Translating SNOMED CT Terminology into a Minor Language

Olatz Perez-de-Viñaspre and Maite Oronoz
Outline

1 Introduction

2 SNOMED CT

3 Translation Algorithm
   • Phase 1: Lexical Resources
   • Phase 2: Finite State Transducers and Biomedical Affixes

4 Results

5 Conclusions
Outline

1 Introduction

2 SNOMED CT

3 Translation Algorithm
   - Phase 1: Lexical Resources
   - Phase 2: Finite State Transducers and Biomedical Affixes

4 Results

5 Conclusions
Introduction. SNOMED CT and Basque

- **SNOMED Clinical Terms (SNOMED CT)**
  - Comprehensive, multilingual clinical healthcare terminology
  - Enables consistent representation of meaning in EHRs

- **Basque**
  - Spoken by the 27% of Basques (714,136 out of 2,648,998)
    - 663,035 in the Spanish part
    - 51,100 in the French part
  - Basque is a minority language in its standardization process and persists between two powerful languages, Spanish and French
  - Nowadays, co-official in some parts, during centuries out of educational systems, media, and industrial environments
  - Written use of the Basque Language in the bio-sanitary system and in EHRs is low but co-official

![Classification of dialects in Basque (Koldo Zuazo, 2008)](image-url)

Figure 1: Classification of dialects in Basque (Koldo Zuazo, 2008)
Introduction. SNOMED CT and Basque

**SNOMED Clinical Terms (SNOMED CT)**
- Comprehensive, multilingual **clinical healthcare terminology**
- Enables consistent representation of meaning in EHRs

**Basque**
- Spoken by the 27% of Basques (714,136 out of 2,648,998)
  - 663,035 in the Spanish part
  - 51,100 in the French part

- Basque is a minority language in its standardization process and persists between two powerful languages, Spanish and French
- Nowadays, co-official in some parts, during centuries out of educational systems, media, and industrial environments

Written use of the Basque Language in the bio-sanitary system and in EHRs is low but co-official

**Figure 1:** Classification of dialects in Basque
(Koldo Zuazo, 2008)
Introduction. SNOMED CT and Basque

SNOMED Clinical Terms (SNOMED CT)
- Comprehensive, multilingual clinical healthcare terminology
- Enables consistent representation of meaning in EHRs

Basque
- Spoken by the 27% of Basques (714,136 out of 2,648,998)
  - 663,035 in the Spanish part
  - 51,100 in the French part

- Basque is a minority language in its standardization process and persists between two powerful languages, Spanish and French
- Nowadays, co-official in some parts, during centuries out of educational systems, media, and industrial environments

Written use of the Basque Language in the bio-sanitary system and in EHRs is low but co-official

Figure 1: Classification of dialects in Basque
(Koldo Zuazo, 2008)
Introduction. SNOMED CT and Basque

- **SNOMED Clinical Terms (SNOMED CT)**
  - Comprehensive, multilingual *clinical healthcare terminology*
  - Enables consistent representation of meaning in EHRs

- **Basque**
  - Spoken by the 27% of Basques (714,136 out of 2,648,998)
    - 663,035 in the Spanish part
    - 51,100 in the French part

  - Basque is a *minority language* in its *standardization* process and persists between two powerful languages, Spanish and French
  - Nowadays, co-official in some parts, during centuries out of educational systems, media, and industrial environments

  Written use of the Basque Language in the bio-sanitary system and in EHRs is low but co-official

*Figure 1: Classification of dialects in Basque (Koldo Zuazo, 2008)*
Introduction. SNOMED CT and Basque

- **SNOMED Clinical Terms (SNOMED CT)**
  - Comprehensive, multilingual *clinical healthcare terminology*
  - Enables consistent representation of meaning in EHRs

- **Basque**
  - Spoken by the 27% of Basques (714,136 out of 2,648,998)
    - 663,035 in the Spanish part
    - 51,100 in the French part
  - Basque is a *minority language* in its *standardization* process and persists between two powerful languages, Spanish and French
  - Nowadays, co-official in some parts, during centuries out of educational systems, media, and industrial environments

*Written use of the Basque Language in the bio-sanitary system and in EHRs is low but co-official*

---

*Figure 1:* Classification of dialects in Basque
(Koldo Zuazo, 2008)
Introduction. Goals

Goals

- 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
- To be able to access multilingual medical resources in Basque language

How?

