EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

PhD thesis summary

Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque

Olatz Perez de Viñaspre Garralda

2017

EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque

This summary is a shortened and translated version of the dissertation entitled "Osasun-arloko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea", written by Olatz Perez de Viñaspre Garraldak under the supervision of Dr Maite Oronoz Anchordoqui and Dr Jon D. Patrick. It also includes the papers which the candidate has published on the research presented here.

Donostia 2017.

Acknowledgements

The department of Education, Universities and Research of the Basque Government who awarded a pre-doctoral fellowship (BFI-2011-389) to the author of this PhD dissertation to conduct this research.

Contents

1 Presentation of the PhD project

1.1 Introduction and motivation

In the Basque Autonomous Community (*Euskal Autonomia Erkidegoa* or EAE in Basque) Basque and Spanish are co-official languages, and even if sometimes we can speak in Basque with the health workers (specially nurses and physicians), they usually write any report in Spanish. The situation on the rest of regions of the Basque Country is not better, and Spanish and French are the main written languages. Sometimes, this is caused by the lack of habit due to different reasons (language used in education, knowledge of health related terminology...); but some other times, they feel forced to write them in Spanish: for instance, Osakidetza (the Basque Health Service in the Autonomous Community) has a centralised system so any professional that takes care of a patient can access all of her (or his) clinical records written by any other worker. If the following readers are not Basque speakers, the security of the patient can be endangered.

We checked the communication between health workers and patients in other bilingual (or multilingual) countries. In Canada, in the areas where bilingualism is official, patients set the language they want to communicate with (Desjardins, 2003), and so all their health reports are written in the same language. However, in Belgium, each of the language communities has its own non-centralised health service system, and in the case of Brussels (that is a bilingual area), both health services are offered: French and Flemish (Gerkens and Merkur, 2010). Finally in Luxembourg, being as small as

diverse country, even if German, French, Italian, English and Portuguese are widely used languages, they use French as *lingua franca* (European Observatory on Health Care Systems, 1999).

In the Basque Country, the language communities are merged, and the health service is unified inside the Autonomous Community. In the south area of the Basque Country Spanish is used as *lingua franca* and in the north area French, but that does not guarantee the linguistic rights of the Basque speaking patients. Even if Basque is an official language, not all the health workers are able to speak or understand it. The ones that can do it, they speak in Basque to the patients, and take notes in Spanish, translating everything at the moment. In any case, the patient can not get the whole clinical attention in Basque. This work wants to make some steps in this direction, so the clinical attention can be given completely in Basque.

In order to collect the habits and problems they have to work in Basque, we conducted a survey on 45 health workers: we wanted to know how they use Basque in their daily work, and regarding the writing of clinical reports in any language, we wanted to perceive the difficulties they have. Also, in the case of having the choice of writing a clinical record in Basque, we needed to identify the tools they would like to integrate them in a writing assistant tool, such as a spell checker, a health domain terminology dictionary in Basque language, templates...

In the following lines, we will enumerate the conclusions obtained from that survey. Regarding the use of languages, 41 workers said they use Basque to speak to the patients, but only 5 of them used it to write reports. We invited the participants to write a clinical record in Basque and after writing it, those are the difficulties they pointed out: i) lack of specialised terminology in Basque (32 people out of 45), ii) lack of models or examples (24 people) and iii) lack of habit (15 people). We also asked them to enumerate the resources they would appreciate in order to write a clinical record in Basque, and they gave priority to: i) having a unified and complete terminology of the health domain (39 people out of 45), ii) having an in-domain spell checker (30 people) and iii) having models or templates of clinical reports (25 people).

The results of the survey remark the importance of having a unified and complete terminology of the health domain in Basque, in order to enforce the use of Basque among health workers and patients. That has been our main motivation, *to give the first steps in order to have health-science terminology in Basque, in order to enforce the use of Basque in this domain.*

We analysed the terminological systems available for the clinical domain

to select one of them to translate it into Basque. The chosen one has been Systematised Nomenclature Of Medicine – Clinical Terms or SNOMED CT. SNOMED CT is the most comprehensive resource in English, Spanish and other languages that gathers clinical terminology and, in the same way, it is considered the most complete multilingual terminology available until now. Somehow, it is a normalised dictionary that allows the automatic interpretation of clinical records written in different languages and systems, and it defines relationships among its entries. It collects concepts, descriptions and relationships related to clinical records. The original language in SNOMED CT is English and it contains more than 300,000 concepts and more than 1,000,000 descriptions.

The main advantage of SNOMED CT is its wide recall. It gathers concepts used in the clinical domain, including the terminology used in most of the medical specialities. In addition, it specifies relations between concepts, on the one hand relationships that give the hierarchical structure to SNOMED CT and, on the other hand, relationships that extend the semantic information, such as causative-agent, finding site or associated morphology. Besides the internal relationships in SNOMED CT, nowadays SNOMED CT concepts are linked to many other resources. In fact, SNOMED CT is part of the metathesaurus of the Unified Medical Language System or UMLS (Bodenreider, 2004), and, thanks to it, it is linked to the rest of the resources included in that metathesaurus. Having a Basque version of SNOMED CT we will be able to access any technology involving those resources for Basque.

In this PhD project, we designed and inplemented an algorithm to create a Basque version of SNOMED CT automatically. The algorithm includes four steps: i) we use the bilingual and multilingual lexical resources available for Basque of the biomedical domain to translate SNOMED CT terms, ii) we translate English neoclassical terms by means of affix equivalences and transliteration rules (NeoTerm), and as far as the number of words of the terms increases, iii) we translate complex terms by applying some translation patterns based on nested terms (KabiTerm), and iv) we adapt the rule based machine translation system Matxin (Mayor *et al.*, 2011) to the medical domain creating MatxinMed. In addition, in order to fill out the work done, we adapt Xuxen, a spell checker for Basque to the medical domain (XuxenMed), and we developed a prototype to write clinical records in Basque, that can be the basis for the automatic translation of them. In the 1.3 section we will expose in which chapter we explain each of the contributions.

1.2 Objectives

The main objective of this PhD project is *to generate resources to automatically process text from the health domain in Basque*. As seen in the motivation, the basic resource for the processing of specialised texts is terminology. Thus, the main task is to obtain a wide and unified terminology in Basque, using automatic techniques. In any case, this is not the unique objective, and in the following lines we enumerate the secondary objectives:

- **Get to know SNOMED CT in deep**:

  We decided to translate into Basque the terminology of SNOMED CT and thus, it is unavoidable to find out about it, in order to be able to take advantage of its benefits as well as to design its translation to Basque.

- **Unify and get to know the medical lexical resources available for Basque**:

  As in many other domains, in the health science domain there are specialised dictionaries available. One of the concerns of the health workers is the lack of an unified terminology, as having different resources the terminology of reference is confusing. In addition, in order to translate SNOMED CT into Basque we want to reuse all the resources available.

- **Develop systems to automatically create medical terms in Basque**:

  We want to perform the translation regarding the complexity of the terms, that is, we want to start from the one-word terms (simple terms), and use them in order to translate the complex ones. Thus, we want to:

  - **Develop a system to get Basque equivalences from neoclassical terms**:

    Neoclassical terms are composed of morphemes with Greek and Latin origin. Some examples of those neoclassical terms are *hypoglucemia* and *photodermatitis*. Considering their nature (they barely change from one language to other), we want to develop a system to get their Basque equivalences.

  - **Analyze the nature of complex terms and to develop a system to get their Basque equivalent terms**:

    We want to develop a system that will be able to get Basque equivalences for terms of more than one word (complex terms) further from Machine Translation systems, to take advantage of the neoclassical equivalences and lexical resources. In order to analyse the structure of the complex terms, we will develop an analyser to examine SNOMED CT's complex terms.

  - **Adapt a Machine Translation system to the health science domain**:

    We want to adapt an already developed Machine Translation (MT) system to the health science domain, using the new terminology generated with systems developed previously. Thus, we want to analyse whether MT systems are useful for the translation of terminology.

- **To develop a system to manage the translation process of SNOMED CT into Basque**:

  The terminological content of SNOMED CT is very extensive, so we will develop a system that will manage the storage of the information needed for the translation and the translation process.

- **To involve the Basque health science community in the translation process by means of the evaluation**:

  In order to involve the Basque health science community, we will run some evaluation dynamics so we can activate people.

1.3 Outline of this report

This report is the summary of the dissertation entitled "Osasun-arloko terminosorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea". In the following lines we will enumerate the chapters presented in this report as well as a short summary of them.

- 1st – Presentation of the PhD project:

In this first chapter, we present the PhD project's subject, as well as our motivation and objectives. Finally, we enumerate the papers published related to this PhD project.

- 2nd chapter – Analysis of SNOMED CT:

  In order to design the translation of SNOMED CT into Basque, we did a quantitative analysis of SNOMED CT. We analysed the English international version and we pay attention to its hierarchical structure, its terminological richness (number of synonyms) and the number of tokens of each term. Finally, we remark some deficiencies the Spanish version has.

- 3rd chapter – Design of EuSnomed:

  We present the design of the system that will manage the translation of SNOMED CT into Basque called EuSnomed. We expose the algorithm we designed to perform the translation as well as its four main steps. Besides, we explain the framework we adapted for the storage of information in EuSnomed.

- 4th chapter – Sinple terms: lexical resources and neoclassical terms:

  In this chapter we present the first two steps of the translation algorithm. The first step makes use of bilingual lexical resources available for Basque. For the second one, we developed a system for the generation of Basque equivalences from English neoclassical terms called NeoTerm, that makes use of affix equivalences and transliteration rules.

- 5th chapter – Complex terms: nested terms and automatic translators:

  In order to translate complex terms, we expose the last two steps of the algorithm. On the one hand, we developed a system called KabiTerm that is based on nested terms[1]. On the other hand, we adapted a Machine Translator called Matxin to the health science domain. We evaluated both systems by means of a crowd evaluation campaign we called Medbaluatoia.

- 6th chapter – Conclusions and future work:

[1] A nested term is a term that appears inside a complex term.

First, we summarize the conclusions obtained from the work done in the previous chapters and the contributions of this PhD project. Finally, we enumerate the work that will be carried out in the future.

Besides the chapters summarised in this dissertation, the original dissertation written in Basque has two additional chapters regarding the background of the PhD project (2nd chapter) and some resources we developed during the PhD project to show some applications for the terminology created (7th chapter).

1.4 Publications

During the development of this PhD, we published several papers. In the following lines, we show those publications, classified by their corresponding chapter.

- 3rd chapter – Design of EuSnomed:
  - Perez-de-Viñaspre O., and Oronoz M. **Translating SNOMED CT Terminology into a Minor Language**. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, 38–45. Association for Computational Linguistics. Gothenburg, Sweden, 2014.
  - Perez-de-Viñaspre O., and Oronoz M. **An XML Based TBX Framework to Represent Multilingual SNOMED CT for Translation**. *12th Mexican International Conference on Artificial Intelligence, MICAI 2013*. Lecture Notes in Artificial Intelligence, vol. 8265, 419–429. Springer, ISBN 978-3-642-45113-3. Mexico DF, Mexico. 2013.
- 4th chapter – Sinple terms: lexical resources and neoclassical terms:
  - Perez-de-Viñaspre O., Oronoz M., Agirrezabal M., and Lersundi M. **A finite state approach to translate SNOMED CT terms into Basque using medical prefixes and suffixes**. *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, 99–103. St Andrews, Scotland, 2013.

  - Perez-de-Viñaspre O., and Oronoz M. **SNOMED CT in a language isolate: an algorithm for a semiautomatic translation**. *BMC medical informatics and decision making*, volume 15, number 2, S5. BioMed Central. 2015.

- 5th chapter – Complex terms: nested terms and automatic translator:

  - Perez-de-Viñaspre O., and Oronoz M. **Osasun-zientzietako terminologiaren euskaratze automatikoaren ebaluazioa, osasungintzako euskal komunitatea inplikatuz**. *II. IkerGazte, Nazioarteko Ikerketa Euskaraz*. Udako Euskal Unibertsitatea. IruÃśea, Basque Country, 2017.

Even if these other papers do not fit into any chapter, they are also related to the PhD:

- Perez-de-Viñaspre O., Oronoz M., and Patrick J. **Osasun-txosten elebidunak posible ote?** *I. IkerGazte, Nazioarteko Ikerketa Euskaraz*, 730–738. Udako Euskal Unibertsitatea, ISBN 978-84-8438-539-4. Durango, Basque Country, 2015. IkerGazte Special Award.

- Perez-de-Viñaspre O., and Labaka G. **IXA Biomedical Translation System at WMT16 Biomedical Translation Task**. *Proceedings of the First Conference on Machine Translation (WMT16)*, 477–482. Association for Computational Linguistics. Berlin, Germany, 2016.

2 Analysis of SNOMED CT

This chapter makes a deep analysis of SNOMED CT. First, a brief introduction about SNOMED CT is done in section 2.1. Later, in section 2.2 some jobs related to SNOMED CT are described while section 2.3 is devoted to the quantitative analysis of a concrete version of SNOMED CT, the English version released in November of 2015. To finish, we will summarise in section 2.4 the comparison between two versions of SNOMED CT, one in English and the other one in Spanish, a work presented in the master's thesis named *"SNOMED CT sare semantikoa euskaratzeko aplikazioa"* (Perez-de-Viñaspre, 2013). In fact, thanks to the conclusions of that work we chose the English language as source for the translation into Basque.

2.1 Introduction

The ***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms* (SNOMED CT) (IHTSDO, 2014) is considered the most comprehensive, multilingual clinical healthcare terminology in the world[1]. The use of a standard clinical terminology improves the quality and health care by enabling consistent representation of meaning in an electronic health record.

SNOMED CT provides the core terminology for Electronic Health Records (EHR) and contains more than 296,000 active concepts with their descriptions organised into hierarchies. Humphreys *et al.* (1997) showed that

[1]http://www.snomed.org/snomed-ct/why-should-i-get-snomed-ct (accesed May 9, 2017)

SNOMED CT has an acceptable coverage of the terminology needed to record patient conditions. Concepts were defined by means of description logic axioms and were used also to group terms with the same meaning. Christopher G. Chute defined *terminology* as the language labels attached to a concept (Chute 2000). As we mentioned before, each concept in SNOMED CT has one or more labels or *descriptions* that are generally considered as terms.

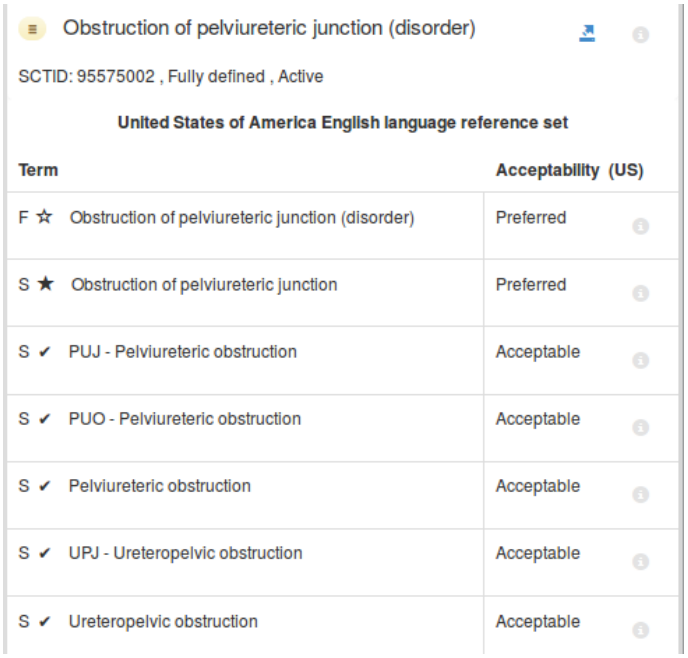
There are two types of descriptions in SNOMED CT: Fully Specified Names (FSN) and Synonyms. Fully Specified Names are the descriptions used to identify the concepts and they have a semantic tag in parenthesis that indicates its semantic type and, consequently, its hierarchy. Synonyms are used to represent a term in a particular language or dialect. Among synonyms there are also Preferred Terms (PT), that is, a particular Synonym that is marked as preferred in such language or dialect. An example is shown in figure 2.1. In the figure, the descriptions for the concept *"95575002 - Obstruction of pelviureteric junction (disorder)"* are shown. The Fully Specified Name is market with an "F" in the first column, the Preferred Term with an "S⋆" in the first column and the word "preferred" in the second column, and the rest of the descriptions are Acceptable Synonyms ("S✓" in the first column and the word "acceptable" in the second column). On the top of the figure, information related to the concept is given, such as the concept identifier ("SCTID: 95575002").

Chute (2000) defined *clinical terminology* as follows:

> *"Standardized terms and their synonyms which record patient findings, circumstances, events, and interventions with sufficient detail to support clinical care, decision support, outcomes research, and quality improvement; and can be efficiently mapped to broader classifications for administrative, regulatory, oversight, and fiscal requirements."*

2.2 Bibliographic Analysis of SNOMED CT

The use of SNOMED CT is recognized in the "Data Analytics with SNOMED CT - Case Studies" report (IHTDSO SNOMED CT, 2015) to *"support data capture, retrieval, and subsequent reused for [...] patient-based queries to operational reporting, public health reporting, strategic planning, predictive medicine and clinical research."* That is, SNOMED CT is presented as a



Obstruction of pelviureteric junction (disorder)

SCTID: 95575002 , Fully defined , Active

United States of America English language reference set

Term	Acceptability (US)
F ☆ Obstruction of pelviureteric junction (disorder)	Preferred
S ★ Obstruction of pelviureteric junction	Preferred
S ✔ PUJ - Pelviureteric obstruction	Acceptable
S ✔ PUO - Pelviureteric obstruction	Acceptable
S ✔ Pelviureteric obstruction	Acceptable
S ✔ UPJ - Ureteropelvic obstruction	Acceptable
S ✔ Ureteropelvic obstruction	Acceptable

Figure 2.1 – Descriptions in SNOMED CT for the concept: *95575002 - Obstruction of pelviureteric junction*. Screen obtained from the SNOMED CT Browser (http://browser.ihtsdotools.org/)

resource that can be used for encoding information and, in consequence, for data analytics (several projects and commercial tools are presented in the report). EHRs are coded using SNOMED CT in several languages.

Several methods have been defined for the construction, maintenance, alignment and evaluation of biomedical ontologies (Yu, 2006). Yu, 2006 explained that these ontologies are useful for i) terminology management, ii) integration, interoperability, and sharing of data and, iii) knowledge reuse and decision support. In our case, after having performed the translation and the manual checking, we aimed to use SNOMED CT mainly to manage and fix the medical terminology in Basque. This is, in our, opinion the first step to face the development of clinical data analysis and decision support tools.

In the literature, terminology systems have been analysed in several ways: from a theoretical point of view, taking as a basis the system itself or comparing it to data extracted from several sources. In Bakhshi-Raiez *et al.*

(2008) a framework is defined towards the standardization of the maintenance of medical terminological systems. It is applicable to all kinds of medical terminologies including SNOMED CT. Campbell *et al.* (2014) investigated whether the terms in SNOMED CT properly represent diagnostic tissue morphologies and notable tissue architectures typically found within pathologists examination and, in this way they tried to identify gaps in expressivity (*Terminology Auditing*). Some desiderata for domain reference ontologies in biomedicine, mainly regarding knowledge processing, were defined in Burgun 2006 by means of an analysis of the *Foundational Model of Anatomy* (FMA) and *Chemical Entities of Biological Interest* (ChEBI) ontologies.

Descriptions about technical and non-technical (e.g. benefits and challenges) aspects of SNOMED CT have been reported in the literature. Regarding the technical ones, Lee *et al.* (2011) reviewed the versioning in SNOMED CT and identified four types of changes that occur over time with new SNOMED CT releases. Descriptions about the use of SNOMED CT in medical "subdomains" have been carried out too. Silva *et al.* (2011) concluded that SNOMED CT has a satisfactory level of representation for its use in the computer tomography domain. In addition, Maheronnaghsh *et al.* (2011) used SNOMED CT in a decision support system that supports decisions about low back pain. Non-technical impressions of direct users about the coverage, concept details and quality of SNOMED CT are collected by means of a survey in Elhanan *et al.* (2011). The 42% of the users perceived that the coverage in SNOMED CT is at least 85% complete and the 60% of the responders were satisfied with its quality. Despite of this, users indicated a desire to improve consistency, quality and completeness of the conceptual representations and an expansion of the coverage.

The analysis of terminology systems lead to the detection of some errors. Indeed, there are works that intentionally apply auditing methods to detect different kinds of errors in medical terminological systems. Jiang and Chute (2009) developed and evaluated an approach that uses a formal concept analysis based model for auditing the semantic completeness of SNOMED CT. They studied the anonymous nodes, nodes without a label for its own object in a concept lattice, as the authors considered that they are useful to identify missing concepts and semantic inconsistencies within a domain. Syntactic regularities and, in consequence, irregularities in SNOMED CT were analysed by Mikroyannidi *et al.* (2012) using the framework RIO for clustering. The authors concluded that there are "design defects" or incomplete

descriptions in SNOMED CT.

SNOMED CT is available in many languages, such as US English, UK English, Spanish, Danish and Swedish. Translations into French, Lithuanian, and several other languages are currently taking place[2]. As referenced in the SNOMED International web-page[3], the translation of SNOMED CT to other languages has been already performed using different techniques. These translations were done using exclusively automatic translation helping systems (this is the case of French Abdoune *et al.* 2011), combining automatic translation and manual work (that is the case of Chinese Zhu *et al.* 2012), or manually (in Danish language, for example Petersen 2011). Schulz *et al.* (2013) compared three kinds of translations from English to German of a set of 500 SNOMED CT terms are compared: i) one translation was performed by professional medical translators, ii) another one used Google Translate[4] and, finally iii) medical students translated the same group of terms.

When two large health care reference terminologies, SNOMED RT and Clinical Terms version 3, were merged to build SNOMED CT, a group of 30 clinical editors were involved in the mapping and the developers review both terminologies to come to an agreement about anatomy models, synonym usage, preferred terms, and so on (Stearns *et al.*, 2001). For example, to represent anatomical entities SNOMED CT uses the "Structure-Entire-Part" modeling approach or SEP. Using this approach the *eye* entity (see figure 2.2) uses the *structure of eye proper* to refer in a general way to the organ, *entire eye* to refer to the complete eye and *eye part* to refer to a part of the organ (explicitly, it does not refer to the entire organ).

2.3 Analysis

The data used for this analysis is obtained from the SNOMED CT International Release Files and we used the English Edition dated on July 2015. This is the version that has been translated in chapters 4 and 5.

In the next section we show the results obtained after making the quantitative analysis of SNOMED CT. The analysed facts are the following ones:

[2]http://www.snomed.org/snomed-ct/snomed-ct-worldwide/translations-of-snomed-ct (accesed May 9, 2017)

[3]http://www.snomed.org/ (accesed May 9, 2017)

[4]http://translate.google.com/ (accesed May 9, 2017)

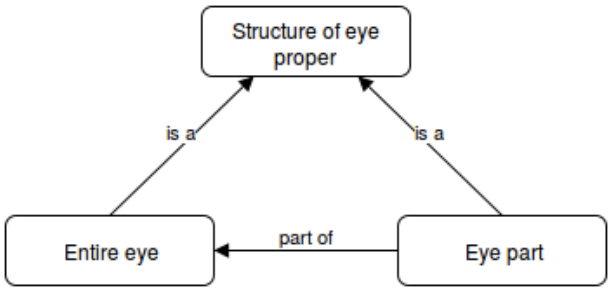


Figure 2.2 – An example of the "Structure-Entire-Part" modelin approach for the "eye" organ.

1. *Hierarchical Structure*: we measure the population in terms of concecpts on each hierarchy and semantic class.
2. *Terminological Richness*: the average number of synonyms that each concept has.
3. *Term Descriptiveness*: we analyse the number of tokens that each description has.
4. *Nested Term Structure*: the structure that nested terms have inside the description is analysed, paying special attention to the most frequent and interesting ones.

2.3.1 Hierarchical Structure

SNOMED CT is organized using a generic concept system. That is, the concepts are linked based on generic relationships (IS-A relationships), and so, the subordinate concept differs from its superordinate concept by some special characteristic. In this way, the most generic concepts are at the top levels and the more specific ones at the lower levels. Thus, a hierarchical structure is given to the concept of SNOMED CT. For example, *"myocardial infarction"* is-a *"myocardial disease"* inside the "Clinical finding" hierarchy.

There are 19 top level hierarchies in SNOMED CT on which all the concepts are allocated. Those hierarchies cover areas of medical information such as "Clinical finding", "Procedure", "Body structure" and so on. But not only medical concepts are represented. SNOMED CT includes hierarchies as "Event", "Environment or geographical location", "Social context",

etc. Thus concepts related to non-clinical information that are relevant in clinical records are also covered.

As mentioned in the Introduction section, each concept in SNOMED CT is uniquely described by a Fully Specified Name. Those special descriptions end with a "semantic tag" in parentheses which indicates the semantic category of the concept. So, the "semantic tag" is used to disambiguate concepts that share the same description and also to infer the hierarchy to which the concept belongs. For example, *"Cyst (disorder)"* is the FSN of the clinical diagnosis when a person has a *"cyst"* whereas *"Cyst (morphologic abnormality)"* is the FSN of the concept representing the *"cyst"* itself.

Table 2.1 shows the population, as regards the number of concepts, of each of the top hierarchies as well as their corresponding semantic tags ordered from the most populated to the less one. The root element is also included in the table by means of the *"SNOMED CT Concept"* hierarchy but we will exclude it from the analysis. The "Concepts" column gives the number of concepts that each semantic tag has, and it is calculated by means of the FSN. The column "Total" refers to the total amount of concepts that each hierarchy has and has been determined by the "is-a" relationships between concepts.

The most populated hierarchy is "Clinical finding" as the table shows, including almost a third part of the total of concepts. This hierarchy comprises concepts that represent the result of a clinical observation, assessment or judment and it divides the concepts between "disorders" and "findings". Being those such a big sub-hierarchies we will distinguish them in the remaining of the paper.

For the 19 hierarchies there are 42 semantic tags nowadays, giving more precise information than the hierarchy does. For instance, the "Body structure" hierarchy has four semantic classes: "body structure", "morphologic abnormality", "cell" and "cell structure". Even so, there are 11 hierarchies (out of 19) where the semantic tag does not give any extra information.

Following with table 2.1, in 4 cases the semantic tag only contains a single element. Those are concepts used to link branches of different "semantic tag". For instance, the "environment / location" semantic tag corresponds to the *"Environment or geographical location (environment / location)"* FSN that is the root concept of the hierarchy with the same name. From this node, two branches are derived: "environment" on the one hand and "geographic location" on the other.

As the reader could notice there are some incoherences. In the row "Phar-

Hierarchy	**Semantic tag**	**Concepts**	**Total**
Clinical finding	disorder	68,839	103,227
	finding	34,388	
Procedure	procedure	52,545	55,128
	regime/therapy	2,583	
Organism	organism	33,036	33,036
Body structure	body structure	25,045	30,871
	morphologic abnormality	4,695	
	cell	627	
	cell structure	504	
Substance	substance	25,828	25,828
Pharmaceutical / biologic product	product	16,931	16,930
Physical object	physical object	14,392	14,393
Qualifier value	qualifier value	9,388	9,388
Observable entity	observable entity	8,410	8,410
Social context	occupation	3,751	4,712
	person	425	
	ethnic group	270	
	religion/philosophy	203	
	social concept	23	
	life style	21	
	racial group	19	
Situation with explicit context	situation	4,159	4,159
Event	event	3,606	3,606
Environment or geographical location	environment	1,197	1,815
	geographic location	617	
	environment / location	1	
Specimen	specimen	1,620	1,620
SNOMED CT Model Component	attribute	1,136	1,547
	namespace concept	185	
	foundation metadata concept	183	
	core metadata concept	33	
	link assertion	8	
	metadata	1	
	linkage concept	1	
Staging and scales	assessment scale	1,110	1,341
	tumor staging	215	
	staging scale	16	
Special concept	navigational concept	640	649
	inactive concept	8	
	special concept	1	
Record artifact	record artifact	225	225
Physical force	physical force	171	171
SNOMED CT Concept	*SNOMED RT+CTV3*	1	1
Total	-	317,057	317,057

Table 2.1 – English hierarchical structure.

maceutical / biologic product" there are 16,931 concepts in its unique semantic tag ("product") and 16,930 concepts in the hierarchy. The opposite occurs with the "Physical object" hierarchy. That exception is related to the concept identifier 440245005. Following the SNOMED CT "is-a" relationships, this concept belongs to the "Physical object" hierarchy (as its parents do) but the semantic tag represented on its FSN is "product"[5].

2.3.2 Terminological Richness

One of the goals of SNOMED CT is to cover most of the clinical concepts used in clinical records. As they claim in the Starter Guide (IHTSDO, 2014) SNOMED CT has *"to ensure understandability, reproduceability and usability"*. Thus, the descriptions should be understandable, acceptable and clinically relevant to the clinicians.

Accordingly, it is important for SNOMED CT terminology to cover the widest range of descriptions possible, so all the terms that describe a concept must be included on it. Humphreys *et al.* (1997) made a study to evaluate the coverage of health data terminologies, and as they highlighted, SNOMED CT has one of the highest scores of exact meanings found: more than 60% of the terms. Yes, the coverage is far from the 100%.

In table 2.2 we offer a general view of the terminological variability of SNOMED CT Synonyms. More specifically, we show the number of Concepts (first column), the number of Synonyms attached to this Concepts (second column), the mean of Synonyms for each Concept (third column) and the median (last column). In this table we did not count the FSN as it is a Description to give an unambiguous way to name a concept.

It might result surprising that, on average, each concept of SNOMED CT has only one Synonym. What is more, the median shows that most of the hierarchies have more than half of the concepts with just one Synonym. "Body structure" is the hierarchy that gets closer to have two synonyms per concept (1.92), and along with "Organism" and "Staging and scales" the median gets a value of 2.

In the case of the "Body structure" hierarchy, we must take into consideration that SNOMED CT uses a modeling approach to represent anatomical entities called "Structure-Entire-Part" triple. This triple sets that "Structure"

[5]http://browser.ihtsdotools.org/?perspective=full&conceptId1=440245005&edition=en-edition&release=v20170131 (accesed May 9, 2017)

concepts are usually named "x structure"; "Entire" concepts "entire x" and "Part" concepts "x part". Thus, the approach generates more Synonyms, as the preferred term should follow the structure set by the model and also the more common way to name the concept. For example, for the anatomical part *"nose"*, the "Structure" concept's Preferred Term is *"Nasal structure"* and *"Nose"* is an Acceptable Synonym. And the same happens to the "Entire" concepts.

As far as the "Organism" hierarchy is concerned, the Fully Specified Name is usually composed of the hierarchy designator, the international taxonomic form and the designation of rank such as Genus, Family, Phylum, ... For instance, for the FSN *"Genus Branchiomyces (organism)"*, *"Branchiomyces"* is the Preferred Term and *"Genus Branchiomyces"* is an Acceptable Synonym. Thus, most of the "Organism" that belong to a class officially recognized as Linnaean taxonomic classes will follow this structure and so will have at least two Synonyms.

Hierarchy	**Concepts**	**Synonyms**	**Mean**	**Median**
Clinical disorder	68,839	114,830	1.67	1
Clinical finding	34,388	52,857	1.54	1
Procedure	55,128	87,104	1.58	1
Organism	33,036	57,582	1.74	2
Body structure	30,871	59,384	1.92	2
Substance	25,828	43,356	1.68	1
Pharmaceutical / biologic product	16,930	25,179	1.49	1
Physical object	14,393	17,838	1.24	1
Qualifier value	9,388	14,440	1.54	1
Observable entity	8,410	13,253	1.58	1
Social context	4,712	5,893	1.25	1
Situation with explicit context	4,159	6,486	1.56	1
Event	3,606	4,404	1.22	1
Environment or geographical location	1,815	2,305	1.27	1
Specimen	1,620	1,982	1.22	1
SNOMED CT Model Component	1,547	1,956	1.26	1
Staging and scales	1,341	2,421	1.81	2
Special concept	649	894	1.38	1
Record artifact	225	284	1.26	1
Physical force	171	272	1.59	1
SNOMED CT Concept	1	4	4.00	4
Total	317,057	512,724	1.62	1

Table 2.2 – English terminological richness.

2.3.3 Term Descriptiveness

The number of tokens on each term may not be a good complexity measurement but it gives us a global idea of the complexity of a description. For example, in terms of complexity, it is not the same a short term like *"Lung cyst"* or a longer one like *"Ruptured emphysematous bleb of lung"*.

What is more, in many cases the number of tokens per term may be related to the descriptiveness of the terms. Very often the more specific the term, the longer it will be, as it will include more characteristics on it. This feature is described in the generic concept system of SNOMED CT. Anyway, we may find differences among synonyms to describe the same concept. For example the concept with the *"Apoptosis (mophologic abnormality)"* FSN has two Synonyms, being *"Apoptosis"* the preferred one (that means *"a falling off"* from Greek[6] and *"Gene-directed cell death"* the Aceptable synonym. As it can be seen, even if the Preferred Term is more concise, the longest one gives more lexical information.

Table 2.3 shows the number of terms with 1 token, 2 tokens,... and more than 8 tokens, repectively. The last three columns give the total amount of Synonyms for each hierarchy, the mean of tokens for each term and the median.

As we can see in table 2.3, on average both the mean and the median are around 4 tokens. From the most populated hierarchies we must underline the ones that stand out from the average: "Organism" and "Substance". They both have a median of two, evidencing the conciseness of the descriptions. Most of the descriptions are formed with between one and seven tokens (more than the 90% of the descriptions). As curiosity, we found two cases on the "Clinical finding" hierarchy with 52 tokens in the description.

Following with table 2.3, it evidences that the group of 2 tokens is the most populated, being almost a quarter of all the descriptions. In any case, if we analyse this data we realize that most of the hierarchies do not follow this pattern, as "Clinical disorder", "Clinical finding" or "Body structure" where the highest column is the corresponding to three tokens, or even "Procedure" with four. The reason why the two token column stands out from the others is the huge difference that the "Organism" hierarchy imposes, as more than the half of its descriptions are formed with two tokens.

[6]https://en.wiktionary.org/wiki/apoptosis (accesed May 9, 2017)

Hierarchy	1	2	3	4	5	6	7	8+	Total	Mean	Median
Clinical disorder	3,863	21,009	25,028	20,732	16,252	10,346	6,748	10,852	114,830	4.37	4
Clinical finding	1,803	8,583	10,953	10,104	8,354	5,187	2,771	5,102	52,857	4.67	4
Procedure	1,996	9,893	15,401	17,049	14,553	10,177	6,776	11,259	87,104	4.90	4
Organism	9,091	32,392	6,346	3,582	1,672	1,453	651	2,395	57,582	2.66	2
Body structure	2,593	10,654	12,700	10,689	9,062	5,978	3,981	3,727	59,384	4.17	4
Substance	8,250	13,917	6,900	6,435	3,274	1,681	1,705	1,194	43,356	3.05	2
Pharmaceutical ...	2,616	2,363	4,987	4,537	3,235	2,379	1,764	3,298	25,179	4.58	4
Physical object	946	3,483	4,228	3,680	2,348	1,340	816	997	17,838	3.97	4
Qualifier value	4,536	4,555	2,760	1,159	717	344	167	202	14,440	2.44	2
Observable entity	459	2,406	3,394	2,739	1,871	1,059	586	739	13,253	4.01	4
Social context	904	2,051	1,179	725	466	256	151	161	5,893	3.03	2
Situation ...	11	401	1,243	1,709	1,272	851	453	546	6,486	4.79	4
Event	67	173	374	485	522	410	389	1,984	4,404	9.89	7
Environment ...	554	752	478	207	93	52	17	152	2,305	2.91	2
Specimen	9	250	572	339	163	167	170	312	1,982	4.81	4
...Model Component	240	522	687	156	252	31	21	47	1,956	3.08	3
Staging and scales	18	119	397	468	411	326	275	407	2,421	5.48	5
Special concept	19	131	220	140	72	122	41	149	894	4.83	4
Record artifact	2	64	53	34	26	8	6	91	284	6.67	4
Physical force	40	127	48	33	17	6	1	0	272	2.57	2
SNOMED CT Concept	0	0	1	0	0	0	0	3	4	16.50	14
Total	38,017	113,845	97,949	85,002	64,632	42,173	27,489	43,617	512,724	4.13	4
Percentage	7.41%	22.20%	19.10%	16.58%	12.61%	8.23%	5.36%	8.51%	100%		

Table 2.3 – English term descriptiveness by means of number of tokens.

2.4 Reasons for choosing the English version as source

When we started with this thesis project, with the aim of choosing the best source language we analysed both the versions in English and in Spanish of SNOMED CT. Two concrete versions were analysed: the English version from the SNOMED CT International Edition of the 31st of November 2012 and the distribution in Spanish from the 30th of April 2012. The comparison was made considering the number of concepts in each hierarchy, the number of preferred terms and the number of concepts without terms (Perez-de-Viñaspre, 2013).

