SNOMED CT concepts in different languages, to manage subsets and its most important function is the version management module that allows files transformation from RF1 format to RF2.

In both of my internship periods, I worked in the OFTSI, the standards and interoperability office under Ariadna Rius supervision. About 60% of my time was spent on ENCODER project, my final thesis which will be presented in this document.

This project has been directed by Ariadna Rius and implemented by Pau Garcia counting on the collaboration of different TIC Salut members. The main objective of this project is to design and implement a search engine. This will serve to study and test several natural language processing algorithms to find the best way to encode sentences into ICD-10-CM/PCS, a clinical classification of diagnostics and procedures.

1.1. Fundació TIC Salut Social

Fundació TIC Salut Social is an agency of the Departament de Salut de Catalunya (Catalan Ministry of Health), which works to promote the development and use of ICTs[1] (Information and communication technologies) in the fields of health. Other functions of the company are to keep abreast of new trends and emerging initiatives, to innovate and promote new projects and to offer an approval and

[1] *TIC* in Catalan and French

accreditation to health products. It is also responsible for providing standardization of Catalan health products.

The Foundation mission is to be a facilitator of the transformation of the health and social care model through ICTs. Its vision is to be a benchmark in the health sector boosting innovation with the use of ICT as a tool for the transformation of the healthcare model. TIC Salut Social is guided by values of transparency, sustainability, commitment to the sector, global and local innovation and management autonomy among others.

1.1.1. OFTSI: Standards and Interoperability Office

I worked in the OFTSI, which is a department of Fundació TIC Salut Social. The department is responsible for managing and distributing some of the controlled vocabularies, terminologies and classifications that are used in the Catalan health system.

The OFTSI works to guarantee the different levels of interoperability between all the healthcare information systems. On the other hand, it promotes standards and defines documents that serve as references to carrying out interoperability.

1.2. Terminologies

In order to contextualize the project functionalities, in this section, the main clinical terminologies used in ENCODER will be presented. They are ICD-10, also named CIM-10 in Catalan and French and SNOMED CT which is the clinical terminology of greater breadth, precision, and importance developed until now.

1.2.1. ICD-10 (CIM-10)

According to the World Health Organization (WHO), the ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems. It is the international standard for reporting diseases and health conditions and it also serves as a diagnostic tool for epidemiology. Uses include interoperability of systems, monitoring of the incidence and study of diseases evolution and distribution.

ICD is managed by the WHO that provides all the necessary documentation and tools to use it. All of them are available online[2] and they are free access for all

[2] http://www.who.int/classifications/icd/en/

users. It contains codes for diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases.

1.2.2. ICD-10-CM / PCS (CIM-10-MC / SCP)

ICD-10-CM/PCS (International Classification of Diseases, 10th revision, Clinical modification/Procedures coding system) are two subsets of ICD-10 that contain the codes for clinical diagnostics (CM) and procedures (PCS).

CM refers to "Clinical Modification" and it contains all the diagnostics codified in ICD-10. On the other hand, PCS refers to "Procedure Code System" and contains all ICD-10 codes corresponding to medical procedures.

The objective of this project implies the use of this both codifications, not all the ICD-10. The expected result is a search engine that takes a natural language fragment and codifies it into either CM or PCS.

1.2.3. SNOMED CT

SNOMED CT or SNOMED Clinical Terms is a systematically organized electronic collection of medical terms providing codes, terms, synonyms, and definitions used in clinical documentation and reports.

SNOMED CT is considered one of the most comprehensive multilingual clinical terminology. The main objective of SNOMED CT is to encode most of the clinical terms used in health information systems and to support effective clinical data reporting to improve the patient care.

This terminology includes clinical outcomes, symptoms, diagnostics, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices, and specimens. SNOMED CT is maintained and distributed by SNOMED International, an international non-profit organization. SNOMED International is the trade name of the International Health Terminology Standards Development Organization (IHTSDO), established in 2007.

2. Framework and objectives

I have expended the most part of my time in TIC Salut Social developing the project ENCODER. Apart from this commission, I gave support to the Standards and Interoperability Office and I did some minor tasks that TIC Salut ask me to do.

In order to understand the need of ENCODER, this section will start contextualizing the law changes in Catalonia that concern to ICD-10. Then, the objectives of the project will be exposed and the last subsection will explain some of the other missions I had in the company.

2.1. Justification of ENCODER

In January 2016, by royal decree (Real Decreto 69/2015), the Ministry of Health, Consumption and Social Welfare[3] of the Spanish Government implemented and launched ICD-10 throughout the Spanish territory except for Catalonia. Due to an independent functioning of the invoice report, in Catalonia, it was applied two years later.

From January 2018, it is mandatory that all the healthcare centers report its economic activity in ICD-10 to CatSalut thus leaving ICD-9. For this reason, all the institutions of SISCAT have to change its information system in order to respond to the new requirements.

ICD-10 incorporates many new codes and important structural changes for both classifications: diagnostics (CM) and procedures (PCS). ICD-10-CM has grown from 14,025 concept codes to 71,486. This increase is due to the incorporation of laterality[4] concepts, new codes for medical and surgical complications, the combination codes[5] and a greater specificity.

On the other hand, ICD-10-PCS has grown from 3,838 to 71,924 concept codes. Some of the most important changes have been the elimination of eponyms and combination codes, the clinical terminology update, the character meaningful position, a better flexibility for new extensions and a greater specificity.

Most doctors and documentarists encode quickly because they know most common codes thanks to their experience. Due to this report procedure changes, all the medical staff has to change their way of coding and learn the new encoding method.

[3] *Ministerio de Sanidad, Consumo y Bienestar Social*

[4] Indicates the part of the body side (e.g. right/left arm).

[5] It corresponds to those pathologies that need more than one code to be identified.

Apart from the exposed changes, often the concept descriptions are written using too technical language which makes difficult to find a code and can provoke coding errors. The ENCODER project is born to satisfy this encoding difficulty in order to find the best way to solve the problem.

2.2. Objective of ENCODER

ENCODER aims to find an artificial intelligence algorithm that serves as the basis of an ICD-10-CM / PCS code search engine. In this way, the task of finding a code in ICD-10-CM / PCS will be simplified since the search will be done from more familiar words by doctors.

The artificial intelligence of the search engine will be able to "understand" the concept although the words introduced do not coincide directly with the description. This tool will propose a variable number of codes sorted by similarity with the user input.

This user, normally medical staff, can use those suggestions to find the best code to encoding the correspondent diagnostic or procedure. The user will be able to write the description on its preferred language because ENCODER will translate from any language to English with a language auto-detection. Furthermore, even if the user makes spelling mistakes, ENCODER incorporates an auto-corrector that will correct all sentences before encoding.

On the other hand, ENCODER matches descriptions with all the synonyms used in SNOMED CT which are most familiar for doctors, what will improve its matching rating.

2.3. Other missions in the company

Apart from ENCODER, I had other missions in the Fundació TIC Salut Social. For example, I did other tasks such as give support to some events, to improve the last year project called SCATManager, to write some articles and so on.

For example, on June 29, TIC Salut Social organized a workshop called “Bussejant entre dades” [*diving among data*] whose main topic was how artificial intelligence can help us discover and process information. In this workshop, there were several conferences about the use of natural language processing (NLP) to encode unstructured text into ICD-10/CM-PCS.

The invited companies presented some of the innovation tools they developed for semiautomatic encoding and information discovering. Finally, there was an open discussion where assistants could make questions to the speakers.

As it was much related to my final degree work, I applied to be organizer together with Ariadna Rius and communication team. I realized tasks like the workshop schedule, establishing some contacts and giving support on the day of the workshop. Apart from this workshop, I have given support to my coworkers in other ones such as "In Deep", "Hello World", "CIOs Workshop for the Trends Map" [6].

Apart from workshops organized by TIC Salut Social, I assisted in some workshops created by other companies (e.g. Interoperability in the social sector). In addition, I realized some online courses like IBM Blockchain Foundation for Developers in order to become a contact reference for the Blockchain topic in the Fundació TIC Salut Social.

The company also commissioned me to write some articles for its webpage and to give support for writing some charters for new projects taking about Blockchain. On the other hand, I have participated in the writing of a guide titled "Application development guide".

