NORMALIZACIÓN DE LAS ANOTACIONES MORFOSINTÁCTICAS Y SINTÁCTICAS DEL ESPAÑOL: EL CASO DE FREELING

ANTONIO PAREJA LORA
Universidad Complutense de Madrid / ATLAS (UNED)
GUILLERMO CÁRCAMO ESCORZA
Universidad Complutense de Madrid
EMILIO DUOBERT COLLAZOS
Universidad Complutense de Madrid

RESUMEN

El grado de interoperabilidad de las herramientas lingüísticas de anotación y sus resultados (las anotaciones lingüísticas) es, por lo general, realmente bajo. Este pobre nivel de interoperabilidad dificulta enormemente su reutilización, por ejemplo, en arquitecturas complejas ('pipelines') de anotación del lenguaje natural.

Desarrollar de nuevo desde cero estas herramientas para hacerlas más interoperables no es factible (por el alto coste de este tipo de desarrollos). Por ello, hay que encontrar urgentemente métodos eficientes para su reutilización sistemática y para hacer que tanto estas como sus anotaciones sean más interoperables.

La respuesta a este problema ha venido casi siempre de la mano de la normalización. En este artículo presentamos una versión de FreeLing que transforma esta herramienta de anotación lingüística en un servicio web y proporciona anotaciones morfosintácticas y sintácticas normalizadas para el español de acuerdo con las normas y propuestas de normas ISO correspondientes.

Palabras clave: Anotación morfosintáctica, anotación sintáctica, normalización, ISO, MAF, SynAF, servicio web, interoperabilidad, FreeLing

ABSTRACT

Linguistic annotation tools and linguistic annotations are scarcely interoperable. This low level of interoperability makes it difficult to reuse both linguistic annotation tools and their annotations in new scenarios, e.g., in natural language processing (NLP) pipelines.

Developing new linguistic tools from scratch is quite a high time-consuming task that also entails a very high cost. Therefore, cost-effective ways to systematically reuse linguistic tools and annotations and make them more interoperable must be found urgently.

A traditional way to overcome reuse and/or interoperability problems is standardisation. In this paper, we present a web service version of FreeLing that provides standard-compliant morpho-syntactic and syntactic annotations for Spanish, according to several ISO linguistic annotation standards and standard drafts.

Keywords: POS tagging, syntactic annotation, standardisation, ISO, MAF, SynAF, web service, interoperability, FreeLing.

1. INTRODUCCIÓN

Muchos son los conflictos y los problemas que impiden la interoperación de las herramientas de anotación lingüística y sus anotaciones (Pareja-Lora, 2012), así como su reutilización inmediata, por ejemplo, en cadenas de componentes para el procesamiento del lenguaje natural (Buyko et al., 2008). A pesar de ello, desarrollar de nuevo desde cero todas estas herramientas es impensable, dado el alto coste que conlleva este tipo de proyectos. Por tanto, es preciso reutilizar las herramientas lingüísticas ya existentes y encontrar modos económicos y eficientes de interoperación tanto de éstas como de sus anotaciones, para dar resultados interconectados y más avanzados.

En este aspecto, la normalización puede ser una gran ayuda, como ya lo viene siendo, por ejemplo, en la (inter)conexión e interoperación de equipos, dispositivos y componentes eléctricos. Esta es la hipótesis de trabajo principal que impulsa el desarrollo de las distintas normas y propuestas de normas de anotación lingüística de la Organización Internacional para la Normalización (ISO), tales como ISO/MAF (2012), para la anotación morfosintáctica, o ISO/SynAF (2010) e ISO/SynAF–<tiger2/> (Bosch et al., 2012), para la anotación sintáctica.

Esta es asimismo la hipótesis que ha guiado el presente trabajo: que la normalización puede ayudar a que las herramientas y las anotaciones lingüísticas interoperen (Ballesteros-Calvo et al., 2013). En particular, este trabajo intenta demostrar, usando la herramienta de anotación FreeLing, versión 3.0 (Padró y Stanilovsky, 2012) y un subconjunto de sus anotaciones, que la normalización de:

a) el modo de operación de una herramienta de anotación lingüística (en este caso, su transformación en un servicio web, objeto de la recomendación W3C/WS-Arch (2004), del *World Wide Web Consortium*),
b) sus anotaciones morfosintácticas y sintácticas para el español, según las [propuestas de] normas ISO/MAF (2012), ISO/SynAF (2010) e ISO/SynAF–<tiger2/> (Bosch et al., 2012),

puede ayudar a enlazar estas anotaciones entre sí y facilitar su interoperación con otras anotaciones.

