Finding your way among ISO standards for terminology, language and linguistics

Antonio Pareja Lora
DSIC & ILSA - Universidad Complutense de Madrid / ATLAS - UNED
aplora@ucm.es

Resumen

Aunque hacemos un uso constante en nuestra vida diaria de las normas y estándares, esto no parece tan frecuente en el caso de los trabajos que tienen que ver con la terminología, las lenguas y/o la lingüística. Ello puede deberse a que las normas en cuestión son de muy reciente creación y, por tanto, ni ellas ni los beneficios asociados a su uso se conocen aún suficientemente. Por este motivo, el objetivo de este artículo es presentar (i) los beneficios que aporta el uso de normas en general; (ii) el comité técnico de normalización internacional responsable de la elaboración de las normas en esta área (el ISO/TC 37); y (iii) un ejemplo de aplicación de algunas de sus normas (el Servicio Web Normalizado de FreeLing – FreeLing SWS) que muestra uno de los beneficios particulares del uso de las normas del ISO/TC 37, esto es, que facilitan la interoperación de las anotaciones lingüísticas.

Palabras clave: ISO, normas, terminología, lenguas, lingüística.

Abstract

Even though we use standards pervasively and constantly in our lives, this is not so frequent a case in terminology, language and/or linguistics works. This might be due to the fact that the related standards have been created quite recently and, thus, they and the benefits that entail using them are not so well-known so far. Thus, the main goal of this paper is to present (i) the benefits of using standards in general, (ii) the international technical committee responsible for the initiatives of standardisation in this field (that is, ISO/TC 37), and (iii) an example of application of some of the standards developed by this technical committee (e.g. the FreeLing Standardised Web Service – FreeLing SWS), which shows, for instance, one of the main particular benefits of using ISO/TC 37 standards (i.e. they help linguistic annotations interoperate).

Keywords: ISO, standards, terminology, language, linguistics.

1. Introduction: the benefits of standards

Even though we might not be aware this fact, we make use of standards twenty-four hours a day, seven days a week. The items that we most use every day comply with some particular standard, such as (i) the power and/or the voltage of the current that our electronic devices use to work;(ii) the shape, number of pins and other features of all kinds of plugs,sockets and connectors for all these devices;(iii) or the wavelength and the frequency of the Wi-Fi signal we use to connect to the internet with our smartphones, tablets and computers. Thus, using our phone, watching TV, cooking in a microwave, or simply turning on or off a light entails using somehow at least an electrical or electronic standard. This means that, clearly, the utilization of standards is pervasive and constant nowadays.

This is no surprise: it stems from the manifold benefits of standards. Briefly, standards have helped technology progress, by providing best practices and agreed-upon, recommended procedures and product features that prevent us from solving a common problem (that is, from reinventing the wheel) time and time again. Accordingly, standards help us save time, which can then be devoted to solving new problems and developing new technologies. Furthermore, these new technologies can be built on the basis of other sufficiently consolidated (and consensus-based) ones, which usually improves and/or helps assess their quality. In particular, at least in the electrical, electronic and/or computational area, the definition of standardised components (e.g., USB connectors) and their standard-compliant manufacturehas facilitatedcomponent interoperationand integration, thus enabling a faster and easier production of new aggregated components. This, as a side effect, results in lower production costs as well.

The main association for the development of standards is ISO, the International Organization for Standardization (presented in detail below). ISO has been publishing standards for almost 70 years already, and has gained invaluable experience and knowledge about these and many other benefits of (using) standards. According to a report of the ISO Central Secretariat (2014), standards help companies

1. Streamline their internal operations: forexample,

> byreducing the time needed to perform specific activities in the various business functions, decreasing waste, reducing procurement costs and increasing productivity. [...] the contribution of standards to the gross profit of companies ranges between 0.15 % and 5 % of the annual sales revenues. (ISO Central Secretariat, 2014)

2. Innovating and scaling up operations, since they can serve "as the basis for innovating business processes, allowing companies to expand their suppliers' network or to introduce and manage new product lines effectively" (ISO Central Secretariat, 2014).
3. Creating or entering new markets: "Standards have been used as the basis for developing new products, penetrating new markets (both domestic and export), supporting the market uptake of products, and even creating markets" (ISO Central Secretariat, 2014).

These and other economic and social benefits of standards, mentioned in the ISO Membership Manual (International Organization for Standardization, 2013a)are summarised in Table 1.

Table 1: The main economic and social benefits of using standards (extracted from International Organization for Standardization (2013a))

AREA / COMMUNITY	MAIN STANDARD BENEFITS
COMPANIES	— Improving operational efficiency and reducing costs; — Rationalizing processes; — Increasing efficiency and effectiveness; — Creating business opportunities; — Facilitating international exchange of goods and services; — Increasing consumer confidence; — Contributing to company gross profits.
CONSUMERS	— Improving choice and access to goods and services; — Lowering costs; — Protecting health, safety and the environment; — Easing quality and reliability assessment of products and services.
PUBLIC AUTHORITIES	— Helping develop and promote efficiently measures of public utility (for example, on safety, security, and protection of the environment).
TECHNICAL REGULATIONS	— Protecting local industries that produce good quality products from unfair competition by imported and sub-standard goods. — Improving the chances of receiving public and stakeholder acceptance, and meeting World Trade Organisation (WTO) requirements.

Hence, standards seem to be useful and, apparently, they should be used everytime and everywhere, since applying and complying with them result in several benefits, at least in the areas included in Table 1.

However, the knowledge and use of standards for language and linguistics seems a bit limited so far. This is particularly true for the ISO standards recently developed for the field of language resources and their annotation. Most people do not know about the ISO standards for morpho-syntactic, syntactic and semantic annotation, for example. And, accordingly, they do not use them and do not profit from the benefits of using them (e.g., interoperability). The main purpose of this paper, thus, is to introduce the ISO standards for terminology, language and linguistics, already developed or currently under development, and present a case of use of two of them, namely ISO/MAF (ISO 24611:2012) and ISO/SynAF (ISO 24615:2010). This case of use shows how these standards help several linguistic annotations interoperate.

Therefore, this paper has been organized as follows. First, Section 2briefly introduces ISO (the International Organization for Standardization), and provides some figures about its standardisation bodies, standards and projects (Subsection2.1). Second, the ISO technical committee dealing with terminological, language and linguistic issues (ISO/TC 37) is presented, together with its internal structure and subcommittees, in Section 3. Third, Section 4 details how ISO/MAF (ISO 24611:2012) and ISO/SynAF (ISO 24615:2010) have been applied to the standardisation of FreeLing's annotations morpho-syntactic and syntactic annotations[1] (Padró & Stanilovsky, 2012). Fourth, in Section 5, the advantages and disadvantages of (using) standards for terminology, language and/or linguistic works are discussed. Fifth, the conclusions of this researchare presented in Section 6. Finally, Sections 7 and 8 include, respectively, the acknowledgements and the references associated to this paper.

2. The International Organization for Standardization (ISO)

As commented in the ISO statutes (International Organization for Standardization, 2013b),

> ISO (International Organization for Standardization) is the world's largest developer of voluntary International Standards. ISO was founded in 1947, and since then has published more than 19 000[2] International Standards covering almost all aspects of technology and business. In 2013, ISO has members from 164 countries. (3)

[1] http://nlp.lsi.upc.edu/freeling/.

[2] As of 2013; see the current figures below.

This paragraph highlights three main facts about ISO. First, the main goal of ISO is developing *standards for technology and business*. Assuming that the leitmotif of technology and business are, respectively, progressing and making (more) money, this entails that ISO standards seek to help (i) technology advance and (ii) companies' benefit increase. Indeed, as stated also in ISO statutes (International Organization for Standardization, 2013b:7), the object of this organization is "to promote the development of standardization and related activities in the world with a view to facilitating international exchange of goods and services and to developing cooperation in the spheres of intellectual, scientific, technological and economic activity". This is one of the main driving forces of ISO standard development; but also, sometimes, one of its main burdens. Finding the balance between scientific, business and economical factors, for example, is not always easy. This is one of manypossible risks that can endanger a standardisation process, as discussed in Section 5.