It is a good practices guide for the development of health mobile applications which gives recommendations and steps to follow from the beginning of an application project until its publication. Several companies in the sector and universities have participated and it will be published on the TIC Salut Social web in September 2018.

Finally, I worked to improve some functionalities of my last project in the company. Last year, in my first internship in TIC Salut Social, I developed a web application called SCATManager to manage the Catalan termbase of SNOMED CT. This year, OFTSI needed to transform all the database data into version 2 of SNOMED CT (RF2) so they ask me to fix some minor bugs of visualization and displayed information.

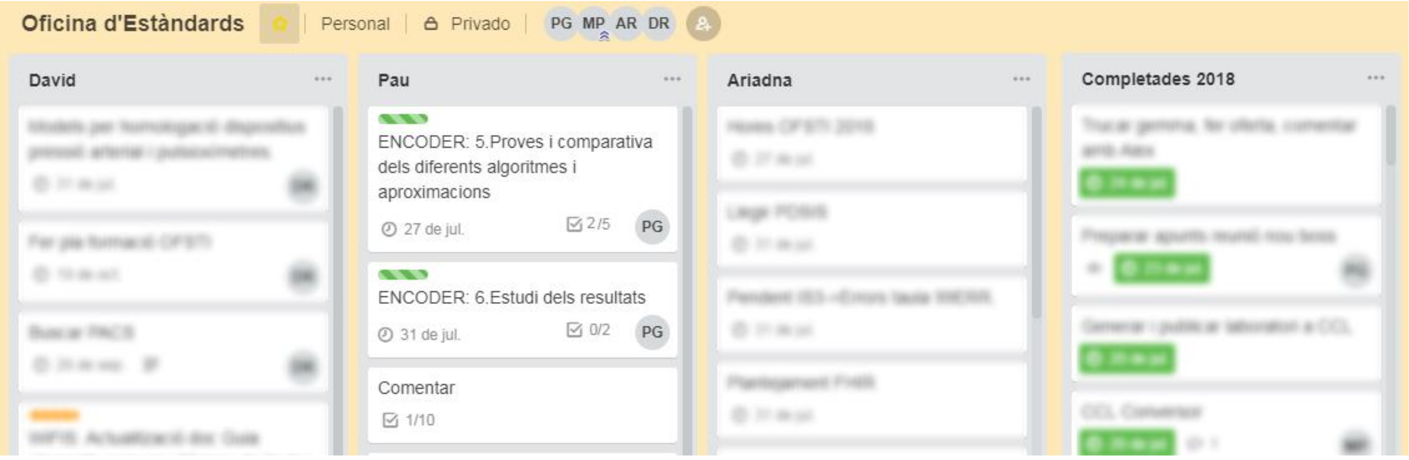
[6] For more information, visit https://ticsalutsocial.cat/

3. Realizations: ENCODER

3.1. Project management

The OFTSI works using Trello which is a to-do online method. On this tool, there are all tasks of the team in form of post-its. In this way, we know what our partners are working with.

For each person, Trello contains a list of to-do tasks and another with done tasks. Each one is responsible for his list and it is important to have it update to keep the team's organization.



Oficina d'Estàndards | Personal | Privado | PG MP AR DR

David

Pau

ENCODER: 5.Proves i comparativa dels diferents algoritmes i aproximacions
27 de jul. 2/5 PG

ENCODER: 6.Estudi dels resultats
31 de jul. 0/2 PG

Comentar
1/10

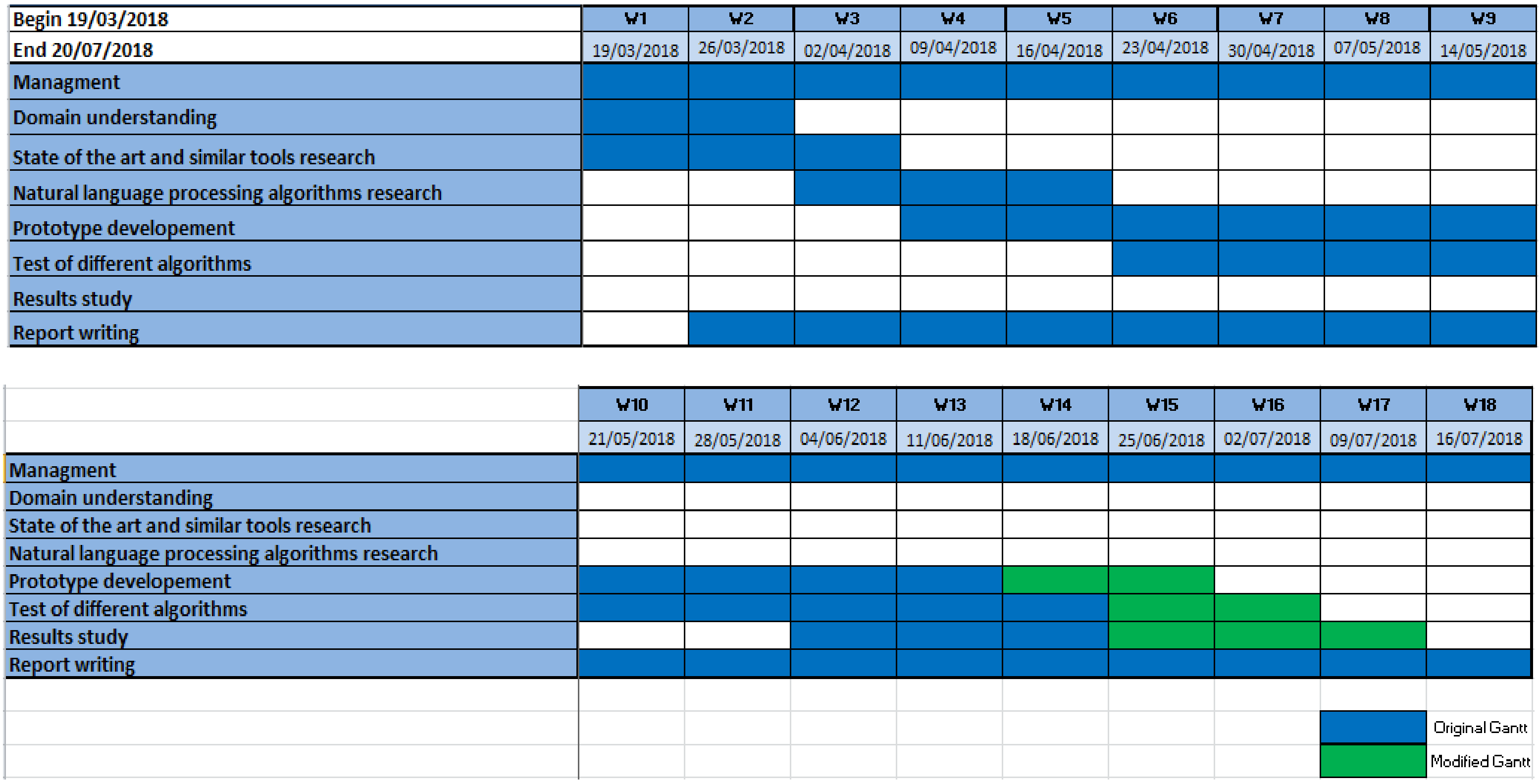
Ariadna

Completades 2018

3.1 Screenshot of the OFTSI Trello's

On the other hand, for each project, we use a Gantt. To begin the project, it is required a Charter[7] which is the project proposal. In this document, there is defined the duration of the task of the project although it is can be modified during the project execution in function of needs.

[7] This document was presented in French by e-mail to Monsieur Soto-Romero at the beginning of the project



Begin 19/03/2018	W1	W2	W3	W4	W5	W6	W7	W8	W9
End 20/07/2018	19/03/2018	26/03/2018	02/04/2018	09/04/2018	16/04/2018	23/04/2018	30/04/2018	07/05/2018	14/05/2018
Managment	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Domain understanding	🟦	🟦							
State of the art and similar tools research	🟦	🟦	🟦						
Natural language processing algorithms research			🟦	🟦	🟦				
Prototype developement				🟦	🟦	🟦	🟦	🟦	🟦
Test of different algorithms						🟦	🟦	🟦	🟦
Results study									
Report writing		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

	W10	W11	W12	W13	W14	W15	W16	W17	W18
	21/05/2018	28/05/2018	04/06/2018	11/06/2018	18/06/2018	25/06/2018	02/07/2018	09/07/2018	16/07/2018
Managment	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Domain understanding									
State of the art and similar tools research									
Natural language processing algorithms research									
Prototype developement	🟦	🟦	🟦	🟦	🟩	🟩			
Test of different algorithms	🟦	🟦	🟦	🟦	🟦	🟩	🟩		
Results study			🟦	🟦	🟦	🟩	🟩	🟩	
Report writing	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
								🟦	Original Gantt
								🟩	Modified Gantt

3.2 Initial (blue) and real (green) Gantt of ENCODER

In the previous image [3.2], it is shown how initial Gantt was modified during the project ENCODER. It is divided by the main tasks of the project, and each one has assigned some period to be realized.