2. POR QUÉ FREELING

FreeLing (Padró & Stanilovsky, 2012) es una herramienta de código abierto para el análisis y la anotación de textos en varios niveles y lenguas (español, catalán, gallego, portugués, italiano, francés, inglés o ruso, entre otros). La versión normalizada que aquí se presenta (*FreeLing SWS*, en adelante, disponible en la URL http://147.96.80.70:8084/ClienteFreeLing/index.jsp) está basada en su versión 3.0. FreeLing 3.0 procesa el texto de entrada y proporciona, por ejemplo, (i) su división en componentes léxicos (*tokens*), (ii) su etiquetado gramatical, (iii) su análisis sintáctico basado en relaciones de constitución y/o de dependencia, (iv) su anotación de entidades con nombre, de acuerdo con la clasificación MUC (Chinchor, 1997), o (v) su anotación del sentido, basada en (Euro)WordNet (Fellbaum, 1998; Vossen, 1998).

Pero, pese a su gran versatilidad (y popularidad), FreeLing tiene aún ciertas limitaciones. Por un lado, su implementación actual no es suficientemente interoperable: por ejemplo, su inclusión junto con otras herramientas de anotación en una cadena (*annotation pipeline*) más compleja, requiere realizar ciertas tareas previas de instalación y programación no necesariamente triviales. Por otro lado, el formato de sus anotaciones no está normalizado: no se generan conforme a un esquema de anotación normalizado, como ISO/MAF (2012) o ISO/SynAF (2010), ni con un lenguaje normalizado, como XML.

3. NORMALIZANDO FREELING

La primera limitación mencionada más arriba se salvó encapsulando FreeLing 3.0 en un servicio web, de acuerdo con las recomendaciones WSDL y SOAP, del *World Wide Web Consortium* (W3C).[1] La principal ventaja de un servicio web (Kashyap et al., 2008) es que proporciona una interfaz de acceso y de ejecución normalizada (y, por tanto, altamente interoperable) para una cierta aplicación informática.

La segunda limitación, que impedía la interconexión y la integración (es decir, la interoperación) de las anotaciones de FreeLing con otras anotaciones, se resolvió aplicando otras tantas

normas ISO de anotación lingüística en la generación de los resultados del nuevo servicio web (FreeLing SWS), tal y como se describe en las siguientes subsecciones.

3.1 FreeLing SWS: Normalización de anotaciones morfosintácticas

El Ejemplo 1 muestra el etiquetado gramatical desambiguado que proporciona FreeLing 3.0 para la entrada 'Mi gato se llama Tiger.'. Como puede verse en el ejemplo, este etiquetado consta de tres campos (o etiquetas) por cada unidad morfosintáctica (*token*): (1) el texto de entrada; (2) su lema; y (3) su etiqueta gramatical, basada en el formato EAGLES (1996), que incluye la anotación de la categoría gramatical y de los atributos morfológicos de la unidad morfosintáctica. También puede obtenerse, en campos adicionales, una anotación de su sentido, basada en (Euro)WordNet (no incluida aquí por razones de espacio).

```
Nos     nos      PP1CP000
vamos   ir       VMIP1P0
a       a        SPS00
el      el       DA0MS0
parque  parque   NCMS000
.       .        Fp
```

Ejemplo 1: Etiquetado gramatical original, con FreeLing 3.0, de 'Nos vamos al parque.'

La gran desventaja de este formato de anotación es que no menciona explícitamente el significado de sus distintas etiquetas. En consecuencia, su semántica queda implícita. Esto complica en gran medida (1) la interpretación automática de estas etiquetas; (2) su comparación con otros etiquetados gramaticales, obtenidos con otras herramientas; y (3) en general, su interoperación con otras anotaciones (Pareja-Lora, 2012). Además, los *tokens* resultantes y sus correspondientes anotaciones no pueden ser reutilizados, por ejemplo, en la anotación sintáctica de esta misma frase, pues no se proporciona una manera de referirse a ellos desde otras anotaciones (usando *URIs*, por ejemplo).[2]