Semi-automatically translating the terminology content of SNOMED CT
Introduction. Goals

Goals

- 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
- To be able to access multilingual medical resources in Basque language

How?

Semi-automatically translating the terminology content of SNOMED CT
Outline

1 Introduction

2 SNOMED CT

3 Translation Algorithm
   • Phase 1: Lexical Resources
   • Phase 2: Finite State Transducers and Biomedical Affixes

4 Results

5 Conclusions
SNOMED CT: core terminology for electronic health records with more than 296,000 active concepts and their corresponding terms (> 1 million terms)

Acceptable coverage of the terminology needed to record patients' conditions (Humphreys et al., 1997)

Description Types:

<table>
<thead>
<tr>
<th>Concept: 95575002 - Obstruction of pelviureteric junction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptions in English</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>Obstruction of pelviureteric junction (disorder)</td>
</tr>
<tr>
<td>Obstruction of pelviureteric junction</td>
</tr>
<tr>
<td>PUJ - Pelviureteric obstruction</td>
</tr>
<tr>
<td>PUO - Pelviureteric obstruction</td>
</tr>
<tr>
<td>Pelviureteric obstruction</td>
</tr>
<tr>
<td>UPJ - Ureteropelvic obstruction</td>
</tr>
<tr>
<td>Ureteropelvic obstruction</td>
</tr>
</tbody>
</table>
SNOMED CT: core terminology for electronic health records with more than 296,000 active concepts and their corresponding terms (> 1 million terms)

Acceptable coverage of the terminology needed to record patients conditions (Humphreys et al., 1997)

Description Types:

<table>
<thead>
<tr>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstruction of pelviureteric junction (disorder)</td>
<td>FSN</td>
</tr>
<tr>
<td>Obstruction of pelviureteric junction</td>
<td>Preferred Term</td>
</tr>
<tr>
<td>PUJ - Pelviureteric obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>PUO - Pelviureteric obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>Pelviureteric obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>UPJ - Ureteropelvic obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>Ureteropelvic obstruction</td>
<td>Synonym</td>
</tr>
</tbody>
</table>
SNOMED CT

- SNOMED CT: core terminology for electronic health records with more than 296,000 active concepts and their corresponding terms (> 1 million terms)
- Acceptable coverage of the terminology needed to record patients' conditions (Humphreys et al., 1997)
- Description Types:

<table>
<thead>
<tr>
<th>Concept: 95575002 - Obstruction of pelviureteric junction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptions in English</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>Obstruction of pelviureteric junction (disorder)</td>
</tr>
<tr>
<td>Obstruction of pelviureteric junction</td>
</tr>
<tr>
<td>PUJ - Pelviureteric obstruction</td>
</tr>
<tr>
<td>PUO - Pelviureteric obstruction</td>
</tr>
<tr>
<td>Pelviureteric obstruction</td>
</tr>
<tr>
<td>UPJ - Ureteropelvic obstruction</td>
</tr>
<tr>
<td>Ureteropelvic obstruction</td>
</tr>
</tbody>
</table>

Louhi 2014 (Gothenburg, Sweden)
SNOMED CT

- SNOMED CT: core terminology for electronic health records with more than 296,000 active concepts and their corresponding terms (> 1 million terms)

- Acceptable coverage of the terminology needed to record patients' conditions (Humphreys et al., 1997)

- Description Types:

<table>
<thead>
<tr>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstruction of pelviureteric junction (disorder)</td>
<td>FSN</td>
</tr>
<tr>
<td>Obstruction of pelviureteric junction</td>
<td>Preferred Term</td>
</tr>
<tr>
<td>PUJ - Pelviureteric obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>PUO - Pelviureteric obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>Pelviureteric obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>UPJ - Ureteropelvic obstruction</td>
<td>Synonym</td>
</tr>
<tr>
<td>Ureteropelvic obstruction</td>
<td>Synonym</td>
</tr>
</tbody>
</table>
## SNOMED CT hierarchies

