Volume 2, Número 3 - Dezembro 2010

linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 2, Número 3 – Dezembro 2010

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Con este quinto número de Linguamática completamos o segundo ano da revista, cun total de 26 contribucións e 4 artigos convidados. Neste breve período de existencia, acadamos a indexación nalgunhas bases de datos ben relevantes, como Latindex, o Directory of Open Access Journals (DOAJ) ou o Google Scholar. Desde hai unhas semanas, todos os artigos publicados por Linguamática son tamén enlazados, resumidos e indexados tematicamente na importante base de datos bibliográfica coñecida como LLBA (Linguistic and Language Behavior Abstracts).

A partir deste número, imos incluír tamén a petición de artigos en lingua vasca grazas á colaboración desinteresada de Zuriñe Folgado, a quen lle queremos agradecer aquí publicamente o seu meritorio labor de tradución. Esperamos que a incorporación desta lingua na petición de artigos favoreza a presentación de orixinais en euskara para publicación na revista.

Desexamos, por último, agradecer o seu interese a todas as persoas que enviaron propostas de artigos para este número de Linguamática, tanto se foron publicadas como se non o foron, e o traballo de revisión fundamental para a revista de todos os membros regulares e convidados do Comité Científico.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís, Universidade de Vigo
Alberto Simões, Universidade do Minho
Aline Villavicencio, Universidade Federal do Rio Grande do Sul
Álvaro Iriarte Sanroman, Universidade do Minho
Ana Frankenberg-Garcia, ISLA e Universidade Nova de Lisboa
Anselmo Peñas, Universidad Nacional de Educación a Distancia
Antón Santamarina, Universidade de Santiago de Compostela
António Teixeira, Universidade de Aveiro
Belinda Maia, Universidade do Porto
Carmen García Mateo, Universidade de Vigo
Diana Santos, SINTEF ICT
Ferran Pla, Universitat Politècnica de València
Gael Harry Dias, Universidade Beira Interior
Gerardo Sierra, Universidad Nacional Autónoma de México
German Rigau, Euskal Herriko Unibertsitatea
Helena de Medeiros Caseli, Universidade Federal de São Carlos
Horacio Saggion, University of Sheffield
Iñaki Alegria, Euskal Herriko Unibertsitatea
Joaquim Llisterri, Universitat Autònoma de Barcelona
José Carlos Medeiros, Porto Editora
José João Almeida, Universidade do Minho
José Paulo Leal, Universidade do Porto
Joseba Abaitua, Universidad de Deusto
Lluís Padró, Universitat Politècnica de Catalunya
Maria das Graças Volpe Nunes, Universidade de São Paulo
Mercè Lorente Casafont, Universitat Pompeu Fabra
Mikel Forcada, Universitat d'Alacant
Pablo Gamallo Otero, Universidade de Santiago de Compostela
Salvador Climent Roca, Universitat Oberta de Catalunya
Susana Afonso Cavadas, University of Sheffield
Tony Berber Sardinha, Pontifícia Universidade Católica de São Paulo
Xavier Gómez Guinovart, Universidade de Vigo

Revisores Convidados

Liliana Ferreira, Universidade de Aveiro
Marcos Garcia, Universidade de Santiago de Compostela
Patrícia França, Universidade do Minho

Artigos de Investigação

La compresión de frases: un recurso para la optimización de resumen automático de documentos

Alejandro Molina
LIA-Université d'Avignon y
GIL-Instituto de Ingeniería UNAM
alejandro.molina@etd.univ-avignon.fr

Iria da Cunha
IULA-Universitat Pompeu Fabra,
LIA-Université d'Avignon y
GIL-Instituto de Ingeniería UNAM
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno
LIA-Université d'Avignon,
École Polytechnique de Montréal y
GIL-Instituto de Ingeniería UNAM
juan-manuel.torres@univ-avignon.fr

Patricia Velázquez-Morales
VM Labs
patricia_velazquez@yahoo.com

Resumen

El objetivo de este trabajo de investigación es confirmar si es adecuado emplear la compresión de frases como recurso para la optimización de sistemas de resumen automático de documentos. Para ello, en primer lugar, creamos un corpus de resúmenes de documentos especializados (artículos médicos) producidos por diversos sistemas de resumen automático. Posteriormente realizamos dos tipos de compresiones de estos resúmenes. Por un lado, llevamos a cabo una compresión manual, siguiendo dos estrategias: la compresión mediante la eliminación intuitiva de algunos elementos de la oración y la compresión mediante la eliminación de ciertos elementos discursivos en el marco de la *Rhetorical Structure Theory* (RST). Por otro lado, realizamos una compresión automática por medio de varias estrategias, basadas en la eliminación de palabras de ciertas categorías gramaticales (adjetivos y adverbios) y una *baseline* de eliminación aleatoria de palabras. Finalmente, comparamos los resúmenes originales con los resúmenes comprimidos, mediante el sistema de evaluación ROUGE. Los resultados muestran que, en ciertas condiciones, utilizar la compresión de frases puede ser beneficioso para mejorar el resumen automático de documentos.

1. Introducción

La compresión de frases es un tema de investigación relativamente reciente. Los métodos sobre compresión de frases están orientados a la eliminación de la información no esencial de las frases de un documento, manteniendo al mismo tiempo su gramaticalidad. Las aplicaciones de la compresión de frases pueden ser muy diversas.

Un ejemplo de ello es la generación automática de títulos. Las agencias de noticias reciben diariamente una gran cantidad de información proveniente de fuentes heterogéneas. Estas agencias cuentan con especialistas encargados de asignar un título a cada una de las informaciones que les llegan y que serán posteriormente convertidas en noticias. (Mittal and Witbrock, 1999) presentan un sistema capaz de generar encabezados de tamaño arbitrario.

Otra aplicación es la generación de subtítulos para medios audiovisuales. Hoy en día, la mayor parte de las películas cuentan con subtítulos, pero la mayoría de las cadenas de televisión todavía ofrecen el subtitulado de manera limitada. Sin embargo, en los últimos años, este tema ha suscitado un gran interés, recibiendo una atención especial. Por un lado, los subtítulos pueden traducir una narración o diálogo que se realiza en un idioma extranjero y, por otro, pueden servir para ayudar a las personas con problemas visuales a recibir la información. (Grefenstette, 1998) presenta un método de reducción de textos que tiene por objetivo disminuir el tiempo de lectura de un sintetizador para ciegos.

Otra de las aplicaciones de la compresión de frases tiene que ver con la telefonía móvil. Actualmente, los dispositivos móviles cuentan con pantallas reducidas donde el número de caracteres mostrados es limitado. La compresión de frases es un método que permitiría reducir la extensión del texto mostrado y, de esta manera, incluir más información en un espacio determinado.

En otra línea de investigación, la compresión de frases podría servir como método para la op-

timización de los sistemas de resumen automático de documentos. El resumen automático es un tema de investigación muy relevante desde hace años y se han realizado estudios para diversos idiomas como el inglés (Marcu, 2000a; Teufel and Moens, 2002), el francés (Torres-Moreno, Velázquez-Morales, and Meunier, 2002; Boudin and Torres-Moreno, 2009), el español (da Cunha and Wanner, 2005; Mateo et al., 2003), el portugués (Salgueiro Pardo and Rino Machado, 2001) y el catalán (Fuentes, González, and Rodríguez, 2004); así como estudios multilingues (Lenci et al., 2002). Recientemente existen estudios sobre resumen de textos especializados en medicina (Afantenos, Karkaletsis, and Stamatopoulos, 2005; da Cunha, Wanner, and Cabré, 2007; Vivaldi et al., 2010), química (Pollock and Zamora, 1975; Boudin, Torres-Moreno, and Velazquez-Morales, 2008; Boudin, Torres-Moreno, and El-Bèze, 2008) y derecho (Farzindar, Lapalme, and Desclés, 2004), e incluso sistemas de resumen de sitios Web (Berger and Mittal, 2000).

Los sistemas de resumen automático, por lo general, siguen el paradigma de la extracción (Edmundson, 1969; Lal and Ruger, 2002), incluyendo las oraciones más relevantes del texto de manera literal. Regenerar automáticamente el texto extraído para crear un resumen por abstracción es sumamente complicado pues se deben incluir los contenidos más relevantes del texto original, pero redactados de manera diferente (Ono, Sumita, and Miike, 1994; Paice, 1990). La compresión de frases puede ser un vínculo en el camino de la extracción a la abstracción, es decir, una forma primaria de paráfrasis. Si partimos de la hipótesis de que, para determinadas tareas, un resumen posee una extensión limitada (como es el caso de los resúmenes de noticias), la compresión de frases conservando su gramaticalidad podría permitir una mayor cantidad de información en el mismo espacio. De confirmarse esta hipótesis, podría emplearse la compresión de frases como recurso para la optimización de sistemas de resumen automático de documentos. El objetivo de este trabajo es precisamente confirmar esta hipótesis.

Como antecedente directo podemos considerar el trabajo de (Lin, 2003), en el cual se comprimen las frases de un sistema de resumen extractivo multi-documento. Las diferencias entre nuestro trabajo y el de Lin son varias: en nuestro caso evaluamos varios sistemas mono-documento, utilizamos diversas estrategias de compresión, utilizamos ROUGE como métrica de evaluación y no empleamos componentes semánticos. Los resultados obtenidos confirman algunas observaciones de Lin, pero también enriquecen las conclusiones con un panorama experimental más amplio.

Nuestra metodología tiene varias etapas. En primer lugar, conformamos un corpus de textos especializados (en concreto, artículos médicos de investigación) acompañados de los resúmenes redactados por los mismos autores de los documentos. En segundo lugar, generamos resúmenes automáticos de los textos del corpus con diversos sistemas de resumen extractivo. En tercer lugar, realizamos una compresión de estos resúmenes, siguiendo tres estrategias diferentes: eliminación manual intuitiva de algunos elementos de la oración, eliminación manual de ciertos elementos discursivos con base en la *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988) y compresión automática por medio de sistemas elementales. Finalmente evaluamos los resultados mediante los sistemas ROUGE (Lin, 2004) y BLEU (Papineni et al., 2002), a fin de verificar si efectivamente los resúmenes comprimidos obtienen mejores resultados al compararlos con los resúmenes del autor.

El artículo está organizado de la siguiente manera: en la sección 2 hacemos una breve presentación del estado del arte de la compresión automática de frases. En la sección 3 detallamos la metodología empleada en nuestro estudio. Los diversos experimentos realizados y los resultados obtenidos son presentados en la sección 4. Para finalizar, en la sección 5 exponemos las conclusiones y algunas perspectivas de trabajo futuro.

2. Estado del arte

La compresión automática de frases ha sido recientemente abordada utilizando tanto métodos simbólicos como estadísticos. A continuación mostramos un breve panorama sobre este tema.

Con respecto a las aproximaciones simbólicas para el idioma inglés, destaca el trabajo de (Cordeiro, Dias, and Brazdil, 2009), donde se propone un sistema completo, no supervisado, que comienza por identificar oraciones similares con alta probabilidad de ser paráfrasis a partir de notas periodísticas de la Web. Posteriormente, estas son alineadas y procesadas por un sistema de programación lógica inductiva (ILP) para deducir una serie de predicados de lógica de primer orden que constituyen las reglas de compresión. Igualmente, (Jing, 2000) describe un complejo sistema que contempla tanto la verificación de la coherencia mediante el análisis sintáctico como la información contextual utilizando WordNet[1]. En (Yousfi-Monod and Prince, 2006; Yousfi-Monod and Prince, 2008) se muestra un método basado

[1] `http://wordnet.princeton.edu`

en reglas de transformación aplicadas a árboles sintácticos de frases en francés.

En la línea de las aproximaciones estadísticas, los trabajos de (Knight and Marcu, 2000; Marcu, 2000b) constituyen quizás los pilares en el estudio de la compresión estadística. Los autores adoptan el modelo de canal ruidoso (*Noisy Channel*) utilizado comúnmente en el área de traducción automática estadística. Aunque este estudio fue realizado para el inglés, la metodología parece resultar lo suficientemente general para ser aplicada a otras lenguas u otros modelos de lengua. (Lin, 2003) confirma que este último método puede resultar interesante en la tarea de resumen automático y posteriormente (Hori and Furui, 2004) muestran que también resulta útil para el resumen del discurso oral (*Speech Summarization*). (Turner and Charniak, 2005) muestran algunos problemas ligados al modelo de *Noisy Channel*, como por ejemplo que este tiende a comprimir muy poco las frases. De manera similar, (Clarke and Lapata, 2006b) indican que, en dicho modelo, la compresión es dependiente del dominio de los corpus de aprendizaje.

En otras direcciones, (Clarke and Lapata, 2006a) presentan un método no supervisado, en el cual se aborda la tarea como un problema de programación lineal. Recientemente, (Fernández and Torres-Moreno, 2009) y (Waszak and Torres-Moreno, 2008) muestran resultados interesantes con métodos diversos basados en la física estadística aplicada a documentos en francés y en inglés.

Por último, cabe mencionar que hasta donde sabemos no existen trabajos sobre la compresión de frases en español, ni tampoco un corpus paralelo (frase/frase comprimida) en esta lengua que pueda utilizarse como referencia para evaluar o entrenar sistemas.

3. Metodología

La metodología empleada en nuestro trabajo incluye las fases principales que se detallan a continuación: 1) conformación del corpus original de documentos especializados, 2) selección de herramientas de resumen automático, 3) compresión manual y automática del corpus y 4) evaluación de resultados.

3.1. Conformación del corpus especializado

En primer lugar, conformamos un corpus especializado del dominio médico. Seleccionamos 40 artículos médicos extraídos de la revista de investigación en español *Medicina Clínica*[2], fundada en 1943[3]. La versión digital de la revista permite acceder a las ediciones electrónicas de años anteriores gratuitamente, posibilitando así la constitución del corpus de estudio.

Cada documento del corpus incluye un apartado de un artículo médico (de aproximadamente 400 palabras): Fundamento, Pacientes y métodos, Resultados y Discusión.

En segundo lugar, obtenemos los resúmenes de los 40 documentos del corpus mediante los diversos sistemas de resumen automático que se detallarán en la sección 3.2.

Además, creamos resúmenes *Baseline* (BL1 o BL-aleatorio) de cada resumen con oraciones seleccionadas aleatoriamente del texto original y otro resúmenes *Baseline* (BL2 o BL-1era frase) a partir de las primeras oraciones del texto original. Todos los resúmenes contienen el mismo número de oraciones, dependiendo del apartado del texto:

- Fundamento (2 oraciones),
- Pacientes y métodos (3 oraciones),
- Resultados (4 oraciones) y
- Discusión (2 oraciones).

Para determinar este número de oraciones se calculó el promedio de las oraciones incluidas en cada apartado de los resúmenes de los autores, ya que estos resúmenes se dividen en cuatro apartados, siguiendo la estructura del artículo original. Posteriormente, se tomó la decisión de incluir una oración adicional, debido a que percibimos que, en gran cantidad de ocasiones, en estos *abstracts* se fusionaron en una sola oración las informaciones de dos o más oraciones de los artículos. Podría decirse que ha sido una decisión empírica con el objetivo de evitar una pérdida de información (da Cunha, 2008).

3.2. Selección de herramientas de resumen autómatico

Los sistemas de resumen automático que hemos empleado en nuestro trabajo se describen a continuación.

1. Cortex (Boudin and Torres-Moreno, 2007; Torres-Moreno, Velázquez-Morales, and Meunier, 2001; Torres-Moreno, Velázquez-Morales, and Meunier, 2002) es un sistema

[2] `http://www.doyma.es/revistas/ctl_servlet?_f=7032&revistaid=2`

[3] *Science Citation Index, Current Contents, Index Medicus* y *Excerpta Medica*

de resumen automático basado en el Modelo de Espacio Vectorial (VSM) (Salton and McGill, 1983). Se trata de un sistema de resumen por extracción mono-documento que combina varias métricas sin aprendizaje. Estas métricas resultan de algoritmos de procesamiento estadísticos y de información sobre la representación vectorial del documento. La idea principal es representar un texto en un espacio vectorial adecuado y aplicar procesamiento estadístico.

2. ENERTEX (Fernández, 2009; Fernández, SanJuan, and Torres-Moreno, 2007; Fernández, SanJuan, and Torres-Moreno, 2008) también es un sistema de resumen automático basado en VSM, pero en este caso se trata de un enfoque de redes de neuronas inspirado en la física estadística. El algoritmo modela los documentos como una red de neuronas de la que se estudia su energía textual. La idea principal es que un documento puede ser procesado como un conjunto de unidades interactivas (las palabras), donde cada unidad se ve afectada por el campo creado por las demás.
3. DISICOSUM (da Cunha, 2008; da Cunha and Wanner, 2005; da Cunha, Wanner, and Cabré, 2007) es un modelo de resumen automático de textos médicos que parte de la idea de que los profesionales de un dominio especializado emplean técnicas concretas para resumir los textos de su ámbito. El algoritmo de DISICOSUM integra criterios basados en la estructura textual, en las unidades léxicas y en la estructura discursiva y sintáctico-comunicativa del texto. El modelo está formado por reglas que se relacionan con estos criterios lingüísticos.
4. RESUMIDOR HÍBRIDO (da Cunha et al., 2007a; da Cunha et al., 2009) consta de varios resumidores autónomos que se combinan de manera equilibrada para formar un único resumidor híbrido. Algunos de los resumidores utilizan métodos numéricos (CORTEX y ENERTEX), otro resumidor tiene un carácter estrictamente lingüístico (DISICOSUM) y en los dos sistemas restantes las métricas estadísticas (de CORTEX y ENERTEX) se combinan con la información lingüística procedente de un extractor de términos (YATE (Vivaldi, 2001; Vivaldi and Rodríguez, 2001; Vivaldi and Rodríguez, 2002)). Las características más relevantes de YATE son: el uso intensivo de información semántica junto con el uso de técnicas de combinación de los resultados obtenidos a partir de diferentes técnicas de extracción. Ha sido desarrollado para el ámbito médico en español, aunque está siendo adaptado con éxito a otros dominios (genómica, derecho, economía, informática y medio ambiente) y otras lenguas (catalán).
5. Dos sistemas de resumen automático relevantes a nivel del estado del arte de esta temática:
   - SWESUM: `http://swesum.nada.kth.se/index-eng.html`
   - OPEN TEXT SUMMARIZER (OTS): `http://libots.sourceforge.net`
6. Dos sistemas de resumen automático comerciales:
   - PERTINENCE SUMMARIZER: `http://www.pertinence.net/index.html`
   - WORD SUMMARIZER

3.3. Herramientas de compresión de frases

Una vez obtenidos los extractos de los sistemas de resumen automático mencionados y las *baselines*, se procedió a su compresión. No se verificó el efecto en el orden inverso, es decir, no se realizó en mi primer lugar la compresión de las frases del texto original para posteriormete realizar un extracto, ya que el objetivo de este trabajo es confirmar si es adecuado emplear la compresión de frases como recurso para la optimización de sistemas de resumen automático. De tal manera que, bajo este enfoque, concebimos la extracción como la primera etapa y la compresión como la segunda etapa.

Para la compresión usamos las siguientes estrategias manuales y automáticas de eliminación de información:

Dos estrategias manuales:

1. Eliminación manual intuitiva
2. Eliminación manual basada en la RST

Cuatro estrategias automáticas:

1. Eliminación adjetival
2. Eliminación adverbial
3. Eliminación adjetival y adverbial
4. Eliminación aleatoria *baseline*

Estos sistemas serán descritos a continuación.

3.3.1. Compresión manual

Con respecto a la compresión manual empleamos dos estrategias:

1. Eliminación intuitiva de elementos no esenciales de la frase, como ciertos artículos, adverbios, elementos parentéticos, aposiciones, locuciones, etc., siguiendo la línea de los trabajos de (Yousfi-Monod and Prince, 2008). Esta estrategia implica cierta subjetividad, ya que pueden existir elementos que un anotador considere presindibles, mientras que otro anotador considere necesarios para el resumen. Para realizar esta tarea, utilizamos el mismo protocolo usado en la construcción del corpus de frases comprimidas en francés[4] del proyecto ANR-RPM2[5] (de Loupy et al., 2010).

   El ejemplo a) del Cuadro 1 muestra una oración original procedente de uno de los resúmenes (resumen del apartado de PACIENTES Y MÉTODOS del resumidor CORTEX) y el ejemplo b) la misma oración final comprimida.

a) "El Servicio de Epidemiología del Instituto Municipal de Salud Pública recoge de manera sistemática los casos de sida notificados por los médicos y, además, los casos procedentes de las altas hospitalarias y del registro de mortalidad."
b) "El Servicio de Epidemiología del Instituto Municipal de Salud Pública recoge casos de sida notificados por médicos y casos procedentes de altas hospitalarias y del registro de mortalidad."