Estos problemas se solucionaron parcialmente mediante la normalización de las anotaciones morfosintácticas (esto es, del

etiquetado gramatical) de FreeLing. Esta normalización se llevó a cabo de acuerdo con la norma ISO/MAF (2012), que proporciona un marco para la anotación de unidades morfosintácticas con su categoría gramatical correspondiente y sus rasgos morfológicos. También recomienda el uso de un esquema XML particularizado, que posibilita su referencia desde otras anotaciones y las hace así más interoperables. Un extracto de la anotación normalizada resultante para la frase del Ejemplo 1 se ha incluido en el Ejemplo 2.

```
<?xml version="1.0" encoding="UTF-8"?>
<maf>
     <token xml:id="t1">Nos</token>
     <token xml:id="t2">vamos</token>
     <token xml:id="t3">a</token>
     <token xml:id="t4" join="left">l</token>
     <token xml:id="t5">parque</token>
     <token xml:id="t6" join="left">.</token>

     <wordForm xml:id="wordForm2" tokens="#t1"
     lemma="nos">
         <fs>
             <f name="pos">
                 <symbol value="PP1CP000"/>
                 <!--Pronombre: Personal, Primera persona,
                 Común, Plural -->
             </f>
         </fs>
     </wordForm>
     <wordForm xml:id="wordForm3" tokens="#t2"
     lemma="ir">
                              ..............
     </wordForm>
</maf>
```

Ejemplo 2: Etiquetado gramatical original, con FreeLing 3.0, de 'Nos vamos al parque.'

Como se ve en el Ejemplo 2, en primer lugar, a cada `token` se le asigna, mediante el atributo `@xml:id`, su correspondiente *URI* dentro del fichero de anotaciones morfosintácticas, que se convierte de esta manera en su propio identificador persistente (PISA, cf. ISO/PISA (2011)). Esto ayuda a construir otras anotaciones a partir de ellos y a

enlazarlas entre sí, tal y como ocurre, por ejemplo, con las anotaciones de las formas de la entrada (`wordForm`), que se refieren a las unidades morfosintácticas que incluyen mediante su PISA local.

En segundo lugar, se usa el atributo normalizado `@join`, con el valor `"left"` (izquierda) para marcar aquellos casos en lo que no hay un espacio de separación entre dos *tokens* (como es el caso de 'a' y 'el' en la contracción 'al', que además se unen después en un mismo `wordForm`). Esto facilita la recuperación del texto de entrada.

Finalmente, los elementos `wordForm` permiten asociar varias unidades morfosintácticas mediante el atributo normalizado `@tokens`, y/o anotarlas con etiquetas cuya semántica está normalizada y queda así explícita. Esto se consigue con (A) el atributo `@lemma`; y (B) su anotación, de acuerdo con la norma ISO/FSR (2006), mediante estructuras de rasgos anidadas (elementos `<fs>`) que encapsulan de manera adecuada la anotación de los rasgos morfológicos (elementos `<f>`) de la forma de entrada. Nótese que también los elementos `wordForm` son designados mediante un identificador (PISA) propio, pudiendo ser referenciados así desde otras anotaciones.

3.2 FreeLing SWS: Normalización de anotaciones sintácticas

De manera similar, las anotaciones sintácticas de FreeLing han sido normalizadas de acuerdo con la norma ISO/SynAF (2010), y usando el esquema de anotación en XML contenido en la propuesta de norma ISO/SynAF-<tiger2/> (Bosch et al., 2012). El Ejemplo 3 incluye un fragmento de la anotación normalizada con dependencias sintácticas de la frase 'Nos vamos al parque.'.

En este ejemplo, (i) las dependencias sintácticas se representan mediante los elementos `<edge>` que se asocian a los nodos terminales (`<t>`) y su atributo normalizado, `@tiger2:target`; y (ii) los nodos terminales referencian los elementos `<wordForm>` de la anotación morfosintáctica (véase el Ejemplo 2) asignando el valor del PISA del `wordForm` en cuestión (definido en el fichero de anotaciones morfosintácticas) al atributo normalizado `@tiger2:corresp`.

```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
<corpus xsi:schemaLocation="http://korpling.german.hu-
berlin.de/tiger2/V2.0.5/ […]"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:tiger2="http://korpling.german.hu-
berlin.de/tiger2/V2.0.5/" […]>
<head> …… </head>

<body>
  <s xml:id="s1">
    <graph discontinuous="false" root="s1_Root">
      <terminals>
        <t
        tiger2:corresp="http://quijote.fdi.ucm.es:8084
        /FreeLingWebService/2014-14-31.11:59:35./
        spanish.example.MAF_dep.maf.xml#wordForm1"
        xml:id="s1_t4" synt-type="patons">
        <!-- Nos -->
            <edge tiger2:target="s1_t7" label="dep"
            tiger2:type="prim.dep"/>
        </t>

        <t
        tiger2:corresp="http://quijote.fdi.ucm.es:8084
        /FreeLingWebService/2014-14-31.11:59:35./
        spanish.example.MAF_dep.maf.xml#wordForm2"
        xml:id="s1_t7" synt-type="grup-verb">
        <!-- vamos -->
            <edge tiger2:target="s1_Root" label="top"
            tiger2:type="prim.dep"/>
        </t>
                                ……
      </terminals>
    </graph>
  </s>
</body>
</corpus>
```