In total, the project has lasted 18 weeks. Green cells are modifications with respect to the initial Gantt. At week 11, we replaned tasks and we extended the time duration of the last three tasks: development, test, and study of the results. As we planed the last weeks with less workload, there were no problems when modifying the initial plan and all tasks were done finally in time.

3.2. Project organization

The team of ENCODER is essentially formed by Ariadna Rius and Pau Garcia but there were different collaborations of members of TIC Salut. Ariadna Rius is the project manager, who defines and plans different tasks and who have a global vision of the project. On the other hand, I am the main programmer analyst of the project. My task is to follow instructions from the project manager and report all the activity in order to carry out a constant feedback.

The first collaboration in the project was Anna Ceresuela that realized some tasks related to ICD-10. She is a documentarist student, with knowledge about medical terminologies, so she helped the team with the ICD-10 structure understanding, providing some of the documents used to import the terminology in the database and so on.

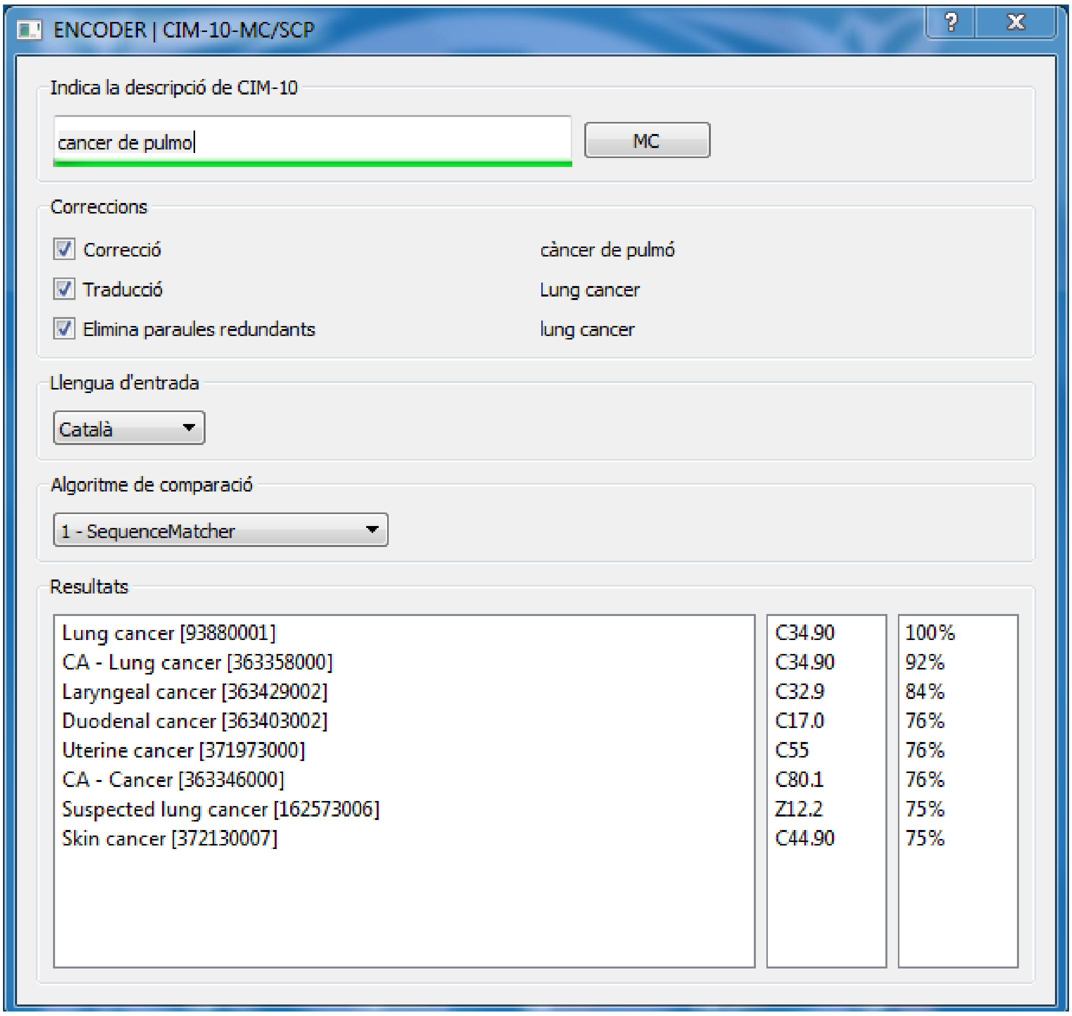
Then, Miquel Martí gave support in the user interface design. He studied videogames design, so he has experience in designing ergonomic and useful interfaces. Finally, Jordi Martinez helped the team writing some diagnostics that we used to test ENCODER. He is the innovation director of TIC Salut and family doctor, so he could write some of most typical diagnostics in medicine. These diagnostics were one of the most important points to test because they were written in real natural language by a doctor.

3.3. Project realization

As a programmer analyst, I did most of the programming part of ENCODER. The result is a python multiplatform program (Windows and Linux) that helps the doctors on the ICD-10 codification task.

Its utilization is pretty simple: the user has to input some ICD-10-CM/PCS code and press a button. After few seconds, ENCODER will search in SNOMED CT and ICD-10 database in order to find the best matches with the input description.

When the search ends, the tool displays a list of results (by default 8) ordered by a matching ratio with the original sentence. Each result is an ICD-10-CM code and for each one, it is shown its SNOMED CT description and ICD-10 description.



Indica la descripció de CIM-10: cancer de pulmo

Correccions

☑ Correcció — càncer de pulmó

☑ Traducció — Lung cancer

☑ Elimina paraules redundants — lung cancer

Llengua d'entrada: Català

Algoritme de comparació: 1 - SequenceMatcher

Resultats

Lung cancer [93880001]	C34.90	100%
CA - Lung cancer [363358000]	C34.90	92%
Laryngeal cancer [363429002]	C32.9	84%
Duodenal cancer [363403002]	C17.0	76%
Uterine cancer [371973000]	C55	76%
CA - Cancer [363346000]	C80.1	76%
Suspected lung cancer [162573006]	Z12.2	75%
Skin cancer [372130007]	C44.90	75%

3.3 Screenshot of ENCODER

The first task was to design the NLP algorithm. Ariadna has a master of artificial intelligence, so the team had periodical meetings to define the best theoretical way of process natural language descriptions written by doctors. In the first meetings, Python was selected as a development language because of its flexibility, simplicity, and speed.

Thanks to her knowledge in artificial intelligence, Ariadna proposed the best way to treat a natural language sentence, proposing the spell checking, translation, truncating, and stopwords treatment. After some researches, we selected the sentences comparison algorithm and she designed how the flow of information should be in ENCODER.

After these meetings, we draw a schema of how the program should process the input text. Then, with a collaboration of another member of the OFTSI (Miquel

Martí), we designed the user interface. He has a degree in video games, so he could help in this task giving ideas of how to make a UI as clear and simple as possible.

At the beginning of the development, I need to find the ICD-10 documents in order to fill the database. At this point, the team had the intervention of Anna Ceresuela. She is a documentarist student that knows well several clinical terminologies, so she was very helpful.

In order to fill the databases with the SNOMED CT elements, we used the data from the work of last year's internship (SCATManager). It was helpful because there are the Catalan, Spanish and international versions in MongoDB format, so it was pretty easy to import to the new project that also uses MongoDB.