Second, a key word in the paragraph cited above is *voluntary*. Indeed, ISO standards are in most cases developed collaboratively byvolunteer experts in the target field (or scope) of the standard, that is, people with a great expertise in the area who,in most cases[3], commit themselves to this duty for free and altruistically.

And third, ISO develops *international standards*. So (1) an international consensus is required in order to publish a standard;and (2) in order to reach this consensus, ISO has had to agree on a set of common languages for the internal discussion, commenting and voting (or balloting, in the ISO terminology) of drafts, and also for the publication of its standards.

Regarding issue (1), international consensus is most often reached in ISO by means of a complex iterative document improvement-ballot process.A schematic view of this iterative process (that is, the route(s) typically followed by ISO drafts, standards and/or deliverables) has been included in Figure 1. It is not further commented here for the sake of space.

In order to proceed to the following phase of improvement of the document, each of its versions has to be approved or rejected in a ballot by a majority of the members of the technical committee responsible for its publication. These members (or member bodies) are, in short, "those national standards bodies most broadly representative of

[3] But not always, since some vested interests may prevail – see Section 5. Fortunately, this does not happen frequently.

standardization in their respective countries" (International Organization for Standardization, 2013b:8)[4].

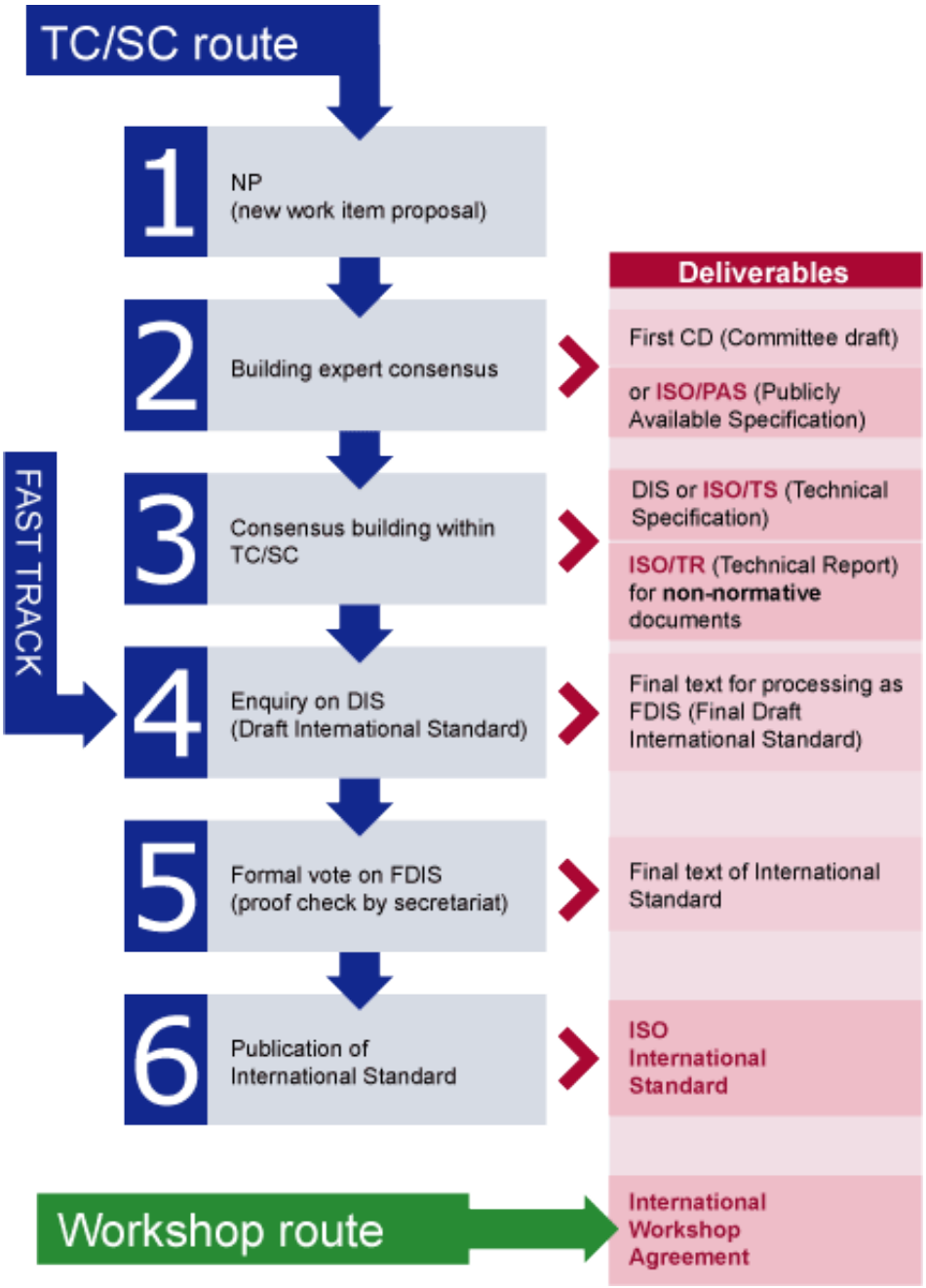

Figure 1: Development routes typically followed by ISO standards and/or deliverables (taken from http://www.iso.org/iso/home/standards_development/deliverables-all.htm)

[4] For example, the national standards bodies most broadly representative of standardization in Spain is AENOR (Asociación Española de Normalización y Certificación, https://www.aenor.es/aenor/inicio/home/home.asp); in France, AFNOR (Association Française de Normalisation, http://www.afnor.org); in Germany, DIN (Deutsches Institut für Normung e.V., http://www.din.de/cmd?level=tpl-home&contextid=din); and ANSI (American National Standards Institute, http://www.ansi.org) in the USA.

Thus, each country has the chance to comment and give its opinion in each of these ballots. Besides, each member body can nominate some national experts before the development of a new standard begins. These nominated experts are incorporated afterwards to the working group in charge of the development of the ISO deliverable, where they can provide a more direct feedback about their country's position for this deliverable.

As for issue (2), the current three official languages of ISO are English, French and Russian[5]. This means that ISO internal documents should be circulated in any of these three languages. However, for convenience, the most used language within ISO is English, and sometimes a bilingual English-French version is produced, but Russian is scarcely brought into play. ISO has already devised a procedure and a set of requirements for the consideration and addition of other official languages (e.g. Spanish). However, in spite of being one of the most extended and/or spoken languages in the world, Spanish is not an official language within ISO yet, since these requirements are hard to fulfil. They are not detailed here for brevity but, basically, there must be an important number of Spanish-speaking national bodies participating in the standardisation processes and asking for the inclusion of Spanish as an official language[6]. This often prevents the Spanish-speaking community from involving more actively into ISO standardisation processes.

From a wider perspective, on the one hand, the need to read, comment and/or discuss documents written in a different language and express opinions in this other language is sometimes a barrier for the participation of relevant experts in standard development processes that should be broken somehow. This is another disadvantage of international standardisation processes, as commented in Section 5.On the other hand, the national standardisation bodies are responsible for the translation, adaptation and adoption of ISO standards (in)totheir country's language(s) and particularities. Consequently, the target users of standards can read them (supposedly) in their own language, which helps extend the use of standards and adopt them at the national level. Nevertheless, the number of translated standards keeps decreasing, mainly due to economical reasons. Unfortunately, this (i) limits the knowledge and the implementation of standards; and (2) usually favours the extension of English and its termsand decreases the lexical and/or terminological productivity of (e.g.) Spanish (cf. Pozzi, 2006).

[5] Taking into account the date of creation of ISO, and that it was born with the under the guidance of the UNO (United Nations Organization), it is more than likely that these three languages were declared official in ISO due to political reasons. Not surprisingly, they correspond to the three main languages of the Allied Forces in World War II.

[6] Clearly, if Russian had to go through a similar process nowadays, it would not be accepted as an ISO official language. In fact, the number of ISO documents and standards currently developed or translated within ISO into Russian is quite small.