As it can be seen in Table 2.4, in the Spanish version there were a lot of concepts without any description or term (the symbol # represents the number of terms). This table shows the *semantic tags* with lost concepts. A total of 34 different semantic tags lose at least one concept being the order if we classify the classes with a higher lack folowing: *"procedure"*, *"disorder"* and *"finding"* (these are the most populated classes).

We want to remark that at the moment that the analysis was performed, the Spanish version was in development. Although nowadays the development stage has improved a lot, we hope to have made a good decision. In the end, the version in Spanish is a translation of the English version and it is usual to have losses in the translation process. That is, the main reason for having opted for the English version as reference and for the Spanish version to feed the version in Basque as we will see in the chapter 4.

2.5 Summary and conclusions

This chapter describes a quantitative analysis of a specific English version of SNOMED CT. The main goal of this analysis is to study the characteristics of SNOMED CT with the aim of obtaining information that will help us in the transalation into Basque.

First, we analysed the hierchical structure of SNOMED CT by giving the population of the 19th main hierarchies and their semantic tags. This way we concluded that the *clinical finding*, the *procedure*, the *organism* and *body structure* are the most populated ones.

Second, we analysed the terminological richness by counting the number

Semantic tag	#	%	**Semantic tag**	#	%
procedure	11,398	16.17 %	*ethnic group*	83	22.68 %
disorder	10,537	11.31 %	*specimen*	69	4.75 %
finding	8,534	18.91 %	*morphologic abnormality*	54	1.06 %
situation	2,931	33.63 %	*administrative concept*	49	61.25 %
occupation	1,792	27.82 %	*assessment scale*	44	3.99 %
regime/therapy	785	22.05 %	*special concept*	29	96.67 %
substance	652	2.55 %	*staging scale*	25	60.98 %
product	484	1.99 %	*tumor staging*	13	4.96 %
qualifier value	483	4.81 %	*attribute*	10	0.87 %
observable entity	408	4.54 %	*religion/philosophy*	10	4.41 %
physical object	381	6.91 %	*navigational concept*	5	0.69 %
event	377	4.21 %	*cell*	3	0.47 %
person	230	34.74 %	*physical force*	3	1.69 %
body structure	157	0.58 %	*racial group*	2	9.52 %
organism	127	0.36 %	*cell structure*	1	0.19 %
environment	90	7.19 %	*social concept*	1	3.70 %
record artifact	84	26.42 %	Without semantic tag	1,013	-

Table 2.4 – Number of lost concepts for each semantic tag, the number of concepts without description or term in the Spanish version of SNOMED CT. In total, there are 40,864 concepts lost.

of synonyms attached to each concept. As the numbers show, the English version has little variability, in most of the hierarchies the concepts have one synonym (1.62 synonyms in average and with a median of 1).

To follow, we quantified the descriptiveness of the terms considering their number of tokens. As we observed, terms mainly have two tokens but terms with three or four tokens are very usual too (the three groups accumulate the 58% of the total). This fact shows that terms are quite synthetic.

To finish, we made a comparison between a version in English and a version in Sapnish (the version at hand by the time the analysis and the choice was made). We shoed the lacks of the version in Spanish with respect to the version in English. This was due to the fact that the version in Spanish was still in a development phase.

Bearing in mind that the aim is to translate SNOMED CT into Basque, the analysis carried out in this chapter led us to make two decisions: the

of choice as the source language and the hierarchies that should be involved to start the translation. We will take the English version as source and regarding the hierarchies we decided to begin with the most populated ones: clinical findings and disorders, procedures and body structures. Although the organisms hierarchy is more populated than the body structure one, we decided not to translate it as it is special. The IHTDSO organisation in its criteria state that the terms in this hierarchy should not be localised; the taxonomic name should be used. We are going to presents mainly results about the mentioned four categories but we are going to use the developed systems to translate terms from all the hierarchies of SNOMED CT.

3

Design of EuSnomed

EuSnomed is the system we designed for the translation of SNOMED CT into Basque. In this chapter we present the main characteristics of the design. We do not enter on details as it is exposed in the *Language Analysis and Processing* Master's thesis (Perez-de-Viñaspre, 2013). Even if nowadays the systems only works for the translation into Basque, it can be easy adapted to any other language.

We present a four step algorithm to get Basque equivalences from SNOMED CT descriptions. The first step takes advantage of the bilingual/multilingual lexical resources to obtain equivalences. The second one translates neoclassical terms, by means of affix equivalences and transliteration rules. The third step is based on the structure that nested terms conforms to define translation patterns. Finally, the fourth step, it takes a general purpose Machine Translator and adapts it to the health science domain.

EuSnomed makes use of the terminological content of SNOMED CT and reuses the resources generated in the process (term-equivalent pairs). Taking that in consideration, we adapted a standar formalism based on XML called Term-Base eXchange (TBX). By means of this formalism, we are able to manage all the information used for the translation process in an ordered and structured way.

We published two papers related to the work done in this chapter: in **Translating SNOMED CT Terminology into a Minor Language** we present the mentioned algorithm and in **An XML Based TBX Framework to Represent Multilingual SNOMED CT for Translation** we published the format we adapted to store the terminological content of SNO-

MED CT as well as the new descriptions created for Basque, and the information to handle the translation.

Translating SNOMED CT Terminology into a Minor Language

Olatz Perez-de-Viñaspre and Maite Oronoz
IXA NLP Group
University of the Basque Country UPV/EHU
Donostia
{olatz.perezdevinaspre, maite.oronoz}@ehu.es

Abstract

This paper presents the first attempt to semi-automatically translate SNOMED CT (***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms*) terminology content to Basque, a less resourced language. Thus, it would be possible to build a new clinical healthcare terminology for Basque. We have designed the translation algorithm and the first two phases of the algorithm that feed the SNOMED CT's Terminology content, have been implemented (it is composed of four phases). The goal of the translation is twofold: the enforcement of the use of Basque in the bio-sanitary area and the access to a rich multilingual resource in our language.

1 Introduction

SNOMED Clinical Terms (SNOMED CT) (IHTSDO, 2014) is considered the most comprehensive, multilingual clinical healthcare terminology in the world. The use of a standard clinical terminology improves the quality and health care by enabling consistent representation of meaning in an electronic health record[1].

Osakidetza, the Basque Sanitary System ought to provide its service in the two co-official languages of the Basque Autonomous Community, in Spanish and in Basque. However, and being Basque a minority language in front of the powerful Spanish language, the use of Basque in the documentation services (for example in the Electronic Medical Records (EMR)) of Osakidetza, is almost zero. One of our goals in this work is to offer a medical terminology in Basque to the bio-medical personnel to try to enforce the use of Basque in the bio-sanitary area and in this way protect the linguistic rights of patients and doctors. Another objective in this work is to be able to access multilingual medical resources in Basque language. To try to reach the mentioned objectives, we want to semi-automatically translate the terminology content of SNOMED CT focusing in some of its main hierarchies.

To achieve our translation goal, we have defined an algorithm that is based on Natural Language Processing (NLP) techniques and that is composed of four phases. In this paper we show the systems and results obtained when developing the first two phases of the algorithm that, in this case, translates English terms into Basque. The first phase of the algorithm is based on the use of multilingual lexical resources, while the second one uses a finite-state approach to obtain Basque equivalent terms using medical affixes and also transcription rules.

In this paper we will leave aside explanations about i) the translation application, ii) the knowledge management and iii) the knowledge representation, and we will focus on term generation. The application framework that manages the terms has been already developed and it is in use. The knowledge representation schema has been designed and implemented and it is also being used (Perez-de-Viñaspre and Oronoz, 2013).

In the rest of the paper after motivating the work and connecting it to other SNOMED CT translations (sections 2 and 3), the algorithm and the material that are needed to implement the first two phases of the translation-algorithm are described (section 4). After that, results are shown and discussed (sections 5 and 6). Finally, some conclusions and future work are listed in the last section (section 7).

2 Background and significance

"Basque is the ancestral language of the Basque people, who inhabit the Basque Country, a region

[1] http://www.ihtsdo.org/snomed-ct/whysnomedct/snomedfeatures/

spanning an area in northeastern Spain and southwestern France. It is spoken by 27% of Basques in all territories (714,136 out of 2,648,998). Of these, 663,035 live in the Spanish part of the Basque country (Basque Country and Navarre) and the remaining 51,100 live in the French part (Pyrénées-Atlantiques)[2]". Basque is a minority language in its standardization process and persists between two powerful languages, Spanish and French. Although today Basque holds co-official language status in the Basque Autonomous Community, during centuries Basque was not an official language; it was out of educational systems, out of media, and out of industrial environments. Due to this features, the use of the Basque Language in the bio-sanitary system is low. One of the reasons for translating SNOMED CT is to try to increase the use of the Basque language in this area.

SNOMED CT is a multilingual resource as its concepts are linked to terms in different languages by means of a concept identifier. Thus, terms in our language will be linked to terms in all the languages in which SNOMED CT is released. Besides, as SNOMED CT is part of the Metathesaurus of UMLS (Unified Medical Language System (Bodenreider, 2004)), Basque speakers will have the possibility of accessing other lexical medical resources (RxNorm, MeSH) containing the concepts of SNOMED CT.

SNOMED CT has been already translated to other languages using different techniques. These translations were done either manually (this is the case of the Danish language (Petersen, 2011)), combining automatic translation with manual work (in Chinese, for example (Zhu et al., 2012)), or using exclusively an automatic translation helping system (that is the case of French (Abdoune et al., 2011)). In the design of the translation task, we have followed the guidelines for the translation of SNOMED CT (Høy, 2010) published by the IHTSDO as it is recommended.

3 SNOMED CT

SNOMED CT provides the core terminology for electronic health records and contains more than 296,000 active concepts with their descriptions organized into hierarchies. (Humphreys et al., 1997) shows that SNOMED CT has an acceptable coverage of the terminology needed to record patient conditions. Concepts are defined by means of description logic axioms and are used also to group terms with the same meaning. Those descriptions are more generally considered as terms.

There are three types of descriptions in SNOMED CT: Fully Specified Names (FSN), Preferred Terms (PT) and Synonyms. Fully Specified Names are the descriptions used to identify the concepts and they usually have a semantic tag in parenthesis that indicates its semantic type and, consequently, its hierarchy. Regarding what we sometimes refer to as "terms" we can distinguish between PTs and Synonyms.

There are 19 hierarchies to organize the content of SNOMED CT (plus 1 hierarchy for metadata). The concepts of SNOMED CT are grouped into hierarchies as *Clinical finding/disorder*, *Organism*, and so on. For translation purposes it is important to deeply analyze these hierarchies as some of them need to translate all the terms while others as *Organism* only admit the translation of the synonyms (the preferred term should be the taxonomic one). The guidelines for the translation of the hierarchies are given in (Høy, 2010). We want to remark that only the terms classified as PTs and synonyms in SNOMED CT have been taken into consideration for the translation purposes, as the structure (relationships, for example) is the ontological core of SNOMED CT.

Considering the lexical resources available in the bio-sanitary domain for Basque and the SNOMED CT language versions released, two source languages can be used for our translation task: English and Spanish. Basque is classified as a language isolate, and in consequence it is not related to English or Spanish and its linguistic characteristics are far away from both of them. For that reason, no English nor Spanish offers any advantage as translation source. Thus, we deeply analyzed both of them to choose the best option. Our starting point was the Release Format 2 (RF2), Snapshot distributions and the versions dated the 31-07-2012 for English and the 30-10-2012 for Spanish. It must be taken into consideration that the Spanish version of SNOMED CT is a manual translation of the English version.

To choose the source version of SNOMED CT that will be translated, we analyzed aspects as i) general numbers of FSNs, PTs and Synonyms, ii) length of the terms in each language and, ii) the lack of elements in each version. These data help

[2] http://en.wikipedia.org/wiki/Basque_language (January 23, 2014)

us to come to a decision:

1. The number of active concepts in both languages is the same (296,433) as the Spanish version uses the English concept file. Nevertheless, the number of terms in Spanish is significantly smaller. In Spanish 15,715 concepts lack of PTs and Synonyms.

2. Regarding the length of the PTs and synonyms, we counted the terms containing one token, two tokens, three tokens, four tokens and those with more than four tokens. In the English version the 6.76% of the terms has one token, the 23.28% two and the 20.70% three tokens. That is, quite simple terms compose the half of the synonyms in the lexicon. In the Spanish version, nevertheless, only the 33.79% of the synonyms has three tokens or less, and there are 66.21% synonyms with four tokens or more.

Considering these data, we can conclude that i) the English version is more complete and consistent than the Spanish one, and that ii) the terms in the English version are shorter in length and, in consequence, simpler to translate than the ones in the Spanish version. Thus, we decided to use the English version of SNOMED CT as the translation source as starting point.

We fix the priority between hierarchies for the translation taking into account the number of terms in each hierarchy. The most populated hierarchies are *Clinical finding/disorder* (139,643 concepts) and *Procedure* (75,078 concepts). The next most populated hierarchies are *Organism* (35,870 concepts) and *Body Structure* (26,960). The translation guidelines indicate that the PTs of the organisms should not be translated. For this reason and being conscious of our limitation to translate this huge terminology, we decided to prioritize the translation of the *Clinical finding/disorder*, the *Procedure* and the *Body Structure* hierarchies.

4 Translation Algorithm

We have defined a general algorithm that tries to achieve the translation with an incremental approach. Although the design is general and the algorithm could be used for any language pair, some linguistic resources for the source and objective languages are necessary. In our implementation, the algorithm takes a term in English as input and obtains one or more equivalent terms in Basque.

The mapping of SNOMED CT with ICD-10 works at concept level. Thus, before executing the implementation of the algorithm the mapping between them should be done (see section 5).

The algorithm is composed of four main phases. The first two phases are already developed and results regarding quantities are given in section 5. The last two phases will be undertaken in the very near future.

We want to remark that all the processes finish in the step numbered as 4 in the algorithm (see Figure 1). The Basque equivalents with their original English terms, and relative information (for instance, the SNOMED CT concept identifier) are stored in an XML document that follows the TermBase eXchange (TBX) (Melby, 2012) international standard (ISO 30042) as exposed in (Perez-de-Viñaspre and Oronoz, 2013). All the lexical resources are stored in another simpler TBX document called ItzulDB (see number 1 in Figure 1). This document is initialized with all the lexical resources available, such as specialized dictionaries and it is enriched with the new translation pairs generated that overcome a confidence threshold with the intention of using them to translate new terms. In this way we achieve feedback.

Let us describe the main phases:

1. *Lexical knowledge*. In this phase of the algorithm (see numbers 1-2-4 in Figure 1), some specialized dictionaries and the English, Spanish and Basque versions of the International Statistical Classification of Diseases and Related Health in its 10th version (ICD-10) are used. ItzulDB is initialized with all the translation pairs (English-Basque) extracted from different dictionaries of the bio-medical domain and the pairs extracted from the ICD-10. For example the input term "abortus" will be stored with all its Basque equivalents *"abortu"*, *"abortatze"* and *"hilaurtze"*. This XML database is enriched with the new elements that are generated when the algorithm is applied (number 4 in Figure 1). Figure 2 shows an example of some translations obtained using *ItzulDB*.

2. *Morphosemantics*. When a simple term (term with a unique token) is not found in *ItzulDB* (number 3 in Figure 1) it is analyzed at word-level, and some generation-rules are used to

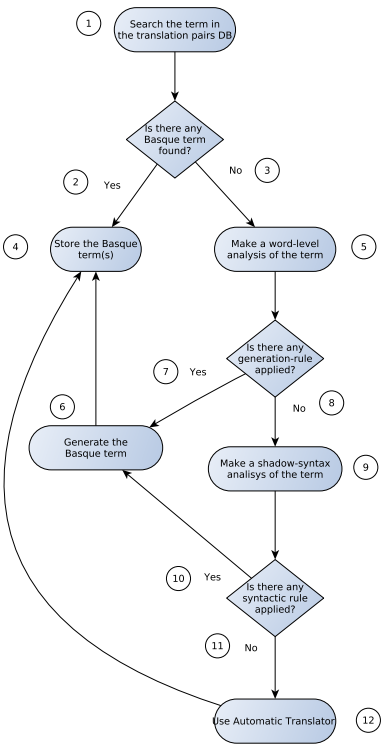


Figure 1: Schema of the Algorithm.

Input term: Deoxyribonuclic acid
Steps in Figure 1 number: 1,2,4
Translation: *Azido desoxirribonukleiko*, *ADN*, *DNA*

Figure 2: Terms obtained from *ItzulDB*.

create the translation. We apply medical suffix and prefix equivalences and morphotactic rules, as well as some transcription rules, for this purpose. This is the case in Figure 3.

Input term: Photodermatitis
Steps in Figure 1 number: 3,5,7,6,4
Applied rules:
Identified parts: photo+dermat+itis
Translated parts: foto+dermat+itis
Translation: *Fotodermatitis*

Figure 3: Terms obtained using generation-rules.

3. *Shallow Syntax*. In the case that the input term does not appear in *ItzulDB* and it can not be generated by word-level rules (number 8 in the algorithm), chunk-level generation rules are used. Our hypothesis is that some chunks of the term will appear in *ItzulDB* with their translation. The application should generate the entire term using the translated components (see example in Figure 4).

Input term: Deoxyribonucleic acid sample
Steps in Figure 1 number: 8, 9, 10, 6, 4
Chunks in *ItzulDB*:
1st chunk: Deoxyribonucleic acid
Basque: *azido desoxirribonukleiko*, *ADN*, *DNA*
2nd chunk: sample
Basque: *lagin*
Translation: *Azido desoxirribonukleikoaren lagin, ADN lagin, DNA lagin*

Figure 4: Terms obtained using chunk-level generation rules.

4. *Machine Translation*. In the last phase, our aim is to use a rule-based automatic translation system called *Matxin* (Mayor et al., 2011) that we want to adapt to the medical domain. Figure 5 shows an attempt of translation with the non adapted translator. For example, *Matxin* translates "colon" as the punctuation mark (*"bi puntu"* or ":") because it lacks the anatomical meaning.

Input term: Partial excision of oesophagus and interposition of colon
Steps in Figure 1 number: 12, 4
Translation: *Esofagoaren zati baten exisiona eta interpositiona bi puntua*

Figure 5: Terms obtained using *Matxin*.

The IHTSDO organization releases a semi-automatic mapping between SNOMED CT and the ICD-10. By identifying the sense of a concept in SNOMED CT, the best semantic space in the ICD-10 for this concept is searched obtaining linked codes. In this way we can obtain the corresponding Basque term for some of the SNOMED CT concepts through ICD-10. Considering that the structures of SNOMED CT and the ICD-10 are quite different, and that the mapping sometimes has "mapping conditions", the use of this

resource has been complex, but fruitful for very specialised terms. Although as we said this mapping is the unique source for obtaining very specialised terms, it should be used carefully as the objectives of SNOMED CT and ICD-10 are different. ICD-10 has classification purposes while SNOMED CT has representation purposes.

A brief description of the first two phases of the algorithm is done in the next subsections (subsections 4.1 and 4.2):

4.1 Phase 1: Lexical Resources

The multilingual specialized dictionaries with English and Basque equivalences that have been used to enrich *ItzulDB* in the first phase of the algorithm are:

- *ZT Dictionary*[3]: This is a dictionary about science and technology that contains areas as medicine, biochemistry, biology... It contains 13,764 English-Basque equivalences.
- *Nursing Dictionary*[4]: It has 5,393 entries in the English-Basque chapter.
- *Glossary of Anatomy*: It contains anatomical terminology (2,578 useful entries) used by University experts in their lectures.
- *ICD-10*[5]: This classification of diseases was translated into Basque in 1996. It is also available in English and in Spanish. The mapping between the different language editions conforming a little dictionary, allowed us to obtain 7,061 equivalences between English and Basque.
- *EuskalTerm*[6]: This terminology bank contains 75,860 entries from wich 26,597 term equivalences are labeled as from the biomedical domain.
- *Elhuyar Dictionary*[7]: This English-Basque dictionary, is a general dictionary that contains 39,164 equivalences from English to Basque.

All these quite different dictionaries have been preprocessed in order to initialize *ItzulDB*. *Elhuyar Dictionary* is a general dictionary that has both not domains pairs but also contains some specialized terminology. This general dictionary will help i) in the translation of not domain terms and ii) also in the translation of the chunks in Phase 3, and thus, on the generation of new terms in Basque.

[3]http://zthiztegia.elhuyar.org
[4]http://www.ehu.es/euskalosasuna/Erizaintza2.pdf
[5]http://www.ehu.es/PAT/Glosarios/GNS10.txt
[6]http://www.euskadi.net/euskalterm
[7]http://hiztegiak.elhuyar.org/en

4.2 Phase 2: Finite State Transducers and Biomedical Affixes

A first approach to this work is presented in (Perez-de-Viñaspre et al., 2013). In that work, finite state transducers described in Foma (Hulden, 2009) are used to automatically identify the affixes in English Medical terms and by means of affix translation pairs, to generate the equivalent terms in Basque. We observed that the behavior of the roots in this type of words is similar to prefixes, so, we will not make distinction between them and we will name them prefixes. A list of 826 prefixes and 143 suffixes with medical meanings was manually translated. An evaluation of the system was performed in a Gold Standard of 885 English-Basque pairs. The Gold Standard was composed of the simple terms that were previously translated in the first phase of the algorithm. A precision of 93% and a recall of 41% were obtained.

In that occasion, only SNOMED CT terms for which all the prefixes and suffixes were identified were translated. For example, terms with the prefix "phat" were not translated as this affix does not appear in the prefixes and suffixes list. For instance, the "hypophosphatemia" term was not translated even though the "hypo", "phos" and "emia" affixes were identified.

We have improved this work by increasing the number of affixes and implementing transcription rules from English/Latin/Greek to Basque.

Figure 6 will help us to get a wider view of the work exposed. The input term "symphysiolysis" is split into the possible affix combination in the first step ("sym+physio+lysis" or "sym+physi+o+lysis"). Then, those affixes are translated by means of its equivalents in Basque ("sim+fisio+lisi" or "sim+fisi+o+lisi"). And finally, by means of morphotactic rules, the well-formed Basque term is composed (in both cases "sinfisiolisi" is generated).

5 Results

Considering the huge size of the descriptions in SNOMED CT and to make the translation pro-

Table 1: Results of the translation.

	Disorder		Finding		Body Structure		Procedure	
	#Synonyms	#Matches	#Synonyms	#Matches	#Synonyms	#Matches	#Synonyms	#Matches
ICD-10 mapping	11,227	-	1,878	-	0	-	0	-
In dictionaries	4,804	3,488	1,836	915	5,896	2,992	778	473
ZT Dictionary	1,104	883	367	311	1,812	1,212	293	253
Nursing Dictionary	437	350	340	245	978	725	199	157
Glossary of Anatomy	3	3	10	8	1,982	1,431	2	2
ICD-10	2,434	2,308	216	195	410	370	5	4
EuskalTerm	906	596	442	306	2,346	1,423	202	155
Elhuyar	299	135	956	300	1,090	367	270	91
Morphosemantics	2,620	2,184	705	578	970	779	1,551	1,362
Total	**17,627**	**5,672**	**4,419**	**1,493**	**6,866**	**3,771**	**2,329**	**1,835**

Input term: symphysiolysis
Identified affixes: sym+physio+lysis, sym+physi+o+lysis
Translation of the affixes: sim+fisio+lisi, sim+fisi+o+lisi
Morphotactics output term: *sinfisiolisi*

Figure 6: Term translated by means of affix equivalences.

cess easy to handle, we have divided it into hierarchies. The *Clinical finding/disorder* hierarchy is specially populated so we have split it considering its semantic tags: *disorders* and *findings*. In addition, the terms from the *Procedure* and *Body Structure* hierarchies have been evaluated too.

Before showing the results, we want to remark some aspects of the evaluation:

- Phase 1: the evaluation has been performed in terms of *quantity*, not of *quality* of the equivalent terms obtained. As the used resources are dictionaries manually generated by lexicographers and domain experts, the quality of the Basque terms is assumed. In any case, and due to the fact that Basque is in its standardization process, the orthographic correctness of the descriptions (see section 6) will be manually checked in the near future.

- Phase 2: the quality of the generated terms could be measured extrapolating the results in the evaluation of the baseline system described in subsection 4.2. That is, 93% precision and 41% recall. The quantity results are shown considering the improvements described in the same subsection.

Table 1 shows the results for the mentioned hierarchies and semantic tags when the translation is performed using both methods: dictionary matching and morphosemantics. Remind that in a previous phase a concept level mapping is completed between SNOMED CT and ICD-10. The first row in Table 1 labeled as "ICD-10 mapping" shows that it is relevant only for the *Clinical disorders and findings* hierarchy, being the *disorder* semantic tag the most benefited one with 11,228 equivalences. The remainder of the results is given at term level.

We made a distinction between the number of obtained Basque terms (1st column, labeled as "#Synonyms") and the number of English terms translated (2nd column, labeled as "#Matches"). Let us see the difference between those two columns looking at the numbers in Table 1. For example, in the *disorder* semantic tag there are 3,488 matches (3,488 original English terms translated), but the number of obtained Basque terms is 4,804 (adding the number of equivalents of all the dictionaries). The reason is that the same input term may have synonyms or even the same equivalent term given by different dictionaries. For example, for the term "allopathy", the same term "alopatia" is obtained in the ZT and Nursing dictionaries (this equivalence will be counted in both ZT and Nursing dictionaries rows).

Table 2 shows the number of tokens in the original English terms. This table refers not to the concepts, but to the terms in the source SNOMED CT in English. The first row shows the number of English terms to which we obtained a Basque equivalent or synonym, the second one the total of English terms and finally, the last row the percentage of translated terms.

Table 3 gives the overall numbers of the translated concepts, in order to take a wide view of the process done.

Let us see the highlights of the results for each

Table 2: Results of the translation regarding the number of tokens of the original term.

		1 token	2 tokens	3 tokens	4 tokens	> 4 tokens	Total
Disorder	**Translated Terms**	3,315	1,114	538	279	426	5,672
	Terms in total	4,066	22,023	24,036	20,005	37,316	107,446
	Percentage	81.53%	5.06%	2.24%	1.40%	1.14%	5.27%
Finding	**Translated Terms**	1,222	158	39	20	54	1,493
	Terms in total	1,830	8,837	10,980	9,814	19,106	50,567
	Percentage	66.78%	1.79%	0.36%	0.20%	0.28%	2.95%
Body Structure	**Translated Terms**	1,942	1,416	334	66	13	3,771
	Terms in total	2,692	11,519	12,575	10,903	21,631	59,320
	Percentage	72.14%	12.29%	2.66%	0.61%	0.06%	6.36%
Procedure	**Translated Terms**	1,741	80	11	2	1	1,835
	Terms in total	1,982	9,966	15,848	16,578	37,695	82,069
	Percentage	87.84%	0.80%	0.07%	0.01%	0.003%	2.24%

Table 3: Overall results.

	Disorder	Finding	Body Structure	Procedure
Translated Concepts	14,125	2,777	3,231	1,502
Concepts in total	65,386	33,204	31,105	82,069
Percentage	21.60%	8.36%	10.39%	1.83%

hierarchy or semantic tag:

- 21.60% of the *disorders* has been translated (see Table 3). This can be considered a very good result. The ICD-10 mapping produces the majority of the translations as it could be expected in this hierarchy (11,227 synonyms obtained). In Table 2 the strength of the morphosemantics phase is evident as the 81.53% of the simple terms is translated.

- The *finding* semantic tag is the most balanced, as no one of the algorithm phase's contribution outlines. The translation of the 8.36% of the concepts is achieved.

- Regarding the results of the *Body Structure* hierarchy, Table 1 shows that the Glossary of Anatomy only contributes in this area. The 10.39% of the concepts get a Basque equivalent.

- In the translation of the *Procedure* hierarchy the dictionaries do not help much as shown in Table 1. In contrast, the mophosemantics contribution allows to translate the 87.84% of the simple terms (see Table 2).

6 Discussion

Some general dictionaries as the ZT dictionary usually contribute in the translation of most of the terms, while more specialized dictionaries only provide translations in the terms related to their domain. For example, both dictionaries, the ZT dictionary and the Nursing dictionary, obtained the Basque terms "mikrozefalia" for "microcephaly" and "metatartso" for "metatarsus". The ICD-10 mapping contributed mainly in the translation of the disorders, and the Glossary of Anatomy in the translation of terms from the Body Structure hierarchy. Sometimes more than an equivalent in Basque is obtained in the translation. For example, for the term "leprosy" we got the equivalents "legen beltz", "legen" and "legenar". Some problems were detected in the Basque terms regarding the standard orthography (the ICD-10 was translated in 1996 and the spelling rules have changed since then) and the form of the word (some obtain the word in finite forms, i.e. "abdomena" for "abdomen" and other in non finite form, "abdomen").

To which the terms generated by finite-state transducers concern, we detected many new affixes from the SNOMED CT terms that do not appear in our lexicon. Even most of those affixes will be correctly transcripted by our transducers, experts insist on enriching the lexicon with new pairs.

7 Conclusions

We have designed a translation algorithm for the multilingual terminology content of SNOMED CT and we have implemented the first two phases. On the one hand, lexical resources feed our database, and on the other hand, Basque equivalents are generated using transducers and medical and biologi-

cal affixes.

Dictionaries provide Basque equivalents of any term length (i.e. unique and multitoken terms) while transducers get as input unique token terms.

In both translation methods results for the most populated hierarchies are shown even though they are applied for all the hierarchies in SNOMED CT. When using lexical resources, results are promising and the contribution of the ICD-10 mapping is remarkable. We obtained the equivalents in Basque of 21.60% of the disorders.

In any case, as we said before, our objective in the future is that specialist in medical terminology can check the quality of the obtained terms and correct them with the help of a domain corpus in Basque. A platform is being developed for this purpose. After the evaluation, and only if it reaches high quality results, our aim is to contact SNOMED CT providers to offer them the result of our work, that at the moment only pertains to the research area.

Regarding the developed systems evaluation, the system used in the first phase extracts English-Basque pairs from dictionaries, so being quite a simple system, does not need of a deep evaluation. A first evaluation of the system that generates terms using medical affixes has been presented. At present, we are evaluating the improvements of this second system with promising results.

In a near future, we want to implement the remainder of the phases in the algorithm: the use of syntax rules for term generation, and the adaptation of the machine translation tool. The promising results in this first approximation encourage us in the way to semi-automatically generate a version in Basque of SNOMED CT.

Acknowledgments

The authors would like to thank Mikel Lersundi for his help. This work was partially supported by the European Commission (325099), the Spanish Ministry of Science and Innovation (TIN2012-38584-C06-02) and the Basque Government (IT344-10 and IE12-333). Olatz Perez-de-Viñaspre's work is funded by a PhD grant from the Basque Governement (BFI-2011-389).

References

Hocine Abdoune, Tayeb Merabti, Stéfan J. Darmoni, and Michel Joubert. 2011. Assisting the Translation of the CORE Subset of SNOMED CT Into French. In Anne Moen, Stig Kjær Andersen, Jos Aarts, and Petter Hurlen, editors, *Studies in Health Technology and Informatics*, volume 169, pages 819–823.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Asta Høy. 2010. Guidelines for Translation of SNOMED CT. Technical Report version 2.0, International Health Terminology Standards Development Organization IHTSDO.

M. Hulden. 2009. Foma: a Finite-State Compiler and Library. In *Proceedings of EACL 2009*, pages 29–32, Stroudsburg, PA, USA.

Betsy L Humphreys, Alexa T McCray, and May L Cheh. 1997. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *Journal of the American Medical Informatics Association*, 4(6):484–500.

International Health Terminology Standards Development Organisation IHTSDO. 2014. SNOMED CT Starter Guide. February 2014. Technical report, International Health Terminology Standards Development Organisation.

Aingeru Mayor, Iñaki Alegria, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an Open-source Rule-based Machine Translation System for Basque. *Machine Translation*, 25:53–82. 10.1007/s10590-011-9092-y.

Alan K. Melby. 2012. Terminology in the Age of Multilingual Corpora. *The Journal of Specialised Translation*, 18:7–29, July.

Olatz Perez-de-Viñaspre and Maite Oronoz. 2013. An XML Based TBX Framework to Represent Multilingual SNOMED CT for Translation. In *Advances in Artificial Intelligence and Its Applications*, pages 419–429. Springer.

Olatz Perez-de-Viñaspre, Maite Oronoz, Manex Agirrezabal, and Mikel Lersundi. 2013. A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes. *Finite State Methods and Natural Language Processing*, page 99.

Palle G. Petersen. 2011. How to Manage the Translation of a Terminology. Presentation at the IHTSDO October 2011 Conference and Showcase, October.

Yanhui Zhu, Huiting Pan, Lei Zhou, Wei Zhao, Ana Chen, Ulrich Andersen, Shuxiang Pan, Lixin Tian, and Jianbo Lei. 2012. Translation and Localization of SNOMED CT in China: A pilot study. *Artificial Intelligence in Medicine*, 54(2):147–149.

An XML Based TBX Framework to Represent Multilingual SNOMED CT for Translation

Olatz Perez-de-Viñaspre
and Maite Oronoz

IXA NLP Group,
University of the Basque Country UPV/EHU
Donostia
operezdevina001@ikasle.ehu.es
maite.oronoz@ehu.es
http://ixa.si.ehu.es

Abstract. In this paper we show an schema to represent the SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) ontology's multilingual terminology. In this case, our objective is the representation of the source SNOMED CT descriptions in English and Spanish, and their translations into the Basque Language. The annotation formalism we defined represents not only the terms but also the metadata needed in order to translate the SNOMED CT descriptions and the information generated from those translations. It has been used to store 276,427 Concepts and 882,003 Descriptions in English, Spanish and Basque. We adapted the TML (Terminological Markup Language) module of the TBX (TermBase eXchange) standard for that purpose. This standard is based on XML.

Keywords: SNOMED CT, XML, Multilingual Medical Terminology, Knowledge Representation

1 Introduction

Correct and suitable representation of data is a necessary task when working in natural language processing in general, and in the gathering of lexical information in particular. In this paper we describe an adaptation of the TermBase eXchange International Standard to represent a clinical terminology called SNOMED CT [3]. According to the multilingual nature of SNOMED CT, each concept has terms in different languages. Thus, we represent multilingual terminological content of the medical domain maintaining the references to its ontological structure. In addition, metadata for describing the information concerning to the translations into Basque is represented too, in the proposed framework.

The "Systematized Nomenclature of Medicine-Clinical Terms" (SNOMED CT) is a comprehensive clinical ontology that provides clinical content for clinical documentation and reporting. It can be used to code, retrieve, and analyze

clinical data. The terminology comprises Concepts, Descriptions and Relationships with the objective of precisely representing clinical information across the scope of health care [6]. SNOMED CT is widely recognized as the leading global clinical terminology for use in Electronic Health Records. It is maintained and developed by an International body: the "International Health Terminology Standards Development Organization" or IHTSDO [5]. Although the SNOMED CT source language is English it has already been translated to other languages like Spanish. There are released guidelines for the translation of it [4].

Being aware of the importance of SNOMED CT for managing and extracting medical information, one of our objectives is the semi-automatic translation of a part of SNOMED CT to our language, Basque. Basque is a minority language spoken in the Basque Country. It is an isolate language, but today holds co-official language status in the Basque Country. It is a highly inflected language with free order of sentence constituents. We know that the general objective of the translation of SNOMED CT is ambitious but at the same time necessary in the process of normalization of the language.

As mentioned before, Basque is a minority language. Thus, the resources to translate SNOMED CT into it are not enough for a manual translation of the Concepts, as recommended in [4]. We propose a semi-automatic translation of its terminology, by means of NLP techniques, so the manual work done by experts will be focused on the validation and correction of the terms generated.

Thus, in this paper we show a formalism to represent the English and Spanish versions of SNOMED CT, as well as the corresponding Basque terms obtained semi-automatically. Furthermore, as our aim is to obtain a multilingual terminology, we use the TermBase eXchange standard, that among others, has been defined for that purpose.

The International Release of SNOMED CT is represented in tab-delimited text. In order to translate its terminological content semi-automatically, additional structured information is needed for the translated Basque terms, such as the translation method or the source term.

SNOMED CT is included in the Metathesaurus of UMLS (***U****nified* ***M****edical* ***L****anguage* ***S****ystem*) [2], and the Methathesaurus, including SNOMED CT, is stored in Rich Release Format (RFF) files or relations of a relational database. Anyway, in order to translate SNOMED CT, we are more interested in its terminology, and less than in the relations between them. Thus, we need a rich representation of SNOMED CT, that allows us to maintain the original structure, but to add structured information for translation purposes.

XML allows to represent data in a semi-structured way. In addition it adds semantics to the data by means of the element tags. In this way, it is possible to adequate the structure of each element to the data related to it. In the near history many standards have been defined in order to represent terminology. It is worth to point out the importance of XML among those standards, being the base of many of them such as XML representation of Lexicons and Terminologies (XLT) [12], Lexical Markup Framework (LMF) (ISO-24613:2008) [7], Dictionary

Markup Language (DML) [10] or TermBase eXchange (TBX) (ISO-30042:2008) [8] [11].

The remainder of the paper is as follows. After this introduction, section 2 describes the main specifications of SNOMED CT and section 3 briefly describes TBX. Section 4 exposes the data model adopted to represent the terminology content of SNOMED CT. Finally, conclusion and future work are exposed in section 5.