In parallel, I planed the development part choosing the development tools and researching the appropriate Python libraries to implement the artificial intelligence part. Apart from this, I chose MongoDB as a database because of its flexibility, speed, and good performance.

Apart from this, I designed the database structure, the Python classes of the program ENCODER and the internal dataflow. However, I always had the help from Ariadna Rius who validated all the decisions and proposed several improvements to the design. Once all these items were defined, we started the development part, where I take charge of all the software implementation having periodical meetings to validate the work.

At the beginning of the project, we realized that there was not an implemented map of ICD-10-PCS to SNOMED CT. This fact difficult a lot the task of encoding because the map is an essential part to obtain the right results. So that, we decided to simplify the objectives and work only the CM part.

3.4. Technical choices and tools

The program has been developed using Python[8] and its main libraries such as sys, os, threading, HTML, JSON and so on. On the other hand, the database of the program has been programmed using MongoDB[9] technologies and Studio 3T[10] as a database manager.

[8] https://www.python.org/

[9] https://www.mongodb.com/

[10] https://studio3t.com/

Another Python library used to implement the user interface part has been QT, specifically PyQt5. To implement the spell checking we have used Language Tool, with the corrector of SoftCatala for the Catalan. For the translation, ENCODER uses the Google Cloud technologies via RESTFUL APIs.

Other libraries used have been NLTK (Natural language toolkit), SKLearn, DiffLib, TextBolb and Jellyfish to make all the natural language processing. This includes the truncating the sentences matching.

3.5. Main flowchart

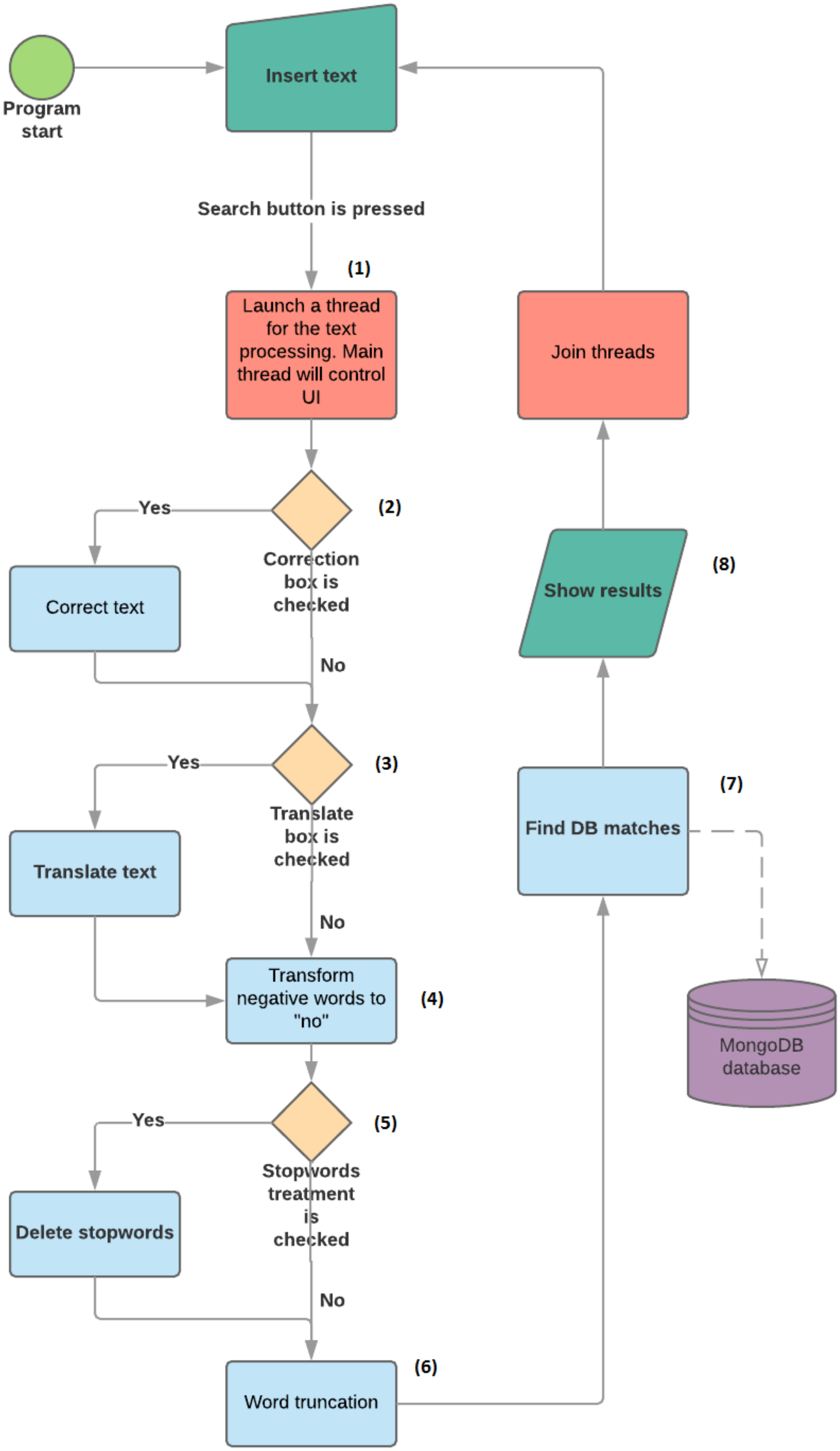
Once ENCODER starts, it waits until user inputs some text in the first field and he clicks to the search button.

When the button is pressed, the program launches a thread [**(1)** *see the image 3.4*] which will take charge of all the calculations to obtain the results. The main thread is responsible for controlling the user interface. This division into threads is used to prevent the program from hanging.

The calculations thread starts verifying if checkboxes are marked. If the correction box is checked **(2)**, ENCODER will correct spelling of the input text using the LanguageTool API provided by a web service.

Next, it checks if it has to translate the text into English **(3)**. If so, it will use the Google Cloud Translation web service to do this task. The following step is to transform all negative words into 'no' **(4)**. This modification is not configurable by the user and it has to be always carried out.

Then, the program will delete all stopwords if the user has checked the corresponding box **(5)**. The last step of text processing is word truncation whereby all sentences have to pass **(6)**.



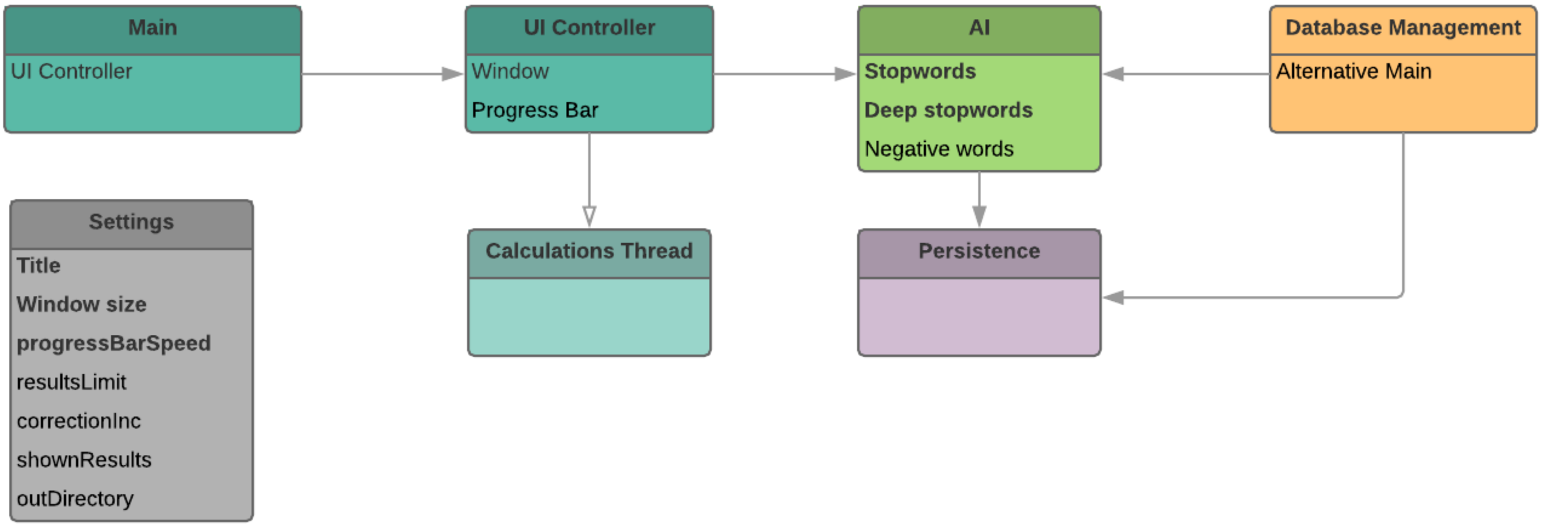
3.4 Main flowchart

Once the text is processed, it is obtained a simplified sentence in English that strictly contains the meaning of the sentence without language complements. The following step is to find SNOMED CT descriptions matching with the processed sentence **(7)**. In this step, the user can choose which of the 5 available algorithms of sentence matching will be used.