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Semantic Tag</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Finding/disorder</td>
<td>disorder</td>
<td>Myocardial infarction</td>
</tr>
<tr>
<td></td>
<td>finding</td>
<td>Hyperalphaglobulinaemia</td>
</tr>
<tr>
<td>Procedure/intervention</td>
<td>procedure</td>
<td>Eye structure transplantation</td>
</tr>
<tr>
<td></td>
<td>regime/therapy</td>
<td>Puased electromagnetic energy to shoulder</td>
</tr>
<tr>
<td>Organism</td>
<td>organism</td>
<td>Pelistega europaea</td>
</tr>
<tr>
<td>Body structure</td>
<td>body structure</td>
<td>Supratentorial brain structure</td>
</tr>
<tr>
<td></td>
<td>morphologic abnormality</td>
<td>Acute erythremia</td>
</tr>
<tr>
<td></td>
<td>cell</td>
<td>Umbrella cell</td>
</tr>
<tr>
<td></td>
<td>cell structure</td>
<td>Viral envelope</td>
</tr>
<tr>
<td>Substance</td>
<td>substance</td>
<td>Bacterial agent</td>
</tr>
<tr>
<td>Pharmaceutical/biologic product</td>
<td>product</td>
<td>Naratriptan</td>
</tr>
<tr>
<td>Qualifier value</td>
<td>qualifier value</td>
<td>Perinatal period</td>
</tr>
<tr>
<td>Observable entity</td>
<td>observable entity</td>
<td>Postvaccination state</td>
</tr>
<tr>
<td>Event</td>
<td>event</td>
<td>Flood</td>
</tr>
<tr>
<td>Situation with explicit context</td>
<td>situation</td>
<td>Mother smokes</td>
</tr>
<tr>
<td>Social context</td>
<td>occupation</td>
<td>Hospital nurse</td>
</tr>
<tr>
<td></td>
<td>person</td>
<td>Homosexual parents (family)</td>
</tr>
<tr>
<td></td>
<td>ethnic group</td>
<td>Irish traveller</td>
</tr>
<tr>
<td></td>
<td>religion/philosophy</td>
<td>Nonconformist religion</td>
</tr>
<tr>
<td></td>
<td>life style</td>
<td>White collar thief</td>
</tr>
<tr>
<td></td>
<td>social concept</td>
<td>Upper class economic status</td>
</tr>
<tr>
<td></td>
<td>racial group</td>
<td>American Indian race</td>
</tr>
<tr>
<td>Physical object</td>
<td>physical object</td>
<td>Cardiac compression board</td>
</tr>
<tr>
<td>Specimen</td>
<td>specimen</td>
<td>Lumpectomy breast sample</td>
</tr>
<tr>
<td>Environment geographical location</td>
<td>environment</td>
<td>Psychiatric intensive care unit</td>
</tr>
<tr>
<td></td>
<td>geographic location</td>
<td>Republic of Serbia</td>
</tr>
<tr>
<td></td>
<td>environment/location</td>
<td>Environment or geographical location</td>
</tr>
<tr>
<td>Linkage concept</td>
<td>attribute</td>
<td>Agent relationship</td>
</tr>
<tr>
<td></td>
<td>link assertion</td>
<td>Has problem member</td>
</tr>
<tr>
<td></td>
<td>linkage concept</td>
<td>Linkage concept</td>
</tr>
<tr>
<td>Staging and scales</td>
<td>assessment scale</td>
<td>Lequesne index</td>
</tr>
<tr>
<td></td>
<td>tumor staging</td>
<td>pM category</td>
</tr>
<tr>
<td></td>
<td>staging scale</td>
<td>Chest pain rating</td>
</tr>
<tr>
<td>Special concept</td>
<td>navigational concept</td>
<td>Enzymes A - L</td>
</tr>
<tr>
<td></td>
<td>namespace concept</td>
<td>Extension Namespace 1000001</td>
</tr>
<tr>
<td></td>
<td>administrative concept</td>
<td>Appointment</td>
</tr>
<tr>
<td></td>
<td>special concept</td>
<td>Special concept</td>
</tr>
<tr>
<td>Record artifact</td>
<td>record artifact</td>
<td>Family history section</td>
</tr>
</tbody>
</table>
Two possible language sources: **English and Spanish**

We analyzed the RF2, Snapshot distributions dated 31-07-2012 (English) and 30-10-2012 (Spanish)