2 SNOMED CT structure

The essential components of SNOMED CT are the Concepts, the Descriptions and the Relationships. A Concept is a clinical idea to which a unique SNOMED CT identifier has been assigned. Each Concept is associated to a set of Descriptions, that have representations such as synonyms and translations to different languages. The Concept is logically defined by its Relationships to other Concepts. All the Concepts in SNOMED CT are organized into 19 Top Level Hierarchies with different semantic tags. The semantic tag indicates the semantic category and hierarchy where the Concept belongs. For example, the hierarchy named "Body structure" groups the semantic tags "body structure", "morphologic abnormality", "cell" and "cell structure".

There are three types of Descriptions: *Fully Specified Name* (FSN), *Preferred Term* (PT) and *Synonyms*. The *Fully Specified Name* makes Concepts readable for humans and they usually finish with a "semantic tag" in parenthesis. In the Technical Implementation Guide [6] it is fixed that "a particular language Reference Set will only contain a single FSN" and it will not be used in a clinical record. *Preferred Terms* are common words or phrases used by clinicians to name that Concept. *Synonyms* are terms that are acceptable alternatives to the Preferred Term as a way of expressing a Concept. As specified in the Technical Implementation Guide, "Synonyms and Preferred Terms (unlike FSNs) are not necessarily unique. More than one concept might share the same Preferred term or Synonym".

Table 1 shows an example in which the Concept with the identifier "95575002" has in its upper side the Descriptions of the English version, being the *Fully Specified Name* "Obstruction of pelviureteric junction (disorder)" with the semantic tag *disorder* and the *Preferred Term* "Obstruction of pelviureteric junction". It has five *Synonyms* in English. In the same table 1 we can see that the number of *Synonyms* for the Spanish version is quantitatively different as it has only two of them. In the case of the synonyms it is easy to understand as each language has its own synonym sets.

The terminological information from SNOMED CT, as shown in Table 1, is stored in a TBX framework described in section 3.

Table 1. An example of the Descriptions of the concept "Obstruction of pelviureteric junction".

Concept: 95575002 - Obstruction of pelviureteric junction	
Descriptions in English	
Description	Type
Obstruction of pelviureteric junction (disorder)	FSN
Obstruction of pelviureteric junction	Preferred Term
PUJ - Pelviureteric obstruction	Synonym
PUO - Pelviureteric obstruction	Synonym
Pelviureteric obstruction	Synonym
UPJ - Ureteropelvic obstruction	Synonym
Ureteropelvic obstruction	Synonym
Descriptions in Spanish	
obstrucción de la unión pelviureteral (trastorno)	FSN
obstrucción de la unión pelviureteral	Preferred Term
obstrucción ureteropelviana	Synonym
obstrucción ureteropélvica	Synonym

3 TermBase eXchange standard

TermBase eXchange (TBX) is an International Standard based on XML that defines a framework with the purpose of representing structured terminological data. It is designed to support different processes, such as analysis or descriptive representation. The main objective of TBX is the interchange of terminological data.

In order to support different types of terminological data used in termbases, TBX includes two modules expressed in XML: a core structure and a formalism to identify the terminological data and their constraints (XCS, eXtensible Constraint Specification). This terminological data is expressed by means of data-categories, that are represented as the value of the "type" attribute in XML. The term TBX implies the result of the interaction of both modules.

The XCS mechanism allows the definition of data-categories to adjust to the requirements of different users. By means of defining the data-categories and the constraints among categories, each user-group can define its own TML (Terminological Markup Language) (ISO-11642:2003). This characteristic of TBX is very useful as it is not usual to find terminology collections that share the exact same data-categories. Anyway, TBX provides a default set of data-categories in order to maximize the interoperability among currently existing terminological data. This set of data-categories is defined and constrained by the default XCS. Thus, TBX provides a blind representation mechanism, so users are able to interpret data without consulting providers.

For getting the Basque translation of SNOMED CT there are available among others some specialized glossaries of the bio-medical domain for Basque. Those have been compiled at the University of the Basque Country and have the aim of

reflecting the terminology used by experts in a real context. All these glossaries are gathered in a system called TZOS [1]. TZOS is an On-Line System for Terminology Service that has been designed as a tool for the creation and spreading of Basque terminology in the University environment. TZOS represents those glossaries following the TBX framework, and that is one of the reasons for adapting the TML used in TZOS to represent SNOMED CT terminology content.

4 Our approach

In this section we describe the data model adopted from TZOS to represent the terminology content of the English and Spanish versions of SNOMED CT as well as to include the additional information needed for translation purposes. That is, having as basis the model used in TZOS, we define and use the new data model exposed in the following lines.

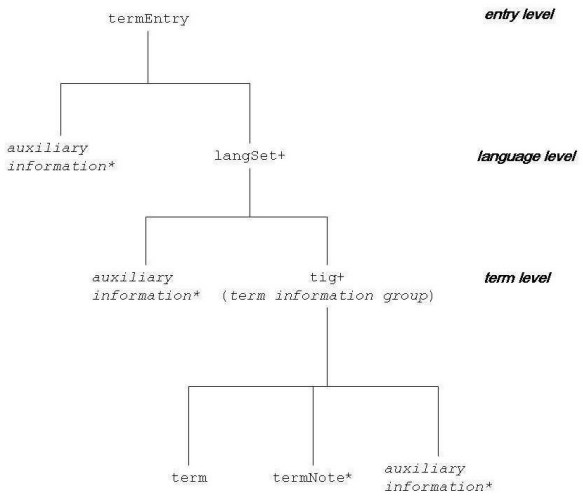


Fig. 1. Structure and levels of a terminological entry in TZOS.

A terminological entry in TZOS has the structure shown in Figure 1. There are three levels having associated its own information:

- *Entry level:* it corresponds to the concept, representing the concept-related information.
- *Language level:* information about the concept is expressed in different languages, that is, a `langSet` for each language (in the figure 1, the symbol + indicates that it can occur once or more times).

– *Term level:* it represents the term itself (`term`) and the associated information by means of a `tig` element. There is a `tig` element for each synonym-term of the concept. By means of the auxiliary information, descriptive features of the term, administrative information or other kind of information can be represented at any level (the symbol * indicates that it can occur 0 or more times).

In the following lines we describe how we reused this structure of TBX in order to represent SNOMED CT Concepts, Descriptions and the corresponding Basque terms.

4.1 Representation of SNOMED CT

For each SNOMED CT Concept we define a `termEntry`. This `termEntry` is complemented by the hierarchy, the semantic tag and the English *Fully Specified Name*. Information regarding the transaction is also represented, such as the person in charge of the generation or edition of the term or the date and time that it was done. This information is helpful in order to manage the database itself.

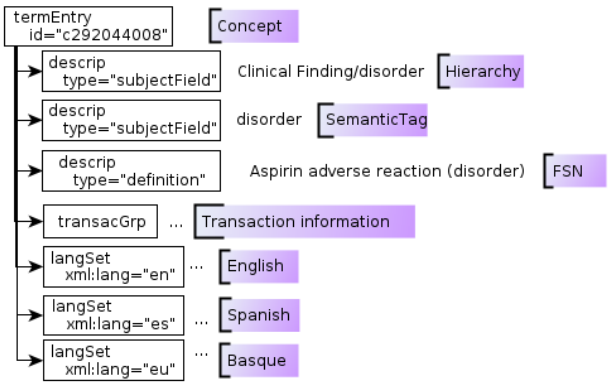


Fig. 2. Structure of a SNOMED CT Concept.

In the Figure 2 we show an example that corresponds to the SNOMED CT Concept with the FSN "Aspirin adverse reaction (disorder)". This information is represented by the data-category *definition* that is the value of the "type" attribute in the `descrip` element. Its Concept identifier is "292044008" (`termEntry` element's id attribute) and the hierarchy to it belongs "Clinical Finding/disorder", being "disorder" the semantic tag (both of them through

the *subjectField* data-category). Regarding the languages, on the one hand, in the English and Spanish language sets (`langSet`) we store the PTs and Synonyms from SNOMED CT; on the other hand, Basque terms are represented with additional descriptive metadata as shown in section 4.2.

We have located the English FSN of the Concept in the entry level instead of in the term level because FSNs are not found in clinical reports and, in consequence, we have decided not to translate them. Furthermore, FSNs are used to identify Concepts so locating them at the entry level is more adequate to SNOMED CT's philosophy.

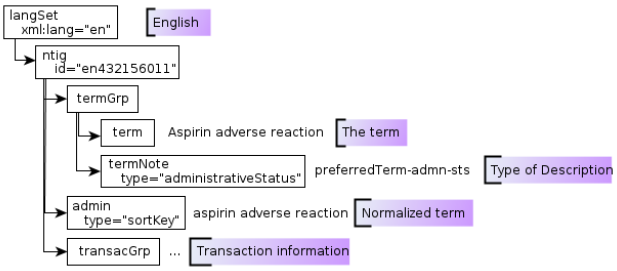


Fig. 3. Structure of an English SNOMED CT term.

To represent a SNOMED CT Description, we store the term itself and also other descriptive metadata. Following the example shown in Figure 3 (that continues the example shown in Figure 2), the term is "Aspirin adverse reaction", its Description type a *Preferred Term* ("preferredTerm-admn-sts" in TBX as the value of the *administrativeStatus* data-category), the normalized term is "aspirin adverse reaction" (through the *sortKey* data-category) and transaction information is also stored.

The design presented in this section allows the representation of the terminology content of SNOMED CT. The structure of SNOMED CT that is obtained by means of Relationships is not represented in TBX but it is accessible using the Concept and Descriptions identifiers in the International Release of SNOMED CT.

4.2 Representation of Basque terms

As mentioned before, our aim is to get Basque terms from the SNOMED CT terminology, so we could obtain a semi-automatic translation to Basque. To represent those terms, we need additional metadata such as the concept source term or the resource from where it has been obtained.

The process to obtain those Basque terms is based on four main resources: i) specialized dictionaries, ii) finite state transducers and word-level morphosemantic rules [13], iii) shadow syntax-level rules and iv) adapted machine translation. The way the term is generated is represented in the *entrySource* data-category. The origin of the SNOMED CT term is represented with the *conceptOrigin* data-category.

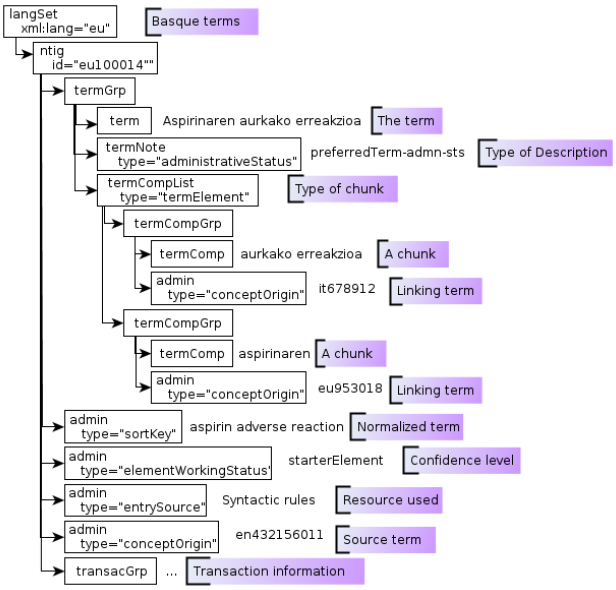


Fig. 4. Structure of a Basque term.

Depending on the resource used, word-level or syntax-level information is also represented. That is, by means of the `termCompList` element, we can split the term into syllables, morphemes, words or chunks. This is useful for the correction of the Basque terms. Let us see the example shown in Figure 4 that complites the previous example (Figures 2 and 3). The Basque word *"aspirinaren"* ("of the aspirin") is linked to the Basque term identified by "eu953018", *"aspirina"*. So, if an expert corrects the Basque term *"aspirina"*, by *"azido azetilsaliziliko"* ("acetylsalicylic acid"), for instance, this is also applied to *"aspirinaren aurkako erreakzioa"* ("aspirin adverse reaction"), obtaining *"azido azetilsalizilikoaren aurkako erreakzioa"* ("acetylsalicylic acid adverse reaction"). This update of changes on the Basque terms is possible thanks to the linking between terms.

In this case, we also need to store the confidence level we give to the obtained term. That is, through the *elementWorkingStatus* data-category we differentiate the Basque term obtained from a specialized dictionary or from word-level or syntax-level rules, giving them a different value. For example, terms extracted from dictionaries have a higher value than the ones generated using finite state transducers because experts have checked them to create the dictionary.

Figure 5 shows the XML representation of the example shown above by the Figures 2, 3 and 4.

5 Conclusions and future work

We defined a formalism to represent SNOMED CT terms through the TBX International Standard. Thus, each SNOMED CT Concept is stored with its English and Spanish terms and also with the Basque terms obtained by a semi-automatic translation process. The role of metadata is essential in order to maintain the relationship to the SNOMED CT ontological graph (we use the Concept and Description identifiers) and also to manage the translation and correction processes (origin of translation, word-level splitting...).

Even the formalism exposed is based on the English and Spanish versions for translating SNOMED CT into Basque, it can be used for any other languages as all the defined adaptations are language independent.

Due to the wide terminology of SNOMED CT the XML generated is so large that we had split it by the hierarchy the concepts belongs to. Thus, we obtain one XML document for each High level SNOMED CT hierarchy. In the case of "Clinical Finding/disorder" the XML document is still too large, so we divided it by the semantic tag ("finding" and "disorder") of the concept's FSN. We have been able to represent 296,427 SNOMED CT Concepts, 476,356 English Descriptions, 379,977 Spanish Descriptions and in the first steps of the translation to Basque, 25,670 terms in Basque.

The first steps towards the semi-automatic translation have been successfully performed obtaining promising results. Regarding the formalism here described, we have probed the appropriateness of it in order to represent SNOMED CT terminology content, as well as to store the Basque terms obtained from the specialized dictionaries.

The XML documents generated following this formalism, are well-formed. Besides, they were validated using both the standard TBX Relax NG schema [9] and the one defined for TZOS. Those two schemes mainly differ in the definition of a new "dateTime" data-category, which allows a better manage of the data generated on-line.

Although the big data already gathered and represented by this formalism asures its robustness, a step forward will be made in a near future by representing the shadow syntax of complex terms.

```
<termEntry id="c292044008">
 <descrip type="subjectField">010</descrip>
 <descrip type="subjectField">011</descrip>
 <descrip type="definition">Aspirin adverse reaction (disorder)</descrip>
 <transacGrp>...</transacGrp>
 <langSet xml:lang="en">
  <ntig id="en432156011">
   <termGrp>
    <term>Aspirin adverse reaction</term>
    <termNote type="administrativeStatus">preferredTerm-admn-sts</termNote>
   </termGrp>
   <admin type="sortKey">aspirin adverse reaction</admin>
   <transacGrp>...</transacGrp>
  </ntig>
 </langSet>
 <langSet xml:lang="es">...</langSet>
 <langSet xml:lang="eu">
  <ntig id="eu100014">
   <termGrp>
    <term>Aspirinaren aurkako erreakzioa</term>
    <termNote type="administrativeStatus">preferredTerm-admn-sts</termNote>
    <termCompList type="termElement">
     <termCompGrp>
      <termComp>aurkako erreakzioa</termComp>
      <admin type="conceptOrigin">it678</admin>
     </termCompGrp>
     <termCompGrp>
      <termComp>aspirinaren</termComp>
      <admin type="conceptOrigin">eu953018</admin>
     </termCompGrp>
    </termCompList>
   </termGrp>
   <admin type="sortKey">aspirinaren aurkako erreakzioa</admin>
   <admin type="elementWorkingStatus">starterElement</admin>
   <admin type="entrySource">102</admin>
   <admin type="conceptOrigin">en432156011</admin>
   <transacGrp>...</transacGrp>
  </ntig>
 </langSet>
</termEntry>
```

Fig. 5. An example of the XML structure.

References

1. Arregi, X., Arruarte, A., Artola, X., Lersundi, M., Zabala, I.: TZOS: An On-Line System for Terminology Service. In: Centro de Lingüística Aplicada, S.d.C. (ed.) Actualizaciones en Comunicacin Social 2013. pp. 400–404 (2013)

2. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32, 267–270 (2004)
3. College of American Pathologists: The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International (1993)
4. Høy, A.: Guidelines for Translation of SNOMED CT. Tech. Rep. version 2.0, International Health Terminology Standards Development Organization IHTSDO (2010)
5. IHTSDO: International Health Terminology Standards Development Organisation (IHTSDO). `http://www.ihtsdo.org/` (2013)
6. International Health Terminology Standards Development Organisation IHTSDO: SNOMED CT Technical Implementation Guide. July 2012 International Release. Tech. rep., International Health Terminology Standards Development Organisation IHTSDO (2012)
7. ISO: Language Resource management – Lexical Markup Framework (LMF). Tech. rep., International Organization for Standarization, Geneva, Switzerland (June 2008)
8. LISA: Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX). Tech. rep., Localization Industry Standards Association (2008)
9. (LISA), L.I.S.A.: TermBase eXchange (TBX). `http://www.ttt.org/oscarstandards/tbx/` (2008)
10. Mangeot, M.: An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language. In: LREC Workshop on International Standards of Terminology and Language Resources Management. pp. 37–44. Las Palmas, Spain (May 2002)
11. Melby, A.K.: Terminology in the age of multilingual corpora. The Journal of Specialised Translation 18, 7–29 (July 2012)
12. SALT project: SALT project – XML representations of Lexicons and Terminologies (XLT) – Default XLT Format (DXLT). Tech. rep., SALT project (2000), reference name of working document: DXLT specification draft 1b
13. de Viaspre, O.P., Oronoz, M., Aguirrezabal, M., Lersundi, M.: A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prexes and Sufxes. In: The 11th International Conference on Finite-State Methods and Natural Language Processing (FSMNLP 2013) (2013)

4

Simple terms: lexical resources and neoclassical terms

In this chapter we explain the techniques developed to generate Basque equivalent terms specially for simple terms (one word terms). On the one hand, we mapped the terms from SNOMED CT with bilingual and multilingual specialised dictionaries from the biomedical domain. In this chapter we include the two papers we published that explain the work done regarding simple terms. In the following lines, we summarise the main conclusions and results about the lexical resources used and the neoclassical term generation approach developed, and right after we attach the corresponding papers. Bare in mind that the original chapter is more extensive, and includes new experiments and results not published yet, that we outline in this summary.

Among the resources used, the ones that gets the best results for the translation of SNOMED CT, evaluated by experts are: i) Science and Technology dictionary (*ZT hiztegia*) with 0.99 of precision, ii) the Basque terminology bank, Euskalterm, with 0.89 of precision and iii) the dictionary of nursing with 0.94 of precision. The Atlas of Human Anatomy did not perform as well as the others in terms of precision, but the contribution made in terms of recall is remarkable with 3,120 new Basque equivalents for the Body Structure hierarchy. The manual evaluation made by experts is not yet published but can be consulted in the original dissertation of the PhD.

On the other hand, we created a system called NeoTerm to translate English neoclassical terms into Basque. We developed three approaches of this system.

The first one is the baseline system, and it is based on the composition of neoclassical affixes. Even if the precision of this approach is high (0.891), the recall is not (0.343) and thus, the priority for the second approach has been the improvement of the recall.

In order to improve the recall, in the second approach we integrated a transliteration module and we extended the dictionaries. Even if we got worse results in terms of precision (8 point less), the recall improved a lot (48 points), and we manage to balance precision and recall obtaining 0.81 of F-measure.

In the last approach, we wanted to improve the identification of neoclassical terms, so NeoTerm can discard better the ones that are not and so avoid errors. We worked on the identification algorithm considering the advises made by the experts. In any case, we did not improve the results got by the second approach and what it is worse, we lost 7 points in recall. Thus, we integrated the second approach of NeoTerm into EuSnomed.

As seen in the results, even if we got a high percentage of translations of simple terms, the numbers regarding complex terms are very high, and so, in the following chapter we will focus on the translation of complex terms.

In the following papers, first of all, we present the work done with the first approach of NeoTerm (**A finite state approach to translate SNOMED CT terms into Basque using medical prefixes and suffixes**). Secondly, we summarise the work done with the lexical resources and the three approaches of NeoTerm, as well as the results obtained from the automatic evaluation (**SNOMED CT in a language isolate: an algorithm for a semiautomatic translation**).

A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes

Olatz Perez-de-Viñaspre, Maite Oronoz, Manex Agirrezabal and Mikel Lersundi
IXA NLP Group
University of the Basque Country UPV/EHU
operezdevina001@ikasle.ehu.es

Abstract

This paper presents a system that generates Basque equivalents to terms that describe disorders in SNOMED CT. This task has been performed using Finite-State transducers and a medical prefixes and suffixes lexicon. This lexicon is composed of English-Basque translation pairs, and it is used both for the identification of the affixes of the English term and for the translation of them into Basque. The translated affixes are composed using morphotactic rules. We evaluated the system with a Gold Standard obtaining promising results (0.93 of precision). This system is part of a more general system which aim is the translation of SNOMED CT into Basque.

1 Introduction

SNOMED Clinical Terms (SNOMED CT) (College of American Pathologists, 1993) is considered the most comprehensive, multilingual clinical healthcare terminology in the world. It does not exist in Basque language, and we think that the semi-automatic translation of SNOMED CT terms into Basque will help to fill the gap of this type of medical terminology in our language. By its translation we have a double objective: i) to offer a medical lexicon in Basque to the bio-medical personnel to try to enforce its use in the bio-sanitary area, and ii) to access multilingual medical resources as the UMLS (***U**nified **M**edical **L**anguage **S**ystem*) (Bodenreider, 2004) in our language.

Basque is a minority language in its standardization process and persists between two powerful languages, Spanish and French. Although today Basque holds co-official language status in the Basque Autonomy Community, during centuries it was out of educational and sanitary systems, media, and industry.

We have defined a general algorithm (see section 2) based on Natural Language Processing (NLP) resources that tries to achieve the translation with an incremental approach. The first step of the algorithm is based on the mapping of some lexical resources and has been already developed. Considering the huge size of SNOMED CT (296,000 active concepts and around 1,000,000 descriptions in the English version dated 31-01-2012) the contribution of the specialized dictionaries has been limited. In the second step that is specified in this paper, we have used Finite State Machines (FSM) in the form of transducers to generate one-word-terms in Basque taking as a basis terms from the English release of SNOMED CT mentioned before. The generation is based on the translation by means of medical suffixes (i.e. *-dipsia*, *-megaly*) and prefixes (i.e. *episio-*, *aesthesi-*) and in their correct composition, considering morphotactic rules. (Lovis et al., 1995) stated that a big group of medical terms can be created by neologisms, that is, concatenations of existing morphosemantic units understood by anybody. This units usually have Greek and Latin origins and their meaning is known by the specialists. (Banay, 1948) specified that about three-fourths of the medical terminology is of Greek origin.

In this work we take advantage of these features to try to translate terms from the *Disorder* subhierarchy of SNOMED CT. This corresponds to one of the 19 top level hierarchies of SNOMED CT, to the one called *Clinical Finding/Disorder*. In our general approach, we prioritized the translation of the most populated hierarchies: *Clinical Finding/Disorder* (139,643 concepts), *Procedure* (75,078 concepts) and *Body Structure* (26,960 concepts). Using lexical resources, we obtained the equivalents in Basque of the 19.32 % of the disorders. In this work we will try to obtain the one-word-terms that are not found in dictionaries.

There are several general-purpose libraries for

the creation of transducers as XFST (Karttunen et al., 1997), Nooj[1] or AT&T's FSM (Mohri et al., 2006). We have used Foma, a free software tool to specify finite-state automata and transducers (Hulden, 2009).

In the rest of the paper the translation algorithm is briefly described in section 2. The use of finite state machines in order to obtain Basque equivalents is explained in section 3. Finally, some conclusions and future work are listed in section 4.

2 Translation of SNOMED CT

The general algorithm (see figure 1) is language-independent. It could be used to translate any term if the linguistic resources for the input and output languages are available.

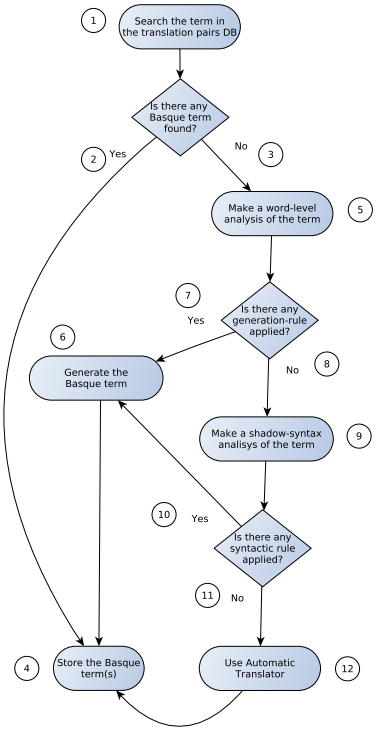


Figure 1: Schema of the Algorithm.

In the first step of the algorithm (see numbers 1-2-4 in Figure 1), some specialized dictionaries and the English, Spanish and Basque versions of the *International Statistical Classification of Diseases and Related Health in its 10th version* (ICD-10) are used. For example for the input term "abortus" all its Basque equivalents *"abortu"*, *"abortatze"* and *"hilaurtze"* are obtained.

The second phase of the algorithm is described in this paper in section 3. When a term is not found in the dictionaries (number 3 in Figure 1) generation-rules are used to create the translation.

In the case that an output is not obtained in the previous phases (number 8 in the algorithm), chunk-level generation rules are used. Our hypothesis is that some chunks of the term will be already translated. The application should generate the entire term using the translated components.

In the last step, we want to adapt a rule-based automatic translation system called *Matxin* (Mayor et al., 2011) to the medical domain.

We want to remark that all the processes finish in the 4th step. That is, we store the generated translations with the intention of using them to translate new terms.

3 Finite-State Models and Translation

This section exposes the system that obtains Basque equivalent terms from English one-word-terms based on FSMs.

3.1 Translation process

The generation of Basque equivalents is performed in two phases: the identification of the affixes first, and the translation and composition of the affixes secondly. All the linguistic information is stored in lexica and 31 rules are written for the process (1 for identification, 1 for translation and 28 for morphotactics).

Figure 2 shows the Finite State Transducer for the identification of the affixes. The lexica of the affixes is loaded (1-6) and then any prefix (the "*" symbol indicates 0 or more times) followed by one unique suffix is identified. The letter "o" may be also identified as it is used to join medical affixes. The "+" symbol is used for splitting the term.

```
1 read lexc prefixes.lex
2 define PREFALL
3 define PREF PREFALL.u ;
4 read lexc suffixes.lex
5 define SUFALL
6 define SUFF SUFALL.u ;
7 regex [[[PREF 0:%+] (o 0:%+)]* SUFF] ;
```

Figure 2: Rules for the affix identification.

The combination of the finite state transducers

[1] http://www.nooj4nlp.net/NooJManual.pdf

for the translation and for the composition using morphotactics is shown in Figure 3. First, the lexica for the translation task is loaded (1-4), then 28 rules for the morphotactics are defined (simplified in the rule numbered 5). The translation rule (shown in rule number 6) is composed of the word-start mark (the ^ symbol), the prefix followed by the optional linking "o" letter zero or more times, and a single compulsory suffix; finally the transducer combines the translation and the morphotactic finite state transducers (7).

```
1 read lexc prefixes.lex
2 define TRANSPRE
3 read lexc suffixes.lex
4 define TRANSSUF
5 define MORPHO ...
6 define TRANS (%^ ) [[[TRANSPRE %+] (o:o %+)]*
TRANSSUF] ;
7 regex TRANS .o.  MORPH ;
```

Figure 3: Rules for the affix translation.

Figure 4 shows the whole process with an example. First, we identify the prefixes and suffixes of the English input term by means of the transducer that marks those affixes (schiz+encephal+y). Then, we obtain the corresponding Basque equivalent for each part and we form the term (eskiz+entzefal+ia).

Input term: schizencephaly
Identified affixes: schiz+encephal+y
Translated affixes: *eskiz+entzefal+ia*
Output. Basque term: *eskizentzefalia*

Figure 4: Basque term generation.

As we said before, in order to obtain a well formed Basque term, we apply different morphotactic rules. For example, in Basque, words starting with the "r" letter are not allowed, and an "e" is needed at the beginning. Figure 5 shows an example where the translated prefix *"radio"* needs of the mentioned rule, obtaining *"erradio"*.

Input term: radionecrosis
Identified affixes: radio+necr+osis
Translated affixes: *radio+nekr+osi*
Basque term: *erradionekrosi*

Figure 5: Morphotactic rule application.

3.2 Resources

In order to identify the English medical suffixes and prefixes we have joined two lists: the "Medical Prefixes, Suffixes, and Combining Forms" from Stedman's Medical Dictionary (Stedman's, 2005) and the "List of medical roots, suffixes and prefixes" from Wikipedia (Wikipedia, 2013). We obtained a list of 826 prefixes and 143 suffixes.

For the translation task, we have manually checked the Basque equivalents of the previously mentioned medical suffixes and prefixes list in specialized dictionaries such as *Zientzia eta Teknologiaren Hiztegi Entziklopedikoa* (Dictionary of Science and Technology) (Elhuyar, 2009), Euskalterm (UZEI, 2004) and *Erizaintzako Hiztegia* (Nursing Dictionary) (EHUko Euskara Zerbitzua and Donostiako Erizaintza Eskola, 2005).

By means of checking the behavior of the prefixes and suffixes in the English and Basque terms we have manually deduced the appropriate Basque equivalent. Table 1 shows an example of obtaining the equivalent of the "encephal" prefix, deducing that *"entzefal"* is the most appropriate equivalent.

English terms	**Basque terms**
echo**encephal**ogram	*eko**entzefal**ograma*
encephalitis	***entzefal**itis*
encephalomyelitis	***entzefal**omielitis*
leuko**encephal**itis	*leuko**entzefal**itis*
...	...

Table 1: The translation of the "encephal" prefix.

From all the prefixes and suffixes listed, we are able to deduce 812 prefixes and 139 suffixes for Basque. Those are currently being supervised by an expert to give them the highest confidence possible. This technique allows the inferring of new medical terms not appearing in dictionaries.

3.3 Results

We selected the one-word-terms of the *Disorder* sub-hierarchy of SNOMED CT. This sub-hierarchy with terms representing disorders or diseases is formed by 107,448 descriptions, being 3,979 one-word-terms. Even this last quantity is low considering the whole sub-hierarchy, we must take into account that the influence of those one-word-terms is very high, appearing around 79,000 times among all the descriptions.

The total one-word-term set has been split into two sets, one for defining and developing the system and another one for evaluating it. The evaluation set is composed of the 885 one-word-terms that have been previously translated in the first

step of the algorithm (see section 2). That is we have the correct English-Basque pairs as Gold Standard. For the development set we have selected the remaining 3,094 one-word-terms.

As mentioned before, in this paper we show the results obtained from the translation of the medical prefixes and suffixes forming the terms. That is, we have only translated the terms that have been completely identified with the medical prefixes and suffixes. For example, terms with the suffix "thorax" have not been translated as it does not appear in the prefixes and suffixes list. That is, the "hydropneumothorax" term has not been translated even though the "hydro" and "pneumo" prefixes have been identified.

In Table 2 we show the quantities and percentages of the terms that have been completely identified in both sets. Our set of the one-word-terms has not been cleaned up to remove the words without any medical affix. Thus, the percentages from the table will never reach 100 per cent.

	Total	**Identified**	**Percent**
Development	3,094	834	26.96%
Evaluation	885	309	34.92%

Table 2: Quantities of completely identified terms.

From the 885 terms in the evaluation set, 728 terms contain at least one medical prefix or suffix, being 309 completely identified. The results obtained in this fist approach are shown in Table 3 by means of True Positives (TP), False Negatives (FN), False Positives (FP), Precision (Prec.), Recall (Rec.) and F-measure (F-M). A recall of 0.41 is obtained (287 correctly identified from 706 TP and FN) and a precision of 0.93 (287 out of 309). The recall will be increased in the future, including not completely identified terms in the system. Thus, we can conclude that the results obtained are very good concerning precision.

Total	**TP**	**FN**	**FP**	**Prec.**	**Rec.**	**F-M**
728	287	419	22	0.93	0.41	0.56

Table 3: Precision and recall of the evaluating set.

Moreover, the quality of the results obtained is also very good. We have been able to give correct equivalents to complex terms such as "hyperprolactinemia", that has five medical prefixes and suffixes ("hyper+pro+lact+in+emia").

We have also analyzed the incorrect results in order to be able to improve the system. For example, the prefix "myc" has been translated as "miz", but we realized that whenever the prefix is followed by an "o", it should be "mik" in order to generate a correct Basque term. Many of the mistakes are easily rectifiable for the final purpose of translating SNOMED CT.

4 Conclusions and future work

We implemented an application that generates Basque terms for diseases in English, by means of finite-state transducers. This application is one of the phases in the way to translate SNOMED CT into Basque. In order to translate the medical prefixes and suffixes, we have manually generated the translation pairs for 951 prefixes and suffixes, obtaining a very useful resource for Basque.

The FSTs exposed in this paper could be easily applicably to other languages whether an affix lexicon with its translation is defined and the morphotactic rules adapted to the target language.

As we have seen in section 3.3, most of the English terms have not been identified completely and that prevented the translation of them. To cope with this problem we have two developing paths: the deduction of new suffixes and prefixes from specialized dictionaries (Hulden et al., 2011); and the implementation of transliteration transformations to those parts (Alegria et al., 2006).

We have only applied the transducers to the *Disorder* sub-hierarchy, and we will have to check the results we can obtain applying it to the *Finding* sub-hierarchy and to the *Procedure* and *Body Structure* hierarchies. We found terms such as "electroencephalography" or "oligomenorrhea" in those hierarchies, formed with medical prefixes and suffixes identified for this task.

The promising results obtained will contribute to the translation of the whole SNOMED CT, but also to the normalization of Basque in the biosanitary domain, as new terms are generated.

References

I. Alegria, N. Ezeiza, and I. Fernandez. 2006. Named Entities Translation Based on Comparable Corpora. In *Multi-Word-Expressions in a Multilingual Context Workshop on EACL06*, pages 1–8.

G. Banay. 1948. An introduction to medical terminology, Greek and Latin derivations. *Bulletin of the Medical Library Association*, 36(1):1–27, Jan.

O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270.

College of American Pathologists. 1993. The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International.

EHUko Euskara Zerbitzua and Donostiako Erizaintza Eskola. 2005. *Erizaintzako Hiztegia*. EHU. Argitalpen Zerbitzua.

Elhuyar. 2009. *Elhuyar Zientzia eta Teknologiaren Hiztegi Entziklopedikoa*. Elhuyar Edizioak & Euskal Herriko Unibertsitatea.

M. Hulden, I. Alegria, I. Etxeberria, and M. Maritxalar. 2011. Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *EMLP 2011: Dialects2011*, pages 39–48.

M. Hulden. 2009. Foma: a Finite-State Compiler and Library. In *Proceedings of EACL 2009*, pages 29–32, Stroudsburg, PA, USA.

L. Karttunen, T. Gaál, and A. Kempe. 1997. *Xerox Finite State Tool*.

C. Lovis, Pa. Michel, R. Baud, and Jr. Scherrer. 1995. Word Segmentation Processing: A Way To Exponentially Extend Medical Dictionaries. *MEDINFO*, 8:28–32.

A. Mayor, I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola. 2011. Matxin, an Open-source Rule-based Machine Translation System for Basque. *Machine Translation*, 25:53–82.

M. Mohri, F. Pereira, M. Riley, and C. Allauzen. 2006. AT & T FSM Library Finite-State Machine Library. Technical report, AT&T Labs-Research, NJ, USA.

Stedman's, 2005. *Stedman's Medical Dictionary*, chapter Medical Prefixes, Suffixes, and Combining Forms. Lippincott Williams & Wilkins, twenty-eighth edition edition.

UZEI. 2004. Euskalterm Terminologia Banku Publikoa. `http://www.euskadi.net/euskalterm`.

Wikipedia. 2013. List of medical roots, suffixes and prefixes – Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=List_of_medical_roots,_suffixes_and_prefixeso`.

RESEARCH

SNOMED CT in a language isolate: an algorithm for a semiautomatic translation

Olatz Perez-de-Viñaspre[1*†] and Maite Oronoz[1†]

*Correspondence:
olatz.perezdevinaspre@ehu.es
[1]IXA NLP Group, University of the Basque Country UPV/EHU , Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: The ***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms* (SNOMED CT) is officially released in English and Spanish. In the Basque Autonomous Community two languages, Spanish and Basque, are official. The first attempt to semi-automatically translate the SNOMED CT terminology content to Basque, a less resourced language is presented in this paper.

Methods: A translation algorithm that has its basis in Natural Language Processing methods has been designed and partially implemented. The algorithm comprises four phases from which the first two have been implemented and quantitatively evaluated.

Results: Results are promising as we obtained the equivalents in Basque of 21.41% of the disorder terms of the English SNOMED CT release. As the methods developed are focused on that hierarchy, the results in other hierarchies are lower (12.57% for body structure descriptions, 8.80% for findings and 3% for procedures).