Cuadro 1: Ejemplo de compresión manual por eliminación intuitiva.

2. Eliminación de satélites de la *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988) del interior de la frase, en la línea de los trabajos de Marcu (Marcu, 1998; Marcu, 2000b). Esta estrategia implica el empleo de una base teórica más marcada. La RST es una teoría descriptiva de organización del texto muy útil para describirlo caracterizando su estructura a partir de las relaciones que mantienen entre sí los elementos discursivos del mismo (Circunstancia, Elaboración, Motivación, Evidencia, Justificación, Causa, Propósito, Antítesis, Condición, entre otras). Estas relaciones pueden ser asimétricas (núcleo-satélite) o simétricas (multinucleares): en las primeras el elemento principal se denomina "núcleo" y el secundario "satélite", mientras que en las segundas todos los elementos son núcleos. Por lo general, los satélites aportan información adicional a sus núcleos. Estos elementos pueden ser oraciones completas, pero también pueden encontrarse a nivel intraoracional, es decir, estar formados por fragmentos del interior de las oraciones. Es en estos casos en los que nos centraremos, ya que, en este trabajo, la compresión de frases se realiza dentro de las oraciones, independientemente de su contexto discursivo en el texto. Aunque existen trabajos sobre análisis discursivo automático para el portugués basados en la RST (Leal, Quaresma, and Chishman, 2006), la compresión de frases mediante esta estrategia se realizó de manera manual, debido a que no existe en la actualidad ningún analizador discursivo completo para el español que pueda detectar núcleos y satélites. Sin embargo, hay un proyecto vigente sobre el tema (da Cunha et al., 2007b; da Cunha et al., 2010), por lo que, en cuanto este analizador discursivo esté operativo, podremos llevar a cabo este tipo de compresión de manera automática.

   En la figura 1 mostramos un árbol discursivo con relaciones de la RST, que incluye una relación multinuclear de Lista y dos relaciones núcleo-satélite, de Concesión y de Elaboración. El ejemplo ha sido extraído de uno de los textos médicos del corpus.

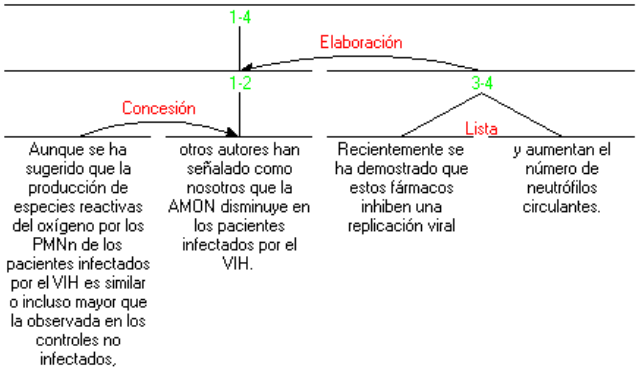

Figura 1: Ejemplo de árbol discursivo con relaciones de la RST.

   El ejemplo a) del Cuadro 2 muestra una oración original de uno de los resúmenes (resumen del apartado de DISCUSIÓN del resumi-

[4] El corpus puede ser recuperado en el sitio web: `http://lia.univ-avignon.fr/rpm2`

[5] `http://labs.sinequa.com/rpm2/`

dor ENERTEX) y el ejemplo b) muestra la oración final comprimida.

a) "No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH."
b) "No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios."

Cuadro 2: Ejemplo de compresión manual por eliminación de satélites.

El fragmento eliminado ("aunque [...] para el VIH".) constituye un satélite de Concesión de la RST, puesto en evidencia mediante el conector discursivo "aunque".

3.3.2. Compresión automática

Con respecto a la compresión automática, hemos desarrollado cuatro sistemas:

1. Sistema de eliminación adjetival (elimADJ). Elimina todas las apariciones de adjetivos dejando los elementos restantes intactos.
2. Sistema de eliminación adverbial (elimADV). Análogo al anterior, pero eliminando adverbios.
3. Sistema de eliminación mixto (elimADJ-ADV). Elimina ambas categorías, adjetivos y adverbios.
4. Sistema de referencia de base (elimALE). Elimina un porcentaje fijo de palabras aleatoriamente (16 % en este caso –de acuerdo con la tasa de compresión promedio de los anotadores humanos–).

El Anexo 1 muestra algunos ejemplos. El ejemplo a) refleja una oración original de uno de los resúmenes (resumen del apartado de DISCUSIÓN del resumidor ENERTEX). El ejemplo b) corresponde a la versión comprimida automática obtenida por el sistema elimADJ. El c) corresponde a la versión comprimida obtenida por el sistema elimADV y el d) corresponde a la versión comprimida del sistema elimADJ-ADV. Finalmente el ejemplo e) corresponde a la salida del sistema de base elimALE. En todos los casos se eliminó el texto entre paréntesis. Estos sistemas se explican en detalle a continuación.

Un análisis estadístico de los elementos eliminados por los anotadores, mediante el protocolo de compresión intuitiva del corpus RPM2 (de Loupy et al., 2010), arrojó resultados interesantes. El Cuadro 3 muestra las cinco secuencias más comúnmente eliminadas mediante este protocolo. Para llevar a cabo este análisis, se extrajeron por separado las secuencias de palabras eliminadas y sus equivalentes en términos de categorías gramaticales. Las categorías gramaticales fueron obtenidas mediante TreeTagger[6]. Elegimos esta herramienta por ser independiente del idioma, además de ser flexible, en el sentido de que es inmediato cambiar de un idioma a otro, lo que nos permitirá emplear, sin complicaciones, la misma metodología en trabajos futuros. Las etiquetas utilizadas en el análisis (que pueden ser consultadas en el sitio Web)[7]) fueron las siguientes: LP (paréntesis izquierdo), RP (paréntesis derecho), CARD (cifras), PERCT (símbolo %), ART (artículo), NP (nombre propio), ADJ (adjetivo) y ADV (adverbio). Observando el Cuadro 3, se puede inferir que la simple extracción de un adjetivo o un adverbio constituye una práctica común en la tarea de compresión. Del total de secuencias eliminadas, el 19.86 % incluyó al menos un adjetivo y el 9.15 % al menos un adverbio. También puede comprobarse que la eliminación del contenido entre paréntesis resulta ineluctable en la tarea de compresión. Del total de secuencias eliminadas, el 36.16 % contiene un texto entre paréntesis y el 31.91 % constituye toda la secuencia eliminada en sí. Otros resultados menos evidentes nos dieron la pauta para nuevas investigaciones al respecto. Por ejemplo, se observó que el 27.45 % de las secuencias contienen al menos una coma y, de estas, aproximadamente en la mitad es el primer símbolo de la secuencia. En sistemas posteriores consideraremos la segmentación de oraciones a partir de delimitadores ortográficos.

El análisis de las secuencias comprimidas nos llevó a construir tres sistemas de compresión elementales: el sistema de eliminación adjetival (elimADJ), el sistema de eliminación adverbial (elimADV) y el sistema de eliminación mixto (elimADJ-ADV). Además se construyó un sistema de referencia (elimALE) que extrae el 16 % de las palabras aleatoriamente –de acuerdo con la tasa de compresión promedio de los anotadores–.

[6] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

[7] `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/spanish-tagset.txt`

Secuencia	**Ejemplos**	**Porcentaje de eliminación**
LP CARD PERCT RP	(33,7 %), (5 %)	7,14 %
ART	el, la, las, los	6,03 %
LP NP RP	(VIH), (ELISA)	4,46 %
ADV	generalmente, probablemente	4,02 %
ADJ	principales, importante	3,79 %

Cuadro 3: Lista de las secuencias más frecuentemente eliminadas en el corpus de resúmenes comprimidos intuitivamente.

En todos los casos se eliminó el contenido entre paréntesis.

4. Evaluación

Todos los resúmenes (comprimidos y sin comprimir) fueron evaluados con el sistema automático ROUGE (Lin, 2004), comparándolos utilizando como referencia los *abstracts* de los autores de los artículos. El protocolo utilizado involucra el uso de resúmenes modelo o de referencia (escritos por personas) y el paquete ROUGE, un sistema de evaluación de resúmenes que se basa en la co-ocurrencia de n-gramas entre resúmenes candidatos (los que se quiere evaluar) y resúmenes modelo. ROUGE mide los máximos, los mínimos y el valor medio (reportado en este artículo) de la intersección de los n-gramas en los resúmenes candidatos y de referencia (por ejemplo, ROUGE-1 compara unigramas, ROUGE-2 compara bigramas, ROUGE-SU4 compara bigramas con huecos, etc.). Las campañas de evaluación del NIST[8] han adoptado este test para medir la relevancia de los resúmenes. Para ser consistentes con la metodología del NIST, adoptamos el mismo protocolo en la evaluación de los resúmenes producidos por nuestro sistema. Los resúmenes fueron previamente truncados a 10, 20, 30 y así consecutivamente hasta 100 palabras automáticamente. Esta proceso garantiza una evaluación en condiciones iguales de tamaño en número de palabras.

Además de la evaluación con ROUGE, decidimos verificar la calidad de las oraciones comprimidas generadas por los sistemas automáticos. Para ello, hemos utilizado BLEU, un método de evaluación semiautomático desarrollado por IBM para la tarea de traducción automática (*Machine Translation* o MT) (Papineni et al., 2002). La idea central en MT es que, a medida que una traducción (hecha por un sistema) se acerca más (comparando la co-ocurrencia de n-gramas) a una referencia hecha por un experto, la traducción es mejor. Hemos optado por utilizar esta herramienta dado que, hasta nuestro conocimiento, no existe aún un método automático de evaluación de oraciones comprimidas. Sin embargo, reconocemos que con este método es posible que aún una frase agramatical obtenga un buen *score* BLEU. La evaluación consistió en tomar como referencia las oraciones comprimidas por los humanos mediante la estrategia intuitiva y la RST, y comparar con las oraciones comprimidas por los sistemas automáticos (elimADJ, elimADV, elimADJ-ADV y elimALE).

La figura 6 del Anexo 2 ilustra la metodología completa empleada en nuestro estudio, detallada en los apartados anteriores.

4.1. Experimentos con compresión manual

Se calculó una media normalizada (en porcentaje) de las compresiones manuales, de la siguiente manera:

$$C = \frac{\langle A \rangle - \langle B \rangle}{\langle A \rangle} \times 100 \qquad (1)$$

donde $\langle A \rangle$ es el número de palabras promedio antes de comprimir y $\langle B \rangle$ el número de palabras promedio después de la compresión. La figura 2 muestra los valores C promedios en cada sección (círculos), que oscilan entre el 13 % y el 24 %. Esta variación indica una cierta independencia del número de frases en la compresión e, inversamente, una fuerte dependencia de la longitud de las mismas. En cuanto a la RST, es importante señalar el comportamiento del porcentaje de compresión de las secciones DISCUSIÓN y RESULTADOS. En la primera, las frases contienen muchos satélites que, al ser eliminados, aumentan la compresión. En la segunda, las frases conservan una estructura mayoritariamente nuclear, que las hace poco candidatas a ser comprimidas.

Para comprobar si los resúmenes comprimidos son mejores que los resúmenes originales de los sistemas de resumen automático y los resúmenes *Baselines*, los evaluamos por separado con ROUGE. En concreto, empleamos ROUGE-2. Como ya hemos comentado, esta medida evalúa la co-ocurrencia de bigramas entre los resúmenes candidatos (es decir, los resúmenes que se desea evaluar) y los resúmenes de referencia o modelos realizados por humanos (es decir, *abstracts* de los autores de los artículos médicos).

Una vez realizada la evaluación de ambos ti-

[8] `http://www.nist.gov/index.html`

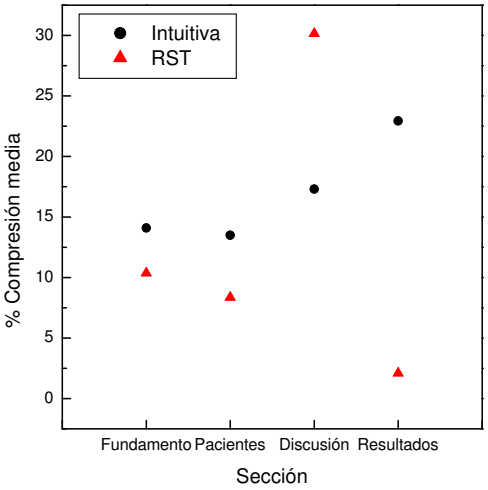

Figura 2: Porcentajes medios de compresión intuitiva y RST por sección.

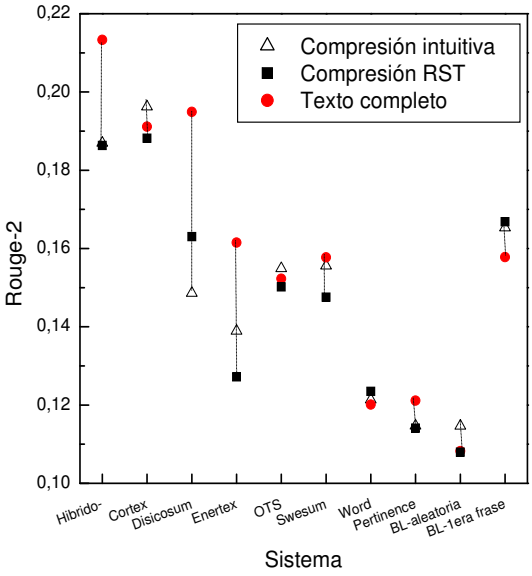

Figura 3: ROUGE-2 para cada sistema: en función del tipo de compresión realizada (truncamiento a 50 palabras) intuitiva, RST y texto completo.

pos de resúmenes (comprimidos y sin comprimir, ambos truncados de 10 a 100 palabras), comparamos el *score* obtenido con ROUGE-2.

En la figura 3 pueden observarse los resultados de ROUGE-2 obtenidos con un truncamiento promedio a 50 palabras, mediante la compresión intuitiva y mediante la compresión RST. El Cuadro 4 incluye los datos numéricos de esta evaluación. Como puede observarse, con este truncamiento, los resúmenes del sistema HÍBRIDO mejoran notablemente después de realizar la compresión mediante la estrategia intuitiva (de 0.18696 a 0.21331), mientras que mantienen una puntuación similar al ser comprimidos mediante la estrategia RST (de 0.18696 a 0.18632). El sistema CORTEX no mejora con la compresión, aunque mediante la compresión con la estrategia intuitiva no pierde excesiva información (disminuye de 0.19624 a 0.19116). DISICOSUM, por su parte, mejora sus resultados con la compresión llevada a cabo mediante ambas estrategias, pasando de 0.14862 a 0.19492 con la estrategia intuitiva y a 0.16303 con la estrategia RST. ENERTEX obtiene valores más elevados después de la compresión intuitiva de sus resúmenes (de 0.13893 a 0.16151).

El sistema OTS no mejora sus resúmenes con ningún tipo de compresión. SWESUM, WORD y PERTINENCE mejoran ligeramente sus resultados con alguno de los tipos de compresión: el primero mediante la compresión intuitiva (de 0.15558 a 0.15773) y el segundo y el tercero mediante la compresión RST (de 0.12136 a 0.12350. y de 0.11471 a 0.12115, respectivamente). Los resúmenes BL-1era frase mejoran ligeramente con la compresión RST. Finalmente, los resúmenes BL-aleatoria no mejoran sus resultados con la compresión, como era de esperarse.

Sistema	ROUGE-2 Texto completo	ROUGE-2 Comp. intuitiva	ROUGE-2 Comp. RST
HÍBRIDO	0.18696	**0.21331**	0.18632
CORTEX	0.19624	0.19116	0.18817
DISICOSUM	0.14862	0.19492	0.16303
ENERTEX	0.13893	0.16151	0.12719
OTS	0.15492	0.15227	0.1502
SWESUM	0.15558	0.15773	0.14756
WORD	0.12136	0.12012	0.1235
PERTINENCE	0.11471	0.12115	0.11408
BL-ALEAT.	0.11466	0.10821	0.10794
BL-1ERA.	0.16533	0.15782	0.16683

Cuadro 4: Resultados de la evaluación ROUGE-2 para cada sistema con truncamiento a 50 palabras.

Los resultados reflejan que algunos de los resúmenes comprimidos de manera intuitiva obtienen mejores resultados que los resúmenes no comprimidos correspondientes, confirmando nuestra hipótesis inicial. Sin embargo, la mejora no es tan significativa como se pensó en un primer momento. Esto puede deberse a que, aunque todos los resúmenes están truncados al mismo número de palabras (50), algunos de ellos pueden incluir menos palabras una vez realizada la compresión. Este hecho puede haber provocado que estos resúmenes obtengan un valor más bajo de ROUGE-2, ya que al contener frases comprimidas ROUGE-2 castigará la falta de co-ocurrencias de bigramas entre resúmenes con frases compri-

midas y los resúmenes de referencia. Asimismo, se observa que, en general, los resúmenes comprimidos mediante la eliminación de satélites de la RST no mejoran demasiado con respecto a los resúmenes no comprimidos. Esta situación puede deberse a que las oraciones de los resúmenes de los textos médicos son breves, porque normalmente reflejan datos o informaciones concretas (sobre todo los resúmenes de los apartados de PACIENTES Y MÉTODOS y RESULTADOS), que generalmente no incluyen satélites.

La figura 4 reporta los resultados de ROUGE-2 obtenidos por cada sistema para resúmenes completos truncados de 10 a 100 palabras. La figura 5 muestra los resultados de ROUGE-2 de todos los sistemas, con resúmenes comprimidos mediante la estrategia intuitiva con un truncamiento de 10 a 100 palabras, además de sin truncamiento. Como puede observarse, el comportamiento de los resúmenes comprimidos intuitivamente con los diferentes niveles de truncamiento (de 10 a 100 palabras) es bastante similar al descrito para los resúmenes truncados a 50 palabras. Los resultados más destacables son la mejora evidente de los resúmenes del Resumidor HÍBRIDO mediante la compresión con un truncamiento de 30, 40, 50 y 60 palabras, la ligera mejora del sistema CORTEX con un truncamiento de 40 palabras, la clara mejora de DISICOSUM con un truncamiento de 30 y 40 palabras y la mejora, también evidente, de los resúmenes de ENERTEX con un truncamiento de 30, 40 y 50 palabras.

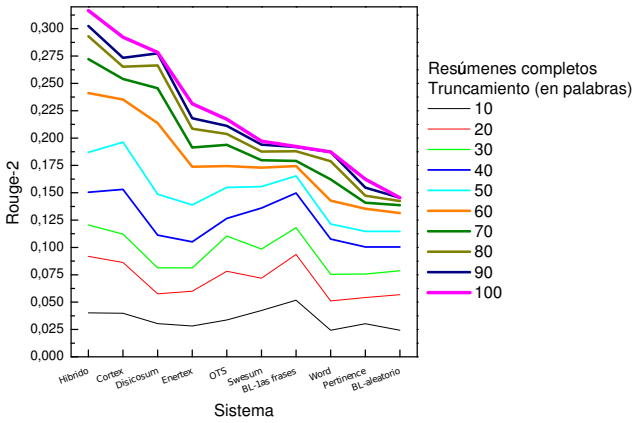

Figura 4: Resultados ROUGE-2 para resúmenes sin compresión con truncamiento de 10 a 100 palabras.

Con respecto al *ranking* de los sistemas, por un lado, al realizar la evaluación de los resúmenes completos, por lo general CORTEX se posiciona en primer lugar, seguido muy de cerca por el Resumidor HÍBRIDO, y posteriormente de la BL-1era frase, OTS, DISICOSUM, ENERTEX, WORD, PERTINENCE y BL-aleatoria.

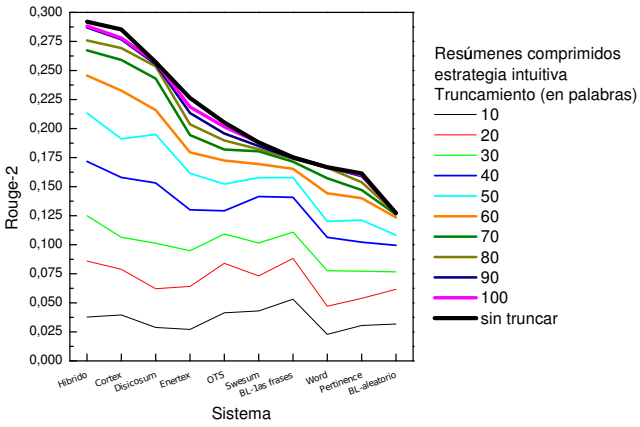

Figura 5: Resultados ROUGE-2 para resúmenes comprimidos mediante la estrategia intuitiva con truncamiento de 10 a 100 palabras y sin truncamiento.