To conclude, the application shows the most similar matchings in the results table (by default the first 8 are displayed) **(8)** and the calculations thread joins the main thread. Then, the user can write another diagnostic and restart the process.

3.6. Class diagram

The source code of ENCODER is divided into seven classes related among them as shown in the class diagram below. ENCODER has two main functions: the first one is found in the Main class and the second one is in Database Management class.



3.5. Class diagram of ENCODER

The class **Database Management** contains an alternative main and three useful functions. One of these functions produces the Processed CM map found in the database.

The other two functions of DataBase Management are used to find the "holes" in the original map. One of these functions checks which SNOMED CT codes have not a correspondence in ICD-10 and vice versa.

On the other hand, the normal execution of the program is the **Main class**. The functions of this class are to create the user interface of ENCODER. Then, the **UI Controller** class will take charge to control it.

The latter creates an instance of **Calculations Thread** that realizes all the hard work of the program (to process text and find matches). This thread calls **AI** class to process the text and this calls **Persistence** in order to obtain the similar descriptions.

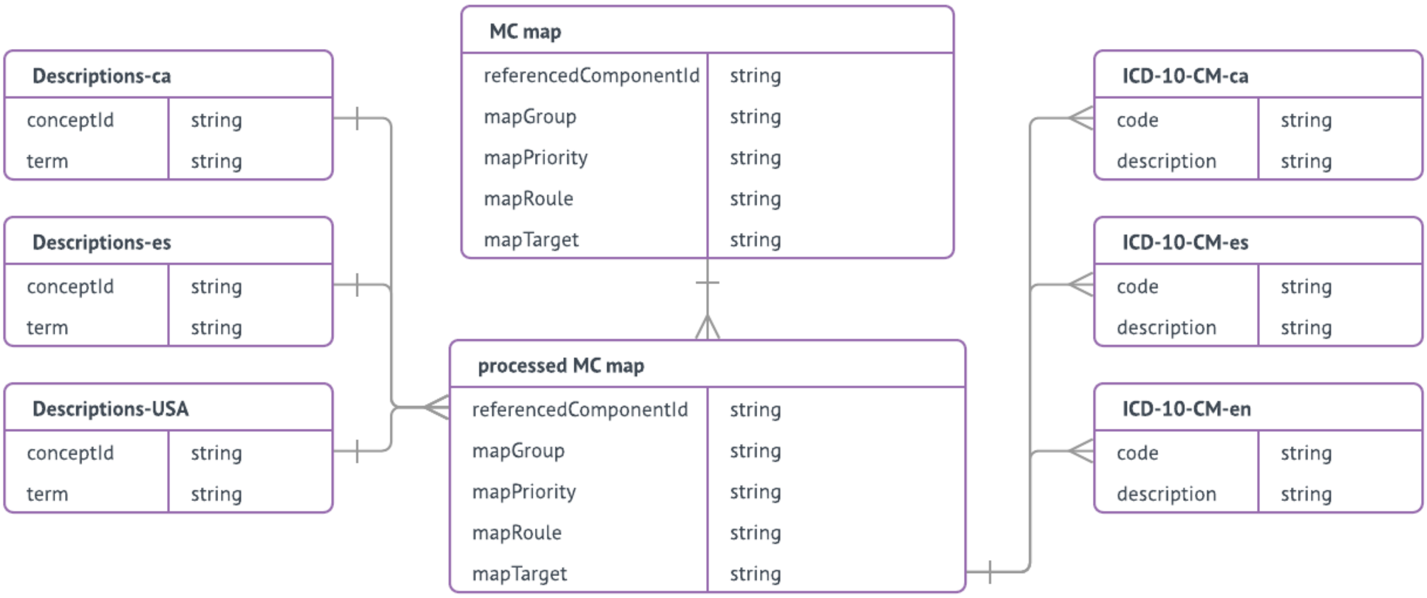
Finally, there is a **Settings** class which contains some variables that can be modified in order to change some parameters of the program such as the number of results shown.

3.7. Database

The database works under MongoDB technology, which is a document-oriented NoSQL database system, developed under the concept of open source. Documents are written codified using JSON format.

The database has 8 collections grouped in 3 categories configured as shown below:

SNOMED CT **Maps** **ICD-10**



3.6. Diagram of ENCODER database structure

The first 3 collections (the first column in the previous diagram) called Descriptions-ca, Descriptions-es, and Descriptions-USA belongs to the first group called SNOMED CT explained in section 3.7.1. The collections in the second column called CM map and processed CM map belongs to group SNOMED CT to ICD-10-CM map and it is described in section 3.7.2. Finally, the last 3 collections belong to the ICD-10-CM group and they are described in section 3.7.3.

3.7.1. SNOMED CT

The classes Descriptions-ca, Descriptions-es and Descriptions-USA contains SNOMED CT descriptions for languages Catalan, Spanish, and English.

Descriptions-ca[11] has the descriptions of Catalan SNOMED CT extension which contains descriptions in Catalan and Spanish.

Descriptions-es[12] has the descriptions of Spanish SNOMED CT extension which contains descriptions in Spanish.

Descriptions-USA[13] has the descriptions of American SNOMED CT extension written in English. As the mapping file uses codes of USA extension, we have had to use the American version instead of the international one.

3.7.2. ICD-10-CM

The classes ICD-10-CM-ca[14], ICD-10-CM-es[15], ICD-10-CM-en[16] contain all ICD-10-CM codes and its description in Catalan, Spanish and English respectively.

3.7.3. SNOMED CT to ICD-10-CM map

The original map[17] (called CM map in the database) does not directly link a SNOMED CT concept to an ICD-10-CM concept. It is constructed by using map rules, which indicates how to get the target code by asking the user at runtime.

[11] Data extracted from TIC Salut Social

[12] Data extracted from http://browser.ihtsdotools.org/

[13] Data extracted from: http://browser.ihtsdotools.org/

[14] Data extracted from: http://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/diagnostics-procediments/cim-10-mc-scp/

[15] Data extracted from: https://eciemaps.msssi.gob.es/ecieMaps/documentation/documentation.html

[16] Data extracted from: http://www.who.int/classifications/icd/en/

[17] File extracted from: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html

For this reason, before use this map, we have had to transform this map in order to make a direct mapping. In a few words, we have processed all the map rules so as to obtain a target code for each source code.

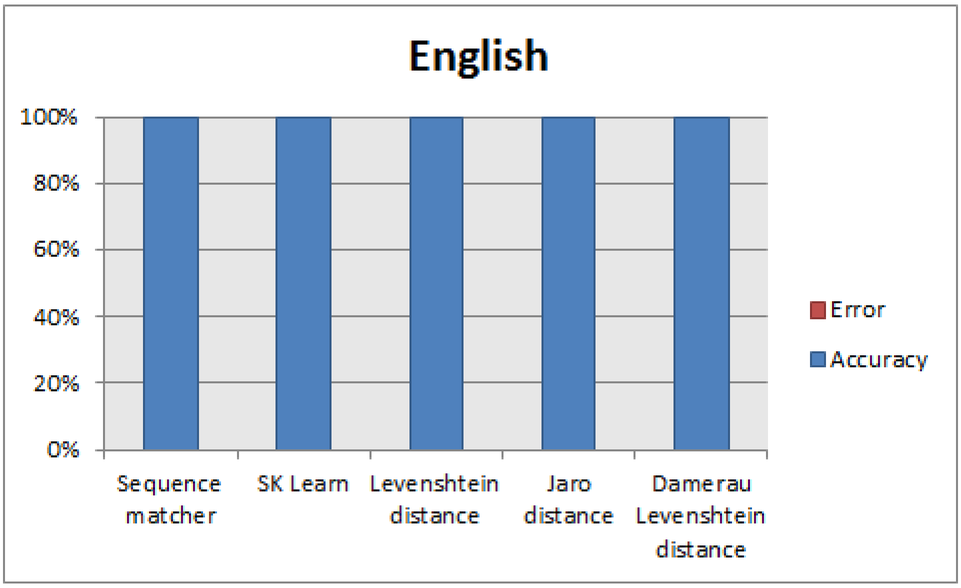
In conclusion, the processed map contains an ICD-10 code for each SNOMED CT code. Apart from this, in order to make ENCODER faster, we have processed all the SNOMED CT descriptions of the American version. We have replaced all negative words by "no", we have truncated the suffix of the words and we have deleted stopwords from the sentences which will simplify the task of comparing one sentence to the entire database.