Conclusions: We are in the way to reach two of our objectives when translating SNOMED CT to Basque: to use our language to access rich multilingual resources and to strengthen the use of the Basque language in the biomedical area.

Keywords: SNOMED CT translation; Basque Language Isolate; Natural Language Processing; Finite State Transducers

Introduction

SNOMED Clinical Terms (SNOMED CT) [1] is widely recognized as the most comprehensive, multilingual clinical health-care lexicon. By using SNOMED CT in electronic health records the consistency of the representation improves, benefiting in this way individuals [2].

The two co-official languages in the Basque Autonomous Community, Spanish and Basque, should be used in Osakidetza (the Basque Sanitary System). Even though in Osakidetza the two languages are used, Spanish is a much stronger language and Basque is hardly used in the documentation services. In 2005 Osakidetza approved its first *Basque Scheme to normalize the use of the Basque language in Osakidetza* for the period 2005-2012. In the evaluation of this plan [3] they concluded that the greatest progress in the use of Basque was done in the area of language profiles (accreditation of language profiles, jobs with mandatory Basque knowledge etc.). For the second scheme (period 2013-2019) [3] one of the hubs that needs to be strengthened is the use of Basque in the documents: “emphasis should be placed on the documents of a care nature through the normalizing

and systematizing of bilingual models of documents, bearing in mind that their adaptation or production in Basque must be facilitated and simplified by professionals. In parallel, in order to have bilingual clinical records available, an in-depth study must be started without further delay and the aspects that influence the process to create and exploit information must be analyzed". However, writing bilingual clinical records can be a tedious work for doctors and what is more, a misuse of their time. The alternative solution, the translation of the records by professional translators could be very expensive.

As far as we know, in other bilingual countries like Canada, the communication language between patient and doctor is established on demand by the patient [4]. This is not possible in our scenario, as some doctors are not able to understand Basque. As the safety of the patients cannot be put at risk, the comprehension of previous or current clinical records is essential for every health professional. Therefore, Spanish is the only language used nowadays for documentation. But this fact produces a complex scenario in which Basque language is isolated as a merely oral interaction tool, and doctors develop the ability to translate from verbal communication in Basque to written notes in Spanish. Nowadays, patients do not have the option of having their medical records in Basque. In a normalized scenario, the clinical notes would be written in the preferred language set by patient-doctor communication.

We agree with the statement that "the summarized clinical history of any patient should be at least in the two co-official languages to assure the security of their assistance". This statement has been made by an Osakidetza committee that has the objective of giving recommendations and analyze whether a bilingual clinical records system is possible or not. In this context, a multilingual version of SNOMED CT including Basque will help to produce such bilingual (or even multilingual) clinical records. That is, by means of a "text to SNOMED" matching tool and a multilingual terminology service based on SNOMED CT, we designed a prototype to help doctors writing clinical records in Basque. By means of a fast and easy disambiguation process of the most relevant medical terms in the record written in Basque, the prototype produces in the present stage of development a minimal Spanish version of the terminological content. The prototype also incorporates a spell-checker adapted to the medical domain based on the specialized terminology from the biomedical domain. The prototype is still in a very early phase and the Basque SNOMED CT terminological content must be completed and manually checked, but it shows i) a use case for the work we present in this paper and, ii) that the creation of a tool to help in the writing of medical records in Basque is feasible. We have presented this prototype to the committee mentioned before with a very positive feedback. The decision about writing medical records in Basque is out of the scope of the scientific discussion and will be taken by the mentioned committee.

In conclusion, one of our goals in this work is to try to enforce the use of Basque in the biomedical area by offering to the medical personnel a standard medical terminology and thus, to safeguard patients and doctors linguistic rights. As mentioned, another goal is to attain multilingual medical resources in the Basque language. These objectives can be reached, in our opinion, by semi-automatically translating the terminology content of SNOMED CT. We will focus on the most populated SNOMED CT hierarchies.

To translate the terminological content of SNOMED CT, we have defined a four phase algorithm that is based on Natural Language Processing (NLP) techniques and that is presented in [6]. In that paper we outlined the main ideas of the translation algorithm and

the implementation of the first two phases (out of four) as well as the Phase 0 about the mapping between SNOMED CT and ICD-10. In the current paper we extend the explanation and we improve the base system. We also expose some new experiments and the corresponding results.

Multilingual lexical resources are the source of information in the implementation of the first phase of the algorithm, while a finite-state approach that uses medical affixes together with transcription rules in order to obtain clinical terms in Basque, is used in the second phase. In both approaches, we use mainly English as source language and in the first phase we also used Spanish-Basque dictionaries to complement the information sources available.

Regarding the third phase which aim is the translation of complex terms, we are analyzing their nature in the English version of SNOMED CT and we found out that many different terms share a specific structure. In Table 3 we show some of the most obvious structures or patterns found from shallow experiments. For instance, there are 1,498 terms with the structure ``[PHARMPRODUCT|SUBSTANCE] + allergy'', that is, a pharmacological product or a substance followed by "allergy", like "urokinase allergy", "cortisone allergy" or "phentolamine allergy". Our hypothesis is based on the evidence that we have already the translations of some chunks within the complex term. In this step the translation application should generate the Basque equivalences using the already translated components and some generation rules.

The fourth and last step will adapt a rule-based automatic translation system called *Matxin* [24] to the medical domain.

Issues as i) the design and implementation of the translation application, ii) the way we manage the terminology and, iii) the representation of the terminological content as metadata (knowledge representation), are not addressed in this paper. Term generation is the main subject of this paper. The translation software framework we use to manage the terms is alredy developed and operative. The schema for knowledge representation is designed and is also in use [5].

The current article is an extended version of the work published in *The Fifth International Workshop on Health Text Mining and Information Analysis (Louhi 2014)* conference [6]. The main novel aspects are, i) an extended introduction and motivation of the work exposed, ii) the inclusion of Spanish-Basque lexical resources, iii) a detailed explanation about new approaches developed for generating simple Basque terms, iv) a detailed description of the finite-state transducers used in the algorithm, v) a more detailed evaluation of the phases already developed of the algorithm and, vi) a table quantifying the number of concepts from SNOMED CT in each of the hierarchies and semantic classes (English and Spanish versions).

The remainder of this paper is arranged as follows: first, a Background section where we justify the work and relate it to other SNOMED CT translations. In the Methods section we focus on the implementation of the first two phases of the translation algorithm. Finally, Results are presented and discussed, and the Conclusions and future lines of this work are listed.

Background

"Basque language, also called Euskara or Euskera, language isolate, the only remnant of the languages spoken in south-western Europe before the region was Romanized in the 2nd through 1st century BCE. The Basque language is predominantly used in an area comprising approximately 3,900 square miles (10,000 square kilometres) in Spain and France"[7].

It is spoken in the Basque Country, a region placed in the northeastern part of Spain and in the southwestern part of France. Basque is a minority language that persists between Spanish and French, two powerful languages. Today Basque is in its standardization process and holds co-official language status in the Basque Autonomous Community but during centuries it was excluded from educational systems, media, and industrial environments. Nowadays, in the Basque Autonomy Community 36.4% of the population knows and uses well the Basque language (30 years ago was 22%); 19.3% is Basque receiver, that is, this percentage of the population understands and reads the language but cannot write or speak it; and, 44.3% do not know the language (30 years ago two thirds of the population was in this situation). There are 749,182 Basque speakers, 318,000 more than in 1981. That is, as mentioned in the *fifth sociolinguistic map of the Basque Autonomous Community* [8] the number of Basque speakers has increased in the working world and the age-range where it has increased the most is in people that are less than 20 years old. Even though the data shows that the use of Basque is increasing, primarily between young people, these people are not in the labor market yet. Due to all these characteristics, the Basque Language in the health system has very low use. With this work we aim at facilitating the use of the Basque language in the biomedical area.

In SNOMED CT concepts are linked to terms in different languages by means of concept identifiers, which makes of SNOMED CT a multilingual resource. With a Basque version of SNOMED CT, we can obtain the terms in our language linked to terms in all the languages represented in SNOMED CT. Besides, SNOMED CT is part of the Metathesaurus of UMLS (Unified Medical Language System [9]), so other lexical medical resources containing SNOMED CT concepts (RxNorm, MeSH...) can be accessed by Basque speakers.

SNOMED CT has been widely used with commercial as well as research purposes. In 2006 a survey on health information technology (HIT) vendors was carried out [10] in order to study the predominance of SNOMED CT in electronic medical records (EHRs). The authors of the study concluded that the respondents who were already working with SNOMED CT increased its use in EHRs for clinical decision support, encoding of healthcare data, health information exchange and patient assessment. Posterior surveys [11] on vendors indicated that although SNOMED CT is highly used in production systems, most of these uses are elementary and do not benefit from the rich semantics of the terminology. In the *SNOMED in Action* initiative [12] several uses of this terminology are listed. Among others it is used to document diagnoses and problems for ambulatory clinic patients, evidence-based medicine, and so on.

One of the strengths of SNOMED CT is its nature as a standard. As it is pointed in [13] "it's software aimed at eliminating potentially dangerous misunderstandings over what medical terms actually mean to different clinicians, researchers, and even to patients". In [14] how SNOMED CT is implemented in 12 health-care organizations across eight countries was studied by means of a survey that took into account design, use and maintenance issues. After this survey they described the advantages of using SNOMED CT as i) clinicians can record the exact diagnosis making use of the large number of synonyms available, ii) via SNOMED CT International Classification of Diseases (ICD) codes can very easily be generated, iii) SNOMED CT offers clinicians the best coverage to describe their use cases and, iv) its standard nature makes patients' records legible.

The paper entitled "The need for SNOMED CT translations" [15] aims at promoting "a discussion about the European wide availability of language-specific SNOMED CT translations" because the authors think that "Language-specific translations of SNOMED CT

are necessary for bringing value-added applications into clinical routine in non-English speaking countries". The authors of the paper recommend the introduction of SNOMED CT across Europe, with special emphasis on German as the largest language group. We agree with the introduction of SNOMED CT but we want to remark that also minority languages should be considered if they want to survive, which is one of the reasons why we are interested in working with Basque.

"Today, SNOMED CT is available in US English, UK English, Spanish, Danish and Swedish. Translations into French, Lithuanian, and several other languages are currently taking place" [16]. As referenced in the IHTSDO web-page, the translation of SNOMED CT to other languages has been already performed using different techniques. These translations were done using exclusively automatic translation helping systems (this is the case of French [19]), combining automatic translation and manual work (that is the case of Chinese [18]), or manually (in Danish language for example [17]). In [20], three kinds of translations from English to German of a set of 500 SNOMED CT terms are compared: i) one translation was performed by professional medical translators, ii) another one used Google Translate [21] and, finally iii) medical students translated the same group of terms. They concluded that machine translation and the employment of student translators are considerable alternatives with "surprisingly" good results, but these methods are not acceptable for the production of terminological standards. However, the authors think that "the combination of machine-translated text with subsequent post-editing by humans could be another translation strategy that reduces time and produces quality translations". This is, in fact, the approach we want to follow in this work.

The guidelines for the translation of SNOMED CT [22] recommended by the IHTSDO have been followed to design the translation task described in this paper.

Spain is a member of the IHTDSO. In May of 2014 this institution presented the *"IHTSDO Policy on Support for Member Country Translation"* proposal, which supports the translation of the CORE of SNOMED CT from English into other languages. If the institutions of the Basque Country obtained this support for the manual translation of SNOMED CT into Basque, the generated corpus of 5,000 manually translated terms would be essential for the evaluation of our system.

Methods

To deal with the translation of SNOMED CT, two strategies can be used: i) the enrichment of the terminology in the SNOMED CT version from Spain (in Spanish) with Basque (as well as with Catalan, or Galician) value sets for the most important concepts and, ii) the creation of an independent SNOMED CT version in Basque. We decided to use the second approach for these reasons: i) We want to collect and create the most extensive terminology possible, not wasting the resources we already have (dictionaries for instance) and, ii) it facilitates the extraction of the most important concepts to enrich the Spanish version.

In this section after describing the analysis of two SNOMED CT releases that led us to choose the source SNOMED CT version for the translation task, we will describe in detail the first two phases of the algorithm as these are the ones already implemented and evaluated.

Analysis to choose the source language in SNOMED CT

SNOMED CT is composed of almost 300,000 active concepts which are represented by descriptions or terms. This terminology corresponds to the core terminology found in elec-

tronic health records and it is organized in hierarchies. SNOMED CT terminological content offers a thorough coverage of the terms used to write the record patient conditions [23]. Concepts are defined by means of description logic axioms and are also used to group terms with the same meaning. In this paper we will refer to these descriptions as terms.

SNOMED CT divides the descriptions in three types (see Table 1): Fully Specified Names or FSN, Preferred Terms or PT and Acceptable Synonyms or Synonyms. The description used to unambiguously describe the concept is called Fully Specified Name. Those descriptions are easily identifiable as they show a semantic tag in parenthesis at the end of the description, e.g. *disorder*, that expresses its semantic category and in consequence, the hierarchy it belongs to, e.g. *Clinical finding/disorder* (even if the hierarchical structure is defined by the relationships between concepts). Regarding the terminology of clinical records, that is, proper "terms" or "descriptions", SNOMED CT distinguishes PTs and Synonyms. PTs are the most common way to name the meaning of the concept according to the IHTSDO. Synonyms are additional terms used to refer to the same concept. Thus, for each SNOMED CT Concept, a language has to define a FSN, a PT and as many Synonyms as there are used to refer that concept (it could have zero to many Synonyms).

Table 2 shows the 18 hierarchies SNOMED CT has its content divided into (plus the metadata hierarchy) and the number of FSNs in each hierarchy and language. We extracted this data from the last version released of the International Release in English, dated on 2014-01-31 and the Spanish version of the International Release, dated on 2014-04-31. As mentioned before, SNOMED CT groups its concepts in hierarchies such as *Clinical finding/disorder*, *Organism*, and so on. These hierarchies differ not only in the content, but also in the requirements for translation. For example, some hierarchies like *Organism* do not require the Preferred Term to be localized, because it corresponds to the taxonomic one. The IHTSDO offers the guidelines for the translation of SNOMED CT in [22], and it describes among others, the recommendations that are important for each hierarchy. The FSN will not be translated, but generated after the validation of the PT, following the rule of creating it by appending a semantic tag to the PT.

We analyzed the multilingual lexical resources available for Basque in the biomedical domain, and the languages in which SNOMED CT is released, and we concluded that two source languages can be used for our translation task: English and Spanish. As Basque is an isolate language, it is not related to either of the mentioned source languages. The linguistic characteristics of Basque differ greatly from those in English and Spanish, so there is no linguistic relatedness reason to choose one of these languages as translation source. Thus, we analyzed both versions of SNOMED CT to choose the best option. The versions we analyzed are dated the 31-07-2012 for English and the 31-10-2012 for Spanish and we focused on the Release Format 2 (RF2) and Snapshot distributions. We must highlight that the Spanish version of SNOMED CT is a manual translation of the English version and at that time the Spanish version was not a complete version.

Even if both languages have the same number of active concepts (296,433 concepts), the Spanish version has a significantly smaller number of terms because the version is at a preview stage: 15,715 concepts in Spanish lack PTs and Synonyms. At a first stage, this data led us to choose as the source for the translation the English version of SNOMED CT but we soon realized that we could not leave aside the already available Basque-Spanish pair resources.

In order to establish a priority between hierarchies for the translation, we counted the number of terms in each hierarchy. The most populated hierarchies both in previous and

current versions are: *Clinical finding/disorder* (99,812 concepts) and *Procedure* (53,629 concepts) followed by *Organism* (33,157 concepts) and *Body Structure* (30,589 concepts). IHTSDO indicates in the translation guidelines that Preferred Terms in the *Organism* hierarchy should not be translated, so we decided to prioritize the translation of the *Clinical finding/disorder*, the *Procedure* and the *Body Structure* hierarchies.

In the next subsection we will describe deeply the first two phases of the algorithm.

Phase 1: lexical resources

The first phase corresponding to the lexical resources has been performed for both language pairs, English-Basque and Spanish-Basque. Although, we decided to take English as source language, we cannot discard the lexical resources available for the Spanish-Basque pair. Thus, we take advantage of the robustness of the English version and of the bigger amount of lexical resources available in the Spanish-Basque pair. These are the multilingual specialized dictionaries used to obtain the Basque equivalences.

- *ZT Dictionary* [25]: a specialized dictionary of science and technology that contains areas included in SNOMED CT as medicine, biochemistry, biology… It contains 10,626 English-Basque equivalences and 10,971 Spanish-Basque equivalences.
- *Nursing Dictionary* [26]: a small dictionary of the nursing domain that has 4,155 entries in the English-Basque chapter and 4,671 entries in the Spanish-Basque one.
- *Glossary of Anatomy*: anatomical terminology used by university experts in their lectures. In its development phase it has 2,818 entries for the English-Basque pair, and 3,940 entries for the Spanish-Basque pair.
- *ICD-10* [27]: The 10th version of the International Classification of Diseases was translated into Basque in 1996. We combined it with the Spanish and English versions and we obtained a dictionary of 6,936 equivalences between English and Basque and 8,842 equivalences between Spanish and Basque.
- *EuskalTerm* [28]: the biggest multilingual terminology bank available for Basque with 75,860 entries. Regarding the domain of biomedicine, the bank contains 32,301 term equivalences. These equivalences are all available for the Spanish-Basque pair, and 10,506 equivalences for the English-Basque pair.
- *Elhuyar Dictionary* [29][30]: a general dictionary that is available for the English-Basque pairs and Spanish-Basque pairs. The English-Basque version contains 39,164 equivalences from English to Basque and the Spanish-Basque version contains 62,215 entries.
- *Dictionary of Sanitary Administration* [31]: a small dictionary that contains 1,799 entries for the Spanish-Basque pair corresponding to the administration of the sanitary domain.

As mentioned before, *Elhuyar Dictionary* is a general dictionary that also contains some specialized terminology. Taking into account the wide variety in SNOMED CT terminology, we decided to use this general dictionary to increase the number of translation pairs only when the source term (English or Spanish) does not exist in the rest of dictionaries. Thus, we limited the big amount of ambiguous equivalent Basque terms of the biomedical domain extracted from *Elhuyar Dictionary*. The use of this dictionary provided i) equivalences of terms not directly related to the biomedical domain (e.g. terms from the "social context" or "qualifier" hierarchies), and also, ii) the equivalences of chunks for the translation of complex terms, and in consequence, the generation of new terms in Basque.

Phase 2: finite state transducers and biomedical affixes

This subsection explains the system that obtains Basque equivalent terms from English simple terms based on Finite State Machines. This approach is based on the idea that a considerable amount of medical terms can be created as neologisms [32], that is, new words and meanings can be created by the concatenation of existing morphosemantic units. These units usually have Greek and Latin origins and their meaning is known by the specialists. In [33] the author specified that about three-fourths of the medical terminology is of Greek origin. Finite State Transducers are appropriate for dealing with the compositional structure of those medical simple terms.

First of all, we will describe the general system for the translation process. Next, we will explain the first approach developed from the baseline system [34]. Finally, we will explain the improvements proposed by experts that have been introduced in the system.

Baseline translation process

The generation of Basque equivalent terms from English terms is performed in three phases: first the identification of the affixes; secondly the translation of the affixes, and finally the composition of the translated affixes. All the linguistic information is stored in lexicons, and rules are written for the process of identification, translation and morphotactics.

Listing 1 shows the Finite State Transducer for the identification of the affixes. The lexica of the affixes is loaded (lines 1-6) and then any prefix (the "*" symbol indicates 0 or more times) followed by one unique suffix is identified. The connecting vowel -o- may be also identified as it is commonly used in connecting two elements of Greek origin. To mark the limits of the affixes the "+" symbol is used. The full explanation about the regular expressions used in Foma is available in [35, 36].

Listing 1 Rules for affix identification.

```
1 read lexc prefixes.lex
2 define PREFALL
3 define PREF PREFALL.u ;
4 read lexc suffixes.lex
5 define SUFALL
6 define SUFF SUFALL.u ;
7 regex [[[PREF 0:+] (o 0:+)]* SUFF] ;
```

In order to reduce the overproduction of the transducer, we fixed the criteria to pick the output with less identified parts. For instance, for the term "photodermatitis" four possible outputs are generated:

- photo+dermat+itis: 3
- photo+derm+at+itis: 4
- phot+o+dermat+itis: 4
- phot+o+derm+at+itis: 5

In this case, the first identification is given to the translation transducer as it contains only three parts.

Following this criterion, even though we can reduce the overproduction we cannot always avoid it. In fact, if we analyze the lexicon of the prefixes we obtain that 93% of the

translation pairs are equal to the ones obtained from transcription rules that will be described in the First approach. In Example 1 we can observe how the equivalence given to "cholecyst" (*"kolezist"* in Basque) is the same as the combination of *"kole"* and *"zist"* so the translation transducer will output the same string. That is to say, in most cases the overproduction is reduced once the translation and the composition FSTs are applied as the output equivalent term will be the same.

Example 1 Some preffix equivalences in our lexicon.

cholecyst:kolezist #;
chole:kole #;
cyst:zist #;

The combination of the Finite State Transducers for the translation and for the composition using morphotactics is shown in Listing 2. First, the lexicons for the translation task are loaded (1-4), and then 28 rules for morphotactics are applied (simplified in the rule numbered 5). Some of these rules were determined empirically by analyzing examples from dictionaries, and others have as a basis the orthographic rules set by the Royal Academy of the Basque Language [37]. The translation rule (shown in rule number 6) is composed of the word-start mark (the ^ symbol), the prefix (named TRANSPRE) followed by the optional linking "o" zero or more times, and a single compulsory suffix (TRANSSUF); finally in the step number 7 the transducer combines the translation (TRANS) and the morphotactic finite state transducers (MORPH) by means of a ".o." composition rule.

Listing 2 Rules for the affix translation.

```
1 read lexc prefixes.lex
2 define TRANSPRE
3 read lexc suffixes.lex
4 define TRANSSUF
5 define MORPHO ...
6 define TRANS (^) [[[TRANSPRE +] (o:o +)]* TRANSSUF] ;
7 regex TRANS .o. MORPH ;
```

We decided to make the suffix compulsory as we discovered that the equivalences of the suffixes are more complex than the equivalences of the prefixes. That is, only 22% of the suffixes follows the transcription rules mentioned before, and what is more, we have not been able to find a pattern based on morphotactics for those endings. Thus, we consider that for this stage of the development the suffix must be compulsory to guarantee a higher precision of the translation. Besides, this condition seems to exclude terms that do not follow a "prefix, root and/or suffix" structure which is the structure this method has been designed for. Example 2 shows the whole process with an example. First, we identify the prefixes and suffixes of the English input term by means of the transducer that marks those affixes (schiz+encephal+y). Then, we obtain the corresponding Basque equivalent for each part and we form the term (eskiz+entzefal+ia).

Example 2 Basque simple term generation.
Input term: schizencephaly
Identified affixes: schiz+encephal+y
Translated affixes: *eskiz+entzefal+ia*
Output. Basque term: *eskizentzefalia*

As we said before, in order to obtain a well formed Basque term, we apply different morphotactic rules. For example, in Basque, there are not words that start with "r" and an "e" is needed at the beginning. Example 3 shows a case where the translated prefix *"radio"* needs of the mentioned rule, obtaining *"erradio"*.

Example 3 Morphotactic rule application.
Input term: radionecrosis
Identified affixes: radio+necr+osis
Translated affixes: *radio+nekr+osi*
Output. Basque term: *erradionekrosi*

In order to identify the English medical suffixes and prefixes we have joined two lists: the "Medical Prefixes, Suffixes, and Combining Forms" from Stedman's Medical Dictionary [38] and the "List of medical roots, suffixes and prefixes" from Wikipedia [39]. From the roots we analyzed, we deduced that their behavior is similar to prefixes when it comes to the composition of words, and so we will label them and include them as prefixes. We manually generated a list of 826 prefixes and 143 suffixes with their Basque equivalents.

To perform the translation task, we manually deduced the appropriate Basque equivalents of the medical affixes. We infer the translation of the affixes from term pairs in specialized dictionaries such as *Zientzia eta Teknologiaren Hiztegi Entziklopedikoa* (Dictionary of Science and Technology) [25], Euskalterm [28] and *Erizaintzako Hiztegia* (Nursing Dictionary) [26]. Table 4 shows an example where the equivalent of the "encephal" prefix is obtained, deducing that *"entzefal"* is the most appropriate equivalent.

From all the prefixes and suffixes listed, we were able to deduce 812 prefixes and 139 suffixes for Basque. They were supervised by an expert so the confidence in the equivalences is high. This technique allows the inference of new medical terms which do not appear in dictionaries.

This baseline approach gave us a precision of 0.94 and a recall of 0.52 as we show in the Results section. Even if the precision is good, the low recall forced us to improve the system, as we will show in the following section.

First approach

In order to improve the very low recall of the Baseline approach, we focused on increasing the number of affixes and implementing transcription rules from English/Latin/Greek to Basque.

To enrich the lexicons of the affixes we included the "Suffix Prefix Dictionary" from Macroevolution [40] and some prefixes from the "Mosby's Medical Dictionary". Thus, we obtained 1,703 prefixes and 630 suffixes manually generated and checked by an expert, and we inferred 40 rules for transcription.

In the Baseline implementation only medical terms fully identified are translated. For example, terms with the prefix "phat" are not translated as this affix does not appear in the

prefixes and suffixes lexicons. In consequence, terms such as “hypophosphatemia” are not translated even though the “hypo”, “phos” and “emia” affixes are identified and appear in the lexicons.

As mentioned before, 93% of the prefixes lexicon were the same as the ones created with transcription rules. Thus, we analyzed the general behavior of the “not identified parts” and prefixes lexicon in the Basque equivalent terms, and we defined 40 transcription rules. For instance, “v” is transcribed as “b” in Basque or “c” is transcribed as “z” whenever is followed by “e”, “i” or “y” and is transcribed as “k” otherwise. The term “diverticulitis” has the Basque equivalent “dibertikulitis” as we find in the ZT Dictionary and in the Nursing Dictionary. This example shows how the rules behave: the “verticul” part is not in our extended lexicon, but we can observe that with the two transcriptions explained earlier we can obtain the proper Basque equivalent: “bertikul” (“v” is “b” and “c” is “k” as it is followed by “u”).

In order to identify the parts that do not appear in the lexicons, we introduced a new rule for the identification task. In Listing 3, line 3 identifies the parts that do not appear in the lexicons writing the # symbol at the end of the part. Line 2 corresponds to the identification explained in Listing 1, and line 4 composes the two previous rules by a priority union (.P. is used).

Listing 3 Rules for the affix identification with parts which do not appear in the lexicons.

```
1  ...
2  define IDEN1 [[[PREF 0:+] (o 0:+)]* SUFF] ;
3  define IDEN2 [(?+ 0:#+) [PREF 0:+]]*(?+ 0:#+) SUFF;
4  regex IDEN1 .P. IDEN2;
```

We had to adapt the criterion to choose among the possible identifications. In this case, whenever a term has non-identified parts, we add to the number of parts the number of characters in the non-identified part. Thus, we choose the term with less non-identified characters, but also with the more robust identification. For example, for the term “diverticulitis” four possible identifications are given (the first number indicates the length of the non-identified part, while the second one gives the number of parts):

- diverticul#+itis: 10 + 2 = 12
- divertic#+ul+itis: 8 + 3 = 11
- di+verticul#+itis: 8 + 3 = 11
- di+vertic#+ul+itis: 6 + 4 = 10

In this case, even if the last option has more parts, it contains less non-identified characters and the addition of the two lengths results in the smallest number. Thus, it will be the one chosen.

Listing 4 shows three of the rules defined for the transcription of the parts. Lines 2 and 3 describe the rule that converts “c” into “k” or “z” depending on the following character and line 4 corresponds to the transcription of “v” into “b”. Lines from 5 to 10 show the rule about the palatalization of the sibilants “z”, “s” and “x”. In Basque, sibilants whenever are followed by a vowel and preceded by “n”, “l”, “r” or “m”, are palatalized by adding a “t” before them.

Listing 4 A few rules for the transcription.

```
...
define C c -> k || [noC] _ [a|o|u|noHC|#],,
          c -> z || [noC] _ [e|i|y];
define define V v -> b;
define Vow [ a | e | i | o | u | y ];
define Sib [ s | z | x ];
define PAL n -> n t || _ Sib Vow,,
           l -> l t || _ Sib Vow,,
           r -> r t || _ Sib Vow,,
           m -> n t || _ Sib Vow;
...
```

By means of these improvements, we are able to translate all the simple terms that contain just a suffix from the suffix lexicon. That is, we still keep the suffix compulsory as mentioned in the Baseline approach. We check whether the term contains any prefix from the translation pair list in order to identify the parts. After the identification, we translate the prefixes and the suffix from the translation pair list and the rest of the parts by means of transcription rules. We finally apply the morphotactic rules from the baseline system to join the translated or transliterated parts and thus create the equivalent Basque term.

Example 4 shows step by step the work carried out. In the first step we take the input term "hypophosphatemia" and we split it into the possible affix combination (in this case "hypo+phos+phat#+emia" or "hypo+phos+phat#+em+ia"). In the second step, we get the Basque equivalences of the affixes ("hipo+fos+fat+emia" or "hipo+fos+fat+em+ia"). Finally, we apply the morphotactic rules to compose the well-formed Basque term (in both cases "hipofosfatemia" is generated).

Example 4 Term translated by means of affix equivalences.
Input term: hypophosphatemia
Identified affixes: hypo+phos+phat+emia, hypo+phos+phat+em+ia
Translation of the affixes: hipo+fos+fat+emia, hipo+fos+fat+em+ia
Morphotactics output term: *hipofosfatemia*

With this improvement the recall of the system increases to 0.826. However, as it is often the case, the precision decreases to 0.813 as shown in the Results section. This loss in the precision led us to analyze the mistakes made by the system with several experts specialized in Basque terminology from the medical domain.

Second approach
Following the advice provided by the experts we restricted the criteria used to choose the terms to be eligible for translation. On the one hand, we reduced the lexicon of the suffixes, excluding the suffixes that are used in common words. That is, suffixes like "-tion" or "-able" have been excluded as they are not exclusive from the biomedical domain, and only suffixes closely related to this specialized terminology were used to conform the lexicon of suffixes. In addition, short prefixes with three characters or less were excluded from the

lexicon of prefixes to eliminate prefixes that could be found within terms. For instance, the prefixes “an-” or “col-” were taken off.

In the following enumeration we list the criteria to identify the components of a term. If we cannot separate the components with the first criterion the second one is tried. If it is not applicable, the last one is attempted.

1. The whole term is identified by means of the extended lexicons (line 8 in Listing 5).
2. The term has the suffix that appears in the reduced lexicon of suffixes (line 9 in Listing 5).
3. The term has the suffix that appears in the extended lexicon of suffixes and contains at least one prefix from the reduced lexicon of prefixes (line 10 in Listing 5).

Listing 5 Rules for the affix identification second approach.

```
1  ...
2  read lexc prefixesReduced.lex
3  define PREFREDUCED
4  define PREFRED PREFREDUCED.u ;
5  read lexc suffixesReduced.lex
6  define SUFFREDUCED
7  define SUFFRED SUFFREDUCED.u ;
8  define IDEN1 [[[PREF 0:\%+] (o 0:+)]* SUFF] ;
9  define IDEN2 [(?+ 0:#+) [PREF 0:+]]*(?+ 0:#+) SUFFRED;
10 define IDEN3 [(?+ 0:#+) [PREFRED 0:+]]+(?+ 0:\#+) SUFF;
11 regex IDEN1 .P. IDEN2 .P. IDEN3;
```

The selection of the suffixes to be excluded has been made by consulting the suffixes in a general dictionary of English suffixes in the Wiktionary [41]. We manually checked the definition of each of the suffixes in the dictionary, so we could exclude the suffixes with a general meaning. We have also excluded the suffixes that are used in non-transcriptable terms like “-hood”. For example, this suffix used in “childhood” or “manhood” have as equivalents in Basque two completely different suffixes: “-aro” in *“haurtzaro”* (“childhood”) and “-tasun” in *“gizontasun”* (“manhood”). We are aware that the manual procedure may be prone to errors, however, we have reviewed the suffixes appearing in the general suffixes list, and so, the most common ones were excluded.

In this process we had to make certain decisions as in the case of the suffix “-on”. Even if its three senses are related to biology or chemistry, it is the ending of many general suffixes as “-tion” or “-isation”, being “-tion” the most popular one. As those suffixes have been excluded from the lexicon, we decided to exclude the suffix “-on”, as by means of including it we will be identifying the terms with the “-tion” suffix in most of the cases.

The exclusion process led us to exclude 71 suffixes and 241 short prefixes, leaving a lexicon of 559 suffixes and 1,462 prefixes.

As we will see in the Results section, the new approach did not improve the results. The precision obtained was 0.813 and the recall 0.747. That is, the precision did not improve, and there was a decrease in the recall.

Results

As mentioned before, we divided SNOMED CT into hierarchies to simplify the translation process. We evaluated the *Clinical finding/disorder*, *Procedure* and *Body Structure* hierarchies, as they are the most populated ones. Since the *Clinical finding/disorder* hierarchy is specially populated we split it according to its semantic tags: *disorders* and *findings*.

Phase 1 results

We want to remark that Phase 1 could not be evaluated in terms of the *quality* of the translations, but of *quantity*. As we used manually generated and checked dictionaries written by lexicographers and domain experts, we assumed the quality of the Basque terms. In any case, Basque is a language in its standardization process and some orthographic rules have been changed, so, the orthographic correctness of the descriptions and its possible disambiguation will be manually checked in the future.

Table 5 shows the evaluation of the Phase 1 regarding the quantities obtained from the different terminology resources. We distinguish the quantity of Basque equivalent terms obtained (column labeled as "#Syn.") and the number of source SNOMED CT concepts translated (column labeled as "#Concepts"). As seen in the table, the same concept may have more than one synonym. For instance, in the *Disorder* sub-hierarchy we have 3,063 SNOMED CT concepts translated and 3,975 Basque terms for the same concepts.

If we consider the Total columns of the table (columns 6 and 7), we can observe that the totals do not match the sum of the previous columns. This is caused by the fact that the same equivalent term may be obtained from the English matching as well as from the Spanish matching, but it is counted only once. For example, the term "drepanozito" is obtained from the source term in Spanish *"drepanocito"* and from the English term "drepanocyte". This equivalence will be counted in both English and Spanish columns, but once in the Total columns.

We can highlight the amount of synonyms obtained in this Phase: 1.86 for each concept. *Body Structure* and *Disorder* hierarchies get the best results in terms of concepts translated (3,295 and 3,275 respectively), but it is remarkable the high amount of synonyms that *Body Structure* has (7,077 synonyms) which can be put down to the very specialized dictionaries devoted to this hierarchies: the Glossary of Anatomy and the ICD-10.

Phase 2 results

In this phase, results are given for the simple terms extracted from the *Disorder*, *Finding*, *Body Structure* and *Procedure* hierarchies. The set of terms from each hierarchy is split into two: i) to define and develop the system and ii) to evaluate it.

The development and test sets comprise the simple terms that have been previously translated in the first phase of the algorithm. That is, we used the correct English-Basque pairs from the dictionaries as Gold Standard. This Gold Standard was manually created by setting a label to each term indicating whether or not the term should be translated by means of this system. That is, the system should not work with terms like "shock" or "dengue" that are not composed of medical roots.

For the evaluation set we took 848 terms from the *Disorder* sub-hierarchy, 375 from *Finding*, 774 from *Body Structure* and 248 from *Procedure*. The remaining 3,114 terms from *Disorder*, 1,446 from *Finding*, 1,838 from *Body Structure* and 1,729 from *Procedure* were used for development.

To measure the results of the experiment True Positives (TP), False Negatives (FN), False Positives (FP) and True Negatives (TN) are defined in the following way:

- True Positives: The term should be translated, it is translated and the translation is correct. That is, at least one of the Basque terms generated matches at least one synonym from the Gold Standard.
- False Negatives: The term should be translated and it is not translated.
- False Positives: The term should not be translated and it is translated, or the term should be translated and the Basque term generated is not correct.
- True Negatives: The term should not be translated and it is not translated.

Table 6 shows the precision, recall and F-Measure of the three approaches detailed in the Methods section. It is worth to mention that we obtain the best results regarding the F-Measure with the first approach. Even if the second approach gives a better precision compared to the first approach, the decrease it generates in the recall is much sharper, and so it is manifested in the F-Measure. Thus, we conclude that the best approach is the first one, and this is the one we use for the evaluation of the whole algorithm.

We must consider that our evaluation does not take into account whether the system overproduces wrong Basque terms if the correct one is also produced. In any case, the overproduction is properly controlled as mentioned in the Methods section, and in average 1.05 Basque equivalents are produced from an English term.

Considering the results obtained in the Baseline approach, the changes made to the system in the first and in the second approaches show a huge improvement of the system. Even if we obtain a small decrease in the precision, the improvement in the recall is remarkable: changes from 0.343 in the baseline to 0.826 in the first approach and 0.747 in the second approach.