Por otro lado, al realizar la evaluación de los resúmenes truncados, el orden del *ranking* cambia ligeramente, posicionándose claramente el Resumidor HÍBRIDO en primer lugar, seguido de CORTEX, DISICOSUM, BL-1era frase, ENERTEX, SWESUM, OTS, WORD, PERTINENCE y BL-aleatoria.

Es destacable el hecho de que, inesperadamente, la BL-1era frase obtiene resultados muy elevados tanto en la evaluación de resúmenes completos como de resúmenes comprimidos, en comparación con algunos otros resumidores. Este hecho puede deberse a que, en el tipo de documentos utilizados (artículos médicos de investigación), las primeras oraciones de cada apartado generalmente contienen las informaciones más relevantes.

4.2. Experimentos con compresión automática

En el Cuadro 5 se comparan los résumenes con frases comprimidas sin truncamiento y utilizando ROUGE-SU4. Bajo estas condiciones se observa que el sistema elimADV da mejores resultados y resulta comparable a la eliminación RST e intuitiva. Sin embargo, una lectura directa de los resúmenes comprimidos muestra que en muchas ocasiones los resúmenes generados por el sistema elimADV perdieron la consistencia debido a la generación de frases agramaticales. En el caso de la compresión intuitiva y la compresión por RST esto no sucede, ya que estas se realizan de manera manual.

El Cuadro 5 muestra los resultados de la evaluación ROUGE para los resúmenes con fra-

Sistema	SU4 RST	SU4 intuitiva	SU4 elimADV	SU4-elimADJ	SU4-elimADJ-ADV	SU4-elimALE
HÍBRIDO	**0.31756**	**0.31333**	**0.31548**	**0.29045**	**0.28276**	0.26315
CORTEX	0.31299	0.30925	0.31136	0.28514	0.27765	**0.27532**
DISICOSUM	0.28297	0.28454	0.28545	0.28297	0.26269	0.23440
ENERTEX	0.25624	0.26229	0.26235	0.24207	0.23519	0.21496
OTS	0.24546	0.24737	0.24521	0.22589	0.22004	0.20397
SWESUM	0.21797	0.22126	0.21993	0.21179	0.20650	0.18582
WORD	0.22048	0.20522	0.20971	0.19629	0.19508	0.17530
PERTINENCE	0.21829	0.21029	0.21268	0.20443	0.20153	0.18305
BL-ALEAT.	0.16933	0.16412	0.17155	0.15777	0.15558	0.14366
BL-1ERA.	0.22766	0.21544	0.21950	0.20225	0.20106	0.18756

Cuadro 5: Evaluación ROUGE-SU4 para resúmenes con frases comprimidas.

ses comprimidas por los sistemas elimADJ, elimADV, elimADJ-ADV y elimALE. El Cuadro 6 muestra el promedio de la evaluación BLEU obtenido por los sistemas de compresión. En tanto que BLEU devuelve valores entre 0 y 1 (1 es asumido como una buena compresión en relación a la referencia), se puede notar que, en general, las heurísticas utilizadas por los sistemas automáticos se correspondieron mejor a la compresión intuitiva que a la compresión RST.

De acuerdo con el Cuadro 6, podría entenderse que la estrategia de eliminación de adverbios se asemeja mucho más al comportamiento intuitivo. Esta última conclusión es engañosa si se considera que, en el corpus original, los adjetivos constituyen la cuarta categoría más frecuente (6,90 %), mientras que los adverbios ocupan el dieciseisavo lugar con apenas un 1,10 %. Es decir, que la heurística de eliminar adjetivos es, en cierto sentido, mucho más arriesgada que aquella de eliminar adverbios por el simple hecho de que estos úlitmos aparecerán con menos frecuencia.

El sistema elimADV tiende a dejar las frases intactas con más frecuencia y el *score* BLEU, en este caso, resulta óptimo por que se compara una frase consigo misma (todos los n-gramas son encontrados intactos).

Sistema	**Referencia**	
	Intuitiva (RPM2)	**Satélites (RST)**
elimALE	0.67408	0.70669
elimADJ-ADV	0.74427	0.67549
elimADJ	0.76857	0.70757
elimADV	**0.82538**	**0.77098**

Cuadro 6: Evaluación BLEU para los cuatro sistemas de compresión automática contra las dos referencias manuales.

5. Conclusiones

En este trabajo hemos explorado la posibilidad de emplear la compresión de frases para la optimización de sistemas de resumen automático de documentos. La metodología empleada consistió en extraer las frases que conformarían el resumen y posteriormente comprimirlas mediante diversas estrategias. Este método nos permitió analizar y evaluar diversas características de ambos procesos por separado. Sin embargo, nuestros trabajos futuros estarán orientados a concebir la selección y la compresión como una tarea conjunta, pues, como se menciona en (Daumé III and Marcu, 2002), este enfoque puede llevar a mejores resultados.

La principal conclusión de nuestros experimentos es que la compresión de frases puede beneficiar a algunos sistemas de resumen automático. Esta mejora parece no ser excesivamente elevada y creemos que se debe a que los resúmenes contienen un cierto número de palabras (de 10 a 100) que después de la compresión disminuye y esto les perjudica en la evaluación ROUGE, pues ésta considera la co-ocurrencia de n-gramas como una buena práctica y es de suponer que algunas de estas co-ocurrencias se pierdan en la compresión. Tenemos razones para creer que esto penaliza injustamente los resúmenes con frases comprimidas. También hemos explorado la implementación de sistemas de compresión que simulen la eliminación humana intuitiva de elementos de la frase para optimizar sistemas de resumen automático. Esta tarea plantea interesantes retos e interrogantes que deben resolverse en el futuro, comenzando por los recursos necesarios para analizar el problema (corpus alineados de frases-frases comprimidas) pues estos son, hasta nuestro conocimiento, escasos y, en algunos casos, como el del español, aún inexistentes. Sin embargo, como parte de este trabajo hemos elaborado de manera semiautomática un corpus alineado experimental para el español. Este corpus está disponible en el sitio web `http://lia.univ-avignon.fr/fileadmin/axes/TALNE/index.html`. También será interesante comprobar, en trabajos futuros, cómo se comporta la compresión en otros géneros, como noticias periodísticas. Tenemos la intuición de que algunos dominios son más sensibles a la compresión que otros.

Los sistemas de compresión descritos aquí son aún prototipos elementales pero nos permitirán contrastar los resultados de sistemas más complejos en un futuro. Por ejemplo, ahora que contamos con un conjunto de secuencias comprimidas, podemos utilizar métodos de aprendizaje supervisado para generar reglas de compresión.

Además, queremos realizar más pruebas de cara a profundizar en los motivos que han hecho que la compresión siguiendo la estrategia de la RST no obtenga resultados demasiado positivos.

Creemos que este hecho ha sido provocado por haber eliminado todos los satélites, independientemente de su tipo. En este tipo de textos científicos, por ejemplo, puede ser que los satélites del apartado RESULTADOS sean relevantes para un resumen.

A su vez, al eliminar los satélites de Condición, se pierde una información necesaria para la compresión del texto.

Finalmente, nos restan por explorar otros experimentos interesantes de compresión contextual de frases: por ejemplo, dada una frase en la posición i, su compresión podría depender del contexto generado por las $i-1$ frases precedentes $j = 1, 2, ..., i-1$. Algoritmos que consideren esta contextualización son actualmente objeto de estudio en nuestro equipo.

Agradecimientos

Parte de este trabajo ha sido financiado mediante una ayuda de movilidad posdoctoral otorgada por el Ministerio de Ciencia e Innovación de España (Programa Nacional de Movilidad de Recursos Humanos de Investigación; Plan Nacional de Investigación Científica, Desarrollo e Innovación 2008-2011) a Iria da Cunha. Asimismo este trabajo fue financiado parcialmente mediante la beca 211963 del CONACYT (México) a Alejandro Molina. El proyecto ha sido además parcialmente financiado por la *Agence Nationale pour la Recherche* (ANR, France), en el marco del proyecto *Resumé Plurimédia Multidocument* (RPM2), concedido a Juan-Manuel Torres-Moreno.

Anexo 1

a) **Oración original**
"Todos presentaron concentraciones de cocaína detectables en la orina, status epiléptico e inestabilidad hemodinámica, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes intactos y uno roto, evolucionando favorablemente y siendo dado de alta (tres de estos casos han sido publicados previamente)."

b) **elimADJ**
"Todos presentaron concentraciones de cocaína en la orina, status e inestabilidad, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes y uno roto, evolucionando favorablemente y siendo dado de alta."

c) **elimADV**
"Todos presentaron concentraciones de cocaína detectables en la orina, status epiléptico e inestabilidad hemodinámica, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes intactos y uno roto, evolucionando y siendo dado de alta."

d) **elimADJ-ADV**
"Todos presentaron concentraciones de cocaína en la orina, status e inestabilidad, falleciendo dos de ellos, el tercero se encuentra en estado de coma vegetativo y el cuarto paciente, una vez estabilizado, fue sometido a laparotomía y se extrajeron 10 paquetes y uno roto, evolucionando y siendo dado de alta."

e) **elimALE**
"Todos presentaron concentraciones de cocaína detectables en la orina, status epiléptico e inestabilidad hemodinamica, falleciendo ellos, el se encuentra en estado coma vegetativo y el cuarto paciente, una vez, fue sometido a laparotomía y se extrajeron 10 paquetes y roto, favorablemente y siendo dado alta."

Anexo 2

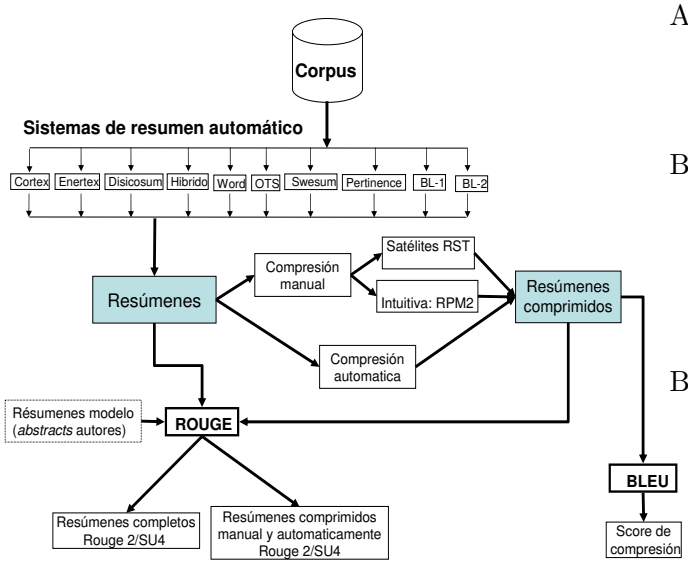

Figura 6: Metodología empleada para la generación de resúmenes, la compresión de frases y sus evaluaciones.

References

Afantenos, S., V. Karkaletsis, and P. Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

Berger, A.L. and V.O. Mittal. 2000. OCELOT: a system for summarizing Web pages. In *Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 144–151. ACM.

Boudin, F. and J.-M. Torres-Moreno. 2007. NEO-CORTEX: A Performant User-Oriented Multi-Document Summarization System. In *Computational Linguistics and Intelligent Text Processing (CICLing'07)*, volume 4394 of *Lecture Notes in Computer Science*, pages 551–562. Springer.

Boudin, F. and J-M. Torres-Moreno. 2009. Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. In *Proceedings of Traitement Automatique de la Langue Naturelle (TALN'09), Senlis.*

Boudin, F., J-M. Torres-Moreno, and M. El-Bèze. 2008. Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. In *Proceedings of the conference CICLing'08, Haifa (Israel), 2008 17-23 February*, pages 334–349. The Springer LNCS 4919.

Boudin, F., J-M. Torres-Moreno, and P. Velazquez-Morales. 2008. An efficient Statistical Approach for Automatic Organic Chemistry Summarization. In *Proceedings of the International Conference on Natural Language Processing (GoTAL), Gothenburg (Sweden)*, pages 89–99. The Springer LNCS 5221.

Clarke, J. and M. Lapata. 2006a. Constraint-based sentence compression: An integer programming approach. In *COLING/ACL 2006 Main Conference Poster Sessions*, pages 144–151.

Clarke, J. and M. Lapata. 2006b. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, page 384. Association for Computational Linguistics.

Cordeiro, J., G. Dias, and P. Brazdil. 2009. Unsupervised induction of sentence compression

rules. In *UCNLG+Sum '09: Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 15–22, Morristown, NJ, USA. Association for Computational Linguistics.

da Cunha, I. 2008. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Ph.D. thesis, IULA-UPF, Barcelona, España.

da Cunha, I., S. Fernández, P. Velázquez Morales, J. Vivaldi, E. SanJuan, and J.-M. Torres-Moreno. 2007a. A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics. In *Lecture Notes in Computer Science, 4827*, pages 872–882. Springer.

da Cunha, I., S. Fernández, P. Velázquez, J. Vivaldi, E. SanJuan, and J.M. Torres-Moreno. 2007b. A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics. In *MICAI 2007: Advances in Artificial Intelligence. Lecture Notes in Computer Science*, pages 872–882. Gelbukh, A. and Kuri Morales, A. F. (eds.), Berlín: Springer.

da Cunha, I., E. SanJuan, J.-M. Torres-Moreno, M. Lloberes, and I. Castellon. 2010. DiSeg : Un segmentador discursivo automatico para el español. *Procesamiento de Lenguaje Natural, ISSN: 1989-7553*, 2010(45).

da Cunha, I., J.-M. Torres-Moreno, P. Velázquez-Morales, and J. Vivaldi. 2009. Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. *Linguamática*, 2(2):67–79.

da Cunha, I. and L. Wanner. 2005. Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria. In *Crossing Barriers in Text Summarization Research (RANLP-2005)*, pages 46–51. Saggion, H. and Minel, J. (eds.), Borovets (Bulgaria): INCOMA Ltd.

da Cunha, I., L. Wanner, and T. Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. *Terminology*, 13(2):249–286.

Daumé III, H. and D. Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 449–456. Association for Computational Linguistics.

de Loupy, C., C. Ayache M. Guigan, S. Seng, and J.-M. Torres-Moreno. 2010. A French Human Reference Corpus for multi-documents summarization and sentence compression. In *International Conference on Language Resources and Evaluation (LREC'10)*.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of ACM*, 16(2):264–285.

Farzindar, A., G. Lapalme, and J.P. Desclés. 2004. Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte: solutions et perspectives*, 45(1):26.

Fernández, S. and J.-M. Torres-Moreno. 2009. Une approche exploratoire de compression automatique de phrases basée sur des critères thermodynamiques. In *Actes de la Conférence sur le Traitement Automatique du Langage Naturel*.

Fernández, S. 2009. *Applications exploratoires des modèles de spins au Traitement Automatique de la Langue*. Ph.D. thesis, Université Henri Poincaré Nancy 2, France.

Fernández, S., E. SanJuan, and J.-M. Torres-Moreno. 2007. Énergie textuelle de mémoires associatives. In *Traitement Automatique des Langues Naturelles*, pages 25–34. Toulouse, France.

Fernández, S., E. SanJuan, and J.-M. Torres-Moreno. 2008. Enertex : un système basé sur l'énergie textuelle. In *Traitement Automatique des Langues Naturelles*, pages 99–108. Avignon, France.

Fuentes, M., E. González, and H. Rodríguez. 2004. Resumidor de noticies en catala del projecte hermes. In *Proceedings of the II Congres d'Enginyeria en Llengua Catalana (CELC04), Andorra*.

Grefenstette, G. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Working notes of the AAAI Spring Symposium on Intelligent Text summarization*, pages 111–118.

Hori, C. and S. Furui. 2004. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE TRANSACTIONS on Information and Systems*, 87:15–25.

Jing, H. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied natural language processing*, pages 310–315. Association for Computational Linguistics.

Knight, K. and D. Marcu. 2000. Statistics-based summarization-step one: Sentence compression. In *National Conference on Artificial Intelligence*, pages 703–710. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Lal, P. and S. Ruger. 2002. Extract-based summarization with simplification. In *Document Understand Conference (DUC'02)*. NIST.

Leal, Ana, Paulo Quaresma, and Rove Chishman. 2006. From syntactical analysis to textual segmentation. In Renata Vieira, Paulo Quaresma, Maria Nunes, Nuno Mamede, ClÃ¡udia Oliveira, and Maria Dias, editors, *Computational Processing of the Portuguese Language*, volume 3960 of *Lecture Notes in Computer Science*, pages 252–255. Springer Berlin / Heidelberg.

Lenci, A., R. Bartolini, N. Calzolari, A. Agua, S. Busemann, E. Cartier, K. Chevreau, and J. Coch. 2002. Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, pages 29–31.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: ACL-04 Workshop*, pages 74–81, Barcelona, July.

Lin, C.Y. 2003. Improving summarization performance by sentence compression-a pilot study. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, pages 1–8.

Mann, W. C. and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, D. 1998. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, Dep. of Computer Science, University of Toronto.

Marcu, D. 2000a. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Marcu, D. 2000b. *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts Institute of Technology, Massachusetts, USA.

Mateo, P.L., J.C. González, J. Villena, and J.L. Martínez. 2003. Un sistema para resumen automático de textos en castellano. *DAEDALUS SA, Madrid, España*.

Mittal, V. O. and M. J. Witbrock. 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. In *SIGIR 9'9: proceedings of 22nd International Conference on Research and Development in Information Retrieval, August 1999*, page 315. University of California, Berkeley.

Ono, K., K. Sumita, and S. Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*, pages 344–348. Association for Computational Linguistics (ACL).

Paice, C.D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Papineni, K., S. Roukos, T. Ward, and W.-j. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Pollock, J.J. and A. Zamora. 1975. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226–232.

Salgueiro Pardo, T.A. and L.H. Rino Machado. 2001. A Summary Planner Based on a Three-Level Discourse Model. In *6th NLPRS - Natural Language Processing Pacific Rim Symposium*, pages 533–538.

Salton, G. and M. McGill. 1983. *Introduction to Modern Information Retrieval*. Computer Science Series, McGraw Hill Publishing, Company.

Teufel, S. and M. Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Torres-Moreno, J.-M., P. Velázquez-Morales, and J.G. Meunier. 2002. Condensés de textes par des méthodes numériques. In *JADT*, volume 2, pages 723–734.

Torres-Moreno, J.-M., P. Velázquez-Morales, and J.G. Meunier. 2001. Cortex : un algorithme pour la condensation automatique des textes. In *ARCo 2001*, pages 65–75. Lyon, France.

Turner, J. and E. Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Association for Computational Linguistics*, volume 43, pages 290–297.

Vivaldi, J. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.

Vivaldi, J., I. da Cunha, J.M. Torres-Moreno, and P. Velázquez-Morales. 2010. Automatic summarization using terminological and semantic resources. In *International Conference on Language Resources and Evaluation (LREC'10)*.

Vivaldi, J. and H. Rodríguez. 2001. Improving term extraction by combining different techniques. *Terminology*, 7(1):31–47.

Vivaldi, J. and H. Rodríguez. 2002. Medical term extraction using the EWN ontology. In *Terminology and Knowledge Engineering*, pages 137–142. Nancy.

Waszak, T. and J.-M. Torres-Moreno. 2008. Compression entropique de phrases contrôlée par un perceptron. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT'08) Lyon*, pages 1163–1173.

Yousfi-Monod, M. and V. Prince. 2006. Compression de phrases par élagage de leur arbre morpho-syntaxique. *Technique et Science Informatiques*, 25:437–468.

Yousfi-Monod, M. and V. Prince. 2008. Sentence Compression as a Step in Summarization or an Alternative Path in Text Shortening. In *Coling'08*.

Avaliação da anotação semântica do PALAVRAS e sua pós-edição manual para o Corpus Summ-it

Élen Cátia Tomazela
etomazela@yahoo.com.br

Cláudia Dias de Barros
claudias84@gmail.com

Lucia Helena Machado Rino
lucia@dc.ufscar.br

Núcleo Interinstitucional de Linguística Computacional
Universidade Federal de São Carlos
São Carlos – SP, Brasil

Resumo

Este artigo apresenta uma avaliação da anotação semântica automática do parser PALAVRAS e sua pós-edição manual para um corpus de textos em português – o Corpus Summ-it. Essa pós-edição visou ao aprimoramento de um modelo linguístico para a sumarização automática de textos e buscou atribuir etiquetas semânticas mais adequadas aos itens lexicais, comparadas às empregadas pelo parser. Essa tarefa foi realizada por linguistas e os casos problemáticos são apresentados neste artigo, os quais levam a considerações sobre o próprio modelo de etiquetagem do PALAVRAS. O corpus revisado estará disponível para a comunidade e poderá ser útil para várias aplicações de Processamento de Línguas Naturais.