2.1. ISO in numbers

According to the data published in ISO's website[7], there are currently 235 active technical bodies within ISO: 221 technical committees and 14 project committees[8]. 14 out of the 221 ISO technical committees are currently STANDBY[9]. Besides, there are 44 additional disbanded technical bodies, such as TC 141 (Powered lawn and garden equipment). Thus, in total, 295 ISO technical bodies have been created so far.

This page states also thatthere are currently 20,386 ISO published standards, while yet another 4,725 ISO standards (or deliverables) are under development. The main driving force of this extraordinary productivity is ISO/IEC JTC 1[10] which, alone, has developed more than 2,800 standards (that is, 13.82% of ISO standards) and is currently working on more than 600 new standardisation projects. Only 17 technical committees are responsible for almost 50% of the published ISO standards. They are included, together with their corresponding number of standards, in Figure 2 (see next page).

3. ISO and the standardization of terminology, language and linguistic issues: the ISO/TC 37

There is a particular ISO technical committee dealing with terminology, language and/or linguistics, namely ISO/TC 37 – the ISO Technical Committee for the Standardization of Terminology and other language and content resources. According to the ISO/TC 37 Business Plan (ISO/TC 37, 2013),

> TC 37 serves all fields and applications, where human-human and human-machine communication is involved. This refers in particular to the language industry (LI) comprising LI products, such as language technology tools and content resources (i.e. terminology and other language and content resources) as well as language services (such as sci-tech writing, technical communication, localization, translation and interpretation, etc.) – and of course the users of the above. It further refers to fields and applications where knowledge is represented in whatever form. Thus TC 37 standards are fundamental for language resource management,

[7] http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees.htm.
[8] As the webpage in the previous footnote states, **project committees** "are established when there is a need for an International Standard on a specific topic that does not fall into the scope of an existing TC. Project committees are disbanded once the standard has been published".
[9] "**STANDBY** refers to technical committees that have no work item in progress or foreseen but that are required to review the ISO International Standards for which they are responsible" (*Idem*).
[10] The ISO-IEC Joint Technical Committee for Information Technology.

knowledge management, [and] content management in most of their guises. (1)

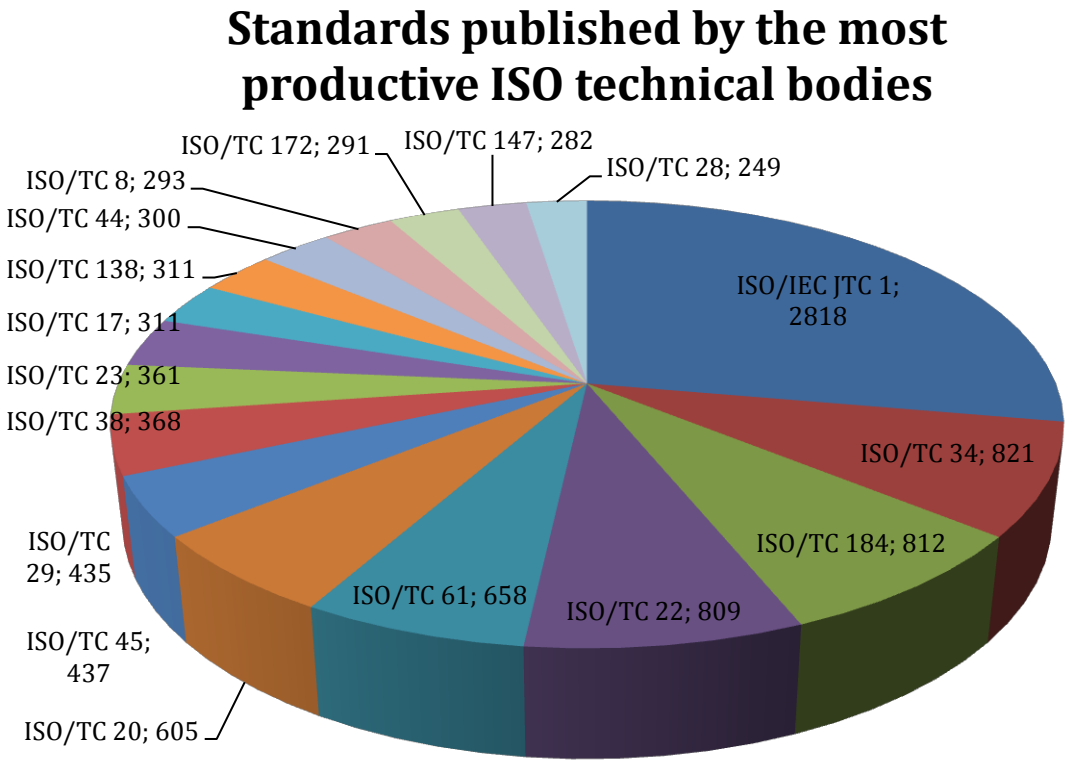

Figure 2: Main ISO technical committees with respect to their standard production (as of April, 2015)

Currently,ISO/TC 37 has five subcommittees (SC 1 to SC 5) and one TC-level active working group (WG 9). Yet another working group was created within ISO/TC 37(WG 9), but it was disbanded in 2014. All of them are summarised in Figure 3.

The five ISO/TC 37 subcommittees are structured internally according to more specific knowledge areas, shown respectively in Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8. All these figures are not further commented here for the sake of space; however, the name of the different subcommittees and working groups is sufficiently unequivocal and should be enough for the reader to comprehend their corresponding scope.

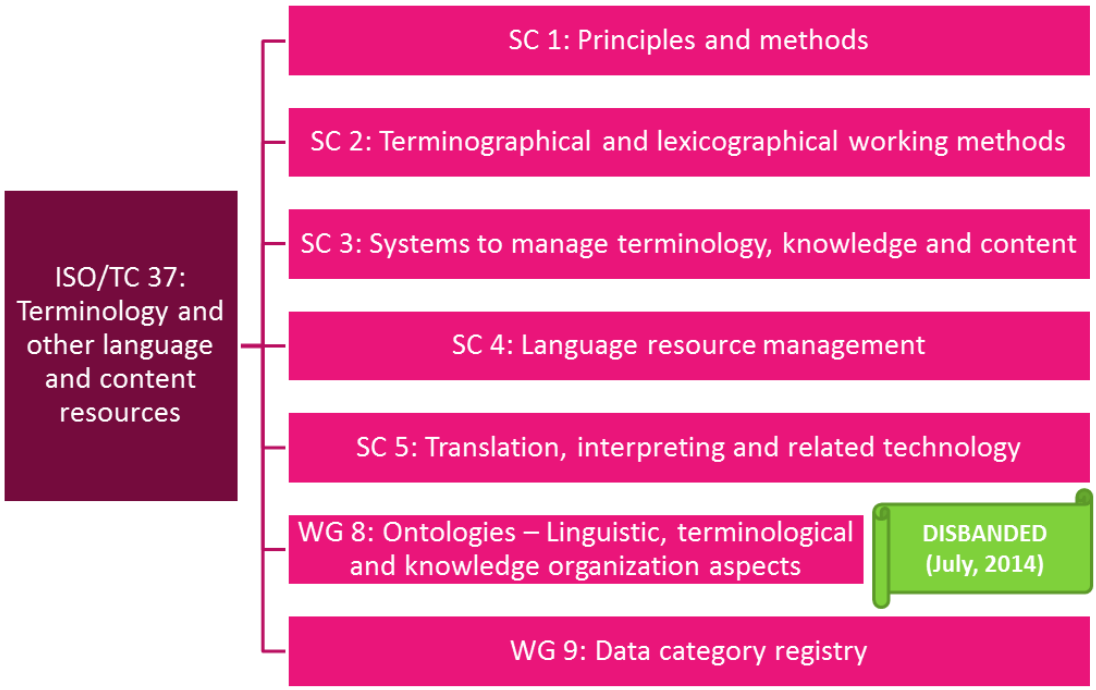

Figure 3: Structure of ISO/TC 37 (as of April, 2015)