We must highlight that we focused the development of the system on the *Disorder* hierarchy as it is the one with more simple terms composed of Latin and Greek roots and affixes. The bias to this sub-hierarchy is evident as the *Disorder* sub-hierarchy obtains the best results.

Overall results

We show the overall results of the translation algorithm in Table 7 regarding the mapping with the ICD-10 classification and the two phases implemented. That is, the table shows the synonyms obtained (named "#Syn." in the table) from the matches ("#Match" in the table) over the ICD-10 mapping, dictionaries and morphosemantics system. The "#Match" columns indicate the number of source terms translated, while the "#Syn." columns show the number of terms obtained. Remember that more than one term could be obtained from a unique source term.

The results labeled as "Phase 0 – ICD-10 mapping" in Table 7 show that the mapping is only relevant in the *Clinical disorder/finding* hierarchy and that the *disorder* semantic tag is the most benefited with 11,224 equivalences. In this case, the mapping does not offer synonyms, but obtains a single term from each mapping.

Table 8 shows the results regarding the number of tokens of the original English descriptions that are included in the source SNOMED CT, and it does not make reference to the number of concepts. The row labeled as *Translated* shows the quantity of English terms for which a translation has been obtained. The second row labeled as *Total* reveals the total amount of English terms, and finally, the last row presents the percentage of the translated terms.

The mentioned Table 8 is useful to measure the progress of the algorithm. That is, the first two phases of the algorithm are focused on single terms, whereas the remaining phases are designed for complex terms. We observe that a high percentage of the single terms is already translated in all the hierarchies, but specially in the hierarchies *Disorder* and *Procedure* (85.51% and 87.84% respectively). It is remarkable that 12.94% of the two tokens terms from *Body Structure* have already been translated from the dictionaries.

In order to give a wider view of the process followed, Table 9 presents the overall numbers of the translated SNOMED CT concepts.

Let us highlight the most promising results for each hierarchy:

- Regarding the *Disorder* sub-hierarchy, we obtained the translation of 21.41% of the terms (see Table 9). Considering that we have focused our work until now mainly on simple terms, we can consider that it is a very good result. The ICD-10 mapping contribution is the major one, producing 11,224 synonyms. In any case, the strength of the morphosemantics phase is noticeable in Table 8, which shows that 85.51% of the simple terms are translated.
- In regards to the *Finding* sub-hierarchy, we can consider it as the most balanced one, as it does not outline any method used. In this case, we achieved the translation of 8.80% of the concepts.
- In the *Body Structure* hierarchy, 12.57% of the concepts get a Basque equivalent, with outstanding results for complex terms (12.94% of two token terms).
- For the *Procedure* hierarchy the dictionaries are of hardly any use (536 Basque terms as seen in Table 7). In contrast, after applying the mophosemantics phase 87.84% of the simple terms are translated (see Table 8). In any case, we only obtain 3.00% of the concepts translated, and this must be an aspect to be improved in the following phases.
- In general, even if the overall numbers seems to be low (22,586 concepts translated over 184,030), it is a solid base to implement the following two phases in an incremental strategy.

Conclusions

In this paper we presented some steps of an algorithm for the translation of the multilingual terminology content of SNOMED CT. We also described the good results obtained on the morphosemantics phase by means of an experiment, and how this phase and the dictionaries contribute on the translation of SNOMED CT by means of quantities.

On the one hand, we take advantage of existing lexical resources, and on the other hand, we use transducers to generate Basque equivalents by means of domain-specific affixes and transcription rules. The implementation can be available on request contacting the authors. It will be publicly accessible once the implementation is concluded.

Even if the specialized dictionaries provide Basque simple and complex terms, in this case the transducers are designed to translate simple terms. Thus, we got the translation of 85.51% of the simple terms in the *Disorder* sub-hierarchy and 87.84% in the *Procedure* hierarchy.

Even if in this paper we only show the results obtained in the most populated hierarchies, we applied the translation algorithm to the whole SNOMED CT terminology. The use of lexical resources is promising as seen in the Results section, and the contribution of the

ICD-10 mapping in the *Disorder* sub-hierarchy is especially remarkable (11,224 matchings). The *Disorder* sub-hierarchy is the largest and here we obtained the equivalents in Basque of 5.21% of the source English terms.

Nevertheless, as we said before, our aim is to check the quality of the Basque SNOMED CT version we are generating. For this evaluation (and correction) we count on the help of specialists of medical terminology such as doctors and terminologists. We consider the *linguistic correctness* of the translation and the *fidelity of the translated content* are appropriate for this evaluation of the translation quality. In addition, we are working in a platform to help the specialists with the evaluation and correction. If the quality of the terminology generated reaches high and solid results, we will contact the SNOMED CT providers to offer them the result of our work, which at the moment is in the field of academic research.

In regard to the evaluation of our systems, the first phase does not require a deep evaluation as it extracts English-Basque and Spanish-Basque pairs from dictionaries. In any case, a deeper evaluation of the approaches based on morphosemantics is presented. We implemented and evaluated three systems for the translation of simple terms using morphosemantic characteristics of the terms.

In the future, we plan to implement the remainder of the algorithm in two ways: on the one hand, to generate the complex terms by means of syntax rules and on the other hand, to adapt the machine translation tool. The promising results obtained up to the present encouraged us to finish the semi-automatically generated version in Basque of SNOMED CT.

List of abbreviations used

SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms. HIT: Health Information Technology. EHR: Electronic Health Record. FSN: Fully Specified Name. PT: Preferred Term. ICD-10: International Statistical Classification of Diseases and Related Health in its 10th version. True Positive: TP. False Positive: FP. False Negative: FN. True Negative: TN. FST: Finite State Transducer.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

OPV performed the implementation of the algorithm. OPV and MO wrote, read and approved the final manuscript.

Acknowledgements

The authors would like to thank Mikel Lersundi and Igone Zabala for their help. This work was partially supported by the European Commission (325099), the Spanish Ministry of Science and Innovation (TIN2012-38584-C06-02) and the Basque Government (IT344-10 and IE12-333). Olatz Perez-de-Viñaspre's work is funded by a PhD grant from the Basque Government (BFI-2011-389).

Declarations

Publication costs for this article were funded by the Basque Government, project number IT344-10 (IXA group, Research Group of type A).

Author details

[1]IXA NLP Group, University of the Basque Country UPV/EHU , Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain. [2]IXA NLP Group, University of the Basque Country UPV/EHU , Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain.

References

1. IHTSDO, I.H.T.S.D.O.: SNOMED CT Starter Guide. February 2014. Technical report, International Health Terminology Standards Development Organisation (2014)
2. SNOMED CT Value Proposition, I.: http://www.ihtsdo.org/snomed-ct/whysnomedct/snomedfeatures/
3. Osakidetza: II Scheme to Normalise the Use of the Basque Language in Osakidetza . Public Health Service of the Basque Autonomous Community, 2013-2019, (2013)
4. Desjardins, L.: Le santé des francophones du Nouveau-Brunswick. Petit-Rocher, Société des Acadiens et des Acadiennes du Nouveau-Brunswick (2003)
5. Perez-de-Viñaspre, O., Oronoz, M.: An XML Based TBX Framework to Represent Multilingual SNOMED CT for Translation. In: Springer (ed.) Advances in Artificial Intelligence and Its Applications, pp. 419–429 (2013)

6. Perez-de-Viñaspre, O., Oronoz, M.: Tanslating SNOMED CT Terminology into a Minor Language. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), pp. 38–45. Association for Computational Linguistics, Gothenburg, Sweden (2014). http://www.aclweb.org/anthology/W14-1106
7. Edition, B.A.: http://www.britannica.com/EBchecked/topic/55366/Basque-language
8. Government, E.J.B.: V. Mapa Soziolinguistikoa, (2011)
9. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic acids research **32**(suppl 1), 267–270 (2004)
10. Giannangelo, K., Fenton, S.: SNOMED CT survey: an assessment of implementation in EMR/EHR applications. Perspectives in Health Information Management **5;7** (2008)
11. Elhanan, G., Perl, Y., Geller, J.: A Survey of Direct Users and Uses of SNOMED CT: 2010 Status. In: AMIA Annual Synposium Proceedings, pp. 207–1011 (2010)
12. SNOMED in Action, IHTSDO: http://snomedinaction.org/sct-table.html
13. Shaw, A.: SNOMED is giving clinicians a common vocabulary. Canadian Healthcare Technology. http://www.canhealth.com (2012)
14. Lee, D., Cornet, R., Lau, F., de Keizer, N.: A survey of SNOMED CT implementations. Journal of Biomedical Informatics **46**, 87–96 (2013)
15. Daumke, P., Ingenerf, J., Daniel, C., Asholm, L., Schulz, S.: The need for SNOMED CT translations. In: et al., M. (ed.) 23rd International Conference of the European Federation for Medical Informatics (2011)
16. Supporting Different Languages, I.: http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages
17. Petersen, P.G.: How to Manage the Translation of a Terminology. Presentation at the IHTSDO October 2011 Conference and Showcase (2011)
18. Zhu, Y., Pan, H., Zhou, L., Zhao, W., Chen, A., Andersen, U., Pan, S., Tian, L., Lei, J.: Translation and Localization of SNOMED CT in China: A pilot study. Artificial Intelligence in Medicine **54**(2), 147–149 (2012)
19. Abdoune, H., Merabti, T., Darmoni, S.J., Joubert, M.: Assisting the Translation of the CORE Subset of SNOMED CT Into French. In: Moen, A., Andersen, S.K., Aarts, J., Hurlen, P. (eds.) Studies in Health Technology and Informatics, vol. 169, pp. 819–823 (2011)
20. Schulz, S., Bernhardt-Melischnig, J., Kreuzthaler, M., Daumke, P., Boeker, M.: Machine vs. Human Translation of SNOMED CT Terms. In: et al., C.U.L. (ed.) MEDINFO 2013, pp. 581–584 (2013)
21. Google Translate: http://translate.google.com/
22. Høy, A.: Guidelines for Translation of SNOMED CT. Technical Report version 2.0, International Health Terminology Standards Development Organization IHTSDO (2010)
23. Humphreys, B.L., McCray, A.T., Cheh, M.L.: Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. Journal of the American Medical Informatics Association **4**(6), 484–500 (1997)
24. Mayor, A., Alegria, I., Diaz de Ilarraza, A., Labaka, G., Lersundi, M., Sarasola, K.: Matxin, an Open-source Rule-based Machine Translation System for Basque. Machine Translation **25**, 53–82 (2011). 10.1007/s10590-011-9092-y
25. Elhuyar: Elhuyar Zientzia Eta Teknologiaren Hiztegi Entziklopedikoa, (2009)
26. Zerbitzua, E.E., Eskola, D.E.: Erizaintzako Hiztegia, (2005)
27. World Health Organization and Euskal Autonomi Elkarteko Administrazioa. Osasun Saila and UZEI: GNS-10 (Gaixotasunen Eta Horiekin Lotutako Osasun-arazoen Nazioarteko Sailkapen Estatistikoa - 10. Berrikuspena), (1996)
28. UZEI: Euskalterm Terminologia Banku Publikoa. `http://www.euskadi.net/euskalterm` (2004)
29. Elhuyar: Elhuyar Hiztegia Euskara/Ingelesa English/Basque, (2007)
30. Elhuyar: Elhuyar Hiztegia Euskara/Gaztelania Castellano/Vasco, (2007)
31. Osakidetza, eta UZEI, E.O.S.: Administrazio Sanitarioko Hiztegia, (1999)
32. Lovis, C., Michel, P., Baud, R., Scherrer, J.: Word Segmentation Processing: A Way To Exponentially Extend Medical Dictionaries. MEDINFO **8**, 28–32 (1995)
33. Banay, G.: An introduction to medical terminology, Greek and Latin derivations. Bulletin of the Medical Library Association **36**(1), 1–27 (1948)
34. Perez-de-Viñaspre, O., Oronoz, M., Agirrezabal, M., Lersundi, M.: A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes. Finite State Methods and Natural Language Processing, 99 (2013)
35. Hulden, M., Alegria, I.: Creating LRs and applications using finite-state morphological grammars. LREC 2010. Tutorial. http://foma.sf.net/lrec2010/" (2010)
36. Karttunen, L., Chanod, J.-P., Grefenstette, G., Schille, A.: Regular expressions for language engineering. Natural Language Engineering **2**(04), 305–328 (1996)
37. Royal Academy of the Basque Language. Luis Mitxelena: 0. rule. Ortografia (1968)
38. Stedman's: Medical Prefixes, Suffixes, and Combining Forms. In: Lippincott Williams & Wilkins (ed.) Stedman's Medical Dictionary, Twenty-eighth edition edn. (2005)
39. Wikipedia: List of medical roots, suffixes and prefixes – Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=List_of_medical_roots,_suffixes_and_prefixeso` (2013)
40. Macroevolution: http://macroevolution.net
41. Wiktinionary: Category:English suffixes – Wiktionary, a wiki-based Open Content dictionary. `http://en.wiktionary.org/wiki/Category:English_suffixes` (2014)

Tables

Table 1 Description types in SNOMED CT for the concept: 95575002 - Obstruction of pelviureteric junction.

Description	**Type**
Obstruction of pelviureteric junction (disorder)	FSN
Obstruction of pelviureteric junction	Preferred Term
PUJ - Pelviureteric obstruction	Synonym
PUO - Pelviureteric obstruction	Synonym
Pelviureteric obstruction	Synonym
UPJ - Ureteropelvic obstruction	Synonym
Ureteropelvic obstruction	Synonym

Table 2 SNOMED CT hierarchies and number of FSNs.

	English version		**Spanish version**	
Hierarchy	**Semantic Tag (ST)**	**# FSN**	**Semantic Tag (ST)**	**# FSN**
Clinical Finding/disorder	**disorder**	**66,239**	trastorno	66,199
Procedure/ intervention	**finding**	**33,573**	hallazgo	33,613
Organism	**procedure**	**51,149**	procedimiento	51,149
Body structure	**regime/therapy**	**2,480**	régimen/terapia	2,480
	organism	**33,157**	organismo	33,157
	body structure	**24,950**	estructura corporal	24,953
	morphologic abnormality	**4,509**	anomalía morfológica	4,509
	cell	**626**	célula	626
	cell structure	**504**	estructura celular	501
Substance	substance	23,845	sustancia	23,845
Pharmaceutical /biologic product	product	16,759	producto	16,759
Qualifier value	qualifier value	8,944	calificador	8,944
Observable entity	observable entity	8,278	entidad observable	8,278
Event	event	3,671	evento	3,670
Situation with explicit context	situation	3,561	situación	3,561
Social context	occupation	3,852	ocupación	3,852
	person	425	persona	425
	ethnic group	262	grupo étnico	262
	religion/philosophy	203	religión/filosofía	203
	life style	21	estilo de vida	21
	social concept	23	contexto social	23
	racial group	19	grupo racial	19
Physical object	physical object	4,513	objeto físico	4,513
Specimen	specimen	1,440	espécimen	1,440
Environment geographical location	environment	1,094	medio ambiente	1,094
	geographic location	617	localización geográfica	617
Staging and scales	assessment scale	1,077	escala de evaluación	1,077
	tumor staging	214	estadificación tumoral	214
	staging scale	16	escala de estadificación	16
Special concept	navigational concept	640	concepto para navegación	640
	namespace concept	169	espacio de nombres	169
	special concept	1	concepto especial	1
Record artifact	record artifact	224	elemento de registro	224
Physical force	physical force	171	fuerza física	171
Metadata	foundation metadata	169	metadato fundacional	169
	core metadata concept	31	metadato del núcleo	32

Table 3 Structures of the SNOMED CT terminology content.

Pattern found	**Quantity**
`[PHARMPRODUCT|SUBSTANCE]+allergy`	1,498
`[PHARMPRODUCT|SUBSTANCE]+adverse+reaction`	1,488
`[PHARMPRODUCT|SUBSTANCE]+poisoning`	847
`[PHARMPRODUCT|SUBSTANCE]+overdose`	567
`[PHARMPRODUCT|SUBSTANCE]+poisoning+of+undetermined+intent`	432
`intentional+[PHARMPRODUCT|SUBSTANCE]+poisoning`	429
`accidental+[PHARMPRODUCT|SUBSTANCE]+poisoning`	428
...	...

Table 4 The translation of the "encephal" prefix.

English terms	Basque terms
echo**encephal**ogram	*eko**entzefal**ograma*
encephalitis	***entzefal**itis*
encephalomyelitis	***entzefal**omielitis*
leuko**encephal**itis	*leuko**entzefal**itis*
...	...

Table 5 Results of the Phase 1.

	English		Spanish		Total	
	#Syn.	#Concepts	#Syn.	#Concepts	#Syn.	#Concepts
Disorder	3,975	3,063	2,231	1,602	4,362	3,275
Finding	1,690	857	1,866	759	2,855	1,018
Body Structure	5,554	2,747	5,076	2,616	7,077	3,295
Procedure	557	405	536	377	775	501

Table 6 Results of the Phase 2.

		TP	FN	FP	TN	Total	Prec.	Recall	F-M
Disorder	**Baseline**	289	451	31	77	848	0.903	0.391	0.545
	1st approach	615	67	108	58	848	0.851	0.902	0.875
	2nd approach	577	104	102	65	848	0.850	0.847	0.849
Finding	**Baseline**	79	171	9	116	375	0.898	0.316	0.467
	1st approach	213	29	41	92	375	0.839	0.880	0.859
	2nd approach	178	63	32	102	375	0.848	0.739	0.789
Body Structure	**Baseline**	121	425	23	205	774	0.840	0.222	0.351
	1st approach	322	174	100	178	774	0.763	0.649	0.702
	2nd approach	284	212	91	187	774	0.757	0.573	0.652
Procedure	**Baseline**	98	77	9	64	248	0.916	0.560	0.695
	1st approach	144	16	49	39	248	0.746	0.900	0.816
	2nd approach	154	5	50	39	248	0.755	0.969	0.848
Total	**Baseline**	587	1,124	72	462	2,245	0.891	0.343	0.495
	1st approach	1,295	286	297	367	2,245	0.813	0.826	0.820
	2nd approach	1,304	275	299	367	2,245	0.813	0.747	0.779

Table 7 Results of the translation algorithm.

	Phase 0 ICD-10 mapping		Phase 1 Lexical resources		Phase 2 Morphosemantics		Total	
	#Syn.	#Match	#Syn.	#Match	#Syn.	#Match	#Syn.	#Match
Disorder	11,224	11,224	4,362	5,029	2,699	2,417	17,912	18,670
Finding	1,871	1,871	2,855	1,771	897	655	5,508	4,297
Body Structure	0	0	7,077	5,843	1,026	861	8,036	6,704
Procedure	0	0	536	835	1,780	1,427	2,490	2,262

Table 8 Results of the translation regarding the number of tokens of the original English term.

		1 token	2 tokens	3 tokens	4 tokens	>4 tokens	Total
Disorder	**Translated**	3,388	1,098	533	275	419	5,713
	Total	3,962	21,830	24,054	20,357	39,501	109,704
	Percentage	85.51%	5.03%	2.22%	1.35%	1.06%	5.21%
Finding	**Translated**	1,290	161	39	19	56	1,565
	Total	1,821	8,850	11,126	10,092	19,689	51,578
	Percentage	70.84%	1.82%	0.35%	0.19%	0.28%	3.03%
Body Structure	**Translated**	1,931	1,460	381	72	15	3,859
	Total	2,612	11,287	12,443	10,793	21,515	58,650
	Percentage	73.93%	12.94%	3.06%	0.67%	0.07%	6.58%
Procedure	**Translated**	1,741	80	11	2	1	1,835
	Total	1,982	9,966	15,848	16,578	37,695	82,069
	Percentage	87.84%	0.80%	0.07%	0.01%	0.003%	2.24%

Table 9 Overall results.

	Disorder	Finding	Body Structure	Procedure
Translated Concepts	14,181	2,953	3,845	1,607
Concepts in total	66,239	33,573	30,589	53,629
Percentage	21.41%	8.80%	12.57%	3.00%

5

Complex terms: nested terms and the adaptation of a Machine Translator

This chapter presents the techniques used to translate complex terms into the Basque language. In Section 5.1 we begin by explaining how complex terms are generated from nested terms. Next, in Section 5.2 we discuss the adaptation of the Matxin Machine Translation system to the medical domain. Section 5.3 focuses on the design of a tool developed to assess the automatic generation of complex terms, and Section 5.4 presents the results of said assessment. Finally, Section 5.5 offers a summary of the chapter and presents a series of conclusions.

5.1 Generating complex terms from nested terms

This section explains the third step of the EuSnomed system. To carry out this third step, we developed a system called KabiTerm. KabiTerm uses other terms that appear within complex terms to translate those complex terms into the Basque language.

The current version of KabiTerm takes a complex English term and, providing the resources are available, proposes a Basque equivalent. Here, resources are understood to mean the equivalents and Basque translation patterns for the nested terms. As explained later on in this chapter, in order to facilitate the work carried out by KabiTerm, we have developed an analyser called AnaMed. This analyser is responsible for searching for the informa-

tion required by KabiTerm. It also identifies and prepares the nested terms, leaving KabiTerm free to focus solely on the translation into Basque.

The AnaMed analyser that has been designed and developed for the KabiTerm system will be outlined in Section 5.1.1 , while Section 5.1.2 will explain the KabiTerm system itself.

5.1.1 AnaMed: Medical Term Analyser

This section outlines the AnaMed medical terminology language analyser. In addition to analysing linguistic information, AnaMed also identifies SNOMED CT terms and eponyms in a given text. It has been developed for English and Basque, and the aim is to adapt it also to the Spanish language in the future.

Initially, only the English version of the AnaMed analyser was developed, in response to our need for a tool to search for the information required by the KabiTerm system outlined in this chapter. In other words, the information gathered by AnaMed is information that may prove necessary for automatic machine translation into Basque. However, since AnaMed can easily be adapted to other languages, we decided to develop a Basque-language version, since we believe this may prove useful for the drafting of medical reports in Basque.

AnaMed is based on an automatic analysis system and integrates the identification of both eponyms and SNOMED CT terms. Eponyms are proper nouns that appear in the designation of certain concepts. The architecture of AnaMed is shown in Figure 5.1.

The Stanford CoreNLP tool (Manning *et al.*, 2014) was used as the starting point for the development of the English analyser, along with the Python wrapper for Stanford CoreNLP, developed by Dustin Smith[1]. The Eustagger (Ezeiza *et al.*, 1998) analyser was used for the Basque version (AnaMed_eu). The morphological tokenisers and taggers from the linguistic analyser were used to identify tokens' lemmas and parts of speech. In addition to this information, token offsets and, in the case of AnaMed-en, named entity tags, were also integrated into the analyser (Figure 5.1 shows the output of the first module).

By adding a second module (see Figure 5.1, module 2) we gave the analyser eponym identification capability. Eponyms are very common in medical

[1]https://github.com/dasmith/stanford-corenlp-python (accesed May 9, 2017)

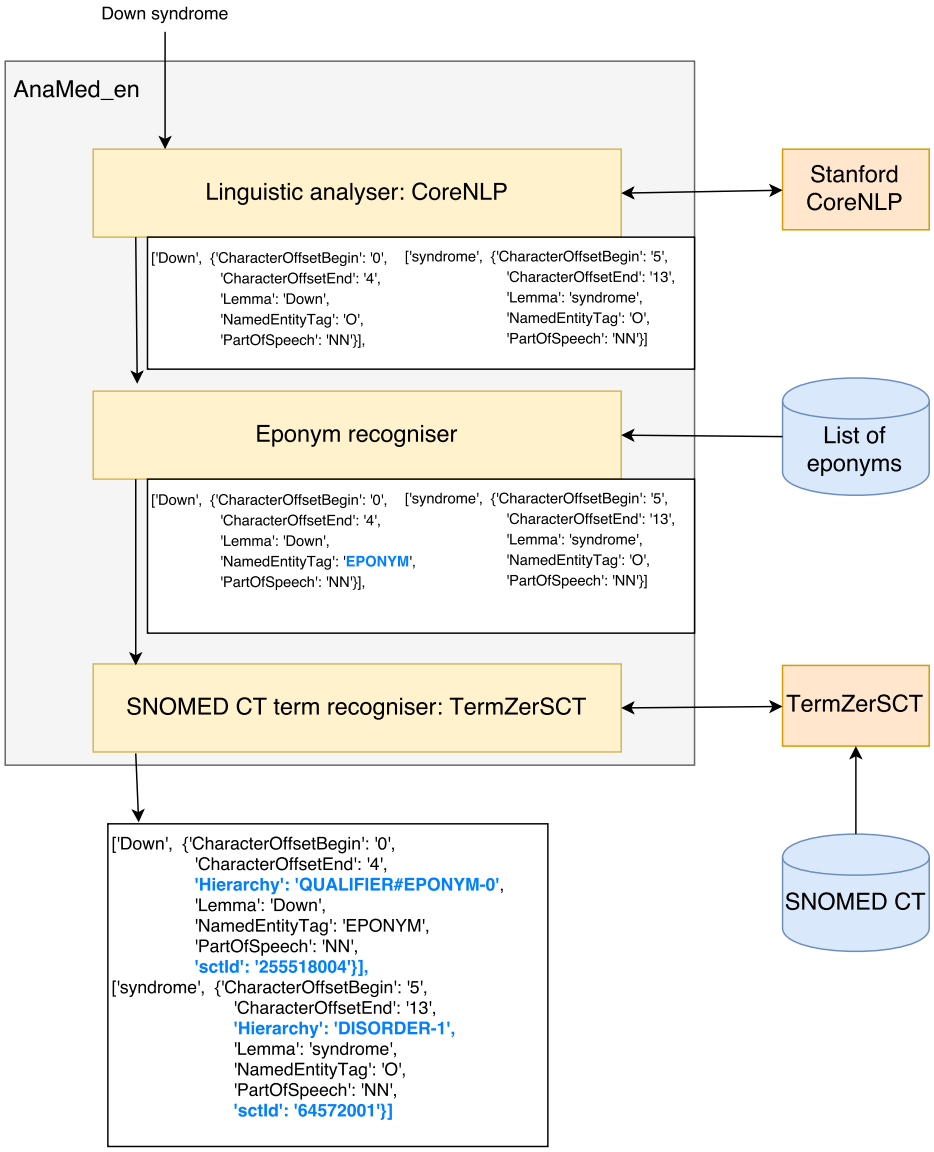


Figure 5.1 – AnaMed analyser architecture.

terminology, particularly in the names of diseases and syndromes. The terms *Down syndrome* and *Alzheimer's disease* are good examples of this[2].

Finally, we also added a SNOMED CT term identifier to the AnaMed

[2]Both terms were extracted from the Euskalterm Public Terminology Database.

analyser (see Figure 5.1, module 3).

No changes were made to the linguistic analyser itself during the course of this thesis project. The subsection below describes the modules generated during this phase.

Eponym recogniser

The most obvious eponyms are found in terms similar to the two examples given above: *Down syndrome* and *Alzheimer's disease*, in which the eponym itself appears explicitly (*Down* and *Alzheimer*, in this case). However, there are also a number of terms that are derivatives of eponyms, such as Daltonism. The term Daltonism was established in honour of the British chemist John Dalton, the first person to describe the condition[3]. However, the eponym recogniser used here does not identify eponym derivatives.

The eponym recogniser was developed with Basque grammar in mind. In other words, the composition of proper nouns (referring to both people and places) differs depending on the declension. No agreement was found regarding the definition of eponyms, and sometimes reference is made to place names also[4,5]. In relation to place names, for example, Stockholm syndrome was named after an event that occurred in the city of Stockholm[6], and as such, the Basque equivalent, *"Stockholmgo sindrome"*, uses the locative genitive case. When referring to the names of specific people, the declension used is the possessive genitive, as is the case with the Weber test, the equivalent of which in Basque is *"Weber-en proba"*.

Before starting work on the eponym identifier, we analysed different named entity recognition systems (Nadeau and Sekine, 2007; Tjong Kim Sang and De Meulder, 2003), testing some of the state of the art systems with SNOMED CT descriptions. In this manual analysis, the best results were obtained by the Stanford CoreNLP named entity recognition tool (Finkel *et al.*, 2005). However, since even with the best available tool the majority of eponyms remained undetected, we conducted an Internet search for lists of the most common eponyms and, on the basis of the results, developed our own eponym recogniser. Thus, both the Stanford CoreNLP named entity recognition tool's persons and the eponyms identified by our system are

[3]https://en.wikipedia.org/wiki/John_Dalton (accesed May 9, 2017)
[4]https://en.wikipedia.org/wiki/Eponym (accesed May 9, 2017)
[5]http://www.dictionary.com/browse/eponym (accesed May 9, 2017)
[6]https://en.wikipedia.org/wiki/Stockholm_syndrome (accesed May 9, 2017)

tagged as eponyms.

The eponym recogniser searches the words contained in the complex term for the eponyms on the list. Sometimes, compound eponyms are used in the names of certain diseases, as in the case of *Verner-Morrison syndrome*, for example. With such cases in mind, when drawing up the list of eponyms we extracted simple eponyms from compound ones. Thus, in the example given above, two eponyms were included on the list. With the aim of broadening the recogniser's coverage, when recognising compound eponyms, the system is designed to identify the whole compound eponym from just one of its components. When compiling the list of eponyms, we analysed all the terms in SNOMED CT, adding all previously unidentified components of compound eponyms to the list. The final list contains around 3,000 proper names for identifying eponyms.

TermZerSCT: SNOMED CT term recogniser

The principal aim of AnaMed is to identify nested terms within terms. Although there are many term extractors currently available, none of them are specifically adapted to the needs of KabiTerm. We are not interested here in identifying general terms, only those included in SNOMED CT, using that system's own hierarchy.

We therefore adapted the TermZerSCT terminology server to identify SNOMED CT terms. SNOMED CT contains a vast amount of terminology (around 300,000 concepts) which takes time to process. TermZerSCT enables faster terminology content management, and when the server is running we receive information about SNOMED CT almost instantly, with only a minimum waiting period.

As stated earlier, the server prepares the terminological content of SNOMED CT in order to provide the client (in this case AnaMed) with the information it requires as efficiently as possible. Among other things, it uses the original SNOMED CT files to classify active concepts into hierarchies. Thus, when the system is given a SNOMED CT concept identifier, in addition to providing that concept's FSN, preferred term and synonyms, it also specifies the hierarchy to which it belongs. This information is added to that provided by AnaMed and the eponym recogniser, as shown in the output section of Figure 5.1.

As we can see in the table below (Table 5.1), we can obtain the SNOMED CT concept identifier for a given term (in this case, *diabetes mellitus*), and

once we have that code, all the information about that concept becomes immediately available, including its fully specified name (FSN), its preferred term (PT) and its synonyms.

Explanation	**Function**	**Result**
Obtain code	`desc2sct`	73211009
Obtain hierarchies	`sct2hie`	DISORDER
Obtain FSN	`sct2fsn`	Diabetes mellitus (disorder)
Obtain PT	`sct2term`	Diabetes mellitus
Obtain synonyms	`sct2syn`	DM - Diabetes mellitus

Table 5.1 – The information that can be obtained regarding the term *diabetes mellitus* using TermZerSCT.

We have developed English, Spanish and Basque versions of the server, since these are the languages in which we have the SNOMED CT terminology, although more work has been carried out on the English and Basque versions, in which, using a lemmatiser (Stanford CoreNLP for the English version and Eustagger for the Basque version), we also offer the option of searching for lemmatised terms. This option will be incorporated also into the Spanish version in the future.

Using the TermZerSCT server, AnaMed identifies the nested terms located within complex terms, enabling us to analyse the structure of said complex terms. Moreover, it also groups nested terms together using underscores ("_"). For example, in the complex term *unstable diabetes mellitus* it identifies two nested terms: the qualifier *unstable* and the disorder *diabetes mellitus*. Thanks to this identification, in addition to providing the complete analysis, AnaMed also gives us the structure (QUALIFIER+DISORDER) and the grouping (*unstable diabetes_ mellitus*), information which is extremely useful for KabiTerm.

5.1.2 KabiTerm: generation of complex terms using nested terms

The section describes the KabiTerm system. KabiTerm is a tool which uses nested terms to generate equivalents of complex terms, with the help of transducers. While in this thesis we present the transducers and application

used for the English-Basque language pair, the system can easily be adapted to work with any two languages.

KabiTerm uses the information provided by the AnaMed analyser outlined in the previous section to identify nested terms within the main term being analysed. While the use of AnaMed is not strictly essential, it does significantly enhance the effectiveness of the KabiTerm tool. Firstly, AnaMed prepares the groupings of the nested terms, thereby simplifying the work to be carried out by the transducers. AnaMed also prepares linguistic information, identifying word form lemmas and providing KabiTerm with the capacity to translate plural nested terms into the Basque language. Therefore, while AnaMed is not entirely indispensable for KabiTerm, it does have a positive impact on both its efficiency and results.

We defined Basque translation patterns using the Foma software program (Hulden, 2009). In other words, we used the Foma tool to generate finite-state transducers (this is same the tool used to develop the NeoTerm system described in the previous chapter). Even though the transducers themselves were generated using Foma, they are combined and managed by an application written in the Python programming language.

Analysing the structure of English terms using AnaMed

As stated earlier, KabiTerm works on the basis of nested terms, or in other words, terms that appear within complex terms.

Thanks to AnaMed, we were able to classify the SNOMED CT terms in accordance with the structure of their nested terms. The table below (Table 5.2) provides a series of examples of such structures. For example, the complex term *malignant neoplasm of renal calyx* is found to contain two principal terms: *malignant neoplasm* from the *disorder* hierarchy and *renal calyx* from the *body structure* hierarchy. It is important to note that AnaMed also identifies other nested terms here, such as the body structure *calyx*, the qualifier *malignant* and the disorder *neoplasm*.

Term	**Grouping**	**Structure**
structure of radial tuberosity	*structure of radial_ tuberosity*	structure+of+BODYSTR
Baelz's disease	*Baelz's disease*	EPONYM+'S+DISORDER
malignant neoplasm of renal calyx	*malignant_ neoplasm of renal_ calix*	DISORDER+of+BODYSTR

Table 5.2 – structures and groupings obtained using AnaMed.

The structures were generated using all nested terms (see the final column

in Table 5.2), and are classified in accordance with number of appearances and number of dependencies. In other words, we counted the number of terms in which each structure appears, as well as the number of other terms in which said term appears in nested form. By way of example, some of these structures are shown in Table 5.3 (the abbreviation *Appear.* refers to the number of times the structure appears and *Depen.* refers to the number of dependencies). For example, 4,469 appearances were found for the QUALIFIER+DISORDER structure; moreover, said structure was found to appear in nested form in 74,208 terms. Since it has a high number of appearances and, moreover, is very important when translating complex terms into Basque, this structure was classified as high priority. The case of the structure QUALIFIER+*neoplasm* is slightly different, since despite only appearing as such in 4 terms, the dependency of other complex terms on these four terms is extremely high (the structure appears in nested form in 28,642 terms).

Structure	**Example**	**Appear.**	**Depen.**
QUALIFIER+DISORDER	*unstable diabetes mellitus*	4,469	74,208
QUALIFIER+*neoplasm*	*malignant neoplasm*	4	28,642
PROCEDURE+of+BODYSTRUCTURE	*amputation of finger*	5,082	33,181
...	...	...	...

Table 5.3 – Appearances and dependencies on other terms of the SNOMED CT term structures.

We gave examples of these structures to two experts so that they could provide us with correct equivalents in Basque. These examples were then used as the basis for defining the Basque translation patterns. Since the knowledge possessed by experts in the field is vital to finding suitable equivalents for terms, the experts responsible for translating these examples into Basque were both physicians. The process for obtaining the examples was divided into two phases, in order to enable the patterns and examples used in the second phase to be selected in accordance with the structures obtained during the first one.

A total of 41 structures were chosen for the first phase, each with at least 3 randomly-selected examples. As shown in Figure 5.2, one expert was given 28 structures and the other 27, with 14 structures being given to both. The output consisted of Basque equivalents for 100 and 97 examples, with 58 being common to both experts.

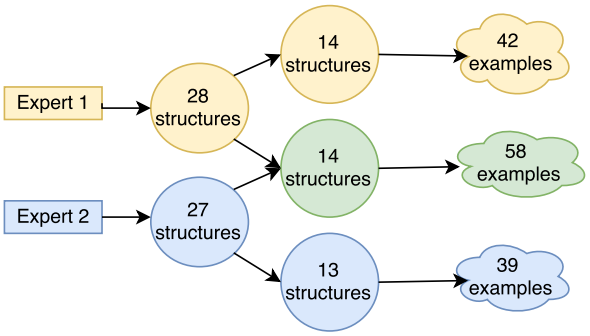


Figure 5.2 – Sample provided to the two experts.

In the case of the examples given to both experts, the level of agreement was generally high, and wherever their opinion diverged an agreement was reached, with both experts employing the same set of criteria. The table below (Table 5.4) shows a number of these examples. In the first one, the two experts initially proposed different equivalents (although an agreement was subsequently reached); in the second one they were in total agreement; and in the third and fourth ones the example was given to only one of the experts. As evident in the examples provided, the terms were far from simple and a thorough knowledge of medicine is required to render them correctly in the Basque language.