3.8. Results

In order to test the accuracy of ENCODER, we have realized several tests in different languages using all the available algorithms. In this way, the results show the precision of encoding into ICD-10-CM.

For language English and Spanish, the test kit is a random subset of 1000 SNOMED CT descriptions. However, the SNOMED CT Catalan extension contains 130 diagnoses, the available ones in the Catalan extension.

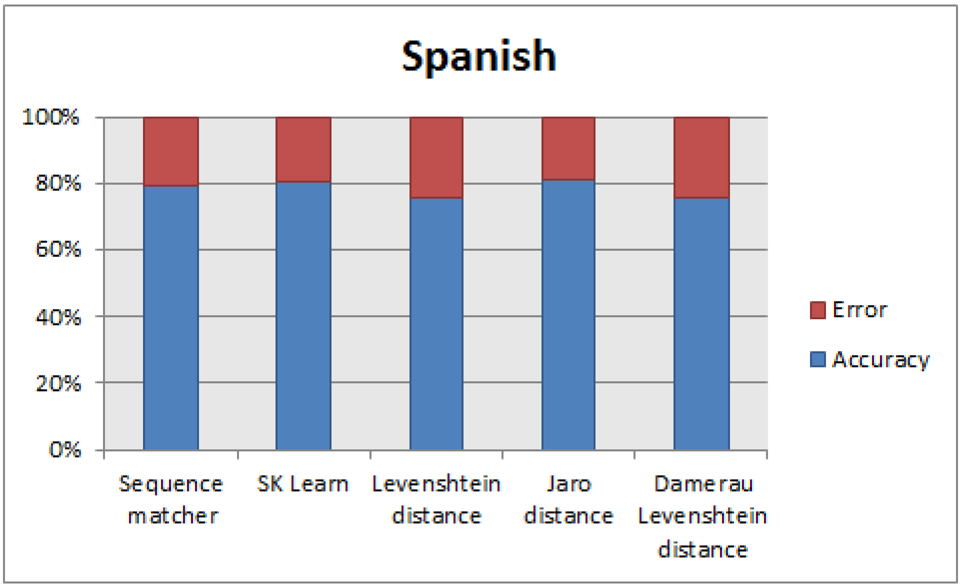
Apart from these three tests, we have asked Dr. Jordi Martinez to write 100 diagnostics. We call this the test of Natural Language because this kit was written by a doctor so it contains typical errors like typos, abbreviations, usage of Catalan and Spanish in the same sentence, etc.



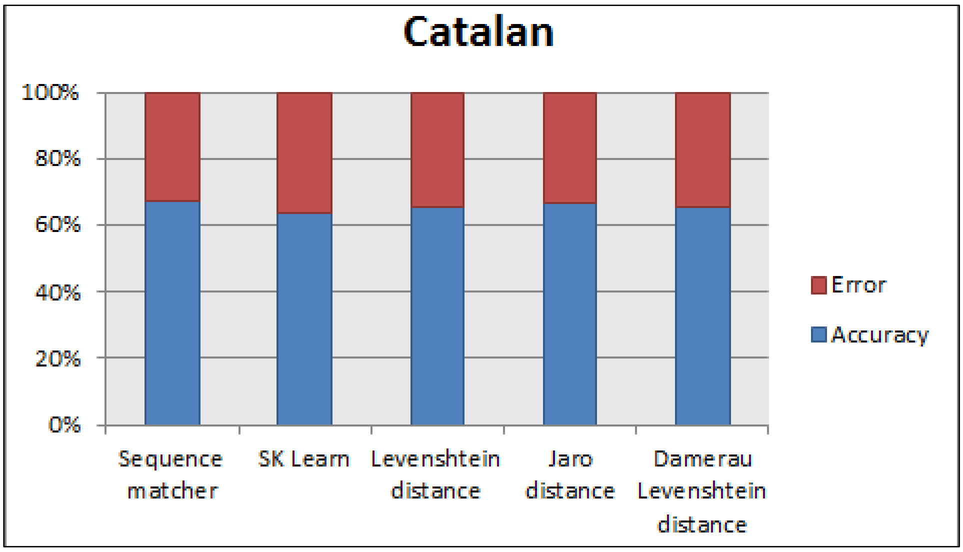
The diagram below shows results obtained using 1000 random English SNOMED CT descriptions. It shows a match for each description, so the accuracy is always

100% what makes an error of 0%. This is due to the usage of SNOMED CT descriptions. The algorithm finds always the correct description because it searches the original description in the SNOMED CT English database, so it matches always.

It is a dummy test because we knew that the result should be of 100% matching. However, we considered necessary to test the good performance of ENCODER. Seen this results, we can accept the correctness of the software.



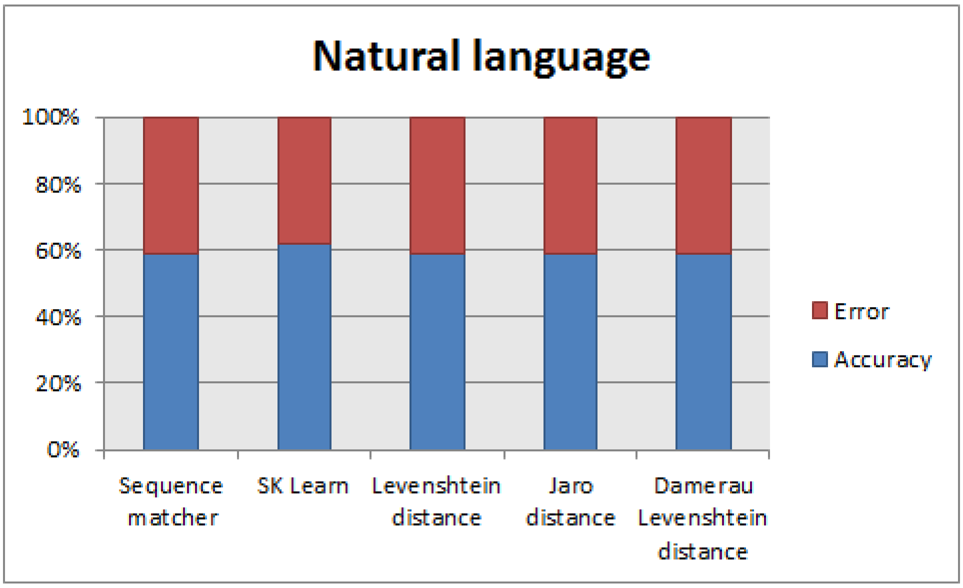
In this test, there are the results for each algorithm using 1000 random SNOMED CT Spanish descriptions as input. We can extract two ideas: the first one is that for each algorithm, the result has been always about 80%, what signifies that ENCODER well-encodes eighth of every ten descriptions. The second thing is that the difference between algorithms is almost negligible and all them gives similar results.



For the Catalan tests, we have obtained worse results than the Spanish ones with all the algorithms. ENCODER applies the same transformations for both inputs, so the only thing that changes is source language and its translation.