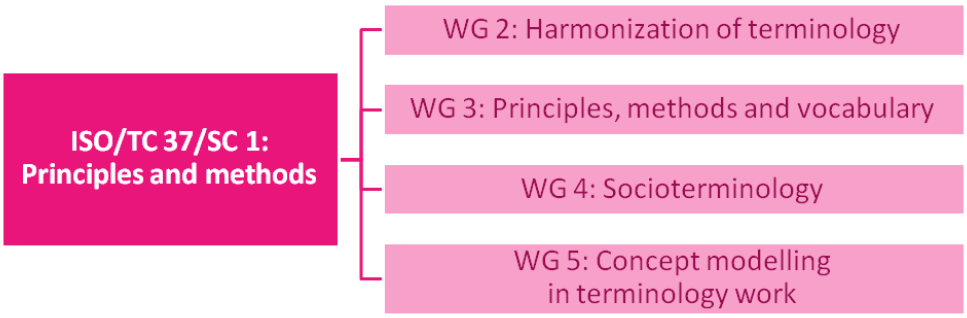

Figure 4: Structure of ISO/TC 37/SC 1 (as of April, 2015)

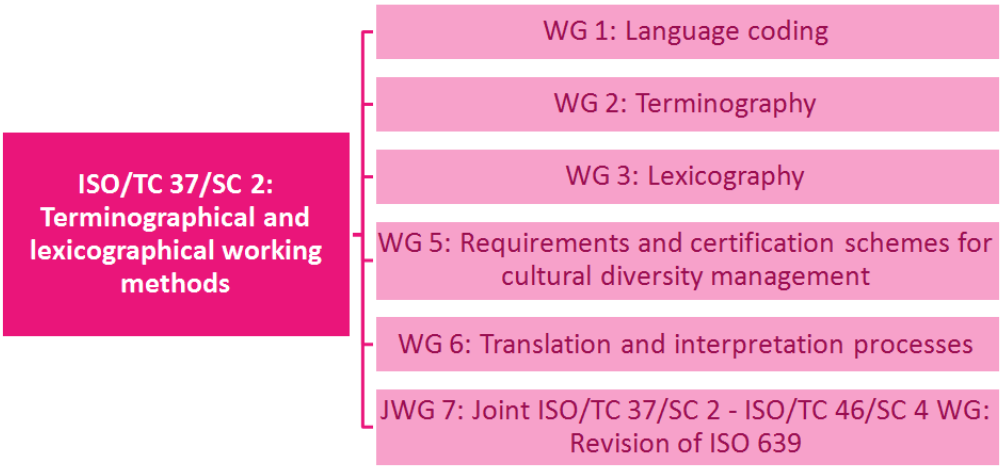

Figure 5: Structure of ISO/TC 37/SC 2 (as of April, 2015)

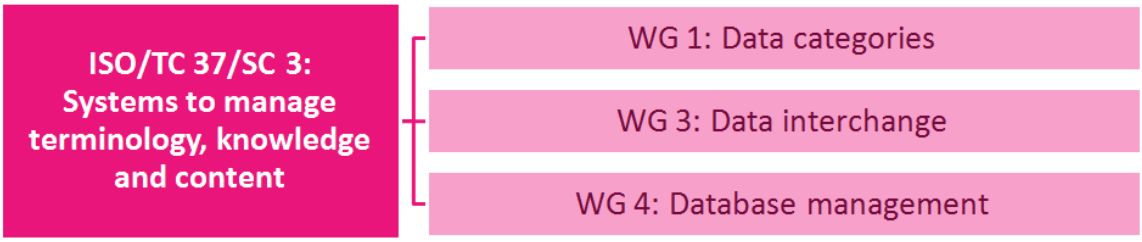

Figure 6: Structure of ISO/TC 37/SC 3 (as of April, 2015)

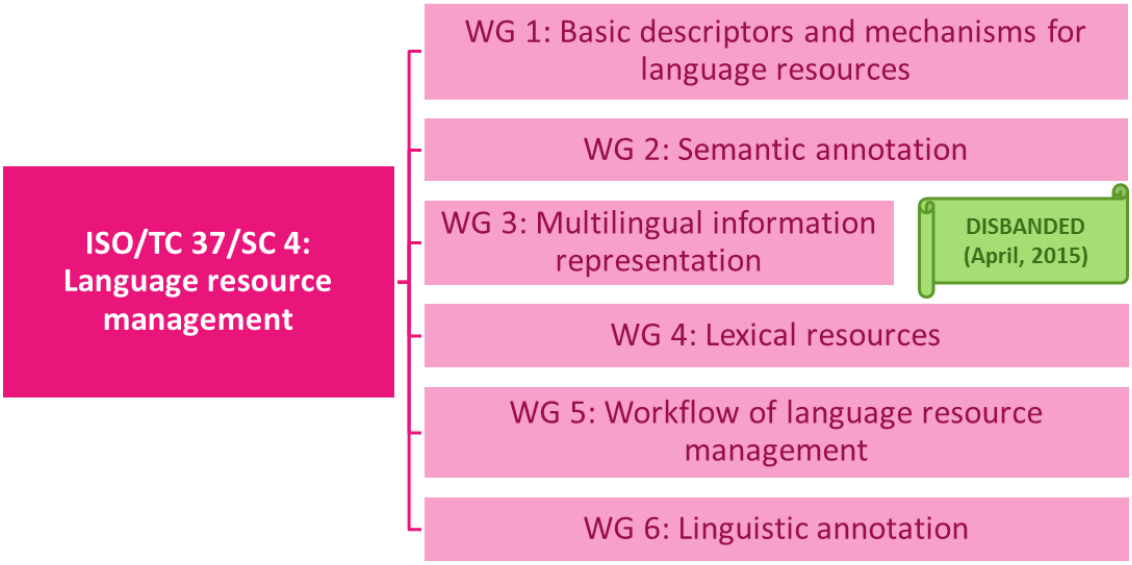

Figure 7: Structure of ISO/TC 37/SC 4 (as of April, 2015)

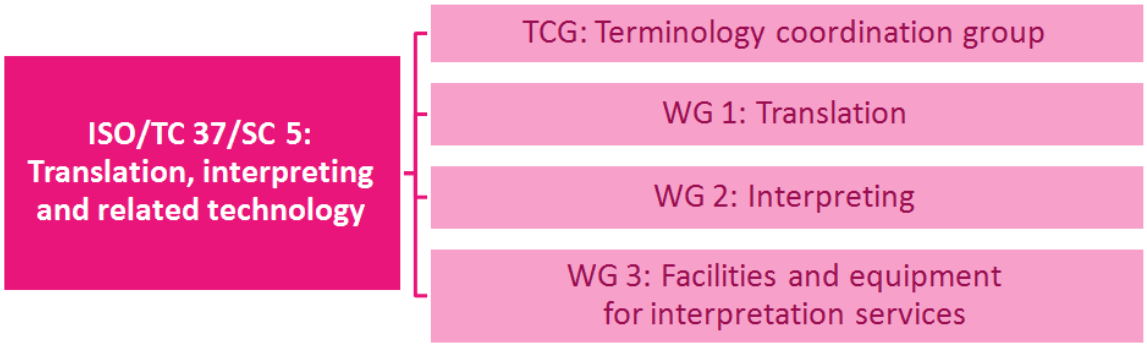

Figure 8: Structure of ISO/TC 37/SC 5 (as of April, 2015)

Finally, Table 2 summarises the particular number of (i) published standards, (ii) standards currently under development, (iii) withdrawn standards so far; and participating countries in ISO/TC 37 (in total and in each of its subcommittees).

The following section presents a particular application of a specific subset of ISO/TC 37 standards developed within SC 4 (mainly ISO/FSR (ISO 24610-1:2006), ISO/PISA (ISO 24619:2011), ISO/MAF (ISO 24611:2012) and ISO/SynAF (ISO 24615:2010)), which has helped show how the use of these standards enables and/or improves the interoperability of linguistic annotations.