1 Introdução

Este artigo tem como finalidade explicitar a avaliação da anotação semântica provida pelo *parser* PALAVRAS (Bick 2000) para os textos que compõem o Corpus Summ-it (Collovini, Carbonel et al. 2007)[1] e o processo de pós-edição manual dessa etiquetagem. Esse corpus foi construído visando à sumarização automática de textos e é utilizado, particularmente, para a modelagem de critérios de decisão do sumarizador automático VeinSum (Carbonel 2007), cujo refinamento foi proposto em (Tomazela 2010).

O corpus possui vários tipos de anotação dos textos: as anotações morfossintática e semântica produzidas pelo *parser,* a anotação de cadeias de correferência (doravante CCRs) e a anotação retórica, esta na forma de estruturas RST (Mann & Thompson 1988). Somente as providas pelo PALAVRAS foram realizadas automaticamente; as demais são resultado de trabalho manual executado por especialistas nas devidas competências.

A anotação semântica é muito relevante para a melhoria do modelo do VeinSum porque é usada para especificar heurísticas de decisão para a escolha de segmentos textuais relevantes aos sumários, os quais são produzidos com foco no fenômeno do encadeamento referencial. Por isso, a anotação de CCRs contemplou os sintagmas nominais (aqui referidos por SNs) e, ainda, ao menos um que fosse expresso por uma descrição definida, por serem estas as construções de interesse para a correferenciação: para a sumarização baseada nas anotações semânticas, outras realizações linguísticas (p.ex., as pronominais) não seriam etiquetadas, a menos que a resolução anafórica fosse realizada automática e previamente, o que não ocorre no PALAVRAS.

O contexto que motiva o uso das anotações semânticas é descrito na Seção 2, o qual motivou a avaliação de desempenho do *parser* e a pós-edição de sua anotação. A descrição do modelo de anotação semântica automática se encontra na Seção 3, seguindo-se o relato das principais características do corpus e da metodologia empregada para a correção de suas etiquetas (Seção 4). Os principais problemas de etiquetagem semântica são descritos na Seção 5, seguindo-se a avaliação do desempenho do *parser* (Seção 6).

Neste artigo, anotação, etiquetagem e etiquetas semânticas são termos adotados para referência ao processo de marcação automática de itens lexicais com suas categorias semânticas, segundo o elenco de etiquetas fornecido pelo PALAVRAS, o que faz dele, além de um gerador de estruturas sintáticas, um *parser* ou etiquetador semântico. A pós-edição refere-se unicamente à revisão e correção das etiquetas semânticas atribuídas pelo sistema a todos os SNs correferentes pertencentes ao corpus.

[1] Disponível no Portal de Corpus do NILC, http://www.nilc.icmc.usp.br:8180/portal/.

2 O modelo de sumarização automática do VeinSum

O VeinSum é um sumarizador automático que segue a abordagem profunda, isto é, ele é baseado em processamento de conhecimento linguístico (Sparck-Jones 1999) para produzir sumários de textos-fonte[2]. Segundo essa abordagem, o sistema recorre a estruturas linguísticas cujos pressupostos teóricos servem para indicar a informação relevante para um sumário e sua organização textual. Essencialmente, essa organização refere-se à preservação da ordem original da informação e não há qualquer reescrita das unidades mínimas de significado tidas como relevantes para compô-lo. Ou seja, essas unidades são meramente *copiadas-e-coladas* do texto-fonte para o sumário. Está nessa etapa de reconhecimento de unidades relevantes para incluir em um sumário, portanto, o maior esforço do sistema para obter resultados satisfatórios. Como a unidade textual mínima é a sentencial, sentenças completas são copiadas nos sumários, resultando nos principais problemas de textualidade já descritos na literatura (p.ex., (Mani 2001)).

Propôs-se resolver no VeinSum um problema particular de textualidade: o de *clareza referencial*. Diz-se que um sumário apresenta clareza referencial quando não há *quebras* de CCRs. Uma quebra de CCR, por sua vez, ocorre quando não é possível, ao leitor, identificar a quem ou a que um determinado pronome ou SN está se referindo (definição da DUC2005 – *Document Understanding Conference*)[3]. Assim, a meta principal do sistema é produzir sumários automáticos que sejam claros referencialmente.

Embora a garantia de textualidade envolva critérios intra e extralinguísticos (Beaugrande & Dressler 1981) e a própria definição de clareza referencial explicite a intervenção do leitor, a modelagem do sumarizador automático contempla somente o nível intratextual, para evitar os demais problemas da referenciação, os quais, até o momento, são intratáveis computacionalmente. Como resultado, somente o aspecto coesivo é considerado, fugindo do escopo da abordagem quaisquer outras considerações relativas à coerência textual, como as apontadas em (Halliday & Hasan 1976), (Marcuschi 1983) ou (Koch & Travaglia 2004). Logo, tratar da clareza referencial para gerar um sumário consiste em determinar automaticamente qual é a sentença com maior probabilidade de conter o antecedente mais completo de um componente anafórico já incluído no sumário.

Uma quebra de clareza referencial é evidenciada no sumário a seguir, gerado automaticamente para o texto-fonte CIENCIA_2001_6410, dado como entrada ao sistema[4]. Esse sumário contém a expressão anafórica '**o pesquisador**' sem que seu antecedente esteja explícito. O excerto do texto-fonte em que essa anáfora se insere segue o sumário. Nota-se que a menção ao antecedente se encontra na sentença imediatamente anterior, a qual foi desconsiderada pelo sumarizador automático, em sua decisão do que incluir no sumário.

> Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. Segundo **o pesquisador**, os contatos via redes de computadores estão na verdade ampliando a socialização das pessoas.

Sumário do texto CIENCIA_2001_6410

> Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. A conclusão é de **Barry Ellman**, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá. Segundo **o pesquisador**, os contatos via redes de computadores estão na verdade ampliando a socialização das pessoas.

Excerto do texto-fonte CIENCIA_2001_6410

Para tratar a clareza referencial de fato seria necessário que o sistema computacional identificasse as CCRs e apontasse seus componentes que resolvem as referências, quer elas sejam anafóricas, quer sejam de qualquer outro tipo descrito na literatura (Coelho, Muller et al. 2006). Esse é o problema que as iniciativas de

[2] A única acepção adotada aqui, para o termo *sumário*, é a de *resumo da fonte de informação.*

[3] http://duc.nist.gov/duc2005/.

[4] Todos os textos ilustrados neste artigo foram extraídos do Corpus Summ-it e seus sumários automáticos, gerados pelo VeinSum.

resolução anafórica automática pretendem resolver. No entanto, as soluções computacionais são aproximações que frequentemente carecem de qualidade, em geral porque os resolvedores anafóricos não conseguem tratar adequadamente esse fenômeno linguístico, já de natureza complexa, que demanda modelos de resolução automática incompletos ou inexatos.

No projeto do VeinSum, optou-se por manter o foco somente na questão de sumarização, evitando aumentar sua complexidade com a agregação de um módulo de resolução anafórica, muito embora a ausência desse processo seja, reconhecidamente, um dos maiores entraves para a Sumarização Automática (Mitkov 1998; Cristea, Postolache et al. 2003) e, em geral, para os sistemas de PLN[5] (Mitkov 2002; Chaves 2007).

A proposta alternativa para buscar a clareza referencial foi a de fazer o sistema manipular as estruturas RST dos textos-fonte. Assim, qualquer texto a sumarizar é, primeiramente, estruturado retoricamente e é a partir de sua estrutura RST que se busca determinar quais as unidades textuais a incorporar aos sumários.

Além de não resolver anáforas explicitamente, o VeinSum sequer é capaz de detectar os termos anafóricos. Na verdade, ele procura delimitar os contextos de possíveis unidades correferentes (os quais incluem as possíveis anáforas e seus antecedentes) somente com base nas estruturas RST, ou seja, na sua posição nas árvores dos textos-fonte. Essa delimitação dos contextos correferenciais fica a cargo da Teoria das Veias, ou VT (Cristea, Ide et al. 1998). Associada à RST, ela o faz com base no *domínio de acessibilidade referencial* (doravante, *acc*) de cada unidade textual da árvore.

O *acc* é, assim, o conjunto de todas as unidades que possam fazer parte da CCR de uma unidade anafórica, a qual também é incluída nesse conjunto. Na ausência da resolução anafórica como tal, o *acc* se constitui, portanto, das sentenças do texto-fonte que, hipoteticamente, são correferentes. Esse é o ponto de partida do VeinSum para buscar manter a clareza referencial dos sumários.

O problema do sistema pode ser descrito, portanto, como o problema de se reconhecer, dentre as N unidades textuais que compõem um texto-fonte e que se encontram relacionadas em sua estrutura RST, quais são as M unidades (M menor que N) que comporão o sumário correspondente, sem que haja quebra da clareza referencial. A VT, juntamente com a RST, fornece todos os *acc*s das sentenças do texto-fonte.

Para indicar quais as M sentenças que serão escolhidas, agrega-se aos dois modelos anteriores o Modelo de Saliência (Marcu 2000), que indica a classificação de saliência das N unidades a partir da qual as M unidades são escolhidas. Ante as restrições de saliência, clareza referencial e taxa de compressão (restrição fundamental da sumarização automática), que são consideradas em conjunto, o sistema finalmente produz o sumário integral.

Um dos motivos da fragilidade dos resultados do VeinSum, como o ilustrado pelo sumário do texto CIENCIA_2001_6410, é que, ao ter que obedecer à taxa de compressão, se necessário o sistema relaxa a restrição de saliência, desprezando sentenças mais salientes para manter integralmente os *acc*s de unidades já escolhidas para compor o sumário. Com isso, informações mais importantes do texto podem ser desprezadas, prejudicando a qualidade do sumário, quando comparado ao seu texto-fonte.

O que gerou a proposta de refinamento de Tomazela (2010) foi a observação de que os *acc*s também poderiam ser reduzidos, pois os contextos de prováveis unidades correferentes apontados pela VT não asseguram, de fato, quais delas são essenciais para a clareza referencial. No melhor caso, bastaria manter, do *acc*, as sentenças que contêm a anáfora e a que contém seu antecedente mais completo.

Assim, na tentativa de reduzir os *acc*s, propôs-se o uso de informações semânticas provindas da anotação do PALAVRAS como coadjuvante dos modelos descritos. Supôs-se, nesse caso, que o problema não estaria na estruturação RST, nem na determinação dos *acc*s de cada componente textual, muito embora tanto a RST quanto a VT tragam reconhecidos problemas para a manipulação de segmentos textuais (Cristea, Postolache et al. 2005; Carbonel 2007; Tomazela & Rino 2009).

Em linhas gerais, buscando selecionar menos sentenças de cada *acc*, o novo sumarizador procede da seguinte forma: uma vez escolhida uma sentença para compor o sumário, as etiquetas semânticas dos núcleos de cada um de seus SNs são usadas para buscar o provável antecedente de uma anáfora hipotética dessa sentença. Esse é apontado como a unidade do *acc* que contenha um ou mais SNs com maior similaridade semântica com os núcleos dos SNs da unidade já escolhida.

É, portanto, a similaridade semântica entre componentes de várias sentenças apontadas no

[5] **P**rocessamento automático de **L**ínguas **N**aturais.

acc que irá indicar a possibilidade de manter a clareza referencial no sumário e, ao mesmo tempo, permitir que a classificação das unidades salientes seja respeitada, para melhor aproximação com a preservação das informações mais relevantes do texto-fonte.

O problema recai, portanto, em como distinguir componentes mais similares – aqueles que possam indicar uma ligação forte de correferência. Isso é feito traçando-se a relação entre as etiquetas semânticas fornecidas pelo PALAVRAS, para os SNs em foco, isto é, os SNs que possivelmente sejam correferentes,. Com base nessa ideia, é que se buscou definir heurísticas para a escolha das unidades relevantes que atendessem aos critérios de similaridade semântica, ditados por um modelo de similaridade baseado na distribuição das etiquetas num corpus (Tomazela 2010). Esse modelo é descrito na próxima seção.

3 O modelo de anotação semântica do PALAVRAS

O processamento semântico do PALAVRAS visa à atribuição de uma etiqueta semântica que indique, *aproximadamente,* o significado de cada item lexical de um texto. Para isso, não se consideram modelos clássicos de semântica lexical, nos quais se buscam significados através de definições dicionarizadas ou por uma classificação ontológica, mas sim, combinações de traços semânticos, os quais fornecem uma identidade ao item lexical. Essa anotação conta com 215 etiquetas semânticas e se baseia em 16 traços, os quais supostamente representam o contexto semântico de quaisquer conceitos usados na produção de uma mensagem (Bick 2000). Note-se que essa concepção implica considerar o modelo semântico independente de língua natural.

Nesse modelo de classificação, são considerados somente os substantivos, entidades nomeadas e alguns adjetivos, para os quais é possível atribuir um valor semântico. As entidades nomeadas, neste trabalho, são o mesmo que entidades mencionadas (Santos 2007), denotadas por nomes próprios que podem indicar nomes de pessoas, organizações, acontecimentos, locais, coisas, obras e conceitos abstratos.

A identificação de itens lexicais similares é atribuída à chamada similaridade prototípica, a qual permite colocar em contexto de uso a configuração semântica, sem que se necessite de coincidências absolutas de significado. Essa medida de similaridade de cada item lexical é proporcional ao número de traços semânticos que compartilham: Bick supõe que, quanto maior esse número, mais similares são os itens lexicais. Daí a possibilidade de agregar, em um único conjunto ou, no caso de interesse para o VeinSum, em uma única heurística, etiquetas semânticas que indiquem itens lexicais possivelmente correferentes.

Foi essa ideia de similaridade semântica baseada nas etiquetas do PALAVRAS que motivou a proposta de se definirem heurísticas para a sumarização automática de textos em português. Porém, ao se analisar a anotação semântica do Corpus Summ-it, descobriram-se vários casos de inadequação da etiquetagem automática, residindo aí a motivação para a sua pós-edição manual e consequente avaliação do *parser* apresentadas neste artigo. O Corpus Summ-it foi o instrumento central para a engenharia do conhecimento visando à formalização de todo o processo.

4 O Corpus Summ-it

O Corpus Summ-it configura-se como o primeiro corpus anotado manualmente com CCRs para textos jornalísticos em português. Foi construído para atender a diversos interesses, dentre os quais os de pesquisa e desenvolvimento de sistemas de sumarização automática de textos, uma das principais áreas de pesquisa do NILC[6]. É composto de 50 textos do caderno de Ciências da Folha de São Paulo, cada um deles de tamanho que varia de 27 a 654 palavras(1/2 a 1 ½ página em formato A4). Os textos contêm de 3 a 24 CCRs, totalizando 589 CCRs no corpus todo. A CCR mais longa contém 16 SNs e a mais curta, apenas 2.

A importância da anotação semântica para o VeinSum se deve ao fato de ele usar os *accs* como conjuntos indicativos do contexto de ocorrência de CCRs e estas terem seus SNs já anotados semanticamente. Isso permite elaborar o processo de determinação dos segmentos a compor um sumário proposto como melhoria do sistema. O fato de os *accs* serem derivados das estruturas RST dos textos-fonte justifica a existência da anotação RST do corpus todo. Entretanto, para o novo processo de minimização dos *accs* ocorrem dois entraves: não se sabe qual o SN anafórico, tampouco qual o SN que poderia ser seu antecedente, daí a busca de heurísticas que possam indicar possíveis contextos correferentes pelas etiquetas semânticas dos componentes dos

[6] Núcleo Interinstitucional de Linguística Computacional, sediado em São Carlos, SP - http://www.nilc.icmc.usp.br/nilc/index.html.

accs. Também por essa razão e pela verificação de problemas na etiquetagem automática, originou-se a necessidade de se revisar os resultados do PALAVRAS. Dessa forma, tentou-se garantir a confiabilidade das heurísticas a incorporar ao VeinSum.

4.1 A necessidade de revisão do corpus

Como o foco é simplesmente a minimização dos *accs,* a pós-edição do corpus Summ-it se restringiu aos SNs que aparecem nas CCRs. Particularmente, foram analisados os substantivos desses SNs, já que eles são os únicos que contêm etiquetas semânticas expressivas, como já mencionado.

Nesta seção relatam-se os principais desvios de anotação semântica do PALAVRAS para os itens lexicais em questão. Foram identificados três problemas significativos: o de segmentação, o de etiquetagem, e o de desambiguação das etiquetas semânticas. Certamente esses problemas são interdependentes: a má segmentação textual interfere nos demais. A desambiguação de sentido é, na verdade, um problema da própria etiquetagem: etiquetas equivocadas podem ser atribuídas por não haver uma determinação clara (ou menos problemática) do significado de algum componente textual. Mesmo a anotação morfossintática (em inglês, *POS tagging*) depende da segmentação, por um lado, e interfere no desempenho semântico, por outro. Ou seja, uma má segmentação textual constitui o primeiro entrave para os demais processos, fato amplamente reconhecido na área de PLN (Pardo & Nunes 2002). Sobretudo no caso de CCRs, julgou-se que a segmentação inadequada pode corromper o encadeamento referencial e levar a problemas sérios para que o modelo de decisão do VeinSum assegure a clareza referencial.

4.2 Os pressupostos para a revisão do corpus

Primeiramente, para manter a tarefa de pós-edição consistente, determinou-se que ela seria feita por duas linguistas especialistas no elenco de etiquetas semânticas do PALAVRAS[7] e que a concordância em suas decisões seria assegurada (Pressuposto 1). Desse modo, o corpus Summ-it revisado pode servir à engenharia de conhecimento na Sumarização Automática (para avaliação ou validação de sistemas), mas também a tarefas que não as de PLN, como as de Linguística de Corpus.

Adotou-se por base as definições originais das etiquetas, o que levou à limitação das possíveis acepções a suas denotações fixas (Pressuposto 2). Esse método de análise semântica está em consonância com as instruções usuais da MUC (*Message Understanding Conference*), ao contrário do que é sugerido em (Santos & Cardoso 2007).

Um exemplo claro da aplicação dessa estratégia no processo de revisão refere-se ao uso metonímico de termos que indicam localizações[8]: no Summ-it, há ocorrências de 'Brasil', cuja etiqueta semântica categorial (ou denotacional) é **<Lciv>** (*Civitas, town, country, county, cidade, país*). Entretanto, o PALAVRAS atribui a esses termos, muitas vezes, a etiqueta **<inst>** *(institution*), claramente metonímica. Todos esses casos foram alterados para <Lciv>.

O terceiro pressuposto foi o de que a revisão em curso não infringiria os próprios pressupostos do PALAVRAS, de que protótipos semânticos podem ser usados para indicar a similaridade (ou dissimilaridade) entre vários itens lexicais. Ao contrário, seria possível usar a prototipagem para fundamentar a revisão – e, certamente, para traçar mecanismos de sumarização automática (Pressuposto 3).

Finalmente, perseguiu-se a perspectiva de que o PALAVRAS se destina ao processamento do português e, assim, tem seu elenco de etiquetas igualmente aplicável e reusável para o processamento dos textos nessa língua, os quais são os objetos de interesse para a sumarização automática em curso.

Vale notar que, exceto pelo uso do elenco de etiquetas semânticas do PALAVRAS, cuja dependência de qualquer língua-fonte pode ser questionada, as demais linguagens de representação adotadas nas anotações do corpus (estruturas RST e estruturas de veias, no caso) são independentes de língua natural.

4.3 A preparação do corpus e a metodologia de revisão

Para evitar que constantes atualizações do PALAVRAS prejudicassem a consistência da tarefa de revisão manual das etiquetas semânticas,

[7] Acessível pela Internet: (http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags).

[8] Conforme discussão de Santos (2007).

adotou-se sua versão de fevereiro de 2007 e, assim, o elenco das 215 etiquetas semânticas, juntamente com suas definições, foi mantido constante[9].

Foram utilizadas diretamente as saídas do sistema para cada um dos 50 textos do corpus: arquivos XML. Esses dados de cada texto foram agrupados em uma única planilha Excel, a qual consistiu o material de trabalho das especialistas linguistas. Como já mencionado, restritos os SNs aos componentes de CCRs, para cada texto somente as anotações semânticas desses dados constam da planilha. As correções das etiquetas foram inseridas também nesse arquivo, de forma que toda síntese numérica (totalizações de casos, estatísticas de ocorrência, etc.) necessária para a análise foi produzida automaticamente, via programação no próprio ambiente da planilha Excel. A partir desse processamento foi possível avaliar o desempenho do PALAVRAS para textos isolados ou em conjunto, resultando na síntese apresentada na Seção 5.