Analyzed aspects:

- **General numbers of FSNs, PTs and Synonyms and their lacks:**
  - The number of active **concepts** is the same: 296,433 (same file)
  - The number of **terms** in Spanish is smaller: 15,715 concepts lack of PTs and Synonyms

- **Length of the terms in each language:**
  - English: 6.76% (1 token), 23.28% (2 tokens) and 20.70% (3 tokens)
  - Spanish version: 33.79% (≤ 3 tokens), 66.21% (≥ 4 tokens)

Conclusions:

- The English version is more complete and consistent than the Spanish one
- The terms in the English version are shorter in length and, in consequence, simpler to translate

We decided to choose the English version as source
Two possible language sources: English and Spanish

We analyzed the RF2, Snapshot distributions dated 31-07-2012 (English) and 30-10-2012 (Spanish)

Analyzed aspects:

▶ General numbers of FSNs, PTs and Synonyms and their lacks:
  ★ The number of active concepts is the same: 296,433 (same file)
  ★ The number of terms in Spanish is smaller: 15,715 concepts lack of PTs and Synonyms

▶ Length of the terms in each language:
  ★ English: 6.76% (1 token), 23.28% (2 tokens) and 20.70% (3 tokens)
  ★ Spanish version: 33.79% (≤ 3 tokens), 66.21% (≥ 4 tokens)

Conclusions:

▶ The English version is more complete and consistent than the Spanish one
▶ The terms in the English version are shorter in length and, in consequence, simpler to translate

We decided to choose the English version as source
Two possible language sources: **English** and **Spanish**

We analyzed the RF2, Snapshot distributions dated 31-07-2012 (English) and 30-10-2012 (Spanish)

**Analyzed aspects:**

- **General numbers of FSNs, PTs and Synonyms and their lacks:**
  - The number of active concepts is the same: 296,433 (same file)
  - The number of terms in Spanish is smaller: 15,715 concepts lack of PTs and Synonyms

- **Length of the terms in each language:**
  - English: 6.76% (1 token), 23.28% (2 tokens) and 20.70% (3 tokens)
  - Spanish version: 33.79% (≤ 3 tokens), 66.21% (≥ 4 tokens)

**Conclusions:**

- The English version is more complete and consistent than the Spanish one
- The terms in the English version are shorter in length and, in consequence, simpler to translate