Ejemplo 3: Anotación normalizada con dependencias sintácticas de 'Nos vamos al parque.'.

Ni este ejemplo ni la normalización de las anotaciones sintácticas con relaciones de constitución se analizan aquí en detalle por falta de espacio.

4. CONCLUSIONES

En este artículo se ha presentado la encapsulación de la herramienta de anotación FreeLing 3.0 en un servicio web normalizado (FreeLing SWS). Esta encapsulación ha ayudado a resolver varios problemas de interoperabilidad que presentan esta herramienta y sus anotaciones sintácticas y morfosintácticas: por ejemplo, (1) su dificultad para integrarse tal cual en cadenas de anotación más complejas; (2) la ausencia de una semántica explícita en sus etiquetados; o (3) la imposibilidad de referenciar sus anotaciones desde otras cualesquiera (véase la Sección 3).

En particular, esta normalización ha permitido la interconexión explícita, efectiva y casi inmediata de las anotaciones morfosintácticas de FreeLing con sus anotaciones sintácticas, sean basadas en relaciones de constitución o de dependencias.

No obstante, aún no se ha conseguido enlazar entre sí estos dos tipos de anotaciones sintácticas, a pesar de que ambas son proporcionadas por FreeLing y están contempladas de forma conveniente en la norma ISO/SynAF (2010). Ello se debe a que estas dos anotaciones son generadas por módulos independientes de FreeLing, que no se comunican entre sí. El proceso de integración de las anotaciones sintácticas requiere un estudio más profundo para su automatización y se abordará en el futuro.

Asimismo, queda pendiente la realización de comprobaciones adicionales del grado real de interoperabilidad del servicio web desarrollado. Estas pruebas incluirán la integración del mismo y de la herramienta original (FreeLing 3.0) en una cadena de anotación más amplia. Esto permitirá, por ejemplo, dar una medida comparada y realista del coste de integración de ambas versiones y de la interoperabilidad de sus resultados.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación, en el contexto del proyecto **SO-CALL-ME** (ref. **FFI2011-29829**): Entorno de aprendizaje móvil y social de lenguas cognitivamente aumentado y basado en una ontología.

También queremos dar las gracias a Alicia Ballesteros Calvo por su inestimable ayuda en el desarrollo de esta versión normalizada de FreeLing 3.0 y al grupo de investigación ATLAS, de la UNED, por su constante inspiración, aliento y apoyo.

NOTAS

[1] WSDL: http://www.w3.org/TR/wsdl; SOAP: http://www.w3.org/TR/soap12-part0/.

[2] *URI: Uniform Resource Identifier*, http://www.w3.org/TR/uri-clarification/.

REFERENCIAS BIBLIOGRÁFICAS

Ballesteros Calvo, A., Cárcamo Escorza, G., Duobert Collazos, E. J. 2013. *Recubrimiento y normalización de recursos lingüísticos: aplicación a las anotaciones morfosintácticas y sintácticas de FreeLing*. Trabajo de Fin de Grado. Facultad de Informática, Universidad Complutense de Madrid. Madrid, Spain.

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

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

EAGLES Consortium. 1996. *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora. EUROPEAN PROJECT DELIVERABLE: EAGLES Document EAG--TCWG—MAC/R*. [http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html (visitado el 17/01/2012)].

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

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

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

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

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

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

Kashyap, V., Bussler, C. & Moran, M. 2008. *The Semantic Web.* Berlin: Springer.

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

Padró, L., Stanilovsky, E. 2012. "FreeLing 3.0: Towards Wider Multilinguality". *Actas de 8th Language Resources and Evaluation Conference (LREC 2012)*. ELRA. Estambul, Turquía. Mayo, 2012.

Pareja-Lora, A. 2012. *Providing Linked Linguistic and Semantic Web Annotations: The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP - LAMBERT Academic Publishing, pp. 1-500.

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