5 A pós-edição do Corpus Summ-it

Considerando a interdependência entre a segmentação e os demais processos do PALAVRAS, relatam-se aqui primeiramente os casos problemáticos de segmentação, para depois apresentar-se os problemas de etiquetagem semântica, propriamente ditos. Maiores detalhes dessa tarefa podem ser encontrados em (Tomazela & Rino 2010).

5.1 Problemas de segmentação

Os casos mais problemáticos de segmentação textual do PALAVRAS residiram na confusa identificação de lexias complexas e de entidades nomeadas. Várias lexias complexas foram consideradas lexias simples, ou seja, foram processadas em componentes separados, com o desmembramento de uma única entidade em vários SNs. Esse padrão foi identificado para as lexias compostas de 'substantivo + adjetivo', como nos seguintes exemplos do corpus:

- 'vaso sanguíneo', sendo 'vaso' etiquetado com **<con>** (*container*) e 'sanguíneo' ignorado. A lexia deveria ser etiquetada com **<an>** (*anatomical noun, umbrela tag - carótida, dorso).*
- 'cadeia evolutiva', sendo 'cadeia' etiquetada com **<inst>** (*institution),* em vez de **<ax>** (Abstract/concept, neither countable nor mass – *endogamia)* e 'evolutiva' ignorada;
- 'batimento cardíaco', sendo 'batimento' etiquetado com **<act>** (*Action, umbrella tag - +CONTROL, PERFECTIVE)*, em vez de **<process>** (process -CONTROL, -PERFECTIVE, cp. <event>, *balcanização, convecção, estagnação*) e 'cardíaco' ignorado.

Esses exemplos evidenciam que haverá prejuízo para a identificação de termos correferentes com a etiquetagem independente: os adjetivos ignorados é que realmente determinam o significado das lexias complexas.

Opostamente a esses casos, a ferramenta aglutinou vários SNs em uma única entidade nomeada, conforme os seguintes exemplos:

- Pesquisadores do Museu Nacional do Rio de Janeiro

  Aqui tem-se o SN 'Pesquisadores' e as entidades nomeadas "Museu Nacional" e "Rio de Janeiro'. O *parser* atribui a etiqueta **<hum>** (*person name)* para todo esse trecho, pois o considera uma única entidade. No entanto, três etiquetas distintas deveriam ter sido atribuídas: 'Pesquisadores', com etiqueta **<Hprof>** (Professional human – *marinheiro*), 'Museu Nacional", com **<org>** (*commercial or non-commercial, non-administrative, non-party organisations*) e 'Rio de Janeiro', com **<civ>** (*civitas - country, town, state, cp. <Lciv>*).

- Organização das Nações Unidas

  O *parser* etiquetou separadamente os seguintes itens lexicais: 'Organização', com **<np-close>**, cuja definição não é encontrada no elenco de etiquetas; 'Nações', com **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora)* e 'Unidas' não recebeu etiqueta semântica alguma. Caso essa entidade nomeada não fosse desmembrada, sua etiqueta deveria ser **<org>.**

Considerou-se que, ao não se reconhecer entidades nomeadas de um texto, a proposta de identificação de elementos correferentes por suas etiquetas semânticas se tornaria mais difícil.

[9] Embora essa preocupação seja procedente, o elenco permanece o mesmo até a presente data.

Entretanto, esta suposição merece uma investigação mais profunda no futuro.

No total, foram identificados 104 casos problemáticos de segmentação (vide seção 5.3), os quais incluem a identificação de lexias complexas e entidades nomeadas.

5.2 Problemas de etiquetação

Primeiramente apresentam-se alguns detalhes sobre o procedimento de verificação da etiquetagem, para depois relatarem-se alguns casos pitorescos do corpus.

5.2.1 Especificidades da revisão

De forma geral, qualquer correção de etiquetas atribuídas pelo sistema aos itens lexicais somente se deu quando a etiqueta apresentou desvios semânticos consideráveis. Nesse caso, optou-se por utilizar etiquetas mais específicas sempre que possível. Entretanto, as etiquetas genéricas produzidas pelo sistema foram mantidas sempre que julgadas apropriadas, buscando não penalizar excessivamente a avaliação de desempenho pretendida. Ou seja, somente foram alterados os casos em que ou a etiqueta era claramente indevida, por ser conflitante com os traços semânticos do componente lexical, ou a etiqueta era tão específica que não correspondia ao seu significado adequado. Nesse caso, adotou-se uma etiqueta referente a um conceito mais geral. Exemplos disso ocorrem para os itens lexicais 'bicho' e 'animal', ambos etiquetados com **<Azo>** (*land animal).* Considerou-se essa etiqueta restritiva porque as acepções desses itens lexicais no corpus em foco abrangem também animais aquáticos. A etiqueta mais genérica atribuída foi, portanto, **<A>** (*Animal, umbrella tag - clone, fêmea, fóssil, parasito, predador)*, a qual, na existência de uma ontologia apropriada, seria considerada um hiperônimo da etiqueta **<Azo>**.

Considerou-se o contexto de ocorrência dos itens lexicais e, assim, recorreu-se aos textos-fonte correspondentes, sobretudo quando se necessitou interpretar itens lexicais anafóricos cujos referentes não estavam acessíveis na planilha Excel.

Também foi necessário verificar os casos de delimitação das entidades nomeadas, para atribuir-lhes uma única etiqueta a partir de sua análise como um todo (afinal, a semântica de um componente desse tipo não é a soma da semântica de suas partes).

Quando não se conseguiu definir a melhor etiqueta para corrigir a automática, recorreu-se ao tópico (ou assunto) do texto, para traçar seu interrelacionamento. Por exemplo, a menção anafórica 'os pesquisados' em uma certa CCR pode se referir a pessoas, animais, medicamentos ou produtos. A partir do tópico principal do texto em que está inserida (CIENCIA_2000_17101), expresso pelo segmento "a alteração da Declaração de Helsinque, na qual os cientistas não se obrigariam a fornecer aos doentes o melhor tratamento conhecido para uma doença", é possível determinar que esse SN se refere a 'os doentes' e, portanto, deve ser etiquetado com **<H>** (*human, umbrella tag).*

Mediante esses casos, vale lembrar que o *parser* não propõe fazer resolução anafórica e, por isso, não tem obrigação de reconhecer esses antecedentes. Porém, ao não fazê-lo, produz etiquetas que podem trazer problemas à clareza referencial dos sumários.

Analisou-se ainda o aspecto dos itens lexicais, particularmente quando indicavam eventos, ações, atividades ou processos. Nesses casos há etiquetas específicas que distinguem a valência (+/-) do traço semântico PERFECTIVE: +PERFECTIVE indica conceito pontual; –PERFECTIVE, conceito progressivo. Distinguiram-se, também, as valências do traço semântico CONTROL, isto é, se os conceitos apresentados eram passíveis ou não de serem controlados. As etiquetas que tratam desses casos são indicadas abaixo:

- **<activity>** (*Activity, umbrella tag - +CONTROL, IMPERFECTIVE, correria, manejo);*
- **<act>** (*Action, umbrella tag - +CONTROL, PERFECTIVE);*
- **<event>** (*event, -CONTROL, PERFECTIVE, milagre, morte)*;
- **<process>** (*process, -CONTROL, –PERFECTIVE, cp. <event>, balcanização, convecção, estagnação)*[10].

Caso as estratégias relativas ao contexto de ocorrência e às definições das etiquetas ainda não fossem suficientes para determinar etiquetas apropriadas, recorreu-se à WordNet (Fellbaum 1998), para buscar seus traços semânticos.

5.2.2 Ocorrências problemáticas no Corpus Summ-it

Destacam-se, aqui, alguns dos problemas de etiquetação mais significativos no corpus:

[10] Entende-se' –PERFECTIVE' como 'IMPERFECTIVE', neste caso.

- Nomes científicos, muito presentes nos textos do corpus em uso, quase sempre são etiquetados ou segmentados erroneamente. '*Tyrannossaurus rex*', p.ex., é etiquetado com **<inst>**, quando deveria receber a etiqueta **<meta>** (meta noun - *tipo*, *espécie*)**.**

- 'células-tronco', quando inicia a oração, recebe etiqueta **<Acell>** (*Cell-animal - bacteria, blood cells: linfócito)*; quando ocorre intraoracionalmente, recebe etiqueta **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora)*, o que contradiz o fato de serem correferentes, já que '*animal celular*' não pode ser correferente a um '*grupo de humanos*'.

- Apesar de sinônimos, alguns itens lexicais correferentes apresentam etiquetas diferentes, como: 'cachorro' **-** etiquetado com **<Azo>** (*Land-animal - raposa)* e 'cão' etiquetado com **<Adom>** (*Domestic animal or big mammal - terneiro, leão/leoa, cachorro)*. O que justifica o fato de 'cachorro' ser animal terrestre e 'cão' ser animal doméstico não é claro.
- Caso análogo ocorre com 'CO2', que recebe etiqueta **<cm-chem>** (*chemical substance, also biological - acetileno, amônio, anilina, bilirrubina)* e 'gás carbônico', etiquetado com **<mat>** (material - *argila, bronze, granito*, cf. <cm>).
- O item lexical 'atmosfera' recebe etiquetas diferentes dependendo da palavra que o segue, como em: 'atmosfera da Terra' e 'atmosfera terrestre', com etiquetas **<Ltop>** (*Geographical, natural place - promontório, pântano)* e **<sit>** (*psychological situation or physical state of affairs - reclusão, arruaça, ilegalidade, more complex & more "locative" than <state> & <state-h>)* respectivamente.

Esses exemplos sugerem que não há tratamento de sinonímia no PALAVRAS, o que também compromete o modelo de busca de itens correferentes. Eles constituem alguns dos exemplos mais problemáticos observados na revisão. A desambiguação de itens lexicais também se mostrou frágil.

5.2.3 A etiquetagem de itens ambíguos

Para determinar o significado adequado, vários fatores entram em perspectiva, sendo dos mais significativos o contexto de ocorrência do item lexical. Se o modelo semântico do *parser* pretende apontar as etiquetas semânticas aproximadas para componentes textuais, ele deveria prover mecanismos para tratar esses fenômenos. Um exemplo dessa deficiência ocorre com o item lexical 'clone', com etiqueta **<H>,** a qual somente se refere a clones humanos. No entanto, os contextos de ocorrência desse item no corpus mostram que esse termo se aplica a clones de animais e, assim, a etiqueta utilizada deveria ser a mais genérica **<A>**.

Já a desambiguação de vários itens lexicais em SNs compostos seria beneficiada se seu interrelacionamento fosse considerado na determinação do significado. O PALAVRAS não parece considerar esse contexto de ocorrência, como ilustram os exemplos a seguir:

- 'as patas e bacia do animal', em que 'bacia' recebe etiqueta **<con>** (*container),* quando deveria receber **<anmov>** (*Movable anatomy - arm, leg, braço, bíceps, cotovelo);*
- 'a física nuclear Eva Maria', em que 'física' recebe etiqueta **<domain>** (*subject matter, profession, cf. <genre>, anatomia, citricultura, datilografia)*, quando deveria ser **<Hprof>** (Professional human - *marinheiro*, also sport, hobby - *alpinista*)*;*
- 'populações de pinguins', em que 'populações' recebe etiqueta **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora),* em vez de **<AA>** (*Group of animals - cardume, enxame, passarada, ninhada*)*;*
- 'esqueleto do navio', em que 'esqueleto' recebe etiqueta **<Hmyth>** (*Humanoid mythical - gods, fairy tale humanoids, curupira, duende),* em vez de **<part-build>** (structural part of building or vehicle - *balustrada, porta, estai*).
- 'filhote' é etiquetado com **<H>** (*Human, umbrella tag)* quando o sentido de animal – etiqueta **<A>** – indicado pelo contexto é ignorado.

Esses exemplos evidenciam a necessidade de um tratamento automático mais elaborado para os casos que envolvem aspectos contextuais.

5.3 Síntese da pós-edição manual do corpus

A Tabela 1 mostra os dados gerais de correção do corpus ('SUBSTs', aqui, é limitado aos substantivos de SNs presentes nas CCRs do corpus). A média de correções de etiquetas semânticas no corpus foi de 41%. A porcentagem de erros de segmentação foi de 4%. Essa baixa porcentagem demonstra que eventuais problemas de etiquetação morfossintática ou semântica não foram causados significativamente pela segmentação automática do PALAVRAS, no corpus Summ-it. Não foi analisada isoladamente a influência da etiquetagem morfossintática na etiquetagem semântica.

Do elenco total de etiquetas (215), somente 115 ocorreram no Corpus Summ-it, segundo a revisão manual aqui relatada. Elas são reproduzidas na Tabela 2.

No tocante aos pressupostos desse trabalho, essa revisão constitui somente o passo inicial para se verificar a adequação da estratégia a outros corpora e, assim, a consistência da revisão aqui apresentada, reafirmando o Pressuposto 1. A limitação da revisão a denotações fixas (Pressuposto 2) certamente é um fator limitante. Porém, considerando-se a perspectiva de se ter um modelo automático, ela representa uma decisão razoável a se adotar, corroborada, inclusive, pelas diretrizes da MUC.

Entretanto, a questão mais polêmica sugerida pela análise aqui descrita diz respeito ao Pressuposto 3, isto é, à incorporação da ideia de protótipos semânticos que propiciem o reconhecimento de entidades similares ou dissimilares. Garantir isso pareceu impossível, dada a especificidade da classificação proposta por Bick (inclusive o fato de ela se basear em corpora de textos), ao fato de ela se inserir no contexto de tradução automática e, até, à necessidade, em alguns casos, de se buscar os vínculos em contexto para se determinar as etiquetas mais adequadas.

Particularmente, buscar a base teórica para a definição dos protótipos semânticos do *parser* foi uma tarefa difícil. Mesmo a forma como Bick propõe obter as categorias baseadas nos protótipos não está clara: foi feita com base em corpora, visando especialmente a tradução automática, com o norueguês e o inglês como línguas interagentes. Entretanto, o *parser* está disponível para anotação de textos em português.

Essas limitações pareceram bastante severas para o reuso das etiquetas e também para a interpretação de sua definição ante a tarefa de revisão. Vale ressaltar que a opção de se escolher sempre uma etiqueta mais genérica (opção plausível em diversas aplicações) não esteve em foco porque ela não permitira alcançar o objetivo de distinguir elementos correferentes. Além disso, o contexto de sumarização automática em foco pode introduzir variações do contexto original ou, até, ser inadequado para se buscar similaridades semânticas pela diferenciação de equivalentes de tradução (Santos 1990). Esta é uma questão ainda em aberto.

Por fim, a geração das heurísticas baseadas na idéia de proximidade semântica das etiquetas que pudessem indicar elementos correferentes foi dificultada porque não foi possível mediante o Pressuposto 3, traçar uma relação clara com a idéia de prototipagem semântica pelo reconhecimento de equivalentes. Esta questão não está em foco neste texto, mas é abordada em (Tomazela 2010).

Ressalta-se que as heurísticas foram geradas somente depois da pós-edição manual da etiquetagem semântica porque, se assim não fosse (isto é, se elas fossem geradas a partir do corpus diretamente anotado pelo PALAVRAS), elas seriam obviamente inválidas e não serviriam ao propósito deste trabalho, pois não assegurariam a indicação de possíveis itens lexicais correferentes.

6 Avaliação do desempenho do PALAVRAS

Dentre as principais dificuldades encontradas no processo de correção das etiquetas semânticas estão: i) a atribuição de etiquetas para itens lexicais de domínios específicos do conhecimento; ii) a inadequação das definições das etiquetas e de seus exemplos, presentes no PALAVRAS; iii) o reconhecimento de etiquetas muito genéricas, muito específicas ou ainda muito abstratas; iv) a dificuldade de adequação de um item lexical a uma única etiqueta, já que muitos deles podem ser etiquetados de várias formas.

O caso (i) foi particularmente complicado, pois, apesar de o corpus ser de domínio geral, há textos de assuntos muito particulares para algumas áreas da ciência. Para esses, o conhecimento especialista foi crucial e as linguistas precisaram recorrer a especialistas das áreas em foco, para determinar as etiquetas que melhor refletissem a natureza dos itens lexicais.

Texto-fonte	# SUBSTs	# SUBSTs corrigidos	# erros de segmentação	% Correção das etiquetas
CIENCIA_2005_6507	24	16	2	66.67%
CIENCIA_2003_6465	41	27	4	65.85%
CIENCIA_2003_24212	106	65	4	61.32%
CIENCIA_2001_19858	63	38	6	60.32%
CIENCIA_2005_28752	72	43	2	59.72%
CIENCIA_2001_6423	17	10	2	58.82%
CIENCIA_2001_6410	27	15	4	55.56%
CIENCIA_2000_17088	62	34	1	54.84%
CIENCIA_2002_22029	99	52	1	52.53%
CIENCIA_2002_6441	21	11	0	52.38%
CIENCIA_2000_6381	60	31	1	51.67%
CIENCIA_2000_17113	76	39	1	51.32%
CIENCIA_2005_28764	98	50	0	51.02%
CIENCIA_2000_17108	55	28	1	50.91%
CIENCIA_2000_6389	31	15	0	48.39%
CIENCIA_2004_26417	52	25	7	48.08%
CIENCIA_2005_28755	82	38	2	46.34%
CIENCIA_2000_17101	59	27	1	45.76%
CIENCIA_2002_22023	60	27	1	45.00%
CIENCIA_2005_28754	65	29	4	44.62%
CIENCIA_2002_22015	70	31	3	44.29%
CIENCIA_2004_6480	50	22	0	44.00%
CIENCIA_2003_24226	84	36	3	42.86%
CIENCIA_2005_28756	75	31	0	41.33%
CIENCIA_2001_6414	30	12	1	40.00%
CIENCIA_2004_26415	33	13	1	39.39%
CIENCIA_2005_28766	107	42	8	39.25%
CIENCIA_2002_22027	91	35	1	38.46%
CIENCIA_2000_17082	37	14	1	37.84%
CIENCIA_2000_17109	75	28	1	37.33%
CIENCIA_2004_6494	30	11	7	36.67%
CIENCIA_2005_6515	41	15	0	36.59%
CIENCIA_2003_6472	22	8	0	36.36%
CIENCIA_2005_28774	85	30	0	35.29%
CIENCIA_2000_17112	54	18	6	33.33%
CIENCIA_2001_6406	21	7	0	33.33%
CIENCIA_2005_6514	37	12	0	32.43%
CIENCIA_2004_26423	115	37	10	32.17%
CIENCIA_2005_6518	45	14	0	31.11%
CIENCIA_2004_6488	13	4	0	30.77%
CIENCIA_2001_6416	43	13	1	30.23%
CIENCIA_2000_6391	41	12	2	29.27%
CIENCIA_2000_6380	31	9	0	29.03%
CIENCIA_2005_28747	42	12	4	28.57%
CIENCIA_2004_26425	99	23	1	23.23%
CIENCIA_2003_24219	81	17	4	20.99%
CIENCIA_2002_22005	62	12	4	19.35%
CIENCIA_2002_22010	36	6	0	16.67%
CIENCIA_2003_6457	45	6	2	13.33%
CIENCIA_2005_28743	35	1	0	2.86%
TOTAIS	**2800**	**1151**	**104**	**2027.46%**

Tabela 1 – Quadro geral de correção da anotação semântica do Corpus Summ-it

A	Aorn	coll-cc	Hbio	mat	sick
AA	Azo	coll-sem	Hfam	meta	sick-c
absname	B	con	HH	mon	sit
ac	BB	conv	Hideo	month	site
ac-cat	build	cord	Hnat	object	suborg
Acell	Bveg	dir	Hprof	occ	temp
ac-sign	cc	disease	Hsick	org	therapy
act	cc-board	domain	hum	part	tool
act-d	cc-fire	drink	inst	part-build	tube
activity	cc-r	dur	L	party	unit
act-s	cc-rag	event	Labs	per	V
admin	cc-stone	f	Lciv	percep-w	Vair
Adom	civ	f-c	Lcover	pict	virtual
Aent	cm	f-h	Lh	piece	VV
am	cm-chem	food	ling	plan	Vwater
amount	cm-gas	food-h	Lopening	process	
an	cm-liq	f-q	Lstar	pub	
anbo	cm-rem	fruit	Lsurf	sem-c	
anmov	col	H	Ltop	sem-r	
anorg	coll	Hattr	Lwater	sem-s	

Tabela 2 – Etiquetas ocorrentes no corpus

O caso (ii) levou a uma grande dificuldade para a análise semântica, pois nem os exemplos fornecidos com o elenco de etiquetas foram suficientes para deixar claras muitas das definições. Etiquetas diferentes destinam-se a designar objetos semânticos diferentes, porém, quando se analisam os exemplos que acompanham suas definições, elas não parecem se diferenciar em nenhum aspecto. Esse é o caso de **<cc-r>** (*read object - carteira, cupom, bilhete, carta, cf. <sem-r>)* e **<sem-r>** (*read-work - biografia, dissertação, e-mail, ficha cadastral),* que indicam, respectivamente, uma descrição de um objeto de leitura e de um trabalho de leitura. Essas definições sugerem que o que se pretende distinguir é o modo de produção das obras escritas: **<cc-r>** seria relativa àquelas de produção simples, enquanto **<sem-r>**, às de produção complexa. Nesse caso, 'e-mail' e 'ficha cadastral', por requerer produção simples, não deveriam ser exemplos de **<sem-r>.**

Há ainda etiquetas cuja definição se aplica a objetos semanticamente díspares, como **<Adom>** (*Domestic animal or big mammal - terneiro,* leão*/leoa, cachorro),* que, contraditoriamente, trata tanto de animais domésticos quanto de grandes mamíferos. Seria mais conveniente que essa disparidade fosse resolvida com etiquetas mais específicas, que diferenciassem animais domésticos e pequenos mamíferos de animais selvagens ou de grandes mamíferos.