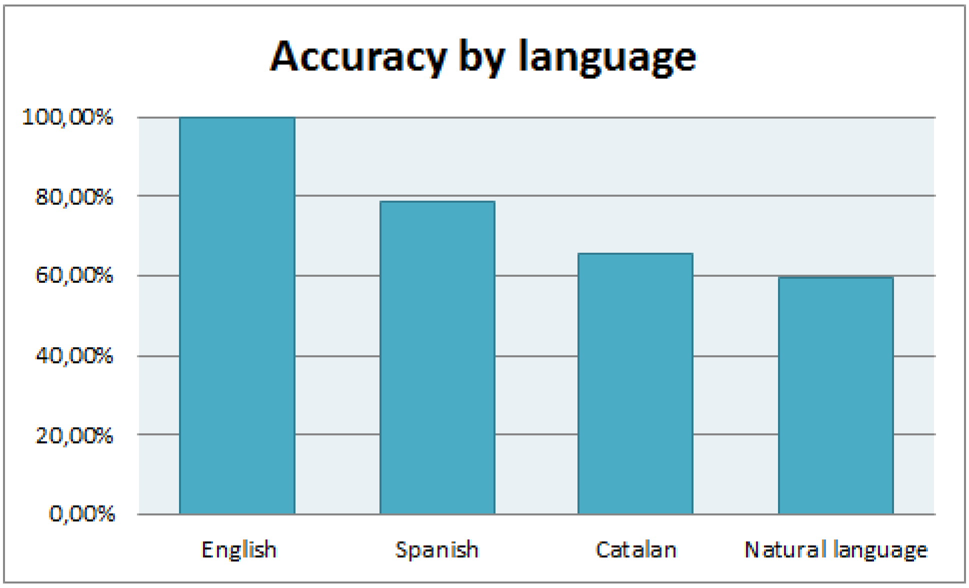
So for this reason, probably these results may be due to a worse translation provided by Google cloud translation. For experience, we know that the Catalan translation is not as accurate as the Spanish one.

This may be because Google uses a machine learning algorithm to translate sentences, so it depends on the number of speakers of the source language. The minor number of Catalan speakers can explain that the algorithm has not been able to "learn" as well as it has done with Spanish.



Finally, in this plot, it is shown results obtained by matching the descriptions written by a real doctor. Using a real natural language, we have obtained a minor accuracy, with all algorithms about 60%. However, they are not so different from the Catalan ones, on average them differ in 5%.

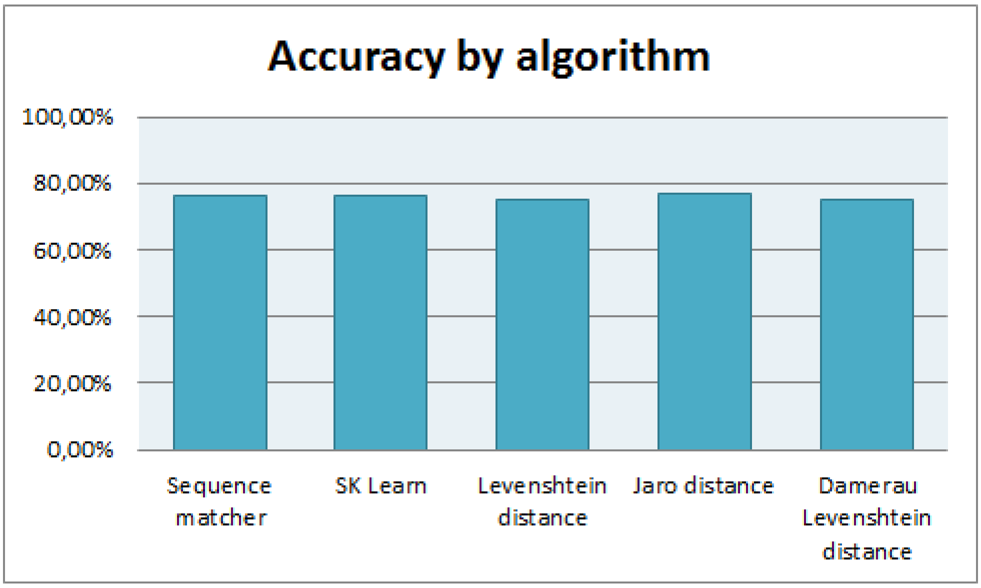
Most of these descriptions were written using informal abbreviations that cannot be matched with a SNOMED CT description and even so results are not so different from the Catalan test. On the other hand, again we see that there is not a significant difference between the results provided by the different algorithms.



In this plot, it is shown the difference between the averages of accuracy for each language. As we see, English is always 100% followed by the Spanish test kit. Finally, in the last group, probably due to a worse translation there are the Catalan and Natural Language test kits.

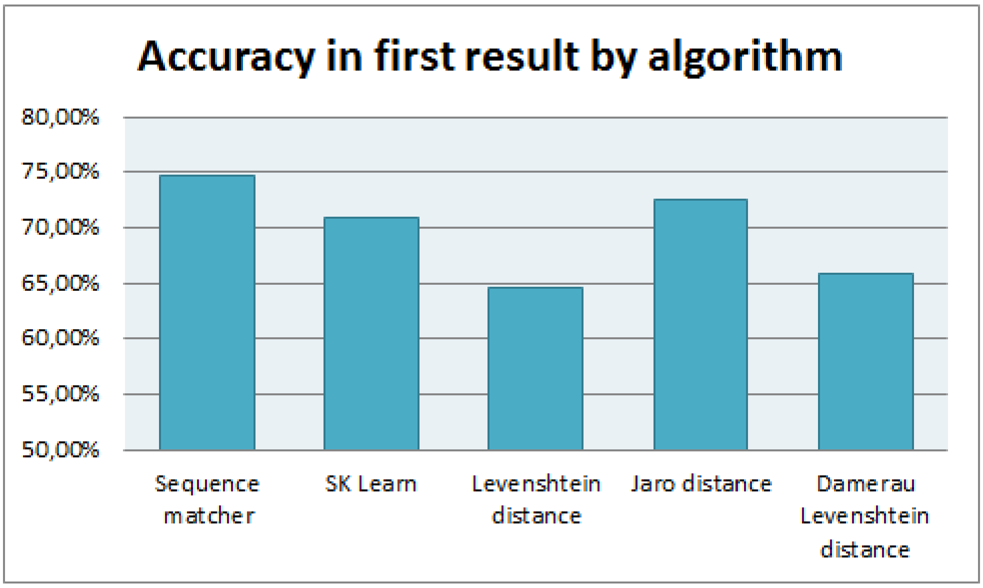
The important point here is to see that the difference between Natural Language results and the Catalan correct ones is about 5%. This means that ENOCER has a well-performance with natural language in comparison to well-written descriptions.

On the other hand, there is a difference of 15% between the Spanish and Catalan descriptions. As it was exposed, considering that the same operations were applied to treat the text, we consider that the problem is found in the translation. So maybe, with a better translation of the Catalan, we could found better results, up to 80%.



Finally, in this plot, there is the difference of accuracy between algorithms. As we see, there is not so much difference between them. Once a major set of descriptions it is provided to the algorithm, the task of this one is to select the 8 most sentences to the original one by using different methods.

From these results, it can be proved that the five algorithms are calculating the similarity pretty well and they differ just in concrete cases.



In order to test the accuracy more meticulously, we have tested which algorithms had provided the solution in the first position. All the other tests were done considering that they had a match if the solution were contained in the first 8 results.

In this plot, it can be appreciated how the algorithms provide the result in the first position. As the results show, the best results are given by Sequence matcher with more than 74% accuracy. This means that the 97% of times that that algorithm shows a correct result, it is found in the first position.

Contrarily, Levenshtein distance shows just the 64% of the right results in the first position. This means that the 86% times it shows the correct answer, it is found in the first position. Even so, we have to take into account that all these algorithms contained the right result among the 8 first positions in almost 80% times.

3.9. Problems found

In general, I have not found organizational problems in the company. I suppose the accommodation was easy for me because I already had worked in TIC Salut in my last internship. Nevertheless, our team found some technical problems during the project development. Our biggest problem probably was to not have the map between SNOMED CT and ICD-PCS.

Due to this fact, we have not been able to develop a program capable of encoding procedures. However, we reached to develop a tool that codifies, with good results, all the diagnostics. Furthermore, it can be easily adapted to codify procedures if ever somebody develops the map SNOMED CT/ICD-10-PCS.