Table 2: Standards, technical documents, projectsand participating and observing countries of ISO/TC 37 technical bodies

ISO Technical Body	PUBLISHED standards and technical documents (including updates)	Standards and projects UNDER DEVELOPMENT	WITHDRAWN standards and projects	Participating countries
ISO/TC 37/ SC 1	7	1	3	26
ISO/TC 37/ SC 2	14	1	3	30
ISO/TC 37/ SC 3	5	2	5	23
ISO/TC 37/ SC 4	17	6	0	22
ISO/TC 37/ SC 5	2	9	0	27
ISO/TC 37 (TOTAL)	**45**	**19**	**11**	**31**

4. Applying ISO/TC 37 standards – an example

As discussed in Pareja-Lora (2012), many conflicts and problems prevent linguistic annotation tools and annotations from interoperating, and make it very difficult to reuse them easily in new scenarios, e.g., in natural language processing (NLP) pipelines (Buyko et al., 2008). However, developing new linguistic tools from scratch is quite a high time-consuming task that also entails a very high cost. Therefore, the need to reuse the existing

linguistic tools and find cost-effective ways to make them and/or their annotations interoperate gets clearer and more urgent every day.

A traditional way to overcome this reuse and/or interoperability problem in several areas (for instance, the (inter)connection of electronic plugs and/or devices) is standardisation. This is one of the main assumptions driving the development of ISO/TC 37/SC 4 standards and standard drafts, such as ISO/MAF[11] (ISO 24611:2012) or ISO/SynAF[12] (ISO 24615:2010) and ISO/SynAF-ISOTiger[13] (Bosch et al., 2012; 2014).

This is also the main hypothesis underlying the work presented in this section: that standardisation can help linguistic tools and annotations interoperate (Ballesteros-Calvo et al., 2013). In particular, this section shows that the standardisation of FreeLing's morpho-syntactic and syntactic annotations for Spanishcan help link them together and make them interoperate.

4.1. Why FreeLing

FreeLing (Padró & Stanilovsky, 2012) is an open source, freely available tool for the analysis and annotation of texts at several levels and layers, written in a number of languages (see below). In particular, FreeLing 3.0 processes the input text and provides, for instance, its (i) token segmentation, (ii) POS tagging, (iii) deep and shallow syntactic constituency parsing; (iv) syntactic dependency parsing, (v) multiword detection, (vi) named entity recognition and classification (according to the MUC classification – Chinchor, 1997), and (vi) (Euro)WordNet-based (Miller, 1995; Fellbaum, 1998; Vossen, 1998) sense tagging.

The input text can be written in Spanish, Catalan, Galician, Portuguese, Italian, French, English or Russian, amongst others. Its openness, availability, versatility and multilingual capabilities have made it a very popular and most widespread tool, for instance, in Spain.

However, FreeLing has a few limitations that require being solved. For example, on the one hand, none of its manifold output annotations comply with the current standards for linguistic annotation (such as ISO/MAF (ISO 24611:2012) or ISO/SynAF (ISO 24615:2010)) and are not even encoded by means of a standard language (such as XML[14]). On the other hand, its current implementation does not allow for its inclusion 'as is' into

[11] ISO/MAF provides a general framework and a set of recommendations for the annotation of morpho-syntactic units with their grammatical category and its morphosyntactic features. It provides also a recommended (not mandatory) XML serialisation for morphosyntactic annotations, which makes them be more syntactically interoperable and referenceable by other annotations.

[12] ISO/SynAF includes a general framework for the annotation of syntactic units and relationships, but contains no particular XML serialization (which has been developed within ISO/SynAF-ISOTiger).

[13] An XML serialization of ISO/SynAF that is based on the TIGER (König & Lezius, 2003) language and format.

[14] http://www.w3.org/TR/2006/REC-xml11-20060816/.

NLP pipelines. These two factors altogether reduce to some extent the interoperability and reusability of this tool.

4.2. Standardizing FreeLing

So, in order to overcome all these problems, first of all, its was decided to transform FreeLing 3.0 into a web service[15], which is a fairly well-known, widespread and standard-based[16] way to improve the interoperability of computer applications (Kashyap et al., 2008)[17].

In spite of this first standardisation step, FreeLing's results (i.e., its annotations) were still encoded in a tool-dependent, non-standard-compliant way. For example, the tokens in its POS annotations were not assigned a URI[18] in order to allow other annotations of the same text to refer to them. This made it difficult to (i) interconnect its annotations together; and (ii) merge them with the annotations performed by other tools. This, in turn, prevented the tool from being sufficiently interoperable. Accordingly, a second step towards the standardisation of FreeLing was required, namely the standardisation of its annotations.

Previous approaches[19] (Poch & Bel, 2011; Morell, Vivaldi & Bel, 2012) had already accomplished the standardisation of FreeLing's (morpho-)syntactic outputs using the Graph Annotation Format (GrAF: Ide & Suderman, 2007). GrAF is an XML serialization of the standard Linguistic Annotation Framework of ISO (ISO/LAF – ISO24612:2012). ISO/LAF and GrAF altogether provide a general annotation framework, pretty suitable for those cases for which no other ISO annotation standard is available. However, it is too general for quite common and useful types of annotations, such as morpho-syntactic and syntactic annotations, for which other specific ISO annotation standards have already been developed (namely ISO/MAF and ISO/SynAF, respectively). While these other ISO standards are also ISO/LAF-compliant, (a) they are also less verbose than GrAF; and (b) provide a further specified (and standardised) vocabulary to encode these particular types of annotations. Thus, in general, these are the ones that should be used for mopho-syntactic and syntactic annotation; nevertheless, they had never been used to encode FreeLing's outputs in a standard-compliant way. Accordingly, ISO/MAF and ISO/SynAF have been used to standardise the corresponding annotations in FreeLing's results.

---

[15] Referred to as FreeLing SWS here.

[16] By using W3C recommendations, such as WSDL (http://www.w3.org/TR/wsdl) and SOAP (http://www.w3.org/TR/soap12-part0/).

[17] FreeLing SWS can be tested using the client available at http://quijote.fdi.ucm.es:8084/ClienteFreeLing/index.jsp.

[18] Uniform Resource Identifier, see http://www.w3.org/TR/uri-clarification/.

[19] Followed within the PANACEA project, http://www.panacea-lr.eu/.

4.2.1. Standardizing FreeLing's morpho-syntactic annotations

Example 1 introduces the native non-standardised FreeLing's POS tagging-only output obtained for the Spanish sentence 'Mi gato se llama Tiger.' ('My cat's name is Tiger.').

```
Mi     mi     DP1CSS
gato   gato   NCMS000
se     se     P00CN000
llama  llamar VMIP3S0
Tiger  tiger  NP00000
.      .      Fp
```

Example 1: FreeLing's native, basic POS tagging of 'Mi gato se llama Tiger.' ('My cat's name is Tiger.')

As Example 1 shows, FreeLing's native, basic POS tagging includes the input text of the token, its associated lemma and its POS tag. They are included in a file, one token per line, one item per column. The first column contains the input text of the token; the second one, its lemma; the third one, an EAGLES (1996)-conformant POS tag[20] that includes essentially both its grammatical category and its morpho-syntactic features.

Unfortunately, this annotation format does not state explicitly what these fields mean. Therefore, the semantics of each tag is implicit. This clearly complicates (1) interpretingannotations automatically; (2) comparing them with other POS annotations; and(3) making them interoperate with other annotations (Pareja-Lora, 2012). Besides, the tokens and their POS annotations cannot be reused and/or referenced, e.g., by a syntactic annotation of the sentence, since no way to link to them is provided.

These problems were partially solved by means of an ISO/MAF based standardisation of FreeLing's morpho-syntactic annotations. The resulting ISO/MAF based annotations of the Spanish sentence 'Mi gato se llama Tiger.',implemented in XML by means of FreeLing SWS, has been included in Example 2 (see next page).

As shown in this example, first, each token element is assigned its corresponding URI within the file, which functions in the annotation as its persistent identifier (PISA: ISO 24619:2011 – ISO/PISA). This PISA is assigned by means of the `@xml:id` attribute, which helps building other annotations (for example, the `wordForm` annotations) on top of this one and linking them together.