**We decided to choose the English version as source**
# SNOMED CT hierarchies

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Semantic Tag (ST)</th>
<th># FSN</th>
<th>Semantic Tag (ST)</th>
<th># FSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Finding/disorder</td>
<td>disorder</td>
<td>94,242</td>
<td>trastorno</td>
<td>82,725</td>
</tr>
<tr>
<td></td>
<td>finding</td>
<td>45,401</td>
<td>hallazgo</td>
<td>36,625</td>
</tr>
<tr>
<td>Procedure/intervention</td>
<td>procedure</td>
<td>75,078</td>
<td>procedimiento</td>
<td>59,411</td>
</tr>
<tr>
<td></td>
<td>regime/therapy</td>
<td>3,573</td>
<td>régimen/terapia</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>régimen/tratamiento</td>
<td>2,773</td>
</tr>
<tr>
<td>Organism</td>
<td>organism</td>
<td>35,870</td>
<td>organismo</td>
<td>35,465</td>
</tr>
<tr>
<td>Body structure</td>
<td>body structure</td>
<td>26,960</td>
<td>estructura corporal</td>
<td>26,747</td>
</tr>
<tr>
<td></td>
<td>morphologic abnormality</td>
<td>5,259</td>
<td>anomalía morfológica</td>
<td>5,082</td>
</tr>
<tr>
<td></td>
<td>cell</td>
<td>645</td>
<td>célula</td>
<td>640</td>
</tr>
<tr>
<td></td>
<td>cell structure</td>
<td>513</td>
<td>estructura celular</td>
<td>509</td>
</tr>
<tr>
<td>Substance</td>
<td>substance</td>
<td>25,834</td>
<td>sustancia</td>
<td>24,918</td>
</tr>
<tr>
<td>Pharmaceutical/biologic product</td>
<td>product</td>
<td>24,379</td>
<td>producto</td>
<td>23,854</td>
</tr>
<tr>
<td>Qualifier value</td>
<td>qualifier value</td>
<td>10,134</td>
<td>calificador</td>
<td>9,570</td>
</tr>
<tr>
<td>Observable entity</td>
<td>observable entity</td>
<td>9,044</td>
<td>entidad observable</td>
<td>8,602</td>
</tr>
<tr>
<td>Event</td>
<td>event</td>
<td>8,959</td>
<td>evento</td>
<td>8,587</td>
</tr>
<tr>
<td>Situation with explicit context</td>
<td>situation</td>
<td>8,716</td>
<td>situación</td>
<td>5,785</td>
</tr>
<tr>
<td>Social context</td>
<td>occupation</td>
<td>6,460</td>
<td>ocupación</td>
<td>4,650</td>
</tr>
<tr>
<td></td>
<td>person</td>
<td>668</td>
<td>persona</td>
<td>432</td>
</tr>
<tr>
<td></td>
<td>ethnic group</td>
<td>366</td>
<td>grupo étnico</td>
<td>283</td>
</tr>
<tr>
<td></td>
<td>religion/philosophy</td>
<td>227</td>
<td>religión/filosofía</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td>life style</td>
<td>30</td>
<td>estilo de vida</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>social concept</td>
<td>27</td>
<td>contexto social</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>racial group</td>
<td>21</td>
<td>grupo racial</td>
<td>19</td>
</tr>
<tr>
<td>Physical object</td>
<td>physical object</td>
<td>5,148</td>
<td>objeto físico</td>
<td>4,747</td>
</tr>
<tr>
<td>Specimen</td>
<td>specimen</td>
<td>1,455</td>
<td>espécnem</td>
<td>1,386</td>
</tr>
<tr>
<td>Environment</td>
<td>environment</td>
<td>1,253</td>
<td>medio ambiente</td>
<td>1,162</td>
</tr>
<tr>
<td></td>
<td>geographic location</td>
<td>619</td>
<td>localización geográfica</td>
<td>619</td>
</tr>
<tr>
<td></td>
<td>environment/location</td>
<td>1</td>
<td>medio ambiente/localización</td>
<td>1</td>
</tr>
<tr>
<td>Linkage concept</td>
<td>attribute</td>
<td>1,157</td>
<td>atributo</td>
<td>1,145</td>
</tr>
<tr>
<td></td>
<td>link assertion</td>
<td>8</td>
<td>relación asertiva</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>linkage concept</td>
<td>1</td>
<td>concepto de enlace</td>
<td>1</td>
</tr>
<tr>
<td>Staging and scales</td>
<td>assessment scale</td>
<td>1,125</td>
<td>escala de evaluación</td>
<td>1081</td>
</tr>
<tr>
<td></td>
<td>tumor staging</td>
<td>261</td>
<td>estadificación tumoral</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>staging scale</td>
<td>41</td>
<td>escala de estadificación</td>
<td>16</td>
</tr>
<tr>
<td>Special concept</td>
<td>navigational concept</td>
<td>732</td>
<td>concepto para navegación</td>
<td>725</td>
</tr>
<tr>
<td></td>
<td>namespace concept</td>
<td>153</td>
<td>espacio de nombres</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>administrative concept</td>
<td>80</td>
<td>concepto administrativo</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>special concept</td>
<td>31</td>
<td>concepto especial</td>
<td>1</td>
</tr>
</tbody>
</table>

Louhi 2014 (Gothenburg, Sweden)  
IXA (http://ixa.si.ehu.es)  
Translating SNOMED CT into Basque 10 / 20
Outline

1 Introduction

2 SNOMED CT

3 Translation Algorithm
   - Phase 1: Lexical Resources
   - Phase 2: Finite State Transducers and Biomedical Affixes

4 Results

5 Conclusions
Translation Algorithm

Incremental approach

The design is for any language pair but some linguistic resources needed for source and objective languages

Our implementation:
- Input: 1 term in English
- Output: ≥ 1 equivalent terms in Basque

The algorithm is applied at term-level (not at concept-level)

Algorithm: 4 phases
- The first 2 phases: developed and evaluated (quantitatively)
- Last 2 phases: in the very near future
Translation Algorithm