Apart from these problems, we have found some "holes" in the ICD-10-CM map used in ENCODER. We have studied this map, and there are some elements that do not have a link to the other terminology. For instance, taking all the diagnostics codified in SNOMED CT, we found that only the 76% have an entry in the map that codifies this diagnostic into ICD-10-CM. That means that if we take some SNOMED CT diagnostic outside of this 76%, the software will not be able to encode it into the target terminology.

Otherwise, checking the ICD-10-CM elements, we found that 50% of the elements in this terminology have not an entry in the map. However, we have to consider that the map has a direction: it allows to transform a SNOMED CT code to an ICD-10-CM one. For this reason, this lack of elements is not as important as the first one because the map is just used in the other direction.

Moreover, this is probably due to the laterality and similar attributes of the codes in the ICD-10. This means that for each diagnostic, there exist different codes that specify the location, the gravity of the diagnostic, the part of the body and so on. For example, there exist two different codes for "*Burn of third degree of* ***right*** *palm*" and "*Burn of third degree of* ***left*** *palm*". In SNOMED CT, this will be encoded to the same code, using another code that codifies just *right* or *left*.

However, we have considered that these difficulties were alien to us, so we do not have the way to solve it because it depends on the terminology form. Furthermore, it does not affect the ENCODER performance because it will be proposed, for example, both codes.

Apart from these problems, we have found some technical problems that have been solved by modifying a fragment of the algorithm or by changing some part of the source code. All these problems were not of great magnitude and it can be easily solved in the periodical meetings with the team.

4. Conclusions and perspectives

4.1. Future work

This tool is now a Python program that needs a specific compilation to be executed. A perspective to ENCODER could be to transform this software into a web-application. In this way, it would be executed on any platform independently of the operating system, facilitating its use.

Moreover, the natural language processing could be improved by applying some different techniques. For instance, an improvement to the algorithm could be a more advanced detection of negations, detecting what negations are denying. Besides, it could be done a treatment of abbreviations and synonyms in order to approach the results to the reality since doctors usually make use of them.

Another improvement could be to study a way to ameliorate the translations of Catalan to get better results. However, this could be difficult as it depends on an external service. In addition, if ever ICD-10-PCS is created, the ENCODER project can be extended to the procedures encode.

Finally, maybe the most important future work to this project can be to study the way of adapting ENCODER to a workstation of doctors and documentarists. It requires some work, as for example to test its correctness more carefully to make sure that the codes provided are strictly always correct. It is important to make this work because it is a health-related tool, so it has to pass a lot of tests after being launched in the health centers.

4.2. Conclusion

As a personal conclusion, this has been a great internship because I have learned a lot of new things about health and computing. On the one hand, I have gained experience on the Python programming and the use its libraries. Moreover, I have learned a lot about artificial intelligence: how to treat the natural language to be able to compare the meaning of two sentences, different algorithms of comparison and so on. Apart from this, I have improved my knowledge in MongoDB which I think that will be very useful for future projects due to its No-SQL paradigm.

On the other hand, I have learned how the two most important health terminologies work internally. I had some knowledge about SNOMED CT thanks to my previous internship, but these 6 months have been very useful to go deeper into this field. Apart from this, I have learned the encoding way of ICD-10 and the manner to pass from one terminology to another using a map.

Apart from the knowledge that I have acquired from doing this project, I have been able to do some tasks such as organizing events, write several articles and know people in the health sector of Catalonia. I have also learned about Blockchain, FHIR, Gantt, and Trello thanks to different minor projects done in the company.

With regard to ENCODER, we have reached the initial objective: to be a tool of support to encode natural language into ICD-10-CM. The objective of encode ICD-10-PCS could not be done due to the inexistence of a map from SNOMED CT to this terminology as it was exposed previously. However, it is easy to implement this new functionality once some institution develops the convenient map.

As the tests have asserted, ENCODER has codified correctly the 80% of the Spanish SNOMED CT diagnostics. Regarding the Catalan ones, it has codified about 65% using the same algorithm. So for this reason, we conclude that this fact is due to a wrong translation provided by Google Cloud translating services.

As it was exposed previously, Catalan is a language with fewer speakers than Spanish. So that, the machine learning algorithms used by Google to translate Catalan is less effective because it has fewer cases to learn from. As a result, the translation is worst coming from Catalan than from Spanish, as it is well-known.

On the other hand, the codifications of real natural language in Catalan (and some in Spanish) have reached the 60% of accuracy. It just differs from 5% of the Catalan results, so it is a good result because it means that it can be useful for doctors to encode diagnostics written in their usual way.

If we regard the results of the algorithms, we have not found a big difference of accuracy. All the five have provided the same average of correct descriptions. Nevertheless, watching it in more detail, it can be appreciated that Sequence matching, SK Learn, and Jaro distance provides more times the correct solution in the first place than Levenshtein distance and Damerau L. distance.

This makes not an important difference because it means that if one of the algorithms provides the good result, probably the other four, will also. However, for a doctor can be useful to select, e.g. Sequence matching, because it will display the correct codification in a higher position than the other ones, so it will be easy to find it.

As a global conclusion, I am satisfied with the work done and I think that it can be a useful tool that the Fundació TIC Salut Social will be able to use in future projects. Moreover, in the future, with some more correctness tests, this tool can be used by medical staff that needs to encode each day a lot of diagnostics into ICD-10.

As it was exposed in the top of this document, the change between ICD-9 and 10 is currently being made. So that, doctors and health personal does not know the ICD-10 codes, so this software can be helpful as a daily workplace tool. Even more, if some institution develops the map between the whole ICD-10, this tool may be a complete program to encode any diagnostic, procedure and so on into this terminology.

5. Bibliography

[1] Suzanne PEREIRA, Aurélie NÉVÉOL, Philippe MASSARI, Michel JOUBERT and Stefan DARMONI. Construction of a semi-automated ICD10 coding help system to optimize medical and economic coding. Technologies for Better Health in Aging Societies A. Hasman et al. (Eds.) IOS Press, 2006.

[2] P. Franz, A. Zaiss, S. Schulz, U. Hahn, and R. Klar. Automated coding of diagnoses--three methods compared. Proc AMIA Symp. 2000: 250–254.

[3] Sue Bowman. Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems. Perspectives in Health Information Management Spring 2005 (May 25, 2005).

[4] Patrick RuchEmail, Julien Gobeill, Christian Lovis, and Antoine Geissbühler. Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Decision Making20088(Suppl 1): S6.

[5] Juan Antonio Goicoechea Salazar, María-Adoración Nieto-García, Antonio Laguna Téllez, Vicente David Canto Casasola, Juliana Rodríguez Herrera, Francisco Murillo Cabezas. Desarrollo de un sistema de codificación automática para recuperar y analizar textos diagnósticos de los registros de servicios de urgencias hospitalarios. Emergencias: Revista de la Sociedad Española de Medicina de Urgencias y Emergencias, ISSN 1137-6821, Vol. 25, Nº. 6 (Diciembre), 2013, págs. 430-436.

[6] S. Nitsuwat, W. Paoin. Development of ICD-10-TM Ontology for a Semi-automated Morbidity Coding System in Thailand. Methods Inf Med 2012; 51(06): 519-528.

[7] Pestana Delgado Roberto, Llanos Zavalaga Luis Fernando, Cabello Morales Emilio Andrés, Lecca García Leonid. Concordance between medical diagnosis and informatics coding, considering ICD 10, at the Hospital Nacional Cayetano Heredia, Lima, Peru. Rev Med Hered v.16 n.4 Lima oct./dic. 2005.

[8] Diccionari Clínic project. Funsació TIC Salut Social.
http://www.ticsalut.cat/estandards/terminologia/diccionari-clinic/

[9] SNOMED International webpage. http://www.snomed.org/

[10] SNOMED CT Browser. http://browser.ihtsdotools.org/

[11] SNOMED CT documentation.
.https://confluence.ihtsdotools.org/display/DOC

[12] MongoDB official webpage where there are information and courses. https://www.mongodb.com/webinars

[13] World Health Organization webpage http://www.who.int/es

[14] WHO ICD-10 release and documentation http://www.who.int/classifications/icd/icdonlineversions/en/

[15] ICD-10 in Catalonia (CIM-10) http://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/diagnostics-procediments/cim-10/

[16] ICD-10-MC/PCS in Catalonia (CIM-10-CM/SCP) http://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/diagnostics-procediments/cim-10-mc-scp/