[20] For instance, the POS tag for 'Mi' ('My') is 'DP1CSS', which means that [A] it is a token whose grammatical category is determinant or pronoun ('DP'); and [B] it has the following morphosyntactic features: first person ('1'), common gender ('C'), singular number ('S') and possessive type (final 'S').

```
<?xml version="1.0" encoding="UTF-8"?>

<maf>
   <tokenxml:id="t1">Mi</token>
   <tokenxml:id="t2">gato</token>
   <tokenxml:id="t3">se</token>
   <tokenxml:id="t4">llama</token>
   <tokenxml:id="t5">Tiger</token>
   <tokenxml:id="t6" join="left">.</token>

   <wordFormxml:id="wordForm1" tokens="#t1" lemma="mi">
      <fs>
         <fname="pos">
            <symbolvalue="DP1CSS"/>
               <!-- Determinante: Posesivo, Primera persona, Común, Número singular,
               Poseedor singular -->
         </f>
      </fs>
   </wordForm>

   <wordFormxml:id="wordForm2" tokens="#t2" lemma="gato">
      <fs>
         <fname="pos">
            <symbolvalue="NCMS000"/>
               <!-- Nombre: Común, Masculino, Singular -->
         </f>
      </fs>
   </wordForm>

   <wordFormxml:id="wordForm3" tokens="#t3" lemma="se">
      <fs>
         <fname="pos">
            <symbolvalue="P00CN000"/>
               <!-- Pronombre: Común, Impersonal/Invariable -->
         </f>
      </fs>
   </wordForm>

   <wordFormxml:id="wordForm4" tokens="#t4" lemma="llamar">
      <fs>
         <fname="pos">
            <symbolvalue="VMIP3S0"/>
               <!-- Verbo: Principal, Indicativo, Presente, Tercera persona,
               Singular -->
         </f>
      </fs>
   </wordForm>

   <wordFormxml:id="wordForm5" tokens="#t5" lemma="tiger">
      <fs>
         <fname="pos">
            <symbolvalue="NP00000"/>
               <!-- Nombre: Propio, Genero Indeterminado, Número Indeterminado -->
         </f>
      </fs>
   </wordForm>

   <wordFormxml:id="wordForm6" tokens="#t6" lemma=".">
      <fs>
         <fname="pos">
            <symbolvalue="Fp"/>
               <!-- Puntuación: Punto Final -->
         </f>
      </fs>
   </wordForm>

</maf>
```

Example 2: MAF-compliant annotation of 'Mi gato se llama Tiger.', obtained with the FreeLing SWS

This way, tokens can be referenced internally (from inside the file), locally (from inside the same [file] system) and globally (from outside the system). For instance, a token can be easily referenced locally by concatenating the identifier of the annotation file where it is included with its token identifier (see an example in Footnote 26).

Second, in order to ease the recoverability of the input text, the `@join` standard attribute of token elements (value: `"left"`) is used to signal those cases in which no space separated two tokens (for example, within contractions).

Finally, a `wordForm` element is attached to each token[21], in order to annotate it and make the semantics of each of its tags explicit. This is achieved by means of (1) the `@lemma` attribute of wordforms and (2) a nested standard-compliant (ISO/FSR – ISO 24610-1:2006) feature structure annotation element (`fs`), which encapsulates the rest of its features[22].

4.2.2. Standardizing FreeLing's syntactic annotations

FreeLing provides both constituency-based and dependency-based syntactic annotations of its inputs, and both of them have already been standardized in FreeLing SWS. However, this section refers mainly to its dependency-based annotations for the sake of space.

The native, non-standardised FreeLing's dependency parsing of the Spanish sentence 'Mi gato se llama Tiger.' ('My cat's name is Tiger.') is shown in Example 3.

```
grup-verb/top/s1_t13 (llama llamar VMIP3S0)s1_nt13 top_[
s1_nt13 [
morfema-verbal/es/s1_t10 (se se P00CN000)
sn/subj/s1_t7 (gatogato NCMS000)s1_nt7 subj_[
      s1_nt7 [
espec-ms/espec/s1_t4 (Mi mi DP1CSS)
      ]]
sn/dobj/s1_t17 (Tiger tiger NP00000)
F-term/term/s1_t19 (. .Fp)
]]
```

Example 3: FreeLing's native dependency parsing of 'Mi gato se llama Tiger.' ('My cat's name is Tiger.')

[21] By means of the standard `@tokens` attribute.

[22] Note that wordforms are assigned their own persistent identifier (by means of the `@xml:id` attribute) as well.

As shown in this example, FreeLing's native dependency-based parser (A) uses its own parenthetical and non-semantically explicit notation to encode syntactic annotations; and (B) includes also a POS tagging of the input[23]. A graphical representation of this dependency parsing (a screenshot of FreeLing's online demo[24]) has been included in Figure 9 for clarity.

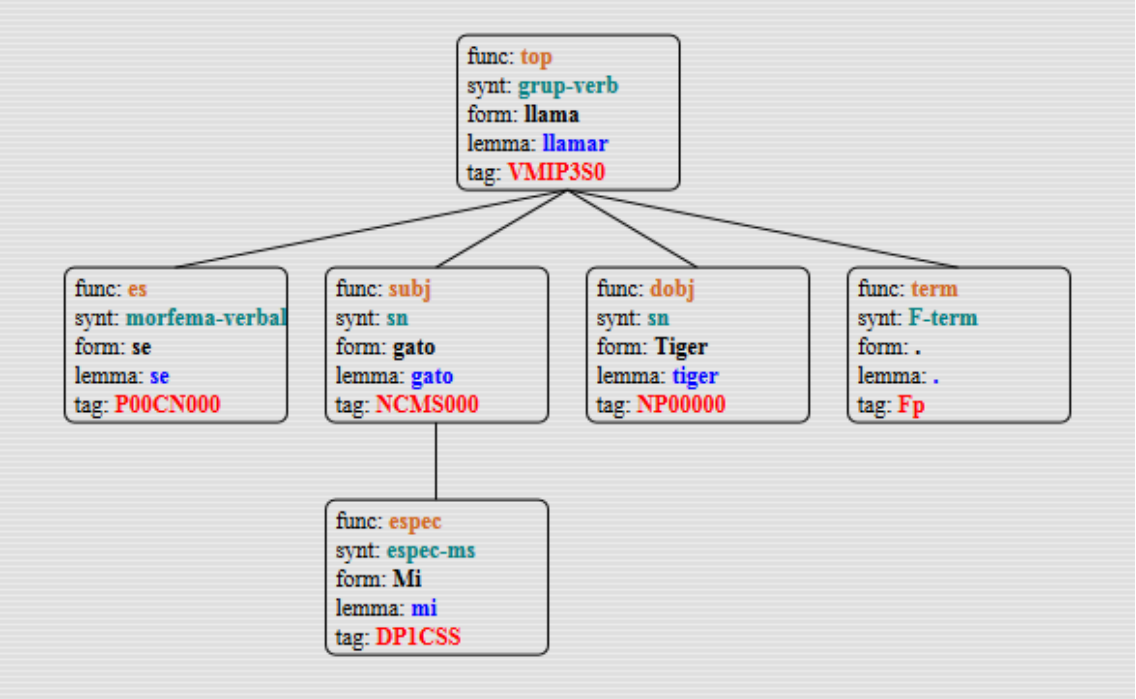

Figure 9: FreeLing's graphical dependency parsing of 'Mi gato se llama Tiger.' ('My cat's name is Tiger.')

On the one hand, (A) clearly complicates interpreting the annotations and making them interoperate; on the other hand, regarding (B), even though having both morpho-syntactic and syntactic annotations together helps making them interoperate, neither FreeLing's native dependency-based (or its constituency-based) annotations nor their nested morpho-syntactic annotations can be referenced from other annotations (e.g. sense tagging).For this reason, FreeLing's morpho-syntactic and syntactic annotations are standardised separately and are interlinked afterwards[25].