English	**Expert 1**	**Expert 2**
cryotherapy to cranial nerve	nerbio kranialaren krioterapia	garezurreko nerbioen krioterapia
calcium regulating agent overdose	kaltzioaren agente erregulatzaileek eragindako gaindosia	kaltzioaren agente erregulatzaileek eragindako gaindosia
open fracture of scaphoid bone of wrist	eskumuturreko eskafoide hezurraren haustura irekia	
adrenergic neurone blocking drug adverse reaction		neurona adrenergikoen blokeatzaileek eragindako kontrako efektua

Table 5.4 – Some example of the Basque translations provided by the two experts.

Once the basic criteria had been established, for the second phase we selected two sets of 25 structures, with each expert receiving 100 examples.

After combining the two phases we obtained around 340 examples on the basis of which to define the Basque translation patterns. As a result of this process we finally defined a total of 53 such patterns.

The design of the KabiTerm system

Having described the process used to obtain the Basque translation patterns, we will now examine how they are used by KabiTerm. KabiTerm's operating process is shown in Figure 5.3:

1. First of all, AnaMed analyses the input term, identifying and grouping any nested term contained within it. In the case of the term *fracture of nasal bones*, *fracture* is a disorder and *nasal_ bone* a body structure. In addition to grouping the nested terms, lemmatisation is also required in this example, since while nasal bone is a SNOMED CT term, *nasal bones* (the plural form) is not.

2. Secondly, the system calls the transducer responsible for identifying the Basque translation patterns and tagging the nested terms, and this transducer uses the appropriate Basque translation pattern to attach the tags required to translate the nested terms into Basque. In the case in question, the transducer identifies the structure DISORDER+of+BODYSTRUCTURE and applies the appropriate Basque translation pattern. It tags *fracture* with "|DIS" because it is a disorder, and it tags *nasal_ bone* with both "|BOD+Eko" and "|BOD+areM" because in addition to being a body part, the term also requires a declension ("+Eko" and "+areM" in this particular case). In addition to all this, the transducer also adds a change of order tag, indicating that the first term should be moved to the end ("&LehenaAzkenera").

3. The next step involves rearranging the nested term, following the instructions provided in the tag added during the previous step ("&LehenaAzkenera"). Thus, *fracture* moves from first to last place in the term.

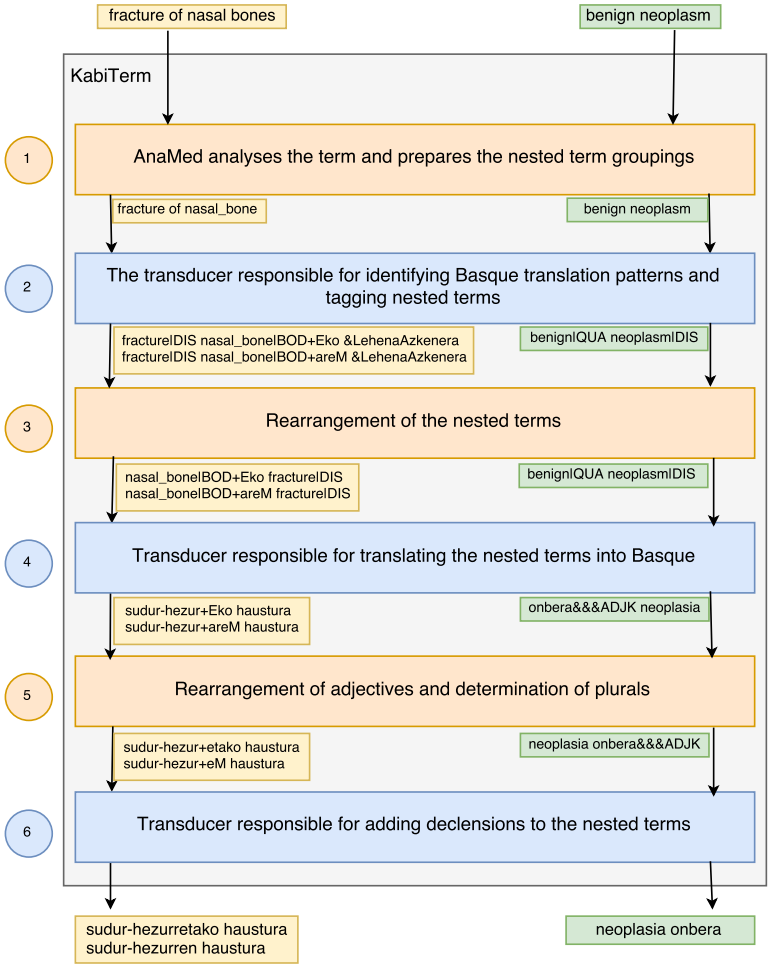


Figure 5.3 – Examples of KabiTerm's architecture and functioning.

4. In the fourth step the system calls up the transducer responsible for translating nested terms into Basque. This transducer provides us with two Basque equivalent terms: "sudur-hezur+Eko haustura" and "sudur-hezur+areM haustura" (the hierarchy tags disappear and the output is the Basque equivalent of each English term).

5. Next, since one of the nested terms was plural in the original English, its declensions are updated to reflect this plural status: "+Eko" becomes "+etako" and "+areM" becomes "+eM". Even though this is not the case in the example being used here, if the term contains an adjective, then said adjective is rearranged also during this step. In this case, the outputs are "sudur-hezur+etako haustura" and "sudur-hezur+eM haustura", since the input term was plural (*fracture of nasal bones*).

6. Finally, the transducer is called up once again to add the declensions to the nested terms, thus obtaining the compound Basque terms "sudur-hezurretako haustura" and "sudur-hezurren haustura".

A number of factors were taken into consideration in the Basque translation process outlined above. In relation to the genitive case, it is often difficult to determine whether the locative genitive or possessive genitive should be used. For example, for the term *abdominal aorta*, Euskalterm uses the locative genitive ("abdomeneko barrunbe"), while Anatomiako Atlas uses the possessive genitve ("abdomenaren barrunbe"). On consulting with our experts, the criterion became clear: when indicating location, the locative genitive should be used; when the aim is to indicate the relationship between the whole and a part of the whole, then the possessive genitive is the correct choice (Zabala *et al.*, 2012). However, the means to automate this criterion is unclear, and expert opinions are vital in order to determine and understand context. In light of this situation, and given that the aim of this thesis was not to generate reference equivalent terms in the Basque language, but rather a series of possible equivalents, we opted to err on the side of overproduction in such cases, offering outputs reflecting both possibilities.

As explained in the figure above, the automatic translation process into Basque is carried out in six separate steps. Step one involves the analysis of the input term and the grouping of the nested terms. Step two involves the identification of structures and the provision of the information required for translating the term into Basque, through the use of tags indicating the hierarchy of the nested terms, declension markers, order changes and the

deletion of certain elements. Step three is the rearrangement of the nested terms. Step four is the automatic translation of those nested terms, using the information obtained in step two. In step five any adjectives are rearranged and if any of the nested terms are plural, their declension tags are updated. Finally, in step six, the correct declensions are added to the translated nested terms.

In steps two and four, bilingual lexicons are used to both identify and translate the nested terms. These lexicons are made up of SNOMED CT terms, or in other words, they are made up of terms for which Basque equivalents have already been established. In order to be enable different equivalent terms in each hierarchy, and because structure identification depends on this system, we generated a separate lexicon for each hierarchy (disorder, body structure, etc.).

This section examines steps two, four and six in more detail. We defined transducers for all three of these steps, with the aim of carrying out the automatic Basque translation process. The remaining three steps (one, three and five) are responsible for preparing and managing this process.

The transducer responsible for identifying Basque translation patterns and tagging nested terms

Each pattern used in this step was established on the basis of four rules. The first identifies the structure; the second adds the tags; the third adds the identifier which corresponds to the pattern; and the fourth combines the previous three, eliminates any English words that need to be eliminated and prepares the tags corresponding to the Basque equivalent term. The figure below (5.4) shows the entire sequence of rules for a single pattern.

```
1  define Dis    ?+ @-> ... {|DIS} || Muga _ Muga ;
2  define Bod    ?+ @-> ... {|BOD} || Muga _ Muga ;
3  define SinGEN ?+ @-> ... {+areM} || Muga _ Muga ;
4  define GEL ?+ @-> ... {+Eko} || Muga _ Muga ;
5  define KenOf " " {of} " " -> " " ;
6  define OrdAldatuLehenaAzkenera ?+ @-> ... " " {&LehenaAzkenera} ;
7  ####################
8  define EzDisOfBod HDIS "\ "\ {of} " " HBOD;
9  define DisOfBod Dis "\ "\ {of} " " (Bod .o. [SinGEN|GEL]);
10 define EtDisOfBod ?+ @-> ... { |pat_or_011};
11 define TrDisOfBod EzDisOfBod .o. DisOfBod .o. KenOf .o.
       OrdAldatuLehenaAzkenera .o. EtDisOfBod;
```

Listing 5.4 – An example of a KabiTerm identification and tagging pattern.

The general rules for adding tags are shown in lines 1 to 6: add the hierarchy tag (lines 1 and 2); add the declension markers (lines 3 and 4); eliminate the English preposition *of* (line 5); and finally, add the change of order tag (line 6). In this case, the tag instructs the system to move the first element to the end of the term.

The rules for identifying and tagging the Basque translation pattern for the DISORDER+of+BODYSTRUCTURE structure are shown in lines 8 to 11. First of all, the rule for identifying the structure is defined under the name `EzDisOfBod`: a term from the disorder lexicon (i.e. a disorder taken from the lexicon saved in the `HDIS` rule), the proposition *of* and finally, a term from the body structure hierarchy (`HBOD`). Next, in line 9, the tags that need to be added to the nested terms are established by the `DisOfBod` rule, and line 10 adds the general pattern tag (`EtDisOfBod`) , which is used to control development and provide results. Finally, in addition to combining the previous three rules, the rule which eliminates the preposition *of* and the ones which change the order of the terms are also combined using the `TrDisOfBod` rule.

Some examples of the Basque translation patterns used in this second step are given in Table 5.5. The first rule comprises an eponym and a disorder. The hierarchy of each term is specified in the pattern: the `Epo` rule adds the eponym marker, and the `Dis` rule adds the disorder marker. Moreover, as evident in the Basque equivalent term, we also add a hyphen between the eponym and the possessive genitive declension. This is done through the `MarGEN` rule. It should be remembered that in Basque, eponyms are generated both with and without declension markers. In other words, the system will generate both "Down-en sindrome" and "Down sindrome", even though in English, terms containing eponyms are usually phrased as appositions[7] (*Down syndrome*). In Spanish, on the other hand, a preposition is used to construct the syntagm ("sindrome de Down"). Given the characteristics of the Basque language, the most natural form would be an apposition ("Down sindrome"), but due to language contamination from Spanish, the other form ("Down-en sindrome") is now also widely used. Moreover, in terminological reference resources such as Euskalterm, the inflected forms appear more frequently. As with the genitive case, here too we opted to err on the side of overproduction and leave the task of selecting the best term and ignoring the

[7]We refer here to structures comprising two nouns, one of which explains or specifies the other.

other alternatives to the experts.

	English	Basque	Rule
1	*Down syndrome*	Down-en sindrome Down sindrome	`(Epo .o. (MarGEN)) " " Dis`
2	*head structure*	buruaren egitura	`(Bod .o. SinGEN) " " Bes`
3	*heroin overdose*	heroinak eragindako gaindosi	`(Phar .o. ERGEra) " " Bes`
4	*fracture of hip*	aldakako haustura aldakaren haustura	`Dis" "{of}" "(Bod.o.[GEL|SinGEN])`
5	*benign neoplasm*	neoplasia onbera	`Qua " " Dis`

Table 5.5 – Some examples of rules used in the identification and tagging step of KabiTerm.

In the second example, we use the`Bod` rule to add the body structure tag, applying also a singular possessive genitive marker using the `SinGEN` rule. The word *structure* does not appear in any hierarchy, and is added to the "others" list using the `Bes` tag. In addition to featuring words or terms that do not appear in any of the hierarchies, this list (`Bes`) also contains terms that are used to define a specific pattern. The third example in the table is an example of this. Even though the English term *overdose* appears in the disorder hierarchy, we created a specialist rule for it here, including it on the "others" list (`Bes`). In this case, as well as adding the ergative case marker, the `ERGEra` rule also attaches the tag for adding the word "eragindako".

In the fourth example (at least) two Basque equivalents are generated using this rule since DISORDER+of+BODYSTRUCTURE is so general that, as explained earlier, both the possessive genitive and the locative genitive are added to the body structure in order to generate the Basque equivalent term.

Finally, in the fifth example we find a qualifier followed by a disorder, and the tags for these are added to the nested terms. In this case, no declension or rearrangement markers are required since, as we will see later on, adjectives are only rearranged when they come after the noun.

Continuing with the examples given in the table above (Table 5.5), Table 5.6 shows the outputs generated by the first phase transducers.

Transducer responsible for translating nested terms into Basque

In the fourth step we use bilingual lexicons to translate the tagged nested terms into Basque. As in the previous step, the lexicons are included in rules such as `HDIS` (disorders) and `HBOD` (body structures) (one lexicon per

	English	Outputs of the 2nd step
1	*Down syndrome*	Down|EPO+-+ReM syndrome|DIS Down|EPO syndrome|DIS
2	*head structure*	head|BOD+areM structure|BES
3	*heroin overdose*	heroin|PHAR+ak_eragindako overdose|BES
4	*fracture of hip*	fracture|DIS hip|BOD+ko &LehenaAzkenera fracture|DIS hip|BOD+areM &LehenaAzkenera
5	*benign neoplasm*	benign|QUA neoplasm|DIS

Table 5.6 – Some examples of the outputs produced by the identification and tagging step of KabiTerm.

hierarchy). The lexicon applied is selected in accordance with each term's assigned tag, as shown in the code appearing between lines 1 and 10 in Figure 5.5. After the terms have been translated into Basque, the tags indicating hierarchy are deleted in line 11. In line 12 we combine all the rules, and finally, in line 13, these combined rules are applied to all nested terms. It should be borne in mind that complex nested terms (i.e. those containing more than one word) must be grouped using underscores (_) in order for Foma to work properly. Therefore, elements separated by a blank space are considered separate entities.

```
1  define IDIS HDIS "|DIS" ;
2  define IFIN HFIN "|FIN" ;
3  define IEPO HEPO2 "|EPO" ;
4  define IBOD HBOD "|BOD" ;
5  define IPROC HPROC "|PROC" ;
6  define IBEST HBEST "|BES" ;
7  define IPHAR HPHAR "|PHAR" ;
8  define IOBV HOBV "|OBV" ;
9  define IQUA HQUA "|QUA" ;
10 define ITZULE  [ IDIS | IEPO | IBOD | IPROC | IBEST | IPHAR | IFIN |
       IOBV | IQUA] (ETIKETAK);
11 define CLEANUP [ "|BOD" | "|EPO" | "|DIS" | "|FIN" | "|BES" | "|PROC" |
       "|PHAR" | "|OBV" | "|QUA" ] -> 0 ;
12 define ITZUL  ITZULE .o. CLEANUP ;
13 regex ITZUL [" " ITZUL]* ;
```

Listing 5.5 – Basque translation of nested terms transducer patterns in KabiTerm.

The table below (Table 5.8) shows the outputs from the previous transducer (identification and tagging) along with the outputs from this step (Basque translation of nested terms). It should be remembered that the

third step (rearrangement of the elements) is carried out in between. If we look at the fourth example in the table, we can see that the order has been changed as a result of this third step. Changing the order of the elements with Foma is extremely complex, which is why the nested terms are rearranged outside the transducer, as explained earlier.

	2nd step outputs	**4th step outputs**
1	Down|EPO+-+ReM syndrome|DIS	Down+-+ReM sindrome
	Down|EPO syndrome|DIS	Down sindrome
2	head|BOD+areM structure|BES	buru+areM egitura
3	heroin|PHAR+ak_eragindako overdose|BES	heroina+ak_eragindako gaindosi
4	fracture|DIS hip|BOD+ko &LehenaAzkenera	aldaka+ko haustura
	fracture|DIS hip|BOD+areM &LehenaAzkenera	aldaka+areM haustura
5	benign|QUA neoplasm|DIS	onbera&&&ADJK neoplasia

Table 5.7 – Some examples of the outputs produced by the Basque translation of nested terms transducer in KabiTerm.

Transducer responsible for adding declensions to the nested terms

Finally, in the sixth and final step, the system uses the declension markers to generate complex terms in the Basque language. The rules for the declension transducer were taken from the Xuxen spellchecker (Agirre *et al.*, 1992) in order to ensure that Basque morphological rules are respected at all times. For example, the "+Eko" marker is used to generate the indefinite possessive genitive. As in the previous example, when the word "hezur" is combined ("hezur+Eko"), the "r" becomes hard and is included in the "e" form, resulting in "hezurreko". On the other hand, when the word "birika" is combined ("birika+Eko"), since the lemma ends in "a", the "e" of the locative genitive disappears, resulting in "birikako".

As before, the order of the adjectives is not managed within the transducer, which is why these elements are rearranged in the 5th step of the process. There are two kinds of adjectives in the Basque language, those that go after the noun (izenondo) and those that go before it (izenlagun). If we look at the fifth example, *benign neoplasm*, we see that phase two produced the nested term "onbera" with the izenondo tag ("&&&ADJK"), and through the main application, the term "onbera" was then moved to after the noun.

	4th step outputs	6th step outputs
1	Down+-+ReM sindrome	Down-en sindrome
	Down sindrome	Down sindrome
2	buru+areM egitura	buruaren egitura
3	heroina+ak_eragindako gaindosi	heroinak_eragindako gaindosi
4	aldaka+ko haustura	aldakako haustura
	aldaka+areM haustura	aldakaren haustura
5	onbera&&&ADJK neoplasia	neoplasia onbera

Table 5.8 – Some examples of the phase in which declensions are added to the nested terms in KabiTerm.

Some details regarding the Basque translation patterns

We have tried to make the patterns as broad-ranging and general as possible. Nevertheless, in some cases, even though we wanted to develop a pattern that would apply to all the terms in a given hierarchy, this proved impossible since not all the terms in said hierarchy translate in the same way into Basque. The following are some of the phenomena we took into consideration:

- The various ways in which prepositions are translated into Basque: there are many English prepositions in the qualifier hierarchy, but not all translate in the same way into the Basque language. For example, the preposition *with* is usually translated using the comitative case, *on* with the locative case and *to* with the allative case. However, in our rules, declension markers are not added through the lexicons and this changes our system considerably. Consequently, we only take into account those cases that appear frequently, such as [PROCEDURE] + to + [BODYSTRUCTURE], [PROCEDURE] + on + [BODYSTRUCTURE] and [DISORDER] + with + [DISORDER].

- The order of adjectives: in the case of the qualifier hierarchy, we initially thought it would be impossible to generalise because adjectives often change position when translated into Basque. Nevertheless, even though it did prove impossible with prepositions, we were able to generalise in the case of other qualifiers, thanks to the difference in Basque between an izenondo (an adjective that goes after the noun) and an izenlagun (an adjective that goes before the noun). However, the majority of words in this hierarchy are adjectives, and in Basque the order

of adjectives can be fairly flexible. As mentioned above, if the word is an izenondo then we move it to after the noun, but we leave all the others where they are, since the order in Basque is the same as in the English term. As in the previous examples, the terms *benign neoplasm* and *congenital cyst* from the [QUALIFIER]+[DISORDER] hierarchy are translated into Basque as "neoplasia onbera" and "jaiotzetiko kiste". Since "onbera" is an izenondo it is moved to after the noun, and since "jaiotzetiko" is an izenlagun it is left where it is. Bearing this in mind, the information regarding whether an adjective is an izenondo or an izenlagun is stored in the lexicons.

- Nouns in the genitive: some Basque translation patterns require the genitive case to be added to a nested term. In such cases, we need to be sure that the term is in fact a noun, since the genitive case marker cannot be added to other parts of speech. When the term in question is not a noun, it is left as it is with no declension marker. For example, in the English term *hypertrophic rhinitis*, *hypertrophic* belongs to the body structure hierarchy (because morphological anomalies are located in this hierarchy), and *rhinits* is a disorder, so the BODYSTRUCTURE+DISORDER structure's equivalent in Basque would be the DISORDER+aren or +ko BODYSTRUCTURE structure. Following this pattern, the Basque equivalent term would be "errinitis hipertrofikoko" or "errinitis hipertrofikoaren", which are incorrect. These types of patterns (i.e. those that require the genitive case) are valid for nouns, not adjectives.

- Plural nested terms: in order to broaden the system's coverage, when the exact forms of the terms are not found in the lexicons, a search is conducted for their lemmas. AnaMed uses the same strategy for term identification. If the term is a plural one, the singular form is sent to the transducer in order to avoid unnecessary duplications in the lexicons. Thus, before calling the third-phase transducers, a plural marker is added to the Basque equivalent of the plural nested term, in order to ensure that the declension is added in the correct way. For instance, in the example given in Figure 5.3, the term *nasal bones* does not appear in SNOMED CT, but its singular form (*nasal bone*) does. As shown in the example, thanks to AnaMed we maintain the singular form of the term until the time comes to add the declension marker.

- Limiting the number of Basque translations: the system's principal problem is overproduction. In order to overcome this problem, we limited the dictionaries by generating a black list and removing any equivalents identified as incorrect during the development of the system, thus rendering it increasingly more effective by teaching it what we ourselves have learned from experience. For example, the term *head* has twelve equivalents in the anatomy dictionary, including "buru", "humeroaren buru" and "falangearen buru", for instance. While the term *head* does indeed have all these meanings, the most common one is the general term "buru", with the body structure that precedes it lending it greater specificity. Bearing this in mind, with the exception of "buru", all the other meanings were relegated to the black list in an effort to limit overproduction. The lexicons were then manually reviewed and any pairs with more than 4 equivalents were revised. The result was that no pair now has more than 2 equivalents, in order to avoid excessive overproduction. For example, without overproduction control, KabiTerm would generate 360 equivalent Basque terms for the English term head of *head of seventh rib structure*, whereas following the changes made to limit overproduction, the system only generates 4 equivalents. Obviously, producing 360 equivalents is counter-productive, even if some of them are correct.

- Ordinal numbers and one-letter terms: the forms used to express ordinal numbers (*seventh*, for example) have three possible equivalents: the ordinal number itself (zazpigarren), the number itself (zazpi) and the fraction (zazpiren). In SNOMED CT, most terms refer to ordinal numbers, which is why only those equivalents were included in the lexicons. Also, no equivalents are sought for one-letter terms, with said terms being left as they are. In SNOMED CT, single-letters generally refer to groups or types (*type C thymoma*, for example) rather than musical notes or other similar equivalents (*C* can also refer to the musical note Do, for example).

Now that we have presented the KabiTerm system, we will now turn our attention to MatxinMed. MatxinMed is the adaptation of the Matxin machine translation tool to the medical domain, based on the terms included in SNOMED CT.

5.2 Adapting Matxin to the medical domain

This section outlines the final step of the algorithm. In this fourth and final step we adapted the Matxin machine translation tool to the medical domain. In the Background section (Section (5.2.1) we provide some general information about machine translation systems and present some of the tools currently available for the English-Basque language pair. Then, in Section 5.2.2 we describe how we adapted the Matxin system.

5.2.1 Background

The aim of machine translation is to produce automatic computer-generated translations from one language to another. Machine translation has always been one of the star applications of the Natural Language Processing (NLP) technique and continues to arouse a great deal of interest as a possible means of ensuring mutual understanding in today's globalised world.

There are two main approaches to the development of effective machine translation systems: Rule-Based Machine Translation (RBMT) and Corpus-Based Machine Translation (CBMT). Rule-based systems use our linguistic knowledge of a given language as the basis of their procedures, whereas the more empirical corpus-based ones use previous translations as their starting point.

While in this section we will use some specific references for certain examples and cases, our main sources were Jurafsky and Martin (2008), Mayor (2007) and Artetxe (2016).

The sub-sections below outline the different machine translation paradigms, along with the systems designed for the specific English-Basque language pair.

Rule-Based Machine Translation

Rule-Based Machine Translation uses linguistic information about the source and target languages to translate from one to the other. The translation process is divided into three phases: analysis, transfer and generation.

During the analysis, the text to be translated (the source text) is analysed using the usual natural language processing chain, which comprises morphological, grammar (*POS tagger*) and syntactic analysers. The outcome of these analyses is an intermediate representation.

During the transfer phase, the intermediate representation in the source language is transferred to the intermediate representation in the target language. There are two types of transfer: syntactic (or superficial) transfer and semantic (or deep) transfer. Syntactic transfers take place at the lexicon or structure level, using bilingual dictionaries and transfer rules. Semantic transfers take place at the semantic level, using complementary structures to express meaning.

Finally, in the generation phase, a translation is obtained from the intermediate representation in the target language, often using morphological dictionaries.

Three different kinds of system can be distinguished in accordance with the information used and the level of abstraction (Hutchins and Somers, 1992). These strategies are often explained using Vauquois' pyramid, which is shown in Figure 5.6: direct translation, transfer-based systems and interlingua-based systems.

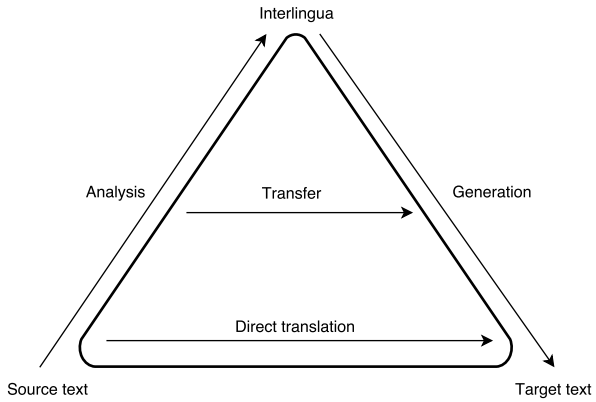


Figure 5.6 – Vauquois' pyramid.

According to Vauquois' pyramid, direct translation at the base of the pyramid does not use analysis, and interlingua-based systems at the apex do not require transfer. As we move towards the apex of the pyramid the intermediate representations become more complex and require deeper analysis.

Corpus-Based Machine Translation

Corpus-based machine translation (CBMT) employs empirical methods to make use of manual translations. Over recent years, thanks to the increased

capacity of today's computers and the huge number of texts available on the Internet, corpus-based techniques have improved dramatically. In relation to machine translation, the main resources are parallel corpora[8].

There are two types of CBMT, Example-Based Machine Translation or EBMT and Statistical Machine Translation (SMT). The EBMT process pairs sentence chunks with examples from the corpus, identifies their translation chunks and recombines these chunks to generate a translation. The SMT process, on the other hand, generates translations from statistical data drawn from large parallel corpora without the use of any explicit linguistic information.

SMT systems are the ones which have been developed most over recent years by the research community. There are three different types of statistical system: word-based systems, phrase-based systems and hierarchical phrase-based systems. Although initially word-based systems were the most common, today phase-based ones are more frequently used[9]. However, thanks to the successful emergence and development of deep learning techniques, traditional SMT systems are increasingly being replaced by Neural Machine Translation systems (Sennrich *et al.*, 2016).

Two models are used in the translation process carried out in phrase-based SMT systems: the translation model itself and the language model. While the bilingual corpus is used to generate the translation model, the monolingual corpus of the target language is used to generate the language model.

The translation model determines the probability of the source language phrase (or string) when it is translated into the target language. To this end the phrase-based model distributes the paired source and target phrases (without taking any linguistic information into account) and calculates the probabilities for that pairing. For further information about these probability calculations, see the paper by Koehn *et al.* (2003).

The language model, on the other hand, determines the probability of the phrases generated in the target language. Although many models have been proposed, most are based on n-grams. Thus, the probability of a phrase is calculated on the basis of the probability of each word appearing in conjunction with its preceding word.

[8]In parallel corpora parallel texts are arranged in pairs at segment level. Pairing is generally carried out at sentence level.

[9]The word phrase does not refer here to linguistic syntagmata, but rather to word sequences.

Of all the statistical systems available, Moses (Koehn *et al.* 2007) is the most widely-used. Moses Statistical Machine Translation is a free software tool which can be used to train and tune SMT systems. Moses makes it much easier to develop a statistical system and minimises the time required to obtain a functional program.

The quality of any corpus-based machine translation system is closely linked to the size of the parallel corpora. This is even more true when the two languages in the language pair have very different characteristics (i.e. are distant languages), as in the case with Basque and English, for example.

English-Basque Machine Translation Systems

Most systems currently available for the English-Basque language pair are rule-based systems, probably due to the distance between the two languages and the lack of large-scale parallel corpora. As stated above, when two languages are fairly distant, an even larger than normal corpus is needed to ensure the quality of the machine translations.

The following is a description of some of the main machine translation systems available for the English-Basque language pair.

- **Matxin** (Mayor *et al.*, 2011) is an open-source system developed by the IXA Group. Even though it was originally developed for the Spanish-Basque language pair, it was later adapted to the English-Basque pair (Aranberri *et al.*, 2015). It is an open-source rule-based machine translation system which follows the classical transfer-based system architecture and comprises three main modules: source language analysis, transfer from source to target language and target language generation (Figure 5.7).

  The analysis module uses the Stanford CoreNLP tool to analyse the English source text. This analysis provides the following data: information about the individual words (POS and morphological inflection), chunks (inter-chunk dependency) and phrase types.

  In the transfer module, two types of information are managed: lexical and structural. The lexical transfer process ensures that the correct Basque equivalents are obtained for the lemmas, using dictionaries, and the structural transfer process focuses on morphosyntactical characteristics, changing the order of chunks and words as required.

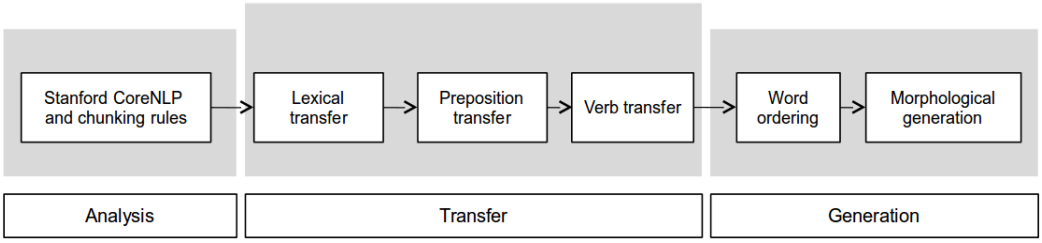


Figure 5.7 – General architecture of the Matxin system. Image taken from the paper by Aranberri (2016).

The generation module is divided into two main steps. In the first step, the word order within the chunks is rearranged, along with the order of the chunks themselves. Moreover, the information obtained at chunk level is shifted to the inflected word (in Basque, this is the last element of the chunk). The second step comprises morphological generation, in which the correct form is obtained from the source lemmas with the help of a morphological dictionary (in this case the Euskararen Datu-Base Lexikala-Basque Lexical Database, EDBL (Aldezabal *et al.*, 2001).

- **TectoMT** is a highly modular rule-based machine translation system (Popel and Žabokrtskỳ, 2010). The system uses syntax as the basis for its transfers and conducts a much deeper analysis than Matxin, obtaining a greater level of abstraction. It therefore uses techogrammar (Hajicová, 2000, which represents language through a deep syntax dependency tree. Despite being a rule-based system, it uses statistical techniques in some of the translation process modules. Although it was originally designed to translate from English to Czech, it was recently adapted to the English-Basque language pair as part of the QTleap[10] project (Aranberri *et al.*, 2016b).

- **Google**'s free translation program, Google Translate, is widely known all over the world. From the time it was first published in 2001 to around 2005-2006, it was based on a rule-based system called Google Translate Systran and initially offered translations between English and

[10]http://qtleap.eu (accesed May 9, 2017)

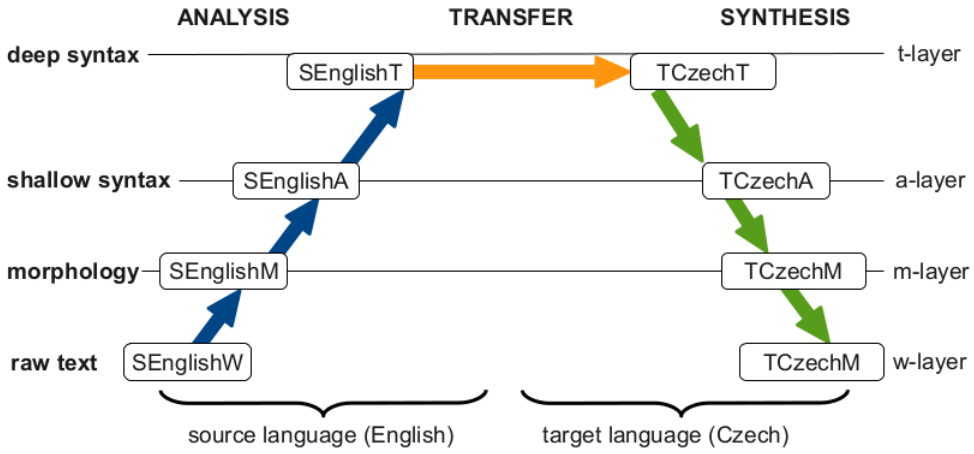


Figure 5.8 – General architecture of the TektoMT system. Image taken from the paper by Popel and Žabokrtskỳ (2010).

eight other languages. In 2005 it began to use statistical systems in order to enable it to expand in the future to include all languages, and that same year won first prize at the *NIST DARPA TIDES Machine Translation Evaluation* competition for its Arabic-English and Chinese-English statistical systems[11]. Statistical translation systems use parallel documents published by the European Union and United Nations, along with parallel data drawn from websites on the Internet. In 2010 the *alpha* version for the Basque language was published and today, the translation system encompasses 90 different languages. When translating between languages with small parallel corpora, the Google Translate tool uses English as a pivot. In other words, it first translates from the source language into English, and then from English into the target language.

Very little is known about how the Google Translate system actually works, and its developers limit themselves to publishing only the latest news and information about its general structure. Therefore, despite being free to use, the source code is not available to enable adaptations.

- The **Lucy** system was developed as the result of a Basque Government

[11]Taken from NIST http://www.itl.nist.gov/iad/mig//tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html (accesed May 9, 2017).

project, first of all for Spanish-Basque and later on for English-Basque. Information about this translation system is hard to come by, since as it is a commercial system, no details have ever been given either at conferences or in specialist journals. Nevertheless, the following can be concluded from the slides of a presentation found on the web (Gieselmann, 2008) and the details published on the HABE (Institution for Basque Adult Literacy and Learning) website[12] i) this rule-based system uses both morphological and syntactical analyses; ii) in addition to lexical and structural transfer, it also carries out contextual transfer; and iii) in addition to morphological generation, it also carries out a of other processes depending on the target language.

- **EuSMT** is a statistical system developed by the IXA Group. It is based on Moses and adopts two approaches to the English-Basque language pair. The first one, known as the baseline approach (**EusSMTo**), is a standard phrase-based SMT system developed using Moses. 85% of the parallel corpus used during the training stage was taken from Elhuyar's translation memories, and the remaining 15% was drawn automatically from the web using the PaCo2 tool (Vicente and Manterola, 2012).

  The second approach adds segmentation to the system (**EuSMTs**). SMT systems work better when the two languages involved are similar, i.e. when the source and target languages share similar grammatical features. In this case, the English language is mainly analytical, with each morpheme having a word; Basque, on the other hand, is an agglutinative language, with morphemes being combined in order to produce words. Segmentation is used to help forge links between the two languages (Al-Haj and Lavie, 2012; Naradowsky and Toutanova, 2011). Segmentation splits the words of the agglutinative language up into morphemes, making it easier to pair with its analytical counterpart.

- **SMatxinT** is a hybrid system developed by the IXA Group which blends the two approaches of the EuSMT system with the Matxin tool, in accordance with that proposed by España Bonet *et al.* (2011) and Labaka *et al.* (2014) . As the developers themselves state, in general, RBMT systems are better at syntactical rearrangement, while CBMT

[12]http://www.habe.euskadi.eus/s23-4728/es/contenidos/noticia/tzulpenautomatikoa_mintegia16/es_def/index.shtml (accesed May 9, 2017)

systems offer better lexicon selection, which is why they decided to develop a hybrid system (see Figure 5.9).

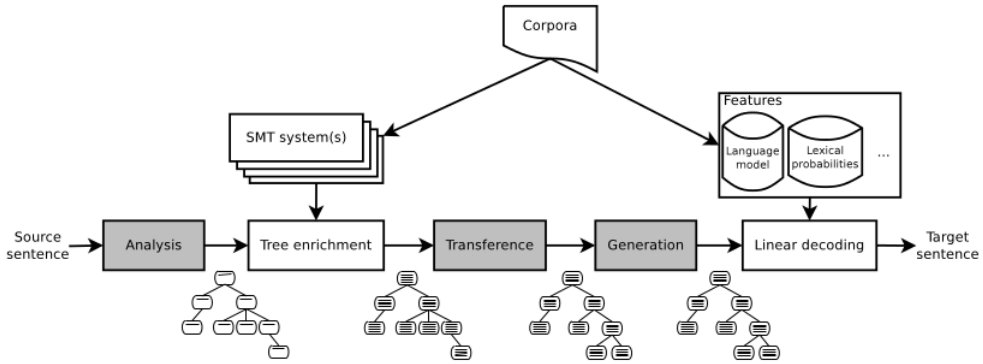


Figure 5.9 – General architecture of SMatxinT, with the RBMT modules marked in grey. Image taken from the paper by España Bonet *et al.* (2011).