Exemplos do caso (iii) são as etiquetas que, de tão específicas, têm pouca utilidade. Esse é o caso de **<anich>** (*Fish anatomy - few: brânquias, siba)* e **<cc-board>** (*flat long object - few: board, plank, lousa, tabla)*, reconhecidas pelo próprio autor da ferramenta (pela palavra "few" em suas definições) como raramente aplicadas aos itens lexicais de qualquer dos *corpora* investigados.

Caso similar ocorreu com as etiquetas de definições muito abstratas, como **<ac-cat>** (*Category Word - latinismo, número* atômico*)*, corroborando o fato de que as especificações providas para o uso desse elenco não são significativamente esclarecedoras.

O fato de algumas etiquetas serem ontologicamente relacionadas[11] dificultou o processo de revisão dos resultados automáticos, já que muitos itens lexicais podiam ser enquadrados em mais de uma etiqueta (caso (iv)). Isso ocorre, p.ex., com **<fruit>** (*fruit, berry, nut - still mostly marked as <food-c>, abricote, amora, avelã,*

[11] Embora o modelo semântico do PALAVRAS não se baseie em uma ontologia (Bick 2000), é inegável a possibilidade de tratar pelo menos parte delas ontologicamente.

cebola) e **<food-c>** (*countable food - few: ovo, dente de alho, most are <fruit> or <food-c-h> culinary countable food - biscoito, enchido, panetone, pastel).* Certamente, as duas etiquetas são apropriadas para alguns itens lexicais, porém optou-se por utilizar a etiqueta mais específica nesses casos.

Além dos casos acima, ocorrências menos significativas, mas não desprezíveis do ponto de vista da proposta semântica do PALAVRAS, foram elencadas. Verificou-se, dentre elas, que o elenco das 215 etiquetas não foi suficiente para descrever alguns itens lexicais comuns. 'vírus', por exemplo, é etiquetado inadequadamente com **<Acell>** - *Cell-animal (bacteria, blood cells: linfócito)*, pois não é um *animal celular*, mas sim "uma partícula proteica que infecta organismos vivos"[12]. A etiqueta mais próxima a ser atribuída a esse item lexical seria **<cc>** - *concrete countable,* porém, por ser muito genérica, ficou difícil determinar, pelo contexto, sua aplicabilidade. Decidiu-se, assim, manter **<Acell>.** Vale ressaltar que esse foi o único caso de manutenção de etiqueta quando claramente imprópria.

Outras etiquetas são classificadas por Bick como *vazias*, como **<cc-h>** (*artifact, umbrella tag - so far empty category in PALAVRAS)* e parecem se associar a casos não previstos (indicação dada pelo termo *umbrella tag*). No entanto, na ausência de etiquetas adequadas, a escolha pelas ditas *vazias* foi considerada.

Há ainda as marcadas como 'Further proposed categories', para as quais não há definições ou não há exemplos, constituindo-se, assim, em etiquetas subespecificadas. **<spice>** é um caso de ausência completa de descrição; **<top>** (*geographical location)* e **<Bveg>** (v*egetable, espargo, funcho*), de subespecificação.

O uso da etiqueta **<meta>** (*meta noun - tipo, espécie)* também não ficou claro. A referência a *tipo* ou *espécie* sugere a possibilidade de se recorrer a uma relação ontológica. Desse modo, ela poderia ser utilizada para itens lexicais que indicam, por exemplo, classe, gênero ou raça (hiperônimos) de 'equinos' ou 'manga-largas' (hipônimos correspondentes). Decidiu-se por utilizá-la para ocorrências de ambos os tipos, já que nenhuma outra etiqueta do elenco seria apropriada para cobrir esses casos.

Os critérios relatados nesta seção foram adotados mediante a necessidade de se buscar etiquetas adequadas a cada caso, restringindo ao máximo as alterações das anotações originais do *parser*. Ressalta-se ainda que todas as etiquetas constantes do elenco foram utilizadas na pós-edição, razão pela qual confirmamos o alto índice de etiquetas não ocorrentes no corpus (100 ocorrências, ou 47% das etiquetas, não ocorrem no Summ-it).

7 Conclusões

Como se demonstrou, o *parser* não dá conta de indicar o conceito semântico adequado para um número significativo de unidades textuais, os quais envolvem, frequentemente, problemas de dependências contextuais e de reconhecimento de entidades nomeadas.

As dificuldades de pós-edição, que implicariam mapeamentos semânticos inadequados dos itens lexicais, foram resolvidas adotando-se vários critérios, dentre os quais o contexto de uso das etiquetas. Evitou-se a opção de adotar etiquetas genéricas quando fosse possível reconhecer alguma mais específica porque essa opção não asseguraria os objetivos do refinamento do VeinSum: ao generalizar etiquetas, a probabilidade de serem indistinguíveis uma unidade textual anafórica e sua antecedente (por suas etiquetas) aumentaria, em vez de diminuir. Assim, embora a etiquetagem semântica de textos de domínios mais genéricos se tenha comprovado menos problemática do que a etiquetagem de textos de domínios mais específicos (que claramente apresentam porcentagem maior de correção), esta opção foi descartada por princípio.

A porcentagem média de correção do corpus (41%) obscurece, certamente, os casos extremos: o texto com menor porcentagem de problemas teve 3% de seus itens lexicais corrigidos; o com maior, aproximadamente 67%. As CCRs referentes a pessoas, as quais, em geral, incluem nomes próprios e profissões, foram as que apresentaram maior índice de acerto.

Considerando-se os vários problemas do *parser* e esses índices de correção da anotação, o corpus pós-editado é um recurso mais rico, pois a atribuição manual de etiquetas foi realizada de forma mais especializada. Resta, assim, sua utilização em tarefas de avaliação ou validação. Particularmente para o modelo de sumarização do VeinSum, será possível validar a revisão das etiquetas verificando se houve melhora da clareza referencial de sumários de outros textos, gerados com base nas heurísticas. Basta compará-los a sumários dos mesmos textos produzidos sem levar em conta as informações semânticas.

[12] http://pt.wikipedia.org/wiki/Vírus (Acesso em 25 jun. 2009).

De modo geral, claro é que, sem uma reengenharia que envolva critérios semânticos mais robustos do que os atuais, qualquer sistema computacional que dependa da etiquetagem continuará muito vinculado a cada corpus em foco (as heurísticas produzidas, afinal, são dependentes da ocorrência de CCRs que envolvem grupos de etiquetas particulares). Será impossível, no entanto, manter a tarefa de pós-edição manual de resultados semânticos automáticos do PALAVRAS, caso se pretenda que o *parser* semânatico seja um dos módulos de sistemas mais complexos, como o VeinSum. Por outro lado, sua ausência certamente comprometerá a qualidade dos resultados finais do sistema principal.

Assim, seria interessante que houvesse também uma reengenharia do próprio *parser,* para verificar se os problemas aqui detectados de fato podem ser evitados com o refinamento do modelo de etiquetagem. Claramente é necessário, antes, garantir que os problemas de etiquetação apresentados de fato são os causadores da maioria dos problemas de clareza referencial de sumários automáticos gerados pelo VeinSum.

Agradecimentos

Agradecemos a valiosa contribuição dos revisores da revista Linguamática a este artigo. Este trabalho contou com o apoio da FAPESP e da CAPES.

Referências Bibliográficas

Beaugrande, R., W. Dressler. 1981. *Introduction to Text Linguistics.* London, UK, Longman.

Bick, E. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Arhus, Arhus University.

Carbonel, T. I. 2007. *Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de correferência em Sumarização Automática.* Dissertação de Mestrado. Departamento de Letras. Agosto. São Carlos, SP, UFSCar.

Chaves, A. R. 2007. *A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov.* Dissertação de Mestrado. Departamento de Computação. Agosto. São Carlos, SP, UFSCar: 116p.

Coelho, J. C. B., Muller, V. M., Abreu, S. C., Vieira, R., Rino, L. H. M. 2006. Resolving Nominal Anaphora. *Lecture Notes in Artificial Intelligence 3960,* pp. 160-169. Springer. Berlin, Germany.

Collovini, S., Carbonel, T. I., Fuchs, J. T., Coelho, J. C., Rino, L. H. M., Vieira, R. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In Violeta Quental, Cláudia Oliveira (eds.), *Proc. of the V Workshop on Information and Human Language Technology* (TIL'2007, CD-ROM). XXVII Congresso da Sociedade Brasileira de Computação (SBC'2007). Rio de Janeiro - RJ.

Cristea, D., Ide, N., Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence. *Proc. of the Coling/ACL 1998*. Montreal, Canada.

Cristea, D., Postolache, O., Pistol, I. 2005. Summarization through Discourse Structure. *Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005*. Mexico City, Mexico, Springer LNSC.

Cristea, D., Postolache, O., Puscasu, G., Ghetu, L. 2003. Summarizing Documents Based on Cue-phrases and References. *Proc. of the International Symposium on Reference Resolution and its Apllications to Questions Answering and Summarization.* Veneza, Itália.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database.* Cambrigde, Massachussets, The MIT Press

Halliday, M. A. K., Hasan, R. 1976. *Cohesion in English.* London, UK, Longman.

Koch, I. G. V., Travaglia, L. C. 2004. *A coerência textual.* São Paulo, SP, Contexto

Mani, I. 2001. *Automatic Summarization.* Amsterdam, John Benjamin's Publishing Company.

Mann, W. C., Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3): 243-281.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA, USA, The MIT Press.

Marcuschi, L. A. 1983. *Linguística de texto: como é e o que se faz.* Universidade Federal de Pernambuco. Recife, PE.

Mitkov, R. 1998. Robust pronoun resolution with limited knowledge. *Proc. of the 18th International Conference on Computational Linguistics Conference* (COLING'98/ACL'98). Montreal, Canada.

Mitkov, R. 2002. *Anaphora Resolution.* London, UK, Longman.

Pardo, T. A. S., Nunes, M. G. V. 2002. Segmentação Textual Automática: Uma

Revisão Bibliográfica. *Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, no. 185* (NILC-TR-03-02). São Carlos, SP, ICMC, Universidade de São Paulo.

Santos, D. 1990. Lexical gaps and idioms in Machine Translation. *Proc. of the 14th International Conference on Computational Linguistics* (COLING'90), pp. 330-335. H. Karlgren. Helsinque.

Santos, D. 2007. O modelo semântico usado no Primeiro HAREM. In D. Santos, N. Cardoso (eds.). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, pp. 43-57, Cap. 4. Linguateca.

Santos, D., Cardoso, N. 2007. Breve introdução ao HAREM. In D. Santos, N. Cardoso (eds.). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, Cap. 1. Linguateca.

Sparck-Jones, K. 1999. Automatic Summarizing: factors and directions. In I. Mani, M. Maybury (eds.), *Advances in automatic text summarization,* pp. 1-12. Cambridge, Massachussets: The MIT Press.

Tomazela, E. C. 2010. *O uso de informações semânticas do PALAVRAS: em busca do aprimoramento da seleção de unidades correferentes na Sumarização Automática.* Dissertação de Mestrado. Departamento de Letras. São Carlos, SP, UFSCar. 115p.

Tomazela, E. C., Rino, L. H. M. 2009. O uso de informações semânticas para tratar a informatividade de sumários automáticos com foco na clareza referencial. In Aline Villavicencio (ed.), *Anais do VII Encontro Nacional de Inteligência Artificial (ENIA 2009)*, pp. 799-808. XXIX Congresso da Sociedade Brasileira de Computação. Bento Gonçalves, RS.

Tomazela, E. C., Rino, L. H. M. 2010. *Correção da etiquetagem semântica do Parser PALAVRAS para o Corpus Summ-it.* Série de Relatórios do NILC. NILC–TR-02-10. São Carlos, SP.

Do termo à estruturação semântica: representação ontológica do domínio da Nanociência e Nanotecnologia utilizando a Estrutura Qualia

Deni Yuzo Kasama
Universidade Estadual Paulista (UNESP)
Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)
deni@me.com

Claudia Zavaglia
Universidade Estadual Paulista (UNESP)
zavaglia@ibilce.unesp.br

Gladis Maria de Barcellos Almeida
Universidade Federal de São Carlos (UFSCar)
gladis@ufscar.br

Resumo

O presente artigo apresenta as etapas de elaboração de uma ontologia do domínio da Nanociência e Nanotecnologia com vistas à sistematização do léxico dessa área de especialidade, por meio de formalismos descritos na Teoria do Léxico Gerativo, com ênfase na Estrutura Qualia e seus quatro papéis semânticos, a saber: Formal, Constitutivo, Agentivo e Télico. A partir de um córpus da área, e valendo-nos de métodos semiautomáticos para a extração de candidatos a termos e identificação de relações semânticas, delineamos um mapeamento semântico partindo de relações de herança conceitual, cuja representação foi feita em linguagem OWL, com o auxílio da ferramenta Protégé.

1. Introdução

No âmbito do Processamento de Línguas Naturais (doravante PLN), o léxico desempenha papel crucial para o eficiente funcionamento de sistemas que visam a tratar automaticamente a língua. Dentre algumas aplicações em PLN, podemos citar a sumarização automática, a mineração de textos, a recuperação de informação e a tradução automática, para os quais um simples elenco de palavras não é suficiente. Segundo o tipo de aplicação, outras informações linguísticas tornam-se necessárias como, por exemplo, um sistema de reconhecimento de fala que necessita de um léxico subjacente que contenha informações do tipo fonológico. Estudos dessa natureza, bem como de dados morfossintáticos, têm sido conduzidos com expressivo sucesso, no que tange a sua correta identificação e categorização por sistemas computacionais. Entretanto, a representação semântica do léxico, seja ele geral ou especializado, é ainda terreno pouco sólido para pesquisas em Linguística Computacional que fazem uso desse tipo de informação. Os formalismos representacionais hoje conhecidos não se mostram eficientes o bastante para expor e tratar a questão da significação lexical com a devida precisão que sistemas de PLN exigem.

Esse caráter pouco domesticável do léxico explica-se por sua estreita relação com a realidade extralínguistica, a qual, segundo Biderman, é demonstrada pelos signos linguísticos ou unidades lexicais "que designam os elementos desse universo segundo o recorte feito pela língua e pela cultura correlatas. Assim, o léxico é o lugar da estocagem da significac□ão e dos conteúdos significantes da linguagem humana" (Biderman, 1996, p.27). Daí o fato de o léxico de uma língua encontrar-se em constante dinamicidade, além de que, para um mesmo significante, podem-se observar múltiplos significados. A tratabilidade dessas informações por máquina depende justamente da eficácia da representação semântica adotada.

Nesse sentido, podemos apontar trabalhos como o de Reeve e Han (2007) que faz uso de relações semântico-lexicais em um sistema de sumarização automática para textos do domínio médico, ou ainda o método desenvolvido por Ercan e Cicekli (2008) que faz uso extensivo de conhecimento semântico-lexical para o funcionamento de um sumarizador. Apontamos ainda para Rino e Pardo (2003) em que são descritos alguns sistemas de sumarização que fazem uso de repositórios lexicais em língua portuguesa; em mineração de textos, Alsumait et al. (2010) apontam para a importância de conhecimento semântico agregado ao léxico para processos de inferência de assuntos tratados em um determinado texto; Fox (1980) trata da importância de relações lexicais para a recuperação da informação; e, por fim, dentro os trabalhos que apontam para uso de repositórios lexicais como estratégia para a condução de tarefas em sistemas de

tradução automática, podemos citar Dorr (1992 e 1993) e Hutchins e Somers (1992).

Os trabalhos acima mencionados fazem uso de léxicos e alguns mencionam explicitamente a melhoria dos resultados quando esse tipo de dado é levado em consideração, uma vez que muitos sistemas que visam ao tratamento de informação textual valem-se apenas de estatísticas de frequência e coocorrência.

A necessidade de um modelo de representação semântica verifica-se já em Katz e Fodor (1963), e Jackendoff (1983) com sua proposta cognitiva que se baseia em uma hipótese ontológica e epistemológica. Os modelos propostos, então, envolviam a decomposição de traços em primitivos semânticos que se mostravam eficientes apenas para uma pequena parte do léxico; mais recentemente, os modelos de representação semântica adotados apontam para um teoria composicional, geralmente utilizada quando faz-se necessário um maior formalismo. Tais fatores nos levaram à adoção de uma teoria que nos permitisse representar o léxico de um domínio por meio de relações semânticas no interior de um conjunto vocabular especializado, bem como de um modelo de representação altamente utilizado para estruturação de um conhecimento, a saber, as ontologias.

O presente artigo subdivide-se da seguinte forma: na seção 2, tratamos do conceito de ontologias, sua utilidade nesta pesquisa e de como seu conceito difere do conceito de mapa conceitual; na seção 3, apresentamos a Teoria do Léxico Gerativo, mais especificamente a Estrutura Qualia e de sua importância para a representação formal de informações semânticas; na seção 4, detalhamos o desenvolvimento da pesquisa: o córpus[1] utilizado, a extração semiautomática[2] de candidatos a termos, a definição de classes e subclasses, o método utilizado para levantamento de relações semânticas e a subsequente implementação dos dados obtidos na ferramenta Protégé; na seção 5, apresentamos alguns dos dados alcançados; na sequência (seção 6), discutimos algumas questões envolvidas em uma pesquisa deste gênero; e na seção 7, apresentamos as conclusões e possíveis desdobramentos futuros.

2. Ontologias

Filósofos, de Aristóteles a Wittgenstein, trataram da existência de categorias lógicas que levariam a uma categorização geral das coisas que existem no mundo, muito embora com visões diferentes (do realismo ao relativismo, respectivamente, passando pelo idealismo kantiano). O termo "ontologia" nasce justamente na filosofia como o estudo da natureza do ser e sua existência, sob uma ótica metafísica e hoje estende-se para áreas como as Ciências da Computação, da Informação e Linguística.

Como já dito na seção anterior, o uso de ontologias tem se mostrado um meio eficiente de representação de conceitos semanticamente relacionados, servindo não só aos propósitos de sistemas de banco de dados, como também para o PLN. Isso porque as ontologias envolvem os formalismos necessários para a descrição de um conhecimento permitindo o uso da lógica e a realização de inferências a partir das informações estruturadas.

Gruber assim define ontologias:

> *"No contexto das ciências da computação e informação, uma ontologia define um conjunto de primitivos representacionais com os quais se modela um domínio do conhecimento ou discurso. Os primitivos representacionais são tipicamente classes (ou conjuntos), atributos (ou propriedades), e relacionamentos (ou relações entre membros das classes). As definições dos primitivos representacionais incluem informações sobre seu significado e restrições sobre sua aplicação consistente de forma lógica".*[3] (Gruber, 2008)

Como forma de estruturar um conhecimento (especializado ou terminológico, neste trabalho), valemo-nos do conceito de ontologias a fim de garantir (i) uma estruturação conceitual baseada em relações de classes e subclasses (ou de hiperônimos e hipônimos) que prevê a herança de conceitos; (ii) um padrão que vem sendo extensivamente utilizado para descrição de domínios; e (iii) um formalismo capaz de garantir o tratamento computacional dos dados linguísticos levantados a partir de um córpus e com recursos à disposição para realizar inferências automáticas a partir de restrições pré-

[1] Adotamos aqui o termo "córpus", tanto para o singular quanto para o plural, grafado com o acento agudo na vogal tônica, em português, em detrimento do latinismo (ou anglicismo) *corpus* e *corpora*. É de nosso conhecimento, entretanto, que, em artigos e livros, encontram-se as duas opções de grafia em vigor, de acordo com a escolha de cada autor.