For this reason, the standardisation of FreeLing's morpho-syntactic annotations is supplemented by the standardisation of FreeLing's syntactic annotations according to and

[23] Both (A) and (B) hold also for FreeLing's native constituency-based annotations.
[24] http://nlp.lsi.upc.edu/freeling/demo/demo.php.
[25] Following the best practices and recommendations discussed in Pareja-Lora (2012).

complying with ISO/SynAF, and using the XML schema included in the ISO/SynAF-ISOTiger standard proposal (Bosch et al., 2012; 2014). This twofold (and separate) standardisation help test the interoperability of both ISO/MAF and ISO/SynAF compliant annotations.

The ISO/SynAF-compliant (and also ISO/SynAF-ISOTiger-compliant) XML annotation of the Spanish sentence 'Mi gato se llama Tiger.', obtained by means of FreeLing SWS, is shown in Example 4 (see next page).

Unfortunately, Example 4 cannot be fully described here for the sake of space; however, it is important to note that (1) the dependencies are represented by means of the `<edge>` elements attached to the terminal nodes (the `<t>` elements) and their standard `@tiger2:target` attribute; and (2) the terminal nodes refer to the morpho-syntactic `wordForm` elements by means of their standard `@tiger2:corresp` attribute[26] being assigned the PISA of the wordforms as value.

```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
<corpus    xsi:schemaLocation=  "http://korpling.german.hu-berlin.de/tiger2/
                                V2.0.5/
                                http://korpling.german.hu-berlin.de/tiger2/
                                V2.0.5/Tiger2.xsd"
           xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xmlns:tiger2="http://korpling.german.hu-berlin.de/tiger2/V2.0.5/"
           xmlns="http://korpling.german.hu-berlin.de/tiger2/V2.0.5/">
    <head>
        <meta>
          <name>http://quijote.fdi.ucm.es:8084/FreeLingWebService/2014-12-
          21.0:08:36./
                example_tiger_dep.standoff.tiger2.xml</name>
          <author>FreeLing SWS</author>
          <date>2014/03/21</date>
          <description>Tiger2 XML syntactic dependency annotations (SynAF-2
          compliant)</description>
          <format>FreeLing Tagset</format>
          <history>version:3.0</history>
        </meta>
        <annotation>
          <externalcorresp="http://quijote.fdi.ucm.es:8084/FreeLingWebService/
          tagsets/MyAnnotations.xml"/>
        </annotation>
    </head>
```

[26]For example, "spanish.example.MAF_dep.maf.xml#wordForm1"is a dereferenceable persistent identifier for the"wordForm1" ('Mi' – 'My') of the "spanish.example.MAF_dep.maf.xml" local file.

```
<body>
  <sxml:id="s1">
    <graphdiscontinuous="false" root="s1_Root">
      <terminals>
        <t tiger2:corresp="http://quijote.fdi.ucm.es:8084/FreeLingWeb
                           Service/2014-12-21.0:08:36./
                           spanish.example.MAF_dep.maf.xml#wordForm1"
                           xml:id="s1_t4" synt-type="espec-ms">
          <!--Mi-->
          <edgetiger2:target="s1_t7" label="espec"
          tiger2:type="prim.dep"/>
        </t>
        <t tiger2:corresp="http://quijote.fdi.ucm.es:8084/FreeLingWeb
                           Service/2014-12-21.0:08:36./
                           spanish.example.MAF_dep.maf.xml#wordForm2"
                           xml:id="s1_t7" synt-type="sn">
          <!--gato-->
          <edgetiger2:target="s1_t13" label="subj"
          tiger2:type="prim.dep"/>
        </t>
        <t tiger2:corresp="http://quijote.fdi.ucm.es:8084/FreeLingWeb
                           Service/2014-12-21.0:08:36./
                           spanish.example.MAF_dep.maf.xml#wordForm3"
                           xml:id="s1_t10" synt-type="morfema-verbal">
          <!-- se -->
          <edgetiger2:target="s1_t13" label="es" tiger2:type="prim.dep"/>
        </t>
        <t tiger2:corresp="http://quijote.fdi.ucm.es:8084/FreeLingWeb
                           Service/2014-12-21.0:08:36./
                           spanish.example.MAF_dep.maf.xml#wordForm4"
                           xml:id="s1_t13" synt-type="grup-verb">
          <!-- llama -->
          <edgetiger2:target="s1_Root" label="top"
          tiger2:type="prim.dep"/>
        </t>
        <t tiger2:corresp="http://quijote.fdi.ucm.es:8084/FreeLingWeb
                           Service/2014-12-21.0:08:36./
                           spanish.example.MAF_dep.maf.xml#wordForm5"
                           xml:id="s1_t17" synt-type="sn">
          <!-- Tiger -->
          <edgetiger2:target="s1_t13" label="dobj"
          tiger2:type="prim.dep"/>
        </t>
        <t tiger2:corresp="http://quijote.fdi.ucm.es:8084/FreeLingWeb
                           Service/2014-12-21.0:08:36./
                           spanish.example.MAF_dep.maf.xml#wordForm6"
                           xml:id="s1_t19" synt-type="F-term">
          <!-- . -->
          <edgetiger2:target="s1_t13" label="term"
          tiger2:type="prim.dep"/>
        </t>
      </terminals>
    </graph>
  </s>
</body>
```

Example 4: FreeLing SWS's SynAF- and SynAF-ISOTiger-compliant annotation of 'Mi gato se llama Tiger.' ('My cat's name is Tiger.')

5. Discussion – Pros and cons of standards for terminological and linguistic works

The previous section has already shown that using some ISO/TC 37/SC 4 standards can help solve several interoperability limitations that linguistic annotation tools (such as FreeLing) and their annotations have, namely

- their low reusability 'as is' in NLP pipelines;
- the lack of semantic explicitness and (inter-)referenceability of its annotations.

Besides, it has been shown that using these ISO/TC 37/SC 4 standards to standardize separately different annotations (in this case, morpho-syntactic and syntactic annotations) helps interlink them successfully and fairly straightforwardly afterwards and make them interoperate.

This is not surprising, since ISO/TC 37/SC 3 and ISO/TC 37/SC 4 (which develop standards that are traversal to the whole ISO/TC 37) are paying special attention to reusability and interoperability issues, and to ensure that the application of the standards they are responsible for produces reusable and interoperable results. For this reason, using ISO/TC 37 standard-compliant resources are usually reusable and interoperable, which is a main concern for e.g. the terminology and the language resources communities.

Using ISO/TC 37 standards can also generate some of the general benefits stated in Section 1, such as improving operational efficiency and reducing costs, or facilitating international exchange of goods and services. However, in order for some of these benefits to be achieved in terminology, language and/or linguistic works and initiatives, they should incorporate standards as soon as possible. Indeed, devising and/or implementing similar standardisation processes to the one presented in Section 4 for FreeLing 3.0 can be high time-consuming tasks. Thus, for instance, developing the standardisation processes required to standardise pre-existing resources and previous resultscan limit the expected reduction of costs temporarily (but not permanently).

In spite of all the benefits stated above, a particular standardmay not be absolutely flawlessand, hence, its application may not produce some of the results stated above. In fact, the amount and the extent of the benefits of using a standard depend directly on its quality and, in turn, the quality of the standard depends on a number of factors, mainly

a) The number of experts participating in its development. On the one hand, as commented in Section 2, experts and/or stakeholders participate in the development of ISO standards on a voluntary basis and for free. However, also standard development is quite a high-consuming task, and not all experts in the area can spend the amount of time required by this task. On the other hand, some relevant experts and/or stakeholder representatives do not master the

English language (the *de facto* language in ISO meetings and ballots) and frequently refuse to read, comment and discuss the standard drafts in their national committees. In both cases, they often decline the invitation to get involved in the process and, consequently, the opinion of relevant experts and/or stakeholders in the field is not integrated in the standard, thus reducing its quality.

b) The degree of disparity in the points of views of the experts that constitute the working group responsible for the development of the standard, their flexibility and their capacity for reaching consensus. Unfortunately, but also fairly logically, some experts (e.g. those representing companies and/or some stubborn academic experts) usually tend to favour their own products, services, theories, ideas and opinions in the standardization process and want to see them conveniently depicted in the final version of the standard. In these situations, as well as when other vested interests get in the way or when the various points of views differ to a great extent, consensus is difficult to reach. This is quite common in terminology, language and linguistic related scenarios, where usually different (and sometimes conflicting) theories and/or approaches come into play. In the end, this can result in the cancellation of the standard development process or in a production of a too fuzzy or not sufficiently detailed standard, which is then useless.