The table below (Table 5.9) provides a summary of the English-Basque language pair systems described above.

	RBMT	CBMT
Matxin	✓	
TectoMT	✓	(✓)
Google		✓
Lucy	✓	
EuSMT		✓
SMatxinT	✓	✓

Table 5.9 – Summary of English-Basque language pair systems.

5.2.2 MatxinMed: adapting the system

Over recent years, the trend towards adapting machine translation systems to specific domains has become increasingly widespread, since it has been proven that when confined to a single domain, the quality of the translations generated in that domain is higher. Although adapting rule-based systems

to a specific domain requires the adaptation of dictionaries and grammars, since this involves an enormous amount of manual labour, said adaptations are usually limited to broadening the scope of the dictionaries in question (Weijnitz *et al.*, 2004). Adapting SMT systems to a specific domain, on the other hand, requires parallel corpora for that domain.

We chose Matxin as the system to be adapted to the medical domain since, given the resources available for the English-Basque language pair, a rule-based system was the most feasible option. Since the parallel corpus for Basque in the medical domain is scarce, it would be extremely difficult at this time to adapt a statistical machine translation system. Moreover, since Matxin was developed by the IXA Group, which is based in the Basque Country, we had a research group nearby able to provide any help required during the adaptation process. Another factor taken into consideration was the ease with which dictionaries can be extended, thus enabling the system to be adapted to the medical domain with a relatively small investment.

Thus, in this section we describe MatxinMed, the adaptation of the Matxin Machine Translation system to the health domain.

It should be remembered that this is the fourth phase of the EuSnomed algorithm, and MatxinMed will only be used when the techniques developed in the previous phases fail. Since machine translation systems are designed to translate phrases, we do not expect the Basque translation of terms to be of particularly high quality, but we believe it is a better option than leaving them untranslated. The aim is to help the experts responsible for developing the Basque version of SNOMED CT as much as possible, since they will now be able to base their selection of terms on the proposed list of Basque equivalents provided.

In the next few paragraphs we will explain how we adapted Matxin to the medical domain, describing in detail the changes made. The adaptation mainly focused on the lexical transfer module. We will first describe Matxin's domain selection function and the dictionary expansion process. We will then focus on the integration of the NeoTerm system, before presenting the language model used to specify the order of the generated equivalents. Finally, we will outline the new rules established for identifying complex terms.

The domain selection function and expansion of the dictionary

As explained in the previous section, Matxin is a Rule-Based Machine Translation tool. The principal resource for selecting the lexicon is the dictionary

that is saved in a specific format by the Matxin program itself. This dictionary contains all the information required for the lexical transfer, including (among others) the translations and POS of specific words.

By way of example, some entries are shown in Figure 5.10. The fourth line contains a place name, *Croatia*, along with the Basque equivalent term "Kroazia", the sense number (`sense`) and the part of speech ("NP_IZE_LIB", the tag "IZE" indicates that it is a noun, and "LIB" indicates that it is a special place name). The fifth line contains the adjective *Croatian* along with its corresponding information ("NP_ADJ_IZO", the tag "ADJ" indicates that it is an adjective and "IZO" indicates that it is an izenondo, i.e. an adjective that goes after the noun) and the sixth line contains the common noun *cross* ("NP_IZE_ARR", the tag "ARR" indicates a common noun).

```
1 <section id="main" type="standard">
2 ...
3 <e><p><l>colon</l><r>bi\_puntu<s n="sense"/>1</r></p>
      <par n="NN_IZE_ARR" /><p><l /><r /></p></e>
4  <e><p><l>croatia</l><r>Kroazia<s n="sense"/>1</r></p>
      <par n="NP_IZE_LIB" /><p><l /><r /></p></e>
5  <e><p><l>croatian</l><r>kroaziar<s n="sense"/>1</r></p>
      <par n="NP_ADJ_IZO" /><p><l /><r /></p></e>
6  <e><p><l>cross</l><r>gurutze<s n="sense"/>1</r></p>
      <par n="NP_IZE_ARR" /><p><l /><r /></p></e>
7  <e><p><l>crown</l><r>koroa<s n="sense"/>1</r></p>
      <par n="NP_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p> </e>
8 ...
9 </section>
```

Listing 5.10 – Some examples of Matxin dictionary entries.

Terms that were added during the expansion of the dictionary as part of the adaptation process can easily be distinguished by means of the tag `dom` , which is a new characteristic added specifically for this purpose. The characteristic indicates domain, and in our case we used the "Med" tag to generate MatxinMed. If we look at line 7 of the previous example (Figure 5.10), we see that it contains the common noun *crown*. Since, in addition to belonging to the general domain, this noun also belongs to the medical domain (crown of a tooth, for example), we have added the domain tag.

Matxin dictionary entries are grouped into *sections* We added the Basque equivalent terms generated thanks to the SNOMED CT Basque translation algorithm to the dictionary. We also revised the existing automatically-generated entries and added the tag corresponding to the domain where appropriate. On other occasions, we added a new section (`section`) to the

dictionary, in this case called "medicine", and included all new pairings there. The example below (Figure 5.11) shows some entries in the *medicine* section. All belong to the medical domain.

```
1  <section id="medicine" type="standard">
2  ...
3   <e><p><l>colon</l><r>kolon<s n="sense" />2</r></p>
        <par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p> </e>
4   <e><p><l>noradrenaline</l><r>noradrenalina<s n="sense"/>1</r></p>
        <par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p> </e>
5   <e><p><l>noradrenaline</l><r>norepinefrina<s n="sense"/>2</r></p>
        <par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p> </e>
6   <e><p><l>Bacteriovoracaceae</l><r>Bakterioborakazeo<s n="sense"/>1</r
        ></p><par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p>
        </e>
7   <e><p><l>steatopygia</l><r>esteatopigia<s n="sense"/>1</r></p>
        <par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p> </e>
8   <e><p><l>sacculotomy</l><r>sakulotomia<s n="sense"/>1</r></p>
        <par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p></e>
9   <e><p><l>cholangiohepatitis</l><r>kolangiohepatitis<s n="sense"/>1</r
        ></p><par n="NN_IZE_ARR" /><p><l /><r><s n="dom"/>Med</r></p>
        </e>
10 ...
11 </section>
```

Listing 5.11 – Some examples of entries in Matxin's specialist dictionary.

As stated above, by expanding the dictionary we enabled Matxin to be adapted to a specific domain. Consequently, the `sense` and `dom` characteristics are extremely important in the new design, since words or terms are selected in accordance with their content.

When selecting the best equivalent for a word, Matxin looks at the sense characteristic of the dictionary entry and chooses the pairing with the lowest value to generate its output. Matxin also enables more complex selections, choosing an equivalent on the basis of context, although specific rules must be established for this.

We therefore defined a new function to enable Matxin to prioritise equivalents in a given domain, once that domain has been specified. The design enables an unlimited number of domains to be specified, and establishes a hierarchy between them. When searching for an equivalent term, if no results are returned for the first domain, then the system looks in the second one, then the third one, etc. If it finds more than one equivalent in a domain, then it selects the one which has the lowest `sense` value.

For example, let us imagine we want to translate a text about paediatrics into Basque. In this case, paediatrics will be the first or principal domain,

and medicine the second one. Therefore, MatxinMed will first search for equivalents in the paediatrics domain, if it fails to find any it will search in the medicine domain, and if no results are forthcoming, it will take all pairings or equivalents into consideration. You can define as many domains as you wish, and MatxinMed will examine all of them in turn, starting with the first one, until it finds an equivalent.

If we look at the dictionary examples (Figures 5.10 and 5.11 we see that there are two senses for the term *colon*. In the general dictionary (Figure 5.10) the term is translated into Basque as "bi puntu", whereas in the medical domain (Figure 5.11), it is translated as "kolon". Thanks to the adaptation of the lexical transfer module, and as shown in the example provided in Table 5.10, MatxinMed is capable of correctly translating phases such as *He has colon cancer*, since it has the resources required to select the best equivalent during the lexical transfer process.

Jatorrizko esaldia	*He has colon cancer.*
Matxinen itzulpena	Hark **bi puntu** minbizia dauka.
MatxinMeden itzulpena	Hark **kolon** minbizia dauka.

Table 5.10 – Example of Matxin and MatxinMed translations.

The NeoTerm integration module

In addition to adding a domain selection function to the lexical transfer module, we also added a module based on transliteration to be used in the medical domain. This module enables the integration of the NeoTerm system described in Chapter 4 of the previous chapter into the Matxin tool. Although we added the terminological contents of SNOMED CT to the MatxinMed dictionary, there may still be some terms that do not appear in SNOMED CT. For those cases, the Basque equivalents are obtained using NeoTerm. By integrating this module, what we did was ensure the possibility of reusing the resources generated.

When Matxin fails to find the equivalent of a word in the dictionary, it leaves the word in its original form and, where necessary, simply adds the appropriate declension. The example below (Table 5.11 shows how when the system fails to find the term *cholangiohepatitis* in the dictionary, it leaves it as it is and adds the definite singular declension ("cholangiohepatitisa"). However, when Matxin finds the equivalent term in one of the dictionaries

(in the example it finds it in the MatxinMed dictionary), it uses it in the output ("kolangiohepatitisa").

Source phrase	*He has cholangiohepatitis.*
Matxin translation	Hark **cholangiohepatitis**a dauka.
MatxinMed translation	Hark **kolangiohepatitis**a dauka.
Source phrase	*He has cholangio**hypo**hepatitis.*
MatxinMed translation	Hark **cholangiohypohepatitis**a dauka.
MatxinMed with NeoTerm	Hark **kolangiohipohepatitis**a dauka.

Table 5.11 – An example of NeoTerm integration.

Even though NeoTerm is not required in this example, it nevertheless shows how Matxin reacts to an unknown word. With this in mind, it is easier to see how integrating NeoTerm means that we can also translate unknown neoclassical terms.

For example, adding the affix *hypo* to the previous example gives us the term *cholangiohypohepatitis*. Even though this term does not exist and makes no sense, it serves our purpose here by acting as an example. Without NeoTerm, MatxinMed would give the same translation as Matxin, but thanks to the integration of this module, it generates an output that is much more Basque-looking, as shown in the second part of Table 5.11.

Language model

As we have seen, some English terms have more than one Basque equivalent (see lines 4 and 5 of Figure 5.11, in which two different Basque terms are given as equivalents of *noradrenaline*: "noradrenalina" and "norepinefrina"). In Matxin's current version, the equivalent term disambiguation tools are fairly underdeveloped, which is why the order we define is extremely important, since the first equivalent term will be the one used. Nevertheless, even though only the first equivalent term is used, we believed it would be a good idea to add all the equivalents to the dictionary so that, in the event of the disambiguation tool being further developed in the future, the dictionary will already have been compiled.

In order to define the order in which the equivalent terms from SNOMED CT appear (indicated by the `sense` tag), we developed a language model. As explained in the previous section (Section 5.2.1), language models are

probabilistic models used in Statistical Machine Translation systems for the target language. They calculate the probability of a string being a viable phrase in the target language, and their main aim is to ensure that the translation is as similar as possible to the target language.

A corpus in the target language is required to generate a language model. In our case, we need a Basque corpus, and moreover we need one in the medical domain. The following sources were used to compile a domain-specific corpus:

- Medical books published by the *Udako Euskal Unibertsitatea* (UEU): over recent decades the UEU has made a concerted effort to publish academic books in the Basque language. A total of 15 books have been published in the medical field, and we were able to extract around 300,000 tokens from that volume of work.

- Notes taken by medical students at the University of the Basque Country: students on the Basque-medium courses run at the Faculty of Medicine aggregated their notes in order to compile a body of material in the Basque language. We gathered the notes and extracted text from them using an automatic system. This process resulted in around 1,200,000 tokens.

- Translation memories generated by the Elhuyar Foundation: Elhuyar has many years of experience translating both books and texts in general into Basque. From the translation memories generated during this process we extracted a parallel subcorpus specific to the medical domain containing 1,000,000 tokens.

- Texts generated by the *Osasungoa Euskalduntzeko Erakundea (OEE)*: The OEE is an association which works to promote and foster the use of the Basque language in the healthcare sector. Every year it organises a Heath Conference to analyse different health-related themes, and these meetings generate a large amount of written material in Basque. We gathered the written material generated by the Conferences held between 1996 and 2014, although we had difficulty extracting the material from certain years due to formatting problems. In total, though, 15 years' worth of conference texts were added to the corpus, obtaining over 400,000 tokens.

- Dissemination reports compiled by *Osakidetza* (the Basque Health Service): over 600,000 tokens were added to the corpus from the 41 reports compiled by the Basque Health Service for general dissemination. The reports focus on a range of different fields, including nursing, ethics in healthcare, healthcare management, occupational health, primary care and mental health, among others.

The final result of the examination of all these sources was a corpus containing over 3,500,000 tokens. Since the corpus was generated automatically, and in some cases we had to extract text from files in .pdf or .docx format, it is possible it may contain some formatting errors. Moreover, and particularly in the case of the notes provided by medical students, we have detected the presence of sections of texts in other languages (both Spanish and English), along with various spelling mistakes. However, it should be remembered that these are unpolished class notes not meant for publication, and such characteristics are typical of this type of text. Moreover, thanks to the large size of the corpus, the impact of these mistakes is fairly small and does not affect its feasibility for the purpose which interests us here, namely the generation of a language model.

It is important to bear in mind the nature of the corpus used. Even though it is a corpus in the medical domain, most of the documents it contains are academic texts (books and student notes). The nature of SNOMED CT, on the other hand, is quite different, since it is mainly comprised of clinical terminology. Nevertheless, we believe our corpus is sufficient to fill the gap which exists in this field until such time as a clinical corpus is compiled.

To generate the language model, the corpus must first be pre-processed. First of all, any empty lines were removed and the whole corpus was shifted to lower case letters. The corpus was also tokenised. The Moses model generator was used to train the model, since this module is specifically designed to prepare language models for use.

We trained two language models, one based on trigrams and another based on five-grams. A superficial analysis of the results of the two models revealed no significant differences, and in the end the decision was made to use the five-gram model. The table below (Table 5.12 presents a comparison of the results obtained using the trigram and the five-gram language models. It should be remembered that, in our case, the language model was developed to help the system choose between different equivalent terms, working on the assumption that the term assigned the highest number by the language model

is the most appropriate. In this case, the best choices are "ezkerreko besoaren haustura" and "ezkerreko besoaren haustura irekiaren infekzio" (note that the numbers in the table are negative). As shown in the table, the numbers assigned by both models are very similar, and for the purposes of the task at hand, the differences in the calculated probabilities is negligible (1.19767 and 1.181446).

	Trigram	**Five-gram**
ezkerreko besoaren haustura	-12.771531	-12.795929
ezkerreko besoko haustura	-13.969201	-13.977375
Difference	1.19767	1.181446
ezkerreko besoaren haustura irekiaren infekzio	-23.259136	-23.283657
ezkerreko besoko haustura irekiaren infekzio	-24.456806	-24.465101
Difference	1.19767	1.181446

Table 5.12 – Comparison between the trigram and five-gram models.

In addition to the domain-specific corpus, we also used a general Basque language corpus. This corpus mainly comprises phrases from the Elhuyar Foundation's translation memories, as well as phrases from the publicly available translation memories generated by the Basque Government[13] and the Gipuzkoa Provincial Council.[14] Other sources include the Basque phrases in the parallel corpus extracted from the Elhuyar Foundation's website (Vicente and Manterola, 2012) and a number of books translated into Basque by the University of the Basque Country. This general corpus contains over 100,000,000 tokens and the aim is to use it to fill in some of the gaps in the medical domain corpus, since it is logical to assume that most of the specific characteristics of the Basque language will appear somewhere in its 100 million tokens.

We trained the (five-gram) language models on both corpora and then interpolated the resulting models. In addition to amalgamating corpora, this interpolation process also enables the model to be optimised in relation to a reference corpus, also known as a *tuning* corpus. In our case, we combined the medical domain corpus and the general Basque corpus, and then optimised the resulting language model using a tuning corpus made up of a selection of SNOMED CT terms.

[13]http://www.ivap.euskadi.eus/ivapeko-itzulpen-zerbitzu-ofizialeko-itzulpen-memoriak/r61-vedorok/eu/ (accesed May 9, 2017)

[14]http://www.gipuzkoa.eus/imemoriak/ (accesed May 9, 2017)

The reference corpora used for the tuning process are usually fairly small. As stated earlier, the aim of this process is to optimise the trained model, assigning weights to the probabilities they generate in order to ensure that the resulting translations are as close as possible to the phrases contained in the tuning corpus. In our case, since we want to generate SNOMED CT terms, of the terms that had already been translated using the lexical resources, we selected only those with one single equivalent for inclusion in the tuning corpus. Given that these terms only have one single equivalent, there is no ambiguity, and since they were extracted from the lexical resources, we were able to guarantee the linguistic correctness of the proposed equivalents. This corpus contains 23,852 Basque terms and 34,583 tokens.

The figure below (Figure 5.12 illustrates the interpolation process. Language models were generated from both the medical domain corpus and the general corpus, and their probabilities (p1 and p2) are fed into the interpolation program. This program then assigns weights to these probabilities in accordance with how optimally it considers them to coincide with the texts contained in the tuning corpus.

Integrating complex terms

Matxin has a special module for recognising Multiword Lexical Units (MLUs). In this module, the information required to recognise MLUs is collected through the use of rules. The figure below (Figure 5.13) shows a couple of the rules established to identify MLUs in the medical field.

These rules were established automatically using one of the applications provided by Matxin for that purpose. The application takes the list of MLUs and establishes the rules on the basis of the linguistic analysis carried out using the Stanford CoreNLP tool. The language analysis establishes the parts of speech, thus enabling the system to determine whether or not a declension should be added to the last word in the MLU.

For example, in the case of the term *no known allergies*, the last word (*allergies*) is a plural noun. In general, when any of the words in a source term are plural nouns, then it is usually necessary for them to appear in the target term in that same form. Therefore, in order to recognise an MLU of this type, all forms must be left unchanged. In the case of the term *Alfentanil allergy*, on the other hand, since the last word is a singular noun, it can (according to Matxin) appear in either singular or plural form. The system therefore searches for the last word's lemma rather than its form.

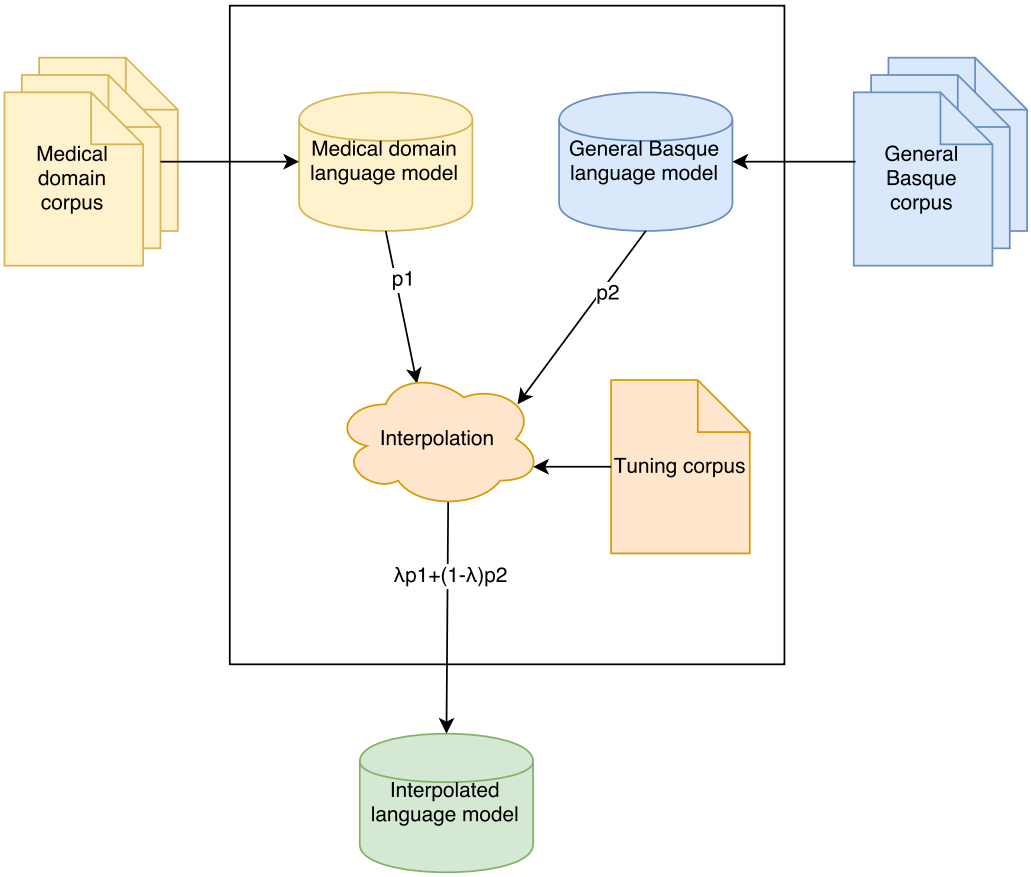


Figure 5.12 – Interpolation process for adapting the language model to SNOMED CT terminology.

The rules that are generated automatically for these examples are shown in Figure 5.13.

```
1 {pattern: (/alfentanil/ [ lemma:"allergy" ]),
      result: Concat("alfentanil_allergy", "|", $0[1].tag)}
2 {pattern: (/no/ /known/ /allergies/),
      result: Concat("no_known_allergies", "|", $0[1].tag)}
```

Listing 5.13 – Rules generated automatically for recognising MLUs in Matxin.

Thanks to these rules, Matxin is able to recognise new MLUs, which in our case are complex terms in the medical field; and as shown in the example

provided in Table 5.13, correctly identifying MLUs means that the system is able to generate better translations of phrases containing complex terms.

Source phrase	*He has Osler syndrome.*
Matxin translation	Hark **sindrome Osler** dauka.
MatxinMed translation	Hark **Osler-en sindromea** dauka.

Table 5.13 – Example of Matxin and MatxinMed translations.

Although the adaptation of Matxin carried out in this thesis is called MatxinMed, it was really a general adaptation which enables Matxin to be adapted in the future to any domain. In other words, the Matxin system can have a number of integrated domains, and the user can decide which one or ones to use by adjusting the parameters. Nevertheless, in order to make the work carried out here easier to understand, we use the name MatxinMed to refer to the Matxin adaptation to the medical domain.

In the following section we will outline the assessment carried out of the MatxinMed and KabiTerm systems. This assessment was conducted through the Medbaluatoia campaign, a crowd evaluation initiative involving the Basque healthcare community.

5.3 Assessment design

A slightly different method was used to assess the final two steps of the project. The expert evaluation of simple terms described in the previous chapter was conducted by four assessors: two linguists and two physicians. The sample assessed comprised 370 concepts and 766 terms, and the process involved a huge amount of manual work by all four experts.

To assess the systems described in this chapter, on the other hand, we used an evaluation system which involved the whole Basque healthcare community. The campaign that was designed and conducted was called Medbaluatoia, and was an adaptation of a previous campaign called Ebaluatoia (Aranberri *et al.*, 2016a). Ebaluatoia was used in 2015 by the IXA Group to classify different machine translation systems, with the participation of the Basque-speaking community. The initiative produced extremely interesting results.

Although automatic tools are the most common method used in the assessment of machine translation systems, with BLEU (Papineni *et al.*, 2002) being the most widespread, human assessment is absolutely essential also if

we wish our system to be used by real people. Consequently, most previous research projects have included a small group of human evaluators (most often made up of members of the research team itself) who have assessed a very small sample of translations.

Ebaluatoia, on the other hand, was a community-based assessment process. Basque speakers constitute an extremely aware, highly dynamic language community, and this method offers a fast means of obtaining a large number of reliable assessments without the need for large-scale investments. Under-resourced language communities are well-known for being extremely willing to participate in activities designed to help ensure the survival of their native tongue.

As stated earlier, in addition to assessing machine translation systems, the Ebaluatoia initiative also aimed to classify them. Five systems were assessed in the 2015 initiative, all focusing on the English-Basque language pair (Aranberri, 2016): EuSMT (baseline), EuSMT with segmentation, Matxin ENEUS, the hybrid SMatxinT system and Google Translate. In all cases, the units assessed were phrases.

In accordance with the findings of Alegria *et al.* (2013), a concerted effort was made to make the design of the Ebaluatoia initiative as simple as possible. To this end, the selected phrases were assessed using the pair-wise comparison method. Participants were shown the source phrase and two machine translations, and were simply asked which of the two they thought was better. This method requires a lesser degree of cognitive effort than other designs and results in a higher level of inter-rater agreement, as evident in the *kappa* scores obtained (between 0.49 and 0.53). The *kappa* scores reported during a similar assessment carried out within the framework of the WMT (Workshop on Statistical Machine Translation) were between 0.075 and 0.324 (Bojar *et al.*, 2014).

We used the methodology developed during the Ebaluatoia initiative to classify alternative Basque equivalents of the same term. Thus, in this case, participants were asked to assess complex terms (containing between 2 and 8 tokens) rather than phrases. Given that the phrases assessed during the Ebaluatoia initiative were mainly complex ones, we anticipated our task being somewhat less difficult, particularly since, in our case, all participants were experts in the field.

In light of the above, we decided to include another task in the Medbaluatoia assessment, namely asking participants to determine whether or not the alternatives given were correct. We believed this would enable us to measure

the correctness of the generated terms.

We also enabled certain fields in order to collect more complete data about participants' profiles. While Ebaluatoia collaborators were people from a wide range of age groups, educational levels and specialist knowledge areas, in our case all participants worked in the healthcare field. And in relation to educational level, all were either university graduates or university students studying at degree or postgraduate level.

Medbaluatoia was used to assess both KabiTerm and MatxinMed, which was why two evaluation groups were established. Three systems were compared to assess KabiTerm: Google Translate (as a baseline system), KabiTerm and MatxinMed-1 (a MatxinMed version without the terms generated by KabiTerm). This group was called the KabiTerm group.

MatxinMed was also assessed. In this case we used the same baseline system, i.e. Google Translate, along with two versions of MatxinMed: MatxinMed-1 (i.e. the same version used for assessing KabiTerm) and MatxinMed-2 (i.e. a MatxinMed version which includes the equivalents generated by KabiTerm). This group was called the MatxinMed group.

The table below (Table 5.14) shows the systems used in both assessments.

	KabiTerm	**Google**	**MatxinMed-1**	**MatxinMed-2**
KabiTerm group	✓	✓	✓	
MatxinMed group		✓	✓	✓

Table 5.14 – The systems assessed in the Medbaluatoia initiative.

As in previous evaluations, the evaluation of this chapter also focused on the disorders, clinical findings, body structures and procedures hierarchies. As regards the number of tokens per term, most terms selected for assessment contained between 2 and 8 tokens, since as described earlier (in Chapter 2) terms containing up to 8 tokens represent over 92% of the entire sample. Consequently, the sampling procedure used here was representative. Nevertheless, the more tokens it contains, the more complex the term, and the more difficult its analysis, generation and assessment.

The sample used for the Medbaluatoia assessment was the same size as that used for the Ebaluatoia one: 500 source terms. It should be remembered that, in our case, two assessments were conducted, and consequently 500 terms were selected from among SNOMED CT's complex terms for each. These terms were then stratified in accordance with the hierarchies and number of tokens in each assessment group (Ripley, 2009). As shown in Table

5.15, we extracted the terms from the English version bearing in mind the proportion of hierarchies and tokens in the entire SNOMED CT sample. To this end we first calculated the percentages of the four hierarchies within the whole sample, and then analysed the proportions of the different numbers of tokens contained in their terms.

	Disorders		**Findings**		**Body Structures**		**Procedures**	
	Prop.	**Num.**	**Prop.**	**Num.**	**Prop.**	**Num.**	**Prop.**	**Num.**
2 tokens	0.20	42	0.18	18	0.20	12	0.13	11
3 tokens	0.24	43	0.23	16	0.23	31	0.20	19
4 tokens	0.20	38	0.21	24	0.19	25	0.22	27
5 tokens	0.16	35	0.18	11	0.16	11	0.19	24
6 tokens	0.10	23	0.11	7	0.11	13	0.13	12
7 tokens	0.07	16	0.06	8	0.07	8	0.09	11
8 tokens	0.04	4	0.03	6	0.03	0	0.05	7
Total	0.37	201	0.17	90	0.19	98	0.28	111

Table 5.15 – The systems assessed in the Medbaluatoia initiative.

Having determined the number of terms to be assessed, English language source terms were then extracted from SNOMED CT. For the KabiTerm assessment group, 500 complex terms were randomly selected from the database, always on the condition that KabiTerm was capable of generating a Basque equivalent. To establish the MatxinMed assessment group, on the other hand, 500 terms were selected from among those KabiTerm was not able to translate. Since MatxinMed will only be used to translate those terms KabiTerm is unable to translate, this same criterion was used when establishing the assessment sample.

In relation to the KabiTerm group, it was important to control overproduction. While other systems generate one single equivalent, KabiTerm tends to provide multiple ones, meaning that the systems would not be competing under the same conditions. We used our Basque medical domain language model to redress this balance. Of all the equivalents suggested by KabiTerm, the system selects the one which to which the language model assigns the highest probability. Unfortunately, this does not guarantee that the best one will always be chosen, but it is the most reliable of all the automatic methods currently available.

A total of 1,000 terms were assessed, with three system-pairs (i.e. three systems) being evaluated in each, meaning that 3,000 individual assessments were carried out. In order to control for subjectivism, each term was as-

sessed by 5 different participants. Consequently, the Medbaluatoia initiative comprised 15,000 individual assessments. Moreover, in order to measure the attention paid by each participant to the task at hand, we also added a number of control terms. These control terms comprised a selection of correct manually-translated and obviously incorrect equivalents. Participants who incorrectly assessed one third or more of the control terms were automatically eliminated.

Control terms are necessary to ensure the reliability of the responses collected, particularly when participants are drawn from such a large pool. It is important to identify (as far as possible) participants with an insufficient knowledge of the language or dishonest performance. As in the Ebaluatoia initiative, the decision was made to exclude participants who assessed the control terms incorrectly. Therefore, one in every five terms was a control term, and any participants who assessed one third of these incorrectly was automatically excluded from the study.

It should be remembered that each participant was permitted to assess up to 1,000 terms, and since one in every five was a control term, 250 control term-equivalent pairs were required. The control terms were taken from the sample compiled by physicians. In other words, they were taken from the list of terms and equivalents we asked our experts to compile for us during the development of KabiTerm. To generate the alternative answers we used the equivalents proposed by expert physicians, along with those generated by the general Matxin system which we then manually changed to render incorrect (using antonyms, antitheses and words out of context. See the examples in Table 5.16).

Source term	*burn of vagina and uterus*
Best equivalent	bagina eta umetokiko erredura
Incorrect equivalent	erre ezazu utero baginako eta
Source term	*excision of fimbrial cyst*
Best equivalent	finbriako kistearen erauzketa
Incorrect equivalent	fimbrial cysteko hanka

Table 5.16 – Two examples of control term-equivalent pairs.

To disseminate news about the Medbaluatoia assessment, on the first day of the campaign we visited every Basque-medium class in the Faculty of Medicine. We also sent out an email to professional practitioners, asking for

their collaboration. Both the Osasungoa Euskalduntzeko Erakundea (OEE) and the Pro Vice-Chancellor for the Basque Language at the University of the Basque Country were key players in the dissemination process.

The interface of the assessment system is shown in the figure below (Figure 5.14). In addition to the English term and two Basque equivalents, the screen also contained a form to enable participants to select the best alternative. Furthermore, beside each equivalent there was a box that could be checked if the participant believes the equivalent to be correct. In addition to the assessment sheet itself, information was also provided regarding the total number of terms and equivalents that had been assessed so far (see the left-hand side of the image in Figure 5.14 and a ranking of the most active collaborators was also posted, designed to incite a degree of healthy competition and encourage greater participation (see the right-hand side of the same image).

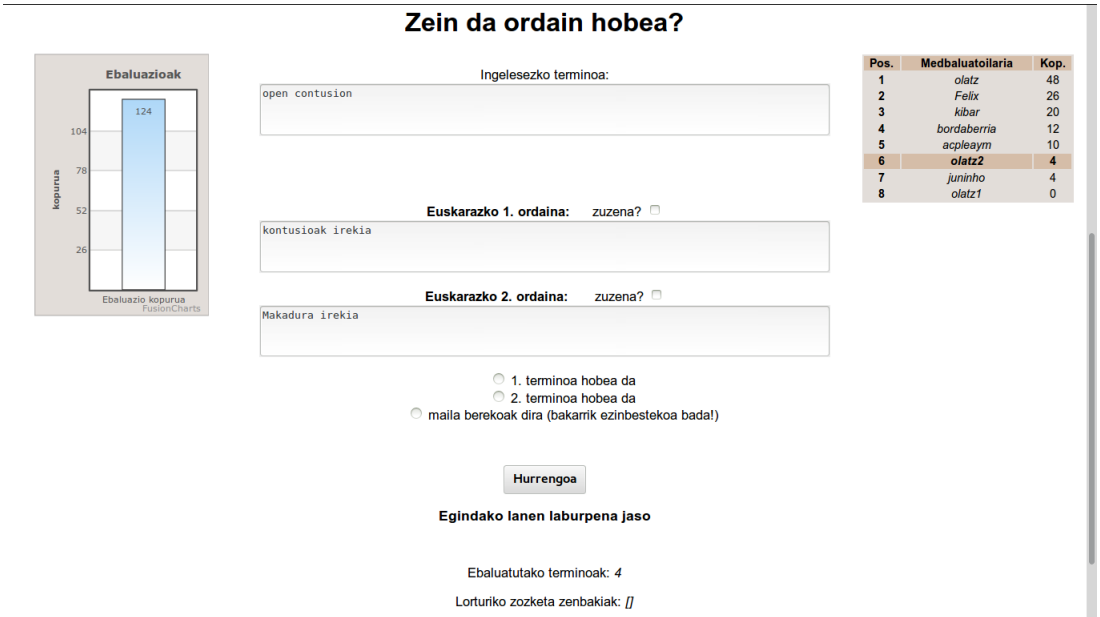


Figure 5.14 – Medbaluatoia assessment interface.

The next section presents the results of the Medbaluatoia assessment, along with inter-rater agreement scores and comparisons between the different systems analysed.

5.4 Results

n this section we present the results obtained from the Medbaluatoia assessment, along with those pertaining to SNOMED CT's coverage.

5.4.1 Medbaluatoia results

First of all, it should be highlighted that the Medbaluatoia campaign was a huge success. The campaign commenced on 10 October 2016 and although it was expected to last three weeks, the entire sample was assessed in just seven days. A total of 217 people participated in the campaign and only 13 were excluded for incorrectly assessing the control terms.

Profile of participants

The table below (Table 5.17) is a summary of the collaborators who participated in the campaign. As you can see, in addition to the 13 who were excluded, another 13 people failed to carry out any assessments, bringing the total number of invalid participants to 26 and the total number of valid participants to 191. The median number of assessments per participant was 24, and the mean was 100.25.

	Number of participants	**%**
Total participants	217	
Excluded	13	5.99
No assessments	13	5.99
Valid participants	191	88.02
	Number of assessments	
Median number per valid participant	24	
Mean number per valid participant	100,25	

Table 5.17 – Summary of participants in Medbaluatoia.

In relation to age range, the majority of participants were young people (Table 5.18), with over 40% being aged between 19 and 25. This is not surprising and is consistent with the results regarding educational level (see Table 5.19), which reveal that over 55% of participants were undergraduates (ranging between years 1 and 6 of their degree course). The second-largest

age group was the 26-45 one (around 40% of participants), which again is consistent with the percentage of graduates in the participant group.

Age range	**Number of participants**	**%**
<18	13	5.99
19-25	94	43.32
26-35	35	16.13
36-45	52	23.96
46-55	19	8.76
56-65	4	1.84
>65	0	0.00

Table 5.18 – Age range of participants in Medbaluatoia.

The fact that the Medbaluatoia campaign was conducted digitally may have affected the nature of the participant group, since this kind of questionnaire is more attractive to people in certain age groups.

Educational level	**Number of participants**	**%**
Year 1	33	15.21
Year 2	11	5.07
Year 3	6	2.77
Year 4	18	8.30
Year 5	19	8.76
Year 6	10	4.61
Residents	15	6.91
Graduates	91	41.94
Others	14	6.45

Table 5.19 – Educational level of participants in Medbaluatoia.

As stated earlier, the aim of Medbaluatoia was to assess machine-generated terms within the medical domain. To this end, it was important to determine whether participants were actually working or studying in the healthcare sector. The results are shown in Table 5.20. Most participants were either practising physicians or medical students (61.75%), although the number of nurses was also very high (21.2%).

Field	Number of participants	%
Nursing	46	21.20
Medicine	134	61.75
Pharmacy	6	2.77
Physiotherapy	17	7.83
Dentistry	0	0.00
Others	14	6.45

Table 5.20 – Specialist field of the participants in Medbaluatoia.