**Phase 0: Mapping of ICD-10**

- Semi-automatic mapping between SNOMED CT and the ICD-10 (IHTSDO)
- By identifying the sense of a concept in SNOMED CT, the best semantic space in the ICD-10 for this concept is searched
- The corresponding Basque term for some of the SNOMED CT concepts is obtained through ICD-10
- To take into account:
  - At concept level, not at term level ⇒ Before executing the algorithm implementation
  - Different purposes: ICD-10 for classification and SNOMED CT for representation
- Fruitful for very specialised terms
Translation Algorithm

1. phase: Lexical Knowledge

- *ItzulDB* (XML): initialized with all the lexical resources available + the pairs generated in the translation process
- Dictionaries of the bio-medical domain and the ICD-10 classification

Example:

**Input term:** Deoxyribonucleic acid
**Steps in figure number:** 1, 2, 4
**Translation:** *Azido desoxirribonukleiko, ADN, DNA*
Translation Algorithm

2. phase: Morphosemantics

- A term is analyzed at word-level and generation-rules are used to create the translation.
- We apply medical suffix and prefix equivalences and morphotactic rules, as well as transcription rules.

Example:

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

3. phase: Shallow Syntax (future)

- Chunk-level generation rules
- Hypothesis: some chunks will appear in ItzuDB

Example:

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

4.phase: Machine Translation (future)

- **Aim:** to adapt a rule-based automatic translation system called *Matxin* (Mayor et al., 2011) to the medical domain

**Example:**

**Input term:** Partial excision of oesophagus and interposition of *colon*

**Steps in figure number:** 12, 4

**Translation:** *Esofagoaren zati baten excisiona eta interpositiona bi puntua*
Translation Algorithm

1. Search the term in the translation pairs DB
2. Is there any Basque term found?
   - Yes → 6
   - No → 3
3. Is there any generation-rule applied?
   - Yes → 7
   - No → 12
4. Store the Basque term(s)
5. Make a word-level analysis of the term
6. Generate the Basque term
7. Is there any generation-rule applied?
   - Yes → 10
   - No → 12
8. Make a shadow-syntax analysis of the term
9. Use Automatic Translator
10. Is there any syntactic rule applied?
   - Yes → 11
   - No → 12
11. Help in the translation of new terms

Feedback

- All the processes finish in step 4
- The Basque equivalents with their original English terms are stored in an XML document that follows the TermBase eXchange
- ItzulDB (lexical resources) is enriched with the translation pairs generated that overcome a confidence threshold
- Help in the translation of new terms
Phase 1: Lexical Resources (English-Basque pairs)

Resources used to initialize ItzulDB

- **ZT Dictionary**: Science and technology (medicine, biochemistry, biology...). 13,764 English-Basque equivalences
- **Nursing Dictionary**: 5,393 entries
- **Glossary of Anatomy**: Anatomical terminology used by University experts in their lectures. 2,578 useful entries
- **ICD-10**: translated into Basque in 1996. Also available in English and in Spanish. 7,061 equivalences
- **EuskalTerm**: Terminology bank contains 75,860 entries from which 26,597 are from the biomedical domain
- **Elhuyar Dictionary**: English-Basque dictionary. 39,164 equivalences
Phase 2: Finite State Transducers and Biomedical Affixes

- FSTs used to identify the affixes in English Medical terms and by means of affix translation pairs, to generate the equivalent terms in Basque

**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

First approach (Perez-de-Viñaspre et al., 2013):
- 826 prefixes and 143 suffixes with medical meanings manually translated
- Evaluation: Gold Standard of 885 English-Basque pairs: precision of 93% and recall of 41%
- Only SNOMED CT terms for which all the prefixes and suffixes were identified were translated
  - For instance, the “hypophosphatemia” was not translated
  - “hypo”, “phos” and “emia” affixes identified
  - But “phat” not identified

Current approach:
- We have increased the number of affixes and transcription rules
- New numbers: 1,703 prefixes and 630 suffixes and 40 rules for transcription
- We are able to translate terms even though all their parts are not identified
- We now translate “hypophosphatemia” into “hipofosfatemia”
Outline

1 Introduction

2 SNOMED CT

3 Translation Algorithm
   - Phase 1: Lexical Resources
   - Phase 2: Finite State Transducers and Biomedical Affixes

4 Results

5 Conclusions
Results in the translation: Dictionary matching and morphosemantics.

- **Phase 1: Dictionary matching**
  - Evaluation in terms of *quantity*, not of *quality*
  - Dictionaries manually generated by lexicographers. The quality is assumed

- **Phase 2: Morphosemantics**
  - 93% precision and 41% recall
  - # Syn: The number of obtained Basque terms
  - # Matches: The number of English terms translated
  - The same input terms may have synonyms or even the same equivalent term given by different dictionaries.
    - Example: “alopatia” obtained in ZT and Nursing.