[2] Advogamos o uso de "semiautomático" uma vez que entendemos ser necessário a intervenção humana em um ou mais etapas do processo.

[3] As citações em língua estrangeira são de tradução dos autores.

determinadas que possibilitam popular classes que atendam tais restrições.

Com efeito, Guarino (1998) relata a existência de três tipos de ontologias: 1. Ontologias genéricas (*top-level ontologies*), 2. Ontologias de domínio (*domain ontologies*) e Ontologias de tarefa (*task ontologies*) e 3. Ontologias de aplicação (*application ontologies*). Este trabalho concentra-se em (2), mais especificamente sobre as ontologias de domínio, definidas pelo autor como o tipo de ontologia que "descreve o vocabulário relacionado a um domínio genérico (como medicina ou automóveis) ou uma tarefa ou atividade genérica (como diagnóstico ou venda), através de uma especialização dos termos introduzidos na ontologia genérica".

O uso de ontologias no processo de criação de produtos terminológicos não é uma etapa necessariamente nova, mas imprescindível quanto a uma possível reutilização em aplicações como aquelas voltadas para a Web Semântica (Berners-Lee et al., 2001, p. 36), por exemplo. Ademais, como aponta Almeida (2000), o papel dos mapas conceituais interfere diretamente na própria pesquisa terminológica, visto:

> *"1) possibilitar um mapeamento mais sistemático de um campo de especialidade; 2) circunscrever a pesquisa, já que todas as ramificações da área-objeto, com seus campos, são previamente mapeadas; 3) delimitar o conjunto terminológico; 4) determinar a pertinência dos termos, pois separando cada grupo de termos pertencentes a um determinado campo, poder-se-á apontar quais termos são relevantes para o trabalho e quais não são; 5) prever os grupos de termos pertencentes à área-objeto, como também os que fazem parte de matérias conexas; 6) definir as unidades terminológicas de maneira sistemática e, finalmente, 7) controlar a rede de remissivas"* (Almeida, 2000, p. 120).

Cabré (1999, p. 144) aponta que os termos mantêm relações (não necessariamente hierárquicas) entre si, compondo dessa forma um mapa conceitual. Ainda para a mesma autora (2003), o lugar que o termo ocupa nesse mapa determina o seu significado, o que denota a importância de tais estruturas no processo de elaboração das definições em um dicionário especializado.

Algumas questões podem ser levantadas quanto ao uso dos termos "ontologia", "mapa conceitual" e "taxonomia". Entendemos haver uma diferenciação entre os conceitos, embora haja uma semelhança evidente, uma vez que, tanto terminólogos quanto ontólogos, trabalham em suas pesquisas com campos conceituais ou nocionais e com listas de unidades lexicais superordenadas em classes. Faz-se necessário, contudo, destacar conceitos como o de hereditariedade semântica e herança múltipla, presentes em ontologias. A esses conceitos agregam-se os de "atributos" e "propriedades", bem como os de "restrições" e "instâncias" ou "membros de classes", conforme citação anterior de Gruber (2008).

Nas Ciências da Computação, mapas conceituais são vistos como uma fase preliminar ao delineamento de uma ontologia, ou ainda, como se pode observar em Graudina (2008), uma reutilização de uma ontologia para fins didáticos:

> *"Levando em consideração similaridades óbvias entre ontologias e mapas conceituais, pesquisas de conversão de ontologia em mapa conceitual foram realizadas. Geração de mapas conceituais a partir de ontologias OWL existentes pode reduzir o trabalho de professores, por exemplo, para avaliação de conhecimentos. A transformação oferece aos professores um mapa conceitual inicial criado automaticamente, e ele só precisa refiná-lo, de acordo com suas necessidades, ampliando ou reduzindo-o".* (Graudina, 2008, p. 80)

Uma vez escolhido o modelo de representação semântica, foi o momento de buscar uma teoria que nos permitisse representar as relações entre os itens lexicais especializados do domínio em questão, bem como a herança conceitual lexical. A escolha recaiu sobre a Estrutura Qualia, uma das facetas do Léxico Gerativo, de James Pustejovsky (1995). O autor realiza uma distinção dicotômica para o estudo e representação da significação lexical: teorias baseadas em primitivos e teorias baseadas em relações. Pottier (1985) é um dos que trataram a semântica lexical com uma teoria de decomposição em primitivos semânticos que se opõem em positivos/negativos (possui ou não possui o sema em questão). Para Pustejovsky, contudo, uma representação semântica deve seguir uma linha composicional (que se enquadraria nas teorias baseadas em relações).

Outros modelos que estabelecem relações de significação entre itens lexicais foram observados, conforme tratado na Introdução deste artigo, contudo, acreditamos que o modelo relacional composicional adotado nos permite uma maior flexibilidade no tratamento das relações e por

estarem divididos em papéis semânticos bem definidos, conforme explicita-se na próxima seção.

3. A Teoria do Léxico Gerativo e a Estrutura Qualia

Uma visão possível para a resolução de questões inerentes ao tratamento semântico-computacional do léxico é a teoria proposta por James Pustejovsky em seu livro *The Generative Lexicon* (1995). Para o autor, os principais problemas para a semântica lexical são:

> "*(a) Explicar a natureza polimórfica da língua; (b) Caracterizar a semanticalidade de sentenças em língua natural; (c) Capturar o uso criativo de palavras em contextos novos; (d) Desenvolver uma representação semântica co-composicional mais rica*". (Pustejovsky, 1995, p. 5)

A maneira puramente morfossintática com que a maioria dos léxicos computacionais é hoje descrita pode explicar os entraves que se observam para que sistemas computacionais que necessitam do léxico funcionem adequadamente. Sem dúvida, a partir do momento que se agrega valor semântico a esses léxicos, obtêm-se resultados muito mais fiáveis e representativos concernentes àquilo que se objetiva a partir de um determinado sistema linguístico-computacional.

Para Pustejovsky, Semântica Lexical é o estudo de como e o que as palavras de uma língua denotam. Para linguistas teóricos e computacionais:

> *"o léxico é um conjunto estático de palavras-sentido, etiquetado com informações do tipo sintáticas, morfológicas e semânticas. Além disso, teorias formais do estudo da semântica de uma língua natural têm dado escassa importância a duas importantes questões: ao uso criativo de palavras em contextos novos e a uma apreciação dos modelos semântico-lexicais baseados na composicionalidade"*. (Zavaglia, 2002, p. 106 e 107)

Os componentes dessa rede de relações são classificados de acordo com o papel que desempenham, divididos da seguinte forma, conforme Pustejovsky (1995, p. 85 e 86):

- **Formal**, papel que faz a distinção do objeto em um domínio maior: i. Orientação, ii. Magnitude, iii. Forma, iv. Dimensionalidade, v. Cor, vi. Posição;
- **Constitutivo** ou **Partes Constituintes**, evidencia a relação entre objeto e suas partes constituintes que lhe são próprias: i. Material, ii. Peso, iii. Partes e elementos componentes";
- **Télico**, mostra o propósito e função do objeto: i. Propósito que um agente tem ao realizar uma ação, ii. Função integrada ou objetivo que especifica certas atividades;
- **Agentivo**, fatores que tratam da origem ou "causas" de um objeto: i. Criador, ii. Artefato, iii. Classe natural, iv. Cadeia causal

Uma abordagem do gênero, i.e. de caráter relacional, elimina entraves de natureza extensiva, pois não se limita, por exemplo, a uma lista exaustiva de traços semânticos e admite uma maior caracterização do léxico pelo próprio léxico. Sobre isso, a Teoria do Léxico Gerativo e, mais especificamente, a Estrutura Qualia, permite que se descreva um léxico valendo-se dos papéis semânticos que atribuem significado a um vocabulário finito e capturam a constituição, função, caracterização e origem dos referentes extralinguísticos que esse léxico representa no interior do sistema linguístico.

4. Metodologia da pesquisa

Antes de detalharmos o delineamento da ontologia em si, acreditamos fazer-se necessário explicitar a composição do córpus de pesquisa, a extração semiautomática dos candidatos a termos que compuseram o mapa ontológico do domínio, o levantamento de classes e subclasses, bem como do método semiautomático utilizado para o levantamento de relações semânticas segundo a Estrutura Qualia e a implementação dos dados na ferramenta Protégé.

4.1 O córpus da pesquisa

O córpus da Nanociência e Nanotecnologia (doravante N&N) foi compilado pelo Grupo de Estudos e Pesquisas em Terminologia, GETerm,[4] e apresenta 2.565.790 palavras (1057 textos, extraídos de 57 fontes diferentes), divididas tipologicamente da seguinte forma:

- Científico: composto por textos extraídos de revistas científicas, do Banco de Teses da

[4] Mais detalhes sobre a compilação do córpus podem ser obtidos em Coleti et al., 2008.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), doadas por CD-ROM;

- Informativo: constituído por jornais, revistas, portais, textos publicados em sites de órgãos de fomento à pesquisa;
- Científico de Divulgação: constituído por documentos extraídos de sites especializados, revistas, da Fundação de Desenvolvimento da Pesquisa (FUNDEP);
- Técnico-Administrativo: textos retirados do portal do Ministério da Ciência e da Tecnologia brasileiro;
- Outros: formado por textos presentes em slides de apresentações, prospectos de empresas e institutos de pesquisas e demais documentos avulsos obtidos em feiras e congressos da área.

A Tabela 1 apresenta a distribuição do número de palavras por tipologia dos textos.

tipos de textos	extensão do córpus
Científico	1.846.763
Informativo	361.607
Científico de divulgação	310.018
Técnico-administrativo	26.877
Outros	20.525

Tabela 1: Número de ocorrências no córpus por tipos textuais.

4.2 Extração semiautomática de candidatos a termos

A partir desse córpus, procedemos à extração semiautomática dos candidatos a termos utilizando-se do pacote *NSP – N-gram Statistics Package* (Banerjee e Pedersen, 2003).

Por meio do pacote NSP, foi possível gerar listas de unigramas, bigramas, trigramas e tetragramas, que correspondem a termos compostos por uma, duas, três ou quatro *tokens*, respectivamente. As listas geradas pelo pacote NSP necessitaram passar por uma limpeza manual, uma vez que, muito do que foi obtido não era necessariamente um termo, como ilustrado no Quadro 1 (os candidatos a termo que foram submetidos ao especialista, neste exemplo, encontram-se em negrito).

DIFRAÇÃO<>DE<>RAIOS<>214 528 31477 436 528 214 436
DAS<>AMOSTRAS<>DE<>209 1684 1438 51641 490 923 672
A<>QUANTIDADE<>DE<>209 20683 470 51641 209 9609 470
O<>NÚMERO<>DE<>209 10266 635 51641 231 5757 613
DENSIDADE<>DE<>CORRENTE<>208 460 31477 580 436 208 373
NA<>FIGURA<>A<>207 4264 2130 9308 318 340 405
DE<>ÓXIDO<>DE<>202 21743 444 51641 249 5655 350
FILME<>DE<>ÓXIDO<>199 590 31477 384 514 199 276
DO<>CAMPO<>ELÉTRICO<>199 7131 1107 507 424 199 485
PARA<>A<>AMOSTRA<>192 5491 9247 861 1928 192 381
DA<>CONCENTRAÇÃO<>DE<>191 7434 724 51641 222 3561 601
AS<>AMOSTRAS<>DE<>190 2149 1438 51641 593 1210 672
DO<>NÚMERO<>DE<>189 7131 635 51641 189 3913 613
A<>FIGURA<>ILUSTRA<>189 20683 2130 189 1077 189 189
CEO<>-AL<>O<>189 189 294 1963 189 189 294
A<>TÉCNICA<>DE<>187 20683 405 51641 187 9609 405
A<>ADIÇÃO<>DE<>187 20683 346 51641 206 9609 327
TAXA<>DE<>CORROSÃO<>187 493 31477 705 493 187 408
TAXA<>DE<>CRESCIMENTO<>93 493 31477 362 493 93 333
CARGA<>E<>DESCARGA<>124 188 2873 168 124 124 124

Quadro 1 – Exemplo de lista de trigrama gerada pelo pacote NSP

No Quadro 1, os *tokens* encontram-se separados pelo sinal "<>", os número que se observam logo após o último sinal "<>" referem-se a frequência no córpus daquele trigrama (neste exemplo, "taxa de corrosão" ocorreu 187 vezes), os demais valores não foram utilizados nesta pesquisa.

Uma vez feita a extração e limpeza das listas geradas, essas foram submetidas à análise do especialista da área, o Prof. Osvaldo Novais de Oliveira Jr. do Instituto de Física da Universidade de São Paulo, que validou os termos e sua pertinência ao domínio em questão.

Os números de candidatos a termos obtidos imediatamente após a utilização do NSP foram muito díspares em relação ao número de termos validados pelo especialista e os que, de fato, figuram na lista final de termos, conforme a Tabela 2. Essa diferença resulta da exclusão de falsos candidatos a termos (do Quadro 1: "das amostras de", "a quantidade de", "o número de" e assim por diante) e de possíveis candidatos a termos enviados ao especialista, mas que não foram confirmados, por ele, como termos da área (é o caso de "carga e descarga" e "taxa de crescimento", do Quadro 1).

É possível afirmar que, geralmente, quanto maior o número de unidades que compõe o termo, maior o número de candidatos que são, efetivamente, termos. Isso porque o pacote NSP não utiliza nenhuma medida de associação para unigramas, apenas a medida de frequência. Nos demais casos, o pacote disponibiliza medidas de Informação Mútua, *log-likelihood* e Coeficiente *Dice* (Banerjee e Pedersen, 2003 e Almeida et al., 2003) entre outras que otimizam os resultados.

	Número de candidatos do NSP	Número final de termos
unigramas	1.081.552	1.795 (0,16%)
bigramas	314.194	587 (0,18%)
trigramas	579.491	591 (1,01%)
tetragramas	123.760	152 (1,22%)
Total	2.098.997	3.125 (0,14%)

Tabela 2: Número de candidatos a termos e número final de termos.

A Tabela 3 apresenta uma parte da lista final de trigramas já validada pelo especialista e da qual partimos para o delineamento da ontologia.

TERMOS	FREQÜÊNCIA	TIPO DE TEXTO
ABSORÇÃO DE RAIOS X	1	TA
ACETATO DE CELULOSE	4	OU
AÇO INOXIDÁVEL DUPLEX	22	CI
AEROSOL EM CHAMA	34	CI
ALARGAMENTO DO PICO	21	CI
ALGINATO DE SÓDIO	6	OU
ALTA RESOLUÇÃO ESPACIAL	22	CI
ALTURA DO PICO	20	CI
ANALISADOR DE ESPECTRO	29	CI
ANALISADOR DE REDE	21	CI
ANÁLISE TÉRMICA DIFERENCIAL	28	CI
ÁREA SUPERFICIAL ESPECÍFICA	111	CI

Tabela 3: Lista de trigramas final.

4.3 Definição de classes e subclasses

Em uma ontologia, a principal relação que se observa é a formal, mais especificamente a relação *é_um*, *é_uma* (do inglês, *is_a)* a qual representa, de maneira objetiva, a herança conceitual de uma classe por sua subclasse. Sendo assim, essa foi a primeira relação que procuramos observar para que a ontologia tivesse uma estrutura hierárquica primária. O exemplo da Figura 1 apresenta uma estrutura indicando relações *é_uma* entre a classe "microscopia eletrônica" e suas subclasses: "microscopia de varredura por sonda" herda os conceitos de "microscopia eletrônica de varredura" que, por sua vez, herda os conceitos de "microscopia eletrônica". Para "microscopia eletrônica de transmissão", esta herda também conceitos de "microscopia eletrônica", mas possui traços diferenciais em relação à "microscopia eletrônica de varredura".

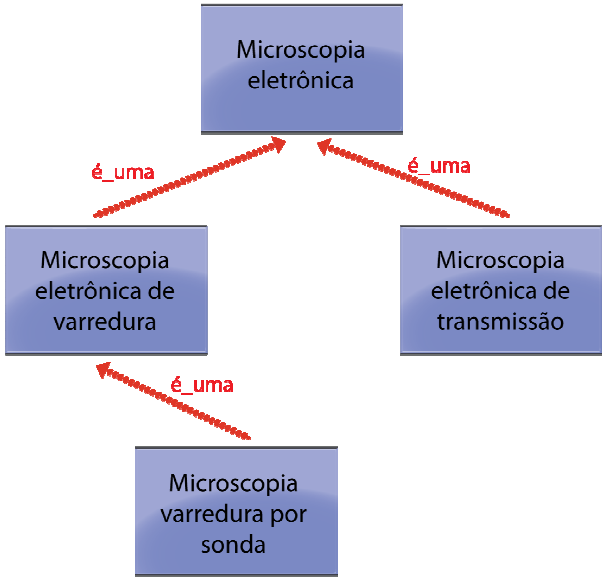

Figura 1: Classe "microscopia eletrônica" e suas subclasses.

A fim de agrupar os termos semanticamente relacionados, baseamo-nos na divisão de subdomínios feita no projeto *Desenvolvimento de uma Estrutura Conceitual (Ontologia) para a Área de Nanociência e Nanotecnologia* (Aluísio et al., 2006), para o qual havia também seis subdomínios principais:

1. "**Synthesis, Processing and Fabrication**";
2. "**Materials**";
3. "**Properties and Characterization techniques**";
4. "**Machines and Devices**";
5. "**Theories and Computational methods**";
6. "**Applications**".

Nesta pesquisa, a divisão foi realizada da seguinte forma:

1. **Aplicações**: Termos relacionados a campos científicos e usos específicos da N&N;
2. **Equipamentos**: Dispositivos utilizados na síntese, processamento e construção de nanomateriais;
3. **Materiais**: Matéria utilizada para a confecção de nanomateriais, os nanomateriais propriamente ditos ou foco de atuação de materiais nanoestruturados;
4. **Métodos e técnicas**: Processos envolvidos na manipulação de nanomatéria;

5. **Propriedades**: Características diversas intrínsecas aos materiais;
6. **Teorias**: Teorias que confluem na manipulação de materiais em nanoescala.

Assim, a classe "microscopia eletrônica", ilustrada acima, faz parte do subdomínio *Métodos e técnicas*.

A nova nomenclatura na divisão foi feita visando a facilitar o agrupamento de conceitos, além de deixar mais claro sobre o que cada subdomínio trata. Nesse sentido, indagou-se como abarcar em um mesmo subdomínio propriedades e técnicas de caracterização. Pareceu-nos que técnicas possuem mais afinidade semântica com métodos de processamento e fabricação, uma vez que, em ambos os casos, tratam-se de processos envolvidos na composição/manipulação dos nanomateriais. E ainda, "Equipamentos" engloba tanto o conceito de "máquinas" quanto o de "dispositivos" utilizados em N&N.

Ademais, a taxonomia em inglês da N&N (desenvolvida no âmbito do projeto acima citado) não corresponde propriamente a uma ontologia formalizada: alguns conceitos encontram-se agrupados em uma mesma classe, como é o caso de "Óxidos e sais", mas não se pode afirmar que as suas subclasses serão, todas elas, um óxido e ao mesmo tempo um sal.

4.4 Levantamento de relações semânticas

Esta etapa consiste na definição de relações semânticas, segundo a Estrutura *Qualia* de James Pustejovsky. Muitas das relações semânticas foram sendo delineadas concomitantemente ao processo de definição de classes e subclasses, uma vez que a observação dos contextos trazidos pelo processador de córpus já evidenciavam tais relações. Contudo, uma forma semiautomática que pudesse destacar tais relações foi útil e proveitosa, na medida em que essas são formadas, em geral, por expressões regulares. Para relações do tipo *Constitutivo*, observamos expressões como *é feito(a) de*, *é constituído(a) de/por*, *tem/têm como parte*, *é composto(a) de/por*, entre outras. Para as relações *Formal*, levantamos diversos termos a partir do subdomínio ao qual pertencem por meio de expressões de busca do tipo *é um equipamento*, *é um material*, *é uma aplicação* e assim sucessivamente para cada subdomínio eleito e elencado na seção anterior.

Visando a facilitar tal trabalho, utilizamos o recurso de *grafos* da ferramenta Unitex,[5] por meio do qual foi possível descrever um conjunto de regras recursivas de busca, permitindo assim um levantamento semiautomático de expressões que pudessem indicar relações semânticas nos quatros tipos descritos pela Estrutura Qualia (seção 3). A avaliação da eficácia do método, comparada ao número de resultados obtidos, pode, a princípio, parecer insatisfatória uma vez que muito do que obtivemos como *output* da ferramenta não foi utilizado; contudo, resultados que efetivamente foram aplicados à ontologia, após nossa análise, não teriam sido facilmente detectados, em uma busca manual, em um córpus de mais de dois milhões de palavras.