As regards their language skills, in order to effectively assess the terms given, participants needed to speak both English and Basque. However, given the task only involved checking terms that had already been translated, no minimum level was required of them in either language. As shown in the tables below (Tables 5.21 and 5.22), while the vast majority of participants had the highest Basque language level (93.09% had level C1-C2), most had only intermediate English (48.85% had level B1-B2), although the highest and lowest level groups were both fairly large (30.87% and 20.28% of participants, respectively).

Basque level	Number of participants	%
A1-A2	2	0.92
B1-B2	13	5.99
C1-C2	202	93.09

Table 5.21 – Basque language level of participants in Medbaluatoia.

English level	Number of participants	%
A1-A2	44	20.28
B1-B2	106	48.85
C1-C2	67	30.87

Table 5.22 – English language level of participants in Medbaluatoia.

Inter-rater agreement

Next we calculated the level of agreement between participants. The most common method for calculating this is the *kappa* (κ) coefficient. As outlined in the previous chapter, Cohen's κ is used to measure agreement between two raters (Cohen, 1960), whereas Fleiss'κ is used to measure agreement between more than two raters (Gwet, 2014, Artstein and Poesio, 2008). It is important to remember that in the case of both Cohen's and Fleiss' *kappas*, it is assumed that all raters have assessed the same sample.

n our case, we had over 200 different raters or participants (191 valid ones), each assessing different samples (as regards both the nature and number of the terms contained). Consequently, and bearing in mind the limitations of the *kappa* measurements, we calculated Cohen's *kappa* in order to be able to compare our results with those reported by other assessments, since this coefficient was calculated in both the previous WMT machine translation campaign (Bojar *et al.*, 2014) and in the Ebaluatoia campaign (Aranberri *et al.*, 2016a).

System pair	***Kappa***
Google - KabiTerm	0.36
Google - MatxinMed-1	0.37
KabiTerm - MatxinMed-1	0.37
hline Google - MatxinMed-2	0.30
Google - MatxinMed-1	0.30
MatxinMed2 - MatxinMed-1	0.48

Table 5.23 – Agreement measured for each system pair (*kappa*).

As shown in the table above (Table 5.23), the *kappa* values obtained oscillate between 0.30 and 0.48. In accordance with the benchmark scale outlined in the previous chapter (Landis and Koch, 1977), in which a value of between 0-0.2 indicates slight agreement, a value of 0.2-0.4 indicates fair agreement, a value of 0.4-0.6 indicates moderate agreement, 0.6-0.8 substantial agreement and 0.8-1.0 almost perfect agreement, we can conclude that our values are indicative of a fair level of agreement.

Although all *kappa* values obtained are within the range reported in the WMT machine translation campaign (Bojar *et al.*, 2014), close to the top, the results were not as good as expected, given that the *kappas* reported by Ebaluatoia were between 0.49 and 0.53. It may be that the additional

task added in this assessment (i.e. asking participants to specify whether they believed the term was correct) had a direct effect on this aspect. Nevertheless, we believe that the agreement values obtained are within acceptable limits.

Number of assessments

In the Medbaluatoia campaign participants were given the source term and two Basque equivalent terms, and were asked to select the one they believed was better. We provided no criteria for selecting the "best" equivalent term, since we wanted the results to reflect each person's own opinion, in accordance with their own criteria.

Our aim was to obtain 5 assessments for each term and each system pair. Nevertheless, due to the configuration of the web applications, on occasions we obtained 7 assessments for some combinations. Given that these additional assessments were perfectly valid, we decided to include them in the results. The table below (Table 5.24 shows the number of assessments obtained for each system.[15] In order to ensure a proper evaluation, 2,500 assessments were required for each system pair (500 terms assessed by 5 separate participants). As shown in the table, more than the required number of assessments were obtained for each pair, with a low level of variability (between 2,523 and 2,540). The results obtained are therefore comparable.

	Google - KabiTerm	**Google - MatxinMed-1**	**KabiTerm - MatxinMed-1**
KabiTerm group	2,529	2,523	2,527
	Google - MatxinMed-2	**Google - MatxinMed-1**	**MatxinMed-2 - MatxinMed-1**
MatxinMed group	2,540	2,535	2,535

Table 5.24 – Total assessments obtained for each system pair.

Results

The method used to determine which system is best, on the basis of the assessment, was as follows: if the difference in votes between two systems is more than two, the one which received more votes is considered the clear

[15]The control terms were not included in the count.

winner (in the table below, clear winners are indicated with the code "X-system++"). If the difference between the number of votes received by each system is 1 or 2, then the system receiving more votes is deemed the winner (in the table, these systems are indicated with the code "X.system+"). If two systems receive the same number of votes, then they are considered equally good. For example, if in the assessment of a term KabiTerm receives 4 votes and Google 1, since the difference between them is 3, then KabiTerm is considered the clear winner and assigned the code "2.system++".

	Google - KabiTerm	**Google - MatxinMed-1**	**KabiTerm - MatxinMed-1**
1.system++	6.8 (34)	13.2 (66)	**46.4(232)**
1.system+	3.2 (16)	9.2 (46)	14.4 (72)
the same	3.2 (16)	7.8 (39)	5.6 (28)
2.system+	13.4 (67)	15.2 (76)	12.4 (62)
2.system++	**73.2(366)**	**54.4(272)**	21.0(105)

Table 5.25 – The results of the Medbaluatoia campaign's assessment system (for the KabiTerm group).

	Google - MatxinMed-2	**Google - MatxinMed-1**	**MatxinMed-2 - MatxinMed-1**
1.system++	19.4 (97)	21.6(108)	6.6 (33)
1.system+	12.0 (60)	16.4 (82)	13.0 (65)
the same	7.0 (35)	8.8 (44)	**62.4(312)**
2.system+	22.0(110)	19.4 (97)	14.6 (73)
2.system++	**39.8(199)**	**34.0(170)**	3.6 (18)

Table 5.26 – The results of the Medbaluatoia campaign's assessment system (for the MatxinMed group).

As evident in the table above (Table 5.25), in the KabiTerm group the KabiTerm system emerged as the clear winner. The difference was particularly notable in comparison with the Google translation system, since in 86.6% of cases assessors voted in favour of KabiTerm (73.2% plus 13.4%). While the difference between KabiTerm and MatxinMed-1 was less pronounced, KabiTerm nevertheless obtained significantly better results (in 46.4% of cases it was the clear winner, whereas MatxinMed-1 was only the clear winner 21% of the time).

As regards the MatxinMed group (Table 5.26), even though the results are not quite so categorical, both versions of MatxinMed performed better than Google Translate. MatxinMed-2 received more votes than Google in 61.8% of cases (22.0% + 39.8%), whereas Google only received more votes 31.4% of the time (19.4% + 12%). When the two versions of MatxinMed were compared, however, the results were even, with both systems being found to have the same quality in 62.4% of cases, MatxinMed-2 being found better in 19.6% of cases (taking both clear wins and wins into account) and MatxinMed-1 being declared winner 18.2% of the time. It should not be forgotten that the only difference between these two systems is the terms generated by KabiTerm. In other words, MatxinMed-2 only generates different equivalents from MatxinMed-1 when it finds nested terms that have been translated by KabiTerm, and in the sample used in the assessment this only occurred very occasionally. Nevertheless, on the basis of the assessment carried out we cannot conclude that one system is better than the other, since both obtained very similar results. It is worth highlighting that the structure of the terms that appear in the KabiTerm group was limited, and in that structure MatxinMed achieved better results than Google Translate.

The new functionality which was added to enable participants to indicate whether or not the term was correct proved unsuccessful. We believe this was because the difference between the terms "correct" and "better" was not made sufficiently clear, and after comparing the equivalents some participants marked their preferred one as "correct", even though it was not. Therefore, given that the results are not meaningful, we have decided not to present them here.

5.4.2 KabiTerm coverage in SNOMED CT

This section presents the data regarding the coverage of the Basque translation of SNOMED CT. In other words, the section specifies the proportion of terms from the disorders, clinical findings, body structures and procedures hierarchies that have been translated into Basque thanks to KabiTerm. Since MatxinMed has the capacity to translate the entire SNOMED CT database, it makes no sense to present the coverage data for that system. However, this information is interesting in relation to KabiTerm. Table 5.27 presents the data for the KabiTerm coverage, specifying how many SNOMED CT terms it has been able to translate and how many equivalents it has generated.

As shown in the table, KabiTerm's coverage in the disorders hierarchy is

	Disorders	Clinical Findings	Body structures	Procedures
Total num. terms	114,830	52,857	87,104	59,384
Into Basque	26,136	4,054	12,497	10,651
% of terms	22.76%	7.67%	21.04%	12.23%
Equivalents	102,724	15,868	43,913	34,232

Table 5.27 – KabiTerm coverage of SNOMED CT terms.

particularly noteworthy, since the system has been able to translate 26,136 terms into the Basque language (out of a total of 114,830). This means that 22.76% of the terms in this hierarchy have been translated, which is, in our opinion, a very positive result. Within the body structure hierarchy the percentage of terms translated is similar (21.04%). The percentages achieved in the clinical findings and procedures hierarchies, on the other hand, were not as high. It is worth noting that most of the structures we focused on were structures used to describe disorders and body structures, since these are the structures that appear most frequently.

Another aspect worth highlighting is KabiTerm's overproduction of equivalent Basque terms. A concerted effort was made to limit this overproduction, with the final mean being 3 to 4 equivalents per source term (for example, 26,136 terms were translated, and 102,724 Basque equivalents generated, giving a mean of 3.9 equivalents per term).

5.5 Summary and conclusions

This chapter presents the work carried out to translate complex terms into the Basque language. Two systems were developed and assessed: KabiTerm and MatxinMed. KabiTerm uses the structure of nested terms to translate complex English terms into Basque. The system is based on the fact that complex terms often contain other terms that appear in nested form. If the nested term (i.e. the term inside the term) has already been translated, then translation patterns can be defined in order to generate the equivalent of the whole complex term. In such cases, the SNOMED CT hierarchies are used to analyse the terms' structures and define the Basque translation patterns.

MatxinMed is an adaptation to the medical domain of a rule-based machine translation system (Matxin). During the process, we added a number

of functionalities to the Matxin program in order to enable it to be adapted to a specific domain. One such functionality was the addition of a new characteristic to the dictionary, as a result of which, during the lexical transfer process, the system selects equivalents in accordance with the specific domain. For the adaptation to the medical domain we enlarged the dictionary by adding SNOMED CT term-equivalent pairs that had already been translated previously. To reduce ambiguity we developed a language model to specify the order of the different equivalents. Moreover, we also incorporated an additional module into which we integrated the NeoTerm system outlined in the previous chapter (Chapter 4). The aim of this module is to enable unknown terms to be translated into the Basque language. Finally, we added a series of rules for identifying complex terms, in order to be able to use their dictionary equivalents during the translation process.

Both techniques were assessed by means of a campaign called Medbaluatoia, which also compared them with the state-of-the-art Google Translate tool. The results were very positive, both from the perspective of the number of people who participated and in relation to the quality of our systems in comparison with Google. The results obtained by KabiTerm are particularly worth highlighting, since the system emerged as the clear winner in comparison with all the other systems analysed.

The Basque-speaking healthcare community's response to the assessment was very positive, and their enthusiasm is a source of motivation for us to continue our research. The experience proved that Medbaluatoia is an extremely useful tool for involving the whole Basque-speaking healthcare community, and we hope to launch similar campaigns in the future in order, for example, to validate and correct the Basque version of SNOMED CT.

6 Conclusions and future work

In this PhD project we designed and developed an algorithm for the automatic generation of term equivalences for a low-resourced language. Thus, taking advantage of lexical resources and without a parallel corpus, we developed two systems that systematically generate terms (NeoTerm and KabiTerm). Besides, we extended the Machine Translator Matxin by means of adding a functionality so it can be adapted to a specific domain. In order to obtain a wide terminology in Basque, we took SNOMED CT as a reference. SNOMED CT is considered the most comprehensive clinical terminology nowadays, it has high recall, and it is used all over the world for coding, extracting or analysing clinical information.

In the following lines, we summarise the general conclusions of the works done (section 6.1) and the main contributions obtained from the development of this PhD project (section 6.2). Finally (section 6.3), we enumerate the works that can be carried out in the future.

6.1 Conclusions

Our main objective has been to generate resources in Basque for the processing of health science texts. In that way, we consider getting a reference terminology in Basque an indispensable step, and that has been the main task of this PhD project. We analysed the state of art regarding automatic term generation, and we mainly found works based on corpus. For low-resourced languages, as it is the case of Basque, those techniques are not applicable, as

there is not any parallel or comparable corpus for the health science domain. Thus, we proposed and developed techniques that avoid the use of corpora.

In the following lines we sill present the main conclusions grouped by subject.

- **Source:** First of all, we carried out a quantitative analysis of SNOMED CT and compared the Spanish and English versions.Being SNOMED CT multilingual, the analysis helped us to choose the source language for the translation into Basque. In this case, we chose the English version, as it is the original version. What is more, in the moment we begun this PhD project the Spanish version was not stable yet and, in consequence, it had some deficiencies. We also chose the hierarchies to run as reference of the translation design: clinical findings (that includes disorders), procedures and body structures. Even if we only give results for those hierarchies, at the end we translated the whole terminological content of SNOMED CT. Although SNOMED CT includes any term that could be relevant in a clinical record, we chose the most populated hierarchies and in the same way, the most relevant to the clinical domain.

- **The system to manage the translation process of SNOMED CT into Basque**: To manage the translation process of SNOMED CT into Basque we designed and developed a system called EuSnomed. We integrated there a four step algorithm we designed to perform the translation process. The first step takes advantage of the bilingual/multilingual lexical resources to obtain equivalences. The second one translates neoclassical terms, by means of affix equivalences and transliteration rules. The third step is based on the structure that nested terms conforms to define translation patterns. Finally, the fourth step takes a general purpose Machine Translator and adapts it to the health science domain. In addition to the implementation of the translation algorithm, EuSnomed also is responsible for the storage of the information and for the reuse of the new Basque terms.

- **Translation of simple terms into Basque**: We designed the first two steps of the algorithm to translate simple terms into Basque. On the one hand, we mapped the terms from SNOMED CT with bilingual and multilingual specialised dictionaries from the biomedical domain. Experts manually evaluated a set of translations of SNOMED CT, and

the resources that get the best results are: i) Science and Technology dictionary (*ZT hiztegia*) with precision of 0.99, ii) the Basque terminology bank, Euskalterm, with precision of 0.89 and, iii) the dictionary of nursing with 0.94 of precision. The Atlas of Human Anatomy did not perform as well as the others in terms of precision, but the contribution made in terms of recall is remarkable. On the other hand, we created a system called NeoTerm to translate English neoclassical terms into Basque. We developed three approaches of this system. The first one is the baseline system, and it is based on the composition of neoclassical affixes. Even if the precision of this approach is high (0.89), the recall is not (0.34) and thus, the priority for the second approach has been the improvement of the recall. In order to improve the recall, in the second approach we integrated a transliteration module and we extended the dictionaries. Even if we got worse results in terms of precision (8 point below), the recall improved a lot (48 points), and we manage to balance precision and recall obtaining an F-measure of 0.81. In the last approach, we aimed to improve the identification of neoclassical terms, in an attempt to make NeoTerm distinguish between neoclassical terms and other terms and, hence, avoid errors. We worked on the identification algorithm considering the pieces of advice given by the experts. In any case, we did not improve the results archieved by the second approach and even worse, we lost 7 points in recall. Thus, we integrated the second approach of NeoTerm into EuSnomed. Overall, the number of simple terms translated into Basque is very high (above than 75% of all the reference hierarchies).

- **Translation of complex terms into Basque**: To get Basque equivalences of complex terms (the last two steps of the algorithm), we developed two systems: KabiTerm and MatxinMed. KabiTerm uses the structure of nested terms to translate complex English terms into Basque. The system is based on the fact that complex terms often contain other terms that appear in a nested form. If the nested term (i.e. the term inside the term) has already been translated, then translation patterns can be defined in order to generate the equivalent of the whole complex term. In such cases, the SNOMED CT hierarchies are used to analyse the terms' structures and define the Basque translation patterns. MatxinMed is an adaptation to the medical domain of a rule-based machine translation system (Matxin). During the process,

we added a number of functionalities to Matxin so it can be adapted to a specific domain. Both systems were assessed by means of a campaign called Medbaluatoia, which also compared them with the state-of-the-art Google Translate tool. This campaign involved the Basque-speaking healthcare community in an open evaluation made online. The results were very positive, both from the perspective of the number of participants and in relation to the quality of our systems in comparison with Google. The results obtained by KabiTerm are particularly worth highlighting, since the system emerged as the clear winner in comparison with the other two analysed systems. The Basque-speaking healthcare community's response to the assessment was very positive, and their enthusiasm is a source of motivation for us to continue our research. The experience proved that Medbaluatoia is an extremely useful tool for involving the whole Basque-speaking healthcare community, and we hope to launch similar campaigns in the future in order, for example, to validate and correct the Basque version of SNOMED CT.

- **Recall of the translation of SNOMED CT**: Table 6.2 shows the general results regarding the recall of EuSnomed (precision has been addressed in their corresponding chapters). As it can be seen, MatxinMed translates all the terms that could not be translated by the previous systems, and thus we get the *alpha* version of SNOMED CT in Basque. Leaving MatxinMed aside, we show the results regarding the number of tokens in Table 6.1. Most of the terms build up by only one token have been translated by means of lexical resources and NeoTerm, and most of the two tokens terms have been translated by means of KabiTerm except for the findings. It is remarkable that most of the translation patterns we have worked for KabiTerm pertain to the disorder and body structure hierarchies, and that is the reason why we get a recall higher than 35% on those hierarchies. In fact, the disorders hierarchy is the most populated one and in the case of body structures, its descriptions have very repetitive structures.

		1token	2token	3token	4token	≥5token	Total
Disorder	**Trans.**	3,265	8,335	9,966	5,614	6,574	33,754
	Total	3,865	21,003	25,038	20,757	44,167	114,830
	Recall	0.845	0.397	0.398	0.271	0.149	0.294
Finding	**Trans.**	1,449	2,442	1,413	568	439	6,311
	Total	1,940	9,737	11,906	11,317	24,640	59,540
	Recall	0.747	0.251	0.119	0.050	0.018	0.106
Body structure	**Trans.**	1,907	4,308	4,631	3,444	3,550	17,840
	Total	2,592	10,863	12,599	10,635	22,695	59,384
	Recall	0.736	0.397	0.368	0.324	0.156	0.300
Procedure	**Trans.**	1,698	3,295	2,744	2,456	2,471	12,664
	Total	1,985	9,892	15,399	17,082	42,746	87,104
	Recall	0.855	0.333	0.178	0.144	0.058	0.145

Table 6.1 – Recall with respect to the number of tokens of the source term.

	Disorders			Findings			Body structures			Procedures		
	Trans.	**Recall**	**Syn.**	**Trans.**	**Recall**	**Syn.**	**Trans.**	**Recall**	**Syn.**	**Trans.**	**Recall**	**Syn.**
ICD10	11,060	-	-	2,555	-	-	-	-	-	-	-	-
Lex. res.	5,750	0.050	6,975	1,635	0.030	2,343	4,667	0.079	6,599	747	0.086	981
NeoTerm	1,868	0.016	1,951	622	0.001	717	676	0.011	726	1,266	0.015	1,308
KabiTerm	26,136	0.228	102,724	4,054	0.077	15,868	12,497	0.210	43,913	10,651	0.122	34,232
MatxinMed	81,076	0.706	81,076	46,539	0.881	46,539	41,535	0.699	41,535	74,436	0.855	74,436
Total	114,830	1.000	180,952	52,857	1.000	95,089	59,384	1.000	92,029	87,104	1.000	109,200

Table 6.2 – Recall of the translation of SNOMED CT.

6.2 Contributions

The main contribution of this PhD project is the development of new systems for the automatic generation of terminology. On the literature, we can find many works about techniques based on bilingual and monolingual corpora, but we can hardly find any that does not make use of corpora. We developed rule-base systems, that can be interesting for low-resourced languages as Basque. In addition, we demonstrated that the methods we developed are useful to help translators and experts in medicine.

In addition to the main contribution, in the following lines we enumerate the rest of the contributions:

- **We developed a system that manages the translation of SNOMED CT into Basque.** (3^{th} chapter)

  We developed the EuSnomed system that implements the translation algorithm. It manages the whole translation process, from the integration of lexical resources to the calculation of the results. It is designed to reuse the new terms generated by the system on the go, so the algorithm can perform in an incremental way. All the code is publicly available on GitHub[1].

- **We developed a system to translate neoclassical terms into Basque called NeoTerm.** (4^{th} chapter)

  NeoTerm is a rule-based system that translates English neoclassical terms into Basque. To that end, we manually created a dictionary of neoclassical affixes that usually appear on health science terms, and we wrote some rules to transliterate them from English into Basque. All the code is publicly available on GitHub[2]. The affix bilingual dictionary (en-eu) and the demo of NeoTerm is available on a webpage[3].

- **Based on nested terms, we developed a system that translates complex terms into Basque called KabiTerm.** (5^{th} chapter)

  To carry out the translation of complex terms, we developed a system called KabiTerm. KabiTerm uses the nested term structure to

[1] https://github.com/olatz87/euSnomed
[2] https://github.com/olatz87/NeoTerm
[3] http://ixa2.si.ehu.es/neoterm/

translate those complex terms into the Basque language. That is, we defined translation patterns based on the structure that nested terms conform within the complex terms. All the code is publicly available on GitHub[4].

- **We extended Matxin by adding a functionality to adapt to domains and we created MatxinMed.** (5^{th} chapter)

  Matxin is a rule-based Machine Translator that translates text from Spanish or English into Basque. We added a number of functionalities to Matxin program in order to make it adaptable to a specific domain. We also created MatxinMed, the adaptation of Matxin to the medical domain by the integration of SNOMED CT's terminology, among others. The code will be available soon on Matxin's GitHub page[5].

- **We involved the Basque-speaking healthcare community on the Medbaluatoia campaign.** (5^{th} chapter)

  For the evaluation of complex terms, we carried out a crowd evaluation involving the Basque healthcare community called Medbaluatoia. This campaign is based on Ebaluatoia. The Basque-speaking healthcare community's response to the assessment was very positive, and showed the need of more initiatives like this.

- **We developed a health science analyzer (AnaMed) and a terminological server based on SNOMED CT (TermZerSCT).** (5^{th} chapter)

  AnaMed is a linguistic analyser for health science that besides analysing linguistic information, it also identifies SNOMED CT's terms and eponyms. TermZerSCT is a terminological server that manages the information related to SNOMED CT, minimising the processing time. All the code is publicly available on GitHub[6,7].

- **We created the *alpha* version of SNOMED CT in Basque.**

  Thanks to the translation algorithm, we created an *alpha* version of SNOMED CT in Basque automatically. Having SNOMED CT in

[4]https://github.com/olatz87/KabiTerm
[5]https://github.com/matxin/matxin
[6]https://github.com/olatz87/anaMed-en
[7]https://github.com/olatz87/TermZerSCT

Basque boosted the interest, specially over the Basque service inside Osakidetza (Basque Health System), and we signed a compromise between Osakidetza and the IXA group to carry out the revised version of SNOMED CT.

- **We created a corpus of the health science domain for Basque.** (5^{th} chapter)

  At the moment we began with the PhD project, there was not any corpus of the health science domain for Basque. For the development of MatxinMed, we were on the need of an in-domain language model for Basque to be able to choose the "best" synonym for each source term. For that, we collected texts from different sources and nature, such as textbooks or students notes. Unfortunately, we can not make the corpus publicly available because of licence and property reasons. Instead, we can publish the language model and it will be available soon.

6.3 Future work

Here we will summarise some of the research lines or works that have not yet been fully addressed by this dissertation, as well as some new research lines that naturally arise from our work:

- **To create a stable Basque version of SNOMED CT, checked by experts**.

  As it is known, the terms generated automatically will not be correct at 100 %. The version of SNOMED CT we generated in this PhD project has to be checked by experts term by term, to accept or correct the automatically generated ones, or even to propose new ones. In addition, in order to obtain an official version of SNOMED CT, experts will have to choose the preferred term among all the synonyms.

- **To measure variability between automatically generated terms and the ones proposed by experts**.

  Once we get a stable version of SNOMED CT in Basque checked by experts, we will be able to measure the variability between the terms and descriptions we generated automatically and the ones chosen or

proposed by experts. This study opens a very interesting research line that could be the final evaluation of the work done in this PhD project.

- **To automatically translate acronyms**.

  We have not worked on the automatic translation of acronyms, and those are specially relevant in this domain. Hence, we want to open the way to automatically translate and expand them.

- **To adapt AnaMed to Spanish**.

  We developed AnaMed for English and Basque during the development of this PhD project, and we consider very interesting to adapt it also to Spanish. In the south part of the Basque Country Spanish is the main language in clinical records, and AnaMed may be of help to process them. As a basis we will have FreelingMed, a linguistic analyser developed in the IXA that is able to identify SNOMED CT terms.

- **To create demos for KabiTerm, MatxinMed and AnaMed.**

  We want to make our tools available for the community by means of demos. The code is already publicly available, but that is not useful for health workers, and by means of demos they shall access our systems and take advantage of them.

- **To adapt EuSMT to the health science domain**.

  By means of bilingual dictionaries and monolingual corpora, we want to adapt the Statistical Machine Translator EuSMT to the health science domain and compare it with MatxinMed. Until we get a bilingual corpus is going to be hard working with statistics models for Machine Translation, but we found very interesting publications based on bilingual dictionaries and monolingual corpora. We must bear in mind that EuSMT was the system that obtained the best results in the Ebaluatoia campaign, and thus, we want to bring those good results to our domain.

- **To adapt our system to translate other terminologies**.

  We want to use the algorithm designed for EuSnomed to translate other terminologies. The Basque Service inside Osakidetza (Basque Health System) has a special interest on translating the 10^{th} ICD into Basque, and we already started working on a project to adapt our system. We

will have to define new translation patters for KabiTerm since the structure of the descriptions of ICD10 is special (it is a classification). For instance, we can find very often descriptions such as *"Salmonella infection, unspecified"*.

- **To develop a system so that the Basque-speaking healthcare community can validate the Basque version of SNOMED CT**.

  We want to adapt Medbaluatoia to validate the terminology of SNOMED CT. In fact, during the campaign we received feedback from some participants stating that they would like to take part in the correction of Basque terms, and that they were determined to evaluate more. Taking into account the willingness of the participants, we are thinking of correcting the whole terminological content of SNOMED CT by means of a similar campaign. Thus, we would be able to collect synonyms involving more people than in a regular procedure to create a gold-standard. The task of choosing a preferred term among synonyms would be taken by a small group of experts, including terminologists and healthcare professionals.

- **To translate clinical records by means of a Machine Translation system**.

  We want to use SNOMED CT as a basis for the automatic translation of clinical records. As mentioned before, healthcare workers usually do not write their clinical records in Basque, as not all the professionals can understand it. That is our motivation to propose the automatic translation of clinical records. We will need to define a controlled language to ensure translation quality, and we will have to train doctors and nurses to learn to use it. We already started working on a prototype, and we took advantage of TermZerSCT and AnaMed to identify terms from the text.

Bibliography

Abdoune H., Merabti T., Darmoni S.J., and Joubert M. Assisting the Translation of the CORE Subset of SNOMED CT Into French. In Moen A., Andersen S.K., Aarts J., and Hurlen P., editors, *Studies in Health Technology and Informatics*, 169 lib., 819–823, 2011.

Agirre E., Alegria I., Arregi X., Artola X., de Ilarraza A.D., Maritxalar M., Sarasola K., and Urkia M. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. *Proceedings of the third conference on Applied natural language processing*, 119–125. Association for Computational Linguistics, 1992.

Al-Haj H. and Lavie A. The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. *Machine translation*, 26(1-2):3–24, 2012.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., and Lersundi M. EDBL: a General Lexical Basis for the Automatic Processing of Basque. *IRCS Workshop on linguistic databases. Philadelphia (USA).*, 2001.

Alegria I., Cabezon U., de Betono U.F., Labaka G., Mayor A., Sarasola K., and Zubiaga A. Reciprocal enrichment between basque Wikipedia and machine translation. *The PeopleâĂŹs Web Meets NLP*, 101–118. Springer, 2013.

Aranberri N. Ebaluatoia: crowd evaluation of English-Basque machine translation. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2016.

Aranberri N., Labaka G., de Ilarraza A.D., and Sarasola K. Exploiting portability to build an rbmt prototype for a new source language. *Proceedings of EAMT*, 2015.

Aranberri N., Labaka G., de Ilarraza A.D., and Sarasola K. Ebaluatoia: crowd evaluation for english–basque machine translation. *Language Resources and Evaluation*, 1–32, 2016a.

Aranberri N., Labaka Intxauspe G., Jauregi O., Díaz de Ilarraza Sánchez A., Alegría Loinaz I., and Agirre Bengoa E. Tectogrammar-based machine translation for English-Spanish and English-Basque. *Procesamiento del Lenguaje Natural*, 73–80, 2016b.

Artetxe M. Distributional semantics and machine learning for statistical machine translation. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2016.

Artstein R. and Poesio M. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Bakhshi-Raiez F., Cornet R., and F. de Keizer N. Development and Application of a Framework for Maintenance of Medical Terminological Systems. *Journal of the American Medical Informatics Association: JAMIA*, 15(5): 687–700, 2008.

Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

Bojar O., Buck C., Federmann C., Haddow B., Koehn P., Leveling J., Monz C., Pecina P., Post M., Saint-Amand H., *et al.*. Findings of the 2014 workshop on statistical machine translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 12–58. Association for Computational Linguistics Baltimore, MD, USA, 2014.

Burgun A. Desiderata for domain reference ontologies in biomedicine. *Journal of Biomedical Informatics*, 39(3):307 – 313, 2006. ISSN 1532-0464. URL `http://www.sciencedirect.com/science/article/pii/S1532046405000997`. Biomedical Ontologies.

Campbell W., Campbell J., West W., McClay J., and Hinrichs S. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. *Journal of the American Medical Informatics Association*, 21 (5):885–892, 2014.

Chute C.G. Clinical Classification and Terminology: Some History and Current Observations. *Journal of the American Medical Informatics Association*, 7(3):298âĂŞ–303, 2000.

Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1), 1960.

Desjardins L. Le santé des francophones du Nouveau-Brunswick. Petit-Rocher, Société des Acadiens et des Acadiennes du Nouveau-Brunswick, 2003.

Elhanan G., Perl Y., and Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. *Journal of the American Medical Informatics Association*, 18(1), 2011.

España Bonet C., Màrquez Villodre L., Labaka G., Díaz de Ilarraza Sánchez A., and Sarasola Gabiola K. Hybrid machine translation guided by a rule-based system. *Machine translation summit XIII: proceedings of the 13th machine translation summit, September 19-23, 2011, Xiamen, China*, 554–561, 2011.

European Observatory on Health Care Systems. Luxembourg: Health system review. *Health Systems in Transition*, 1999.

Ezeiza N., Alegria I., Arriola J.M., Urizar R., and Aduriz I. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 380–384. Association for Computational Linguistics, 1998.

Finkel J.R., Grenager T., and Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370, 2005. URL `http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf`.

Gerkens S. and Merkur S. Belgium: Health system review. *Health Systems in Transition*, 12(5):1–266, 2010.

Gieselmann P. Architecture of the lucy translation system. *Second machine translation marathon, Wandlitz, Berlin*, 28, 2008.

Gwet K.L. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

Hajicová E. Dependency-based underlying-structure tagging of a very large Czech corpus. *TAL. Traitement automatique des langues*, 41(1):57–78, 2000.

Hulden M. Foma: a Finite-State Compiler and Library. *Proceedings of EACL 2009*, 29–32, Stroudsburg, PA, USA, 2009. URL `http://dl.acm.org/citation.cfm?id=1609049.1609057`.

Humphreys B.L., McCray A.T., and Cheh M.L. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *Journal of the American Medical Informatics Association*, 4(6):484–500, 1997.

Hutchins W.J. and Somers H.L. *An introduction to machine translation*, 362 lib. Academic Press London, 1992.

IHTDSO SNOMED CT. Data Analytics with SNOMED CT âĂŞ Case Studies. Barne-txostena, IHTDSO, May 2015.

IHTSDO I.H.T.S.D.O. SNOMED CT Starter Guide. February 2014. Barne-txostena, International Health Terminology Standards Development Organisation, 2014.

Jiang G. and Chute C. Auditing the Semantic Completeness of SNOMED CT Using Formal Concept Analysis. *Journal of the American Medical Informatics Association*, 16(1), 2009.

Jurafsky D. and Martin J.H. *Speech and language processing*. Prentice Hall, 2008.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., *et al.*. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics, 2007.

Koehn P., Och F.J., and Marcu D. Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48–54. Association for Computational Linguistics, 2003.

Labaka G., España-Bonet C., Màrquez L., and Sarasola K. A hybrid machine translation architecture guided by syntax. *Machine translation*, 28(2):91–125, 2014.

Landis J.R. and Koch G.G. The measurement of observer agreement for categorical data. *biometrics*, 159–174, 1977.

Lee D., Cornet R., and Lau F. Implications of SNOMED CT versioning. *International Journal of Medical Informatics*, 80:442–453, 2011.

Maheronnaghsh R., Nezareh S., Sayyah M.K., and Rahimi-Movaghar V. Developing SNOMED-CT for Decision Making and Data Gathering: A Software Prototype for Low Back Pain. *Acta Medica Iranica*, 51(8):548–53, September 9 2011.

Manning C.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J., and McClosky D. The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Mayor A. *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Doktoretza-tesia, Euskal Herriko Unibertsitateko Donostiako Informatika Fakultatea, 2007.

Mayor A., Alegria I., Diaz de Ilarraza A., Labaka G., Lersundi M., and Sarasola K. Matxin, an Open-source Rule-based Machine Translation System for Basque. *Machine Translation*, 25:53–82, 2011. ISSN 0922-6567. URL `http://dx.doi.org/10.1007/s10590-011-9092-y`. 10.1007/s10590-011-9092-y.

Mikroyannidi E., Stevens R., Iannone L., and Rector A. Analysing Syntactic Regularities and Irregularities in SNOMED-CT. *Journal of Biomedical Semantics*, 3(8), 2012.

Nadeau D. and Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Naradowsky J. and Toutanova K. Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 895–904. Association for Computational Linguistics, 2011.

Papineni K., Roukos S., Ward T., and Zhu W.J. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics, 2002.

Perez-de-Viñaspre O. SNOMED CT sare semantikoa euskaratzeko aplikazioa. Doktoretza-ikastaroetako defentsa-lana, Euskal Herriko Unibertsitatea, 2013.

Petersen P.G. How to Manage the Translation of a Terminology. Presentation at the IHTSDO October 2011 Conference and Showcase, October 2011.

Popel M. and Žabokrtskỳ Z. Tectomt: modular nlp framework. *International Conference on Natural Language Processing*, 293–304. Springer, 2010.

Ripley B.D. *Stochastic simulation*, 316 lib. John Wiley & Sons, 2009.

Schulz S., Bernhardt-Melischnig J., Kreuzthaler M., Daumke P., and Boeker M. Machine vs. Human Translation of SNOMED CT Terms. In et al. C.L., editor, *MEDINFO 2013*, 581–584, 2013.

Sennrich R., Haddow B., and Birch A. Edinburgh neural machine translation systems for WMT 16. *arXiv preprint arXiv:1606.02891*, 2016.

Silva T.S.D., MacDonald D., Paterson G., Sikdar K.C., and Cochrane B. Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. *Computer Methods and Programs in Biomedicine*, 101(3):324 – 329, 2011. ISSN 0169-2607. URL `http://www.sciencedirect.com/science/article/pii/S0169260711000125`.

Stearns M., Price C., Spackman K., and Wang A. SNOMED clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium*, page 662âĂŞ666, 2001.

Tjong Kim Sang E.F. and De Meulder F. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 142–147. Association for Computational Linguistics, 2003.

Vicente I.S. and Manterola I. Paco2: A fully automated tool for gathering parallel corpora from the web. In Chair) N.C.C., Choukri K., Declerck T., DoÄ§an M.U., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Weijnitz P., Forsbom E., Gustavii E., Pettersson E., and Tiedemann J. MT Goes Farming: Comparing Two Machine Translation Approaches on a New Domain. *LREC*, 2004.

Yu A.C. Methods in biomedical ontology. *Journal of Biomedical Informatics*, 39(3):252 – 266, 2006. ISSN 1532-0464. URL `http://www.sciencedirect.com/science/article/pii/S1532046405001310`.

Zabala I., San Martin I., Lersundi M., Azkue J.J., and Mendizabal J.L. The Elaboration of Human Anatomy Terminology for the Basque Language: the Contribution of Translators, Linguists and Experts. *Terminàlia*, 15–25, 2012.

Zhu Y., Pan H., Zhou L., Zhao W., Chen A., Andersen U., Pan S., Tian L., and Lei J. Translation and Localization of SNOMED CT in China: A pilot study. *Artificial Intelligence in Medicine*, 54(2):147–149, 2012.