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Finding</th>
<th>Body Structure</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Syn</td>
<td>#Matches</td>
<td>#Syn</td>
<td>#Matches</td>
</tr>
<tr>
<td>ICD-10 mapping</td>
<td>11,227</td>
<td>-</td>
<td>1,878</td>
</tr>
<tr>
<td>In dictionaries</td>
<td>4,804</td>
<td>3,488</td>
<td>1,836</td>
</tr>
<tr>
<td>ZT Dictionary</td>
<td>1,104</td>
<td>883</td>
<td>367</td>
</tr>
<tr>
<td>Nursing Dictionary</td>
<td>437</td>
<td>350</td>
<td>340</td>
</tr>
<tr>
<td>Glossary of Anatomy</td>
<td>3</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>ICD-10</td>
<td>2,434</td>
<td>2,308</td>
<td>216</td>
</tr>
<tr>
<td>EuskalTerm</td>
<td>906</td>
<td>596</td>
<td>442</td>
</tr>
<tr>
<td>Elhuyar</td>
<td>299</td>
<td>135</td>
<td>956</td>
</tr>
<tr>
<td>Morphosemantics</td>
<td>2,620</td>
<td>2,184</td>
<td>705</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>17,627</strong></td>
<td><strong>5,672</strong></td>
<td><strong>4,419</strong></td>
</tr>
</tbody>
</table>

Louhi 2014 (Gothenburg, Sweden)  IXA (http://ixa.si.ehu.es)  Translating SNOMED CT into Basque
Overall Results

<table>
<thead>
<tr>
<th></th>
<th>Disorder</th>
<th>Finding</th>
<th>Body Structure</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translated Concepts</td>
<td>14,125</td>
<td>2,777</td>
<td>3,231</td>
<td>1,502</td>
</tr>
<tr>
<td>Concepts in total</td>
<td>65,386</td>
<td>33,204</td>
<td>31,105</td>
<td>82,069</td>
</tr>
<tr>
<td>Percentage</td>
<td>21.60%</td>
<td>8.36%</td>
<td>10.39%</td>
<td>1.83%</td>
</tr>
</tbody>
</table>

- **Disorder**: 21.60% of the translated. Good. Thanks to the ICD-10 (11,227 synonyms) and morphosemantics (81.53% of the simple terms)
- **Finding**: the most balanced
- **Body Structure**: the Glossary of Anatomy only contributes in this hierarchy (previous table)
- **Procedure**: dictionaries do not help much, in contrast, morphosemantics contribution allows to translate the 87.84% of the simple terms
Introduction

SNOMED CT

Translation Algorithm
  - Phase 1: Lexical Resources
  - Phase 2: Finite State Transducers and Biomedical Affixes

Results

Conclusions
Conclusions

- We have designed a translation algorithm for the multilingual terminology content of SNOMED CT and we have implemented the first two phases:
  1. Lexical resources feed our database
  2. Basque equivalents are generated using transducers and medical and biological affixes

- Dictionaries provide Basque equivalents of any term length while transducers get as input unique token terms

- Results are provided for the most populated hierarchies are shown even though both methods are applied for all the hierarchies in SNOMED CT

- Results are promising. We obtained the equivalents in Basque of 21.60% of the disorders

Future Work:
- Specialist in medical terminology can check the quality of the obtained terms and correct them
- Implement the remainder of the phases in the algorithm: Shallow Syntax and Machine Translation
Translating SNOMED CT Terminology into a Minor Language

Olatz Perez-de-Viñaspre and Maite Oronoz

University of the Basque Country

IXA Taldea

IXA (http://ixa.si.ehu.es)