Apresentamos a seguir os *grafos* utilizados para cada um dos papéis semânticos da Estrutura Qualia, descritos na seção 3.

A Figura 2 apresenta o *grafo* utilizado para as buscas por relações do tipo *Formal*.

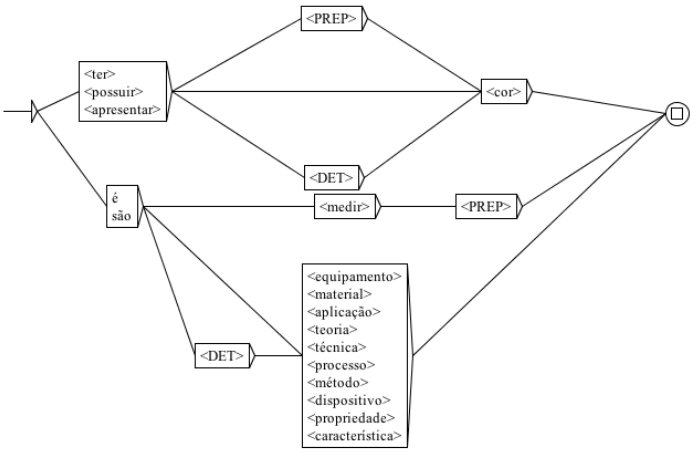

Figura 2: Grafo para busca de relações *Formal*.

O *grafo* representado na Figura 2 permite a realização de buscas que atendam aos seguintes critérios:

[5] Unitex é um sistema de processamento de corpus, baseado na tecnologia autômato-orientada. É um software criado no LADL (Laboratoire d'Automatique Documentaire et Linguistique), sob a direção de Maurice Gross. Com esta ferramenta, tem-se acesso a recursos eletrônicos, tais como dicionários e gramáticas, os quais podem ser aplicados em determinado corpus. O Unitex permite análises nos níveis da morfologia, do léxico e da sintaxe. O programa pode ser obtido gratuitamente em: www-igm.univ-mlv.fr/~unitex/.

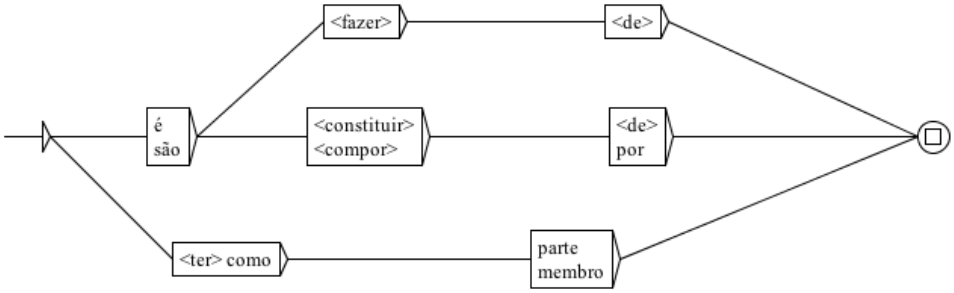

Figura 3: *Grafo* para busca de relações do tipo *Constitutivo*.

1. verbos "ter", "possuir" ou "apresentar" flexionados em qualquer tempo, modo ou pessoa seguidos de uma preposição e esta seguida pela palavra "cor" com flexão;
2. verbos "ter", "possuir" ou "apresentar" flexionados em qualquer tempo, modo ou pessoa seguidos por um determinante qualquer, seguido pela palavra "cor" com flexão;
3. "é" ou "são" seguido pelo verbo "medir" flexionado em qualquer tempo, modo ou pessoa, seguido por uma preposição;
4. "é" ou "são" seguido ou não por um determinante, seguido pelas palavras "equipamento", "material", "aplicação", "teoria", "técnica", "processo", "método", "dispositivo", "propriedade" ou "técnica" incluindo flexões dessas.

Obtivemos com essa busca 293 resultados. A título de exemplo, reproduzimos no Quadro 2 (no Anexo) algumas concordâncias para os critérios descritos no item (4), com os quais foi possível chegar a termos, ausentes até então na ontologia, como: "constante dielétrica", "perfilômetro" e "redução carbotérmica".

A Figura 3 apresenta o *grafo* utilizado para buscar relações do tipo *Constitutivo*.

Em um primeiro momento, o verbo "fazer" estava na mesma caixa dos verbos "constituir" e "compor", contudo constatamos que a combinação "fazer" seguida da preposição "por" não apontava para relações constitutivas (como *feito de*), mas para relações do tipo *Agentivo* (i.e., aquelas envolvidas na origem do objeto), como podemos observar nas concordâncias do Quadro 3 (Anexo).

O *grafo* da Figura 3 permitiu uma busca que retornou 243 resultados e atendeu aos seguintes critérios:

1. "é" ou "são" seguido do verbo "fazer" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "de", contraída com artigo ou não, e com flexão de número;
2. "é" ou "são" seguido dos verbos "constituir" ou "compor" flexionados em qualquer tempo, modo ou pessoa, seguidos da preposição "de", contraída com artigo ou não, e com flexão de número ou da preposição "por";
3. verbo "ter" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "como", seguido das palavras "parte" ou "membro"

Para as relações do tipo *Télico*, estabelecemos os seguintes critérios:

1. verbo "é" ou "são", seguido do verbo "utilizar" ou "usar" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "em" ou "para";
2. verbo "ter" flexionado em qualquer tempo, modo ou pessoa, seguido ou não da preposição "como" ou "a", seguido do substantivo "finalidade", "objetivo" ou "escopo" flexionado em número, seguido ou não da preposição "de";
3. verbo "fazer" flexionado em qualquer tempo, modo ou pessoa, seguido da palavra "uso", seguida da preposição "de";
4. verbo "utilizar" ou "usar" flexionado em qualquer tempo, modo ou pessoa, com próclise ou ênclise do pronome "se", seguido da preposição "de";
5. locução prepositiva "a fim de" ou preposição "para", seguida do verbo "obter" flexionado em qualquer tempo, modo ou pessoa ou seguida do substantivo

"obtenção", seguido ou não da preposição "de";

6. verbo "é" ou "são", seguido do verbo "fazer" flexionado em qualquer tempo, modo ou pessoa, seguido da preposição "para".

Utilizando o método aqui descrito, é possível ter um foco maior nas relações que se busca e que, numa busca manual, poderiam passar despercebidas. Os critérios descritos em (1) nos levaram às concordâncias reproduzidas no Quadro 4 (Anexo), a partir do córpus. Destacamos a última delas, que nos apontou para uma relação *Télica* importante entre os termos "óxido misto" e "coprecipitação", ilustrada na Figura 4.

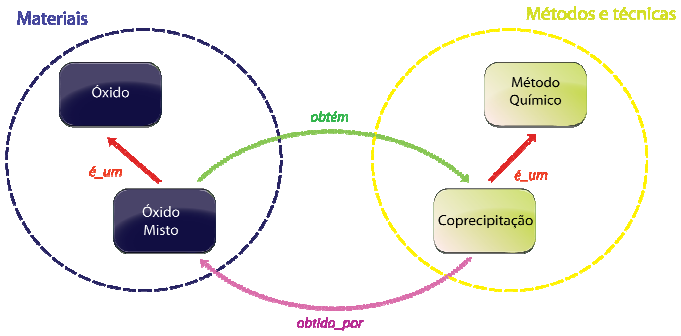

Figura 4: *Relação Télica, obtém.*

Cumpre aqui dizer que, conforme ilustrado na Figura 4, *obtém* e *obtido_por* são relações inversas, sendo *Télica* (função do objeto) e *Agentiva* (origem do objeto), respectivamente.

Da mesma forma, *utilizado_em* e *utiliza* (respectivamente, relações *Télica* e *Agentiva*) são inversas, segundo a Figura 5, em que são representados os termos "nitrogênio" (do subdomínio de "Materiais") e "secagem" (do subdomínio de "Métodos e técnicas").

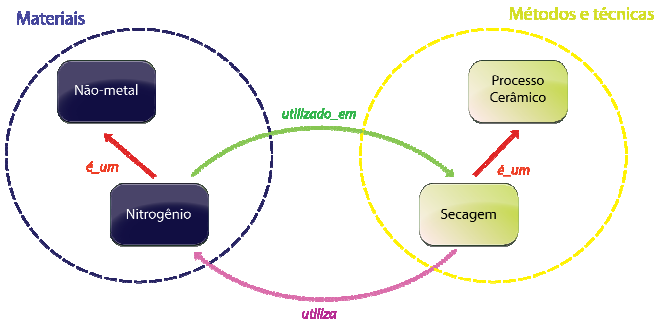

Figura 5: Relações inversas *utilizado_em* e *utiliza*.

4.5 A linguagem OWL

Neste trabalho, a linguagem adotada para a representação do domínio da N&N foi a OWL (*Web Ontology Language*), considerada atualmente, o padrão mais corrente para a representação de informações ontológicas na Web. A OWL (Smith et al., 2004) foi antecedida pelas linguagens RDF (*Resource Description Framework*) e RDFS (*RDF-Schema*), mostrando-se mais potente em termos de descrição e instanciação. Essas duas últimas correspondem a linguagens em que os recursos são descritos como trios de objetos-atributos-valores, semelhantes ao sujeito-verbo-objeto das redes semânticas.

4.6 Implementação dos dados na ferramenta Protégé.

A implementação dos resultados alcançados em uma ferramenta computacional específica para ontologias garante que os formalismos adotados para a representação do domínio escolhido sejam respeitados. Além disso, as possibilidades existentes de reuso de uma ontologia, quando expressa em uma linguagem computacional corrente e atual, são variadas. Nesse sentido, buscamos utilizar um software que possuísse facilidade de uso aliada a potencialidades de funções. A escolha incidiu sobre a ferramenta Protégé[6] (Noy *et al.*, 2000), uma vez que atende a esses quesitos.

Em consonância com os princípios de construção de ontologias, a ferramenta permite que ontologias sejam constantemente alimentadas e representadas em diferentes formatos e linguagens. Segundo Noy *et al.* (2001, p. 62), a ferramenta possui: um "modelo de conhecimento extensível", sendo possível redefinir seus primitivos representacionais; um "formato de arquivo de saída customizável", o que permite gerar arquivos em qualquer linguagem formal; "uma interface com o usuário customizável", possibilitando adaptar os componentes da interface com o usuário para a nova linguagem escolhida; "uma arquitetura extensível que permite integração com outras aplicações", isso torna a ferramenta conectável a módulos semânticos externos.

[6] Desenvolvida pela Divisão de Informática Médica do Departamento de Medicina da Universidade de Stanford, o Protégé foi inicialmente idealizado para modelar o domínio da medicina e traçar relações entre os muitos conceitos que englobam tal campo de especialidade. A ferramenta encontra-se disponível para download gratuitamente em http://protege.stanford.edu/

A representação do conhecimento no Protégé se dá por meio de três entidades básicas:

- *Classes* – define conceitos no domínio;
- *Propriedades* (*Properties*) – define atributos das classes;
- *Facetas* (*Facets*) – define restrições nos valores de classes (por exemplo: tipos, cardinalidade,[7] padrões).

Essa ferramenta permite a definição de propriedades inversas (Figura 6), o que facilita a representação de questões como aquelas ilustradas pelas Figuras 4 e 5.

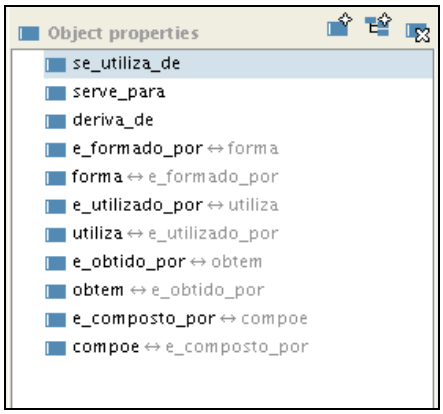

Figura 6: Relações semânticas representadas como Propriedades.

6. Resultados

O subdomínio que apresentou mais classes e subclasses foi o de *Materiais* (230), seguido de *Métodos e Técnicas* (68), *Propriedades* (42), *Equipamentos* (18), *Aplicações* (17) e *Teorias* (3), totalizando 378 classes e subclasses.

Estruturalmente, uma classe como "Material nanoestruturado" (do subdomínio de "Materiais") e suas subclasses podem ser assim representadas, de acordo com a figura 7.

Com o auxílio do *plugin* OWLViz, essa mesma classe e suas subclasses ficam ilustradas, conforme a Figura 8.

Algumas das relações semânticas identificadas ao longo do processo foram implementadas na ferramenta. Elas encontram-se elencadas, quantificadas e exemplificadas na Tabela 4.

7 A cardinalidade diz respeito a um dado expresso em valor numérico ou por um conjunto deles.

Formal		
é_medido_em	43	<Nanoporo> é_medido_em <Nanômetro>
é_um, é_uma	376	<Densidade> é_uma <Grandeza_física>
Constitutivo		
compõe	1	<Carbono> compõe <Nanotubo_de_carbono>
é_composto_por	8	<Vitrocerâmica> é_composto_por <Cálcio>
é_feito_de	4	<Nanotubo de carbono> é_feito_de <Carbono>
é_formado_por	3	<Nanocompósito polimérico> é_formado_por <Borracha>
forma	6	<Quinona> forma <Nanocápsula>
Télico		
obtém	2	<Precursor_polimérico> obtém <Óxido_de_estanho>
produz	2	<Bactéria> produz <Antígeno>
utilizado_em	5	<Vidro> utilizado_em <Vitrocerâmica>
Agentivo		
deriva_de	4	<Plástico> deriva_de <Petróleo>
é_produzido_por	2	<Antígeno> é_produzido_por <Vírus>
obtido_por	7	<Óxido_de_estanho> obtido_por <Precursor_polimérico>
originado_de	1	<Vitrocerâmica> originado_de <Vidro>
utiliza	2	<Fotoalinhadora> utiliza <Luz_ultravioleta>

Tabela 4: Relações individuadas a partir das buscas na ferramenta Unitex.

5. Discussões

A Estrutura Qualia permite-nos ter um maior controle sobre as relações semânticas do domínio da N&N, uma vez que os delimita em quatro papéis funcionais. O método que aqui se motiva é um primeiro passo para trabalhos terminológicos que fazem uso de grandes córpus. As relações pré-estabelecidas podem não cobrir todas as relações que podem figurar no domínio mas já apontam para aquelas fundamentais. Além disso, os *grafos* podem ser ampliados e adaptados de acordo com as necessidades de cada pesquisa. Uma vez individuadas as relações básicas do tipo *Formal* *é_um*, *é_uma* (presentes em qualquer ontologia), as demais podem ser estendidas partindo-se das listas geradas pelo Unitex.

Delineamos, neste trabalho, a área técnico-científica da N&N, uma ciência interdisciplinar e inovadora, cujas técnicas de manipulação de materiais têm obtido investimentos enormes e cujas possibilidades de aplicação são inúmeras. A definição de sua estrutura conceitual permitirá que o produto terminográfico seja coeso e uniforme. Por outro lado, essa mesma estrutura, quando dotada de formalismos, pode também servir como um léxico computacional que sirva para alimentar sistemas de PLN.

A observação dos fenômenos linguísticos por meio de um processador de córpus ressalta a importância desse tipo de ferramenta e a necessidade de automatização das pesquisas em Linguística, de um modo geral. O alto nível de conhecimento do uso dessas ferramentas, por parte do pesquisador, aprimora os resultados da pesquisa e permite uma adaptação dessas ferramentas às necessidades particulares de cada investigação científica.

A importância da utilização de métodos computacionais é grande, na medida em que o volume de informações, com que muitos trabalhos científicos se deparam tem sido cada vez maior. Nesta pesquisa, a extração semiautomática de termos e o levantamento de candidatos a relações semânticas mostraram-se um fim cujos meios para alcançá-los foram enormemente facilitados pelo auxílio de recursos informatizados. Entretanto, a observação cautelosa e criteriosa desses dados por parte do pesquisador foi o elemento-chave para que chegássemos aos resultados esperados.

Os recursos aqui descritos encontram-se disponíveis no Portal de Ontologias OntoLP.[8]

6. Conclusões

A Engenharia Ontológica é um vasto campo a ser explorado por pesquisadores de disciplinas diversas que têm estudado e aplicado, cada vez mais, seus conhecimentos na criação de uma metodologia que permita a criação e reuso de ontologias. Há, nessas disciplinas distintas, conceitos que se interpolam e se confundem, permitindo que se trate de conceitos relativos às ontologias de maneiras diversas e complementares. Aquilo que a Computação entende por ontologias, os formalismos que ela adota para sua criação e manipulação beneficiam o poder de descrição semântica de um dado vocabulário por parte de um lexicólogo/terminólogo, conferindo-lhe também a possibilidade de realizar uma aplicação computacional para seu trabalho, se assim desejar. Logo, podemos afirmar que tais formalismos garantem um processo definitório mais consciente, uma vez que, para o tratamento informático do léxico, as ambiguidades, inconsistências e imprecisões devem ser minimizadas. Para tanto, deve-se ter à disposição um modelo semântico eficiente que estenda a exposição lexical a um nível superior ao da morfologia e da sintaxe fornecendo à máquina condições de inferir e interpretar dados linguísticos. A esse propósito, a Estrutura Qualia representa um método eficaz para uma representação semântica inter-relacional, e garantiu a esta pesquisa meios de estabelecer relações de tipologias diversas ao léxico em questão, permitindo que a sua semântica fosse exposta e computacionalmente tratável.

Embora a N&N seja uma área de especialidade multidisciplinar que se utiliza de conceitos e técnicas da Física, Química, Biologia, Medicina, Engenharia de Materiais e áreas afins, o que percebemos é que pesquisadores em N&N têm criado novos materiais (em sua maioria, aqueles em escala nanométrica) e esses devem ser nomeados. Procuramos, dessa forma, estudar também esses novos termos e os processos aí envolvidos. Destacamos, assim, a partícula *nano-* como formadora desses itens neológicos especializados e os métodos envolvidos nesse levantamento.[9]

Salienta-se ainda que os resultados alcançados podem ser estendidos a partir do modelo proposto. As 361 classes e subclasses apresentadas representam o domínio da N&N, mas não integralmente. Essa delimitação deve-se, em primeiro lugar, à extensão do domínio e, posteriormente, ao grande tempo requerido por tarefas como:

- resgate de conceitos;
- observação das diversas ocorrências de um mesmo termo;
- correlações com termos semelhantes;
- real estatuto de termo de determinadas lexias;
- identificação do equivalente em português, em casos nos quais preferiu-se pelo uso de um termo estrangeiro;
- pertinência de um termo a duas superclasses distintas – em qual delas o termo estaria melhor representado e de qual superclasse há uma herança conceitual mais clara?

[8] Acessível em http://www.inf.pucrs.br/~ontolp/index.php.

[9] Uma análise dos processos de formação neológica no domínio da N&N pode ser encontrada em Kasama et al. (2008).

Esses são alguns exemplos de dificuldades encontradas no desenvolvimento da pesquisa aqui relatada, mas que lhe são inerentes.

As contribuições deste trabalho fazem-se sentir em áreas como:

- a Linguística: por meio do estabelecimento de uma metodologia, fundamentalmente embasada em ferramentas computacionais, que permite a observação de termos em uso e sua estruturação a partir de critérios semânticos;
- as Ciências da Computação: que se beneficia de conceitos linguísticos no seu fazer e pode reaproveitar os resultados obtidos para avaliação e uso real de uma ferramenta computacional que se sirva de informações semânticas;
- a área de N&N: cuja sistematização vocabular permite que pesquisadores da área possuam uma fonte de referência no que tange suas práticas. Ademais, a recolha de termos em língua portuguesa, variante brasileira, contribui para o desenvolvimento da área no país;
- o ensino em geral: seja para alunos de graduação ou pós-graduação em cursos afins à N&N, mas também para alunos de Ensino Médio, tendo em vista que a multidisciplinaridade da N&N promove o conhecimento da Física, da Química e da Biologia.

As possibilidades iniciadas neste trabalho vão além daquilo que obtivemos. Esperamos que a ontologia ora proposta auxilie, de fato, no processo de elaboração do dicionário de N&N em língua portuguesa do Brasil, mas também que possa ter utilidade e aplicação real em sistemas de PLN.

Além da elaboração da ontologia em si, esperamos ter proposto uma metodologia para a elaboração de novas representações do conhecimento valendo-nos de preceitos observados na Linguística de Córpus, na Terminologia e nos formalismos computacionais que buscamos seguir.

Agradecimentos

Agradecemos à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo financiamento da