Nevertheless, ISO published standards can be reviewed anytime, and a particular ballot to identify the need for review of a standard is issued five years after its publication or last review (as a maximum). The standard can then be revised and/or amended, if required. Thus, luckily, the quality of an international standard can improve over time.

6. Conclusions

This paper has presented (a) some benefits of using standards in general; (b) the main standardisation organization (that is, ISO) and some relevant figures about its committees and standards; (c) ISO/TC 37, the "ISO Technical Committee for the Standardization of Terminology and other language and content resources", its structure and some relevant figures; (d) the standardisation of FreeLing 3.0 annotations by complying with some standards (being) developed within ISO/TC 37, namely ISO/MAF (ISO 24611:2012), ISO/SynAF (ISO 24615:2010) and ISO/SynAF-ISOTiger (Bosch et al., 2012; 2014); and (e) a discussion about the pros and cons of using standards in general and ISO/TC 37 standards in particular.

Specifically, this paper has shown that using some ISO/TC 37 standards can help increase the interoperability of linguistic annotation tools and their results (i.e., their annotations), which increases their reusability as well. It has shown also the main benefits that using standards can yield to (i) companies (reducing costs, increasing efficiency and effectiveness, creating business opportunities, etc.), (ii) consumers (e.g. better quality of goods and services), and/or (iii) public authorities (like helping develop and promote efficiently measures of public utility – for example, on safety, security, and protection of the environment). This paper has also mentioned some of the problems that may arise in standard development processes and how they can limit the eventual quality of standards. Further work includes, for example, finding out to what extent each of these problems contribute actually to the quality of standards and/or how they could be solved.

7. Acknowledgements

The research described in this paper has been partly funded by the Spanish Ministry of Science and Innovation, **Grant FFI2011-29829**: *Entorno de aprendizaje móvil y social de lenguas cognitivamente aumentado y basado en una ontología* (**SO-CALL-ME**)[27].

I would like to thank Guillermo Cárcamo-Escorza, Alicia Ballesteros-Calvo and Emilio Duobert-Collazos for their invaluable help in the development of the standard-compliant, web service-based version of FreeLing (FreeLing SWS).

8. References

Ballesteros-Calvo, A., **Cárcamo-Escorza**, G., and **Duobert-Collazos**, E. *Recubrimiento y normalización de recursos lingüísticos: aplicación a las anotaciones morfosintácticas y sintácticas de FreeLing (B.Sc. Degree).* Madrid: Facultad de Informática, Universidad Complutense de Madrid, 2013.

Bosch, S., **Choi**, K.S., **de La Clergerie**, É., V. **Fang**, A. C., **Faaβ**, G., **Lee**, K., **Pareja-Lora**, A., **Romary**, L., **Witt**, A., **Zeldes**, A., and **Zipser**, F. "<tiger2/> as a Standardised Serialisation for ISO 24615 – SynAF." *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*.Lisboa: Ediçôes Colibri, 2012, 37-60.

Bosch, S., **Eckart**, K., **Faaβ**, G., **Heid**, U., **Lee**, K., **Pareja-Lora**, A., **Pretorius**, L., **Romary**, L., **Witt**, A., **Zeldes**, A., and **Zipser**, F. "From <tiger2/> to ISOTiger – Community Driven Developments for Syntax Annotation in SynAF." *Proceedings of the Thirteenth International*

[27] A Social and Ontology-based framework for Cognitively-Augmented Language Learning in Mobile Environments (SO-CALL-ME).

*Workshop on Treebanks and Linguistic Theories (TLT13).*Tübingen, Germany: Department of Linguistics (SfS), University of Tübingen. 2014, 258 – 264.

Buyko, E., **Chiarcos**, Ch., and **Pareja-Lora**, A. "Ontology-Based Interface Specifications for an NLP Pipeline Architecture".*Proceedings of the 6th International Language Resources and Evaluation Conference (LREC'08).* Paris: European Language Resources Association (ELRA), 2008, 847 – 854.

Chinchor, N.*MUC-7 Named entity Task Definition Version 3.5.* 1997. 23/10/2013. http://itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html.

EAGLES Consortium.*EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora. EUROPEAN PROJECT DELIVERABLE: EAGLES Document EAG–TCWG–MAC/R,* 1996.

Fellbaum, Ch. (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

Ide, N., and **Suderman**, K. "GrAF: A Graph-based Format for Linguistic Annotations".*Proceedings of the (First) Linguistic Annotation Workshop,* held in conjunction with ACL'2007. Prague, Czech Republic, 1-8, June, 2007.

International Organization for Standardization. ISO 24610-1:2006,*Language resource management – Feature structures – Part 1: Feature structure representation.* Geneva: International Organization for Standardization, 2006.

International Organization for Standardization. ISO 24615:2010,*Language resource management – Syntactic annotation framework (SynAF)*. Geneva: International Organization for Standardization, 2010.

International Organization for Standardization. ISO 24619:2011,*Language resource management – Persistent identification and sustainable access (PISA).*Geneva: International Organization for Standardization, 2011.

International Organization for Standardization. ISO 24611:2012,*Language resource management – Morpho-syntactic annotation framework (MAF)*. Geneva: International Organization for Standardization, 2012.

International Organization for Standardization. ISO 24612:2012,*Language resource management – Linguistic annotation framework (LAF)*. Geneva: International Organization for Standardization, 2012.

International Organization for Standardization. *ISO Membership Manual*. Geneva: International Organization for Standardization, 2013.

International Organization for Standardization. *ISO Statutes*.17th ed. Geneva: International Organization for Standardization, 2013.

ISO Central Secretariat.International Organization for Standardization. *Economic Benefits of Standards*.Geneva: International Organization for Standardization, 2014.

ISO/TC 37.*ISO/TC 37 Business Plan*. Geneva: International Organization for Standardization, 2013.

Kashyap, V., **Bussler**, Ch., and **Moran**, M. *The Semantic Web*. Berlin: Springer, 2008.

König, E., and **Lezius**, W.*The TIGER language - A Description Language for Syntax Graphs, Formal Definition (Technical report IMS)*.Stuttgart: Universität Stuttgart, 2003.

Miller, G. A. "WordNet: A Lexical Database for English". *Communications of the ACM* 38.11, 1995, 39-41.

Morell, C., **Vivaldi**, J., and **Bel**, N. "Iula2Standoff: a tool for creating standoff documents for the IULACT". Calzolari, N. (et al.) (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Paris: European Language Resources Association (ELRA), 2012.

Padró, Ll, and **Stanilovsky**, E. "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*. Paris: European Language Resources Association (ELRA), 2012.

Pareja-Lora, A. *Providing Linked Linguistic and Semantic Web Annotations: The OntoTag Hybrid Annotation Model*.Saarbrücken: LAP - LAMBERT Academic Publishing, 2012.

Poch, M., and **Bel**, N. "Interoperability and technology for a language resources factory".*Proceedings of the IJCNLP 2011 Workshop on Language Resources, Technology and Services in the Sharing Paradigm*. ChiangMai, Thailand: IJCNLP, November, 2011.

Pozzi, M. "El español en el contexto de la normalización terminológica internacional". *Actas del III Congreso «El Español, Lengua de Traducción» Contacto y contagio.* 2006, 155-204.

Vossen, P. (ed.). *EuroWordNet: a multilingual database with lexical semantic networks*, Norwell, MA: Kluwer Academic Publishers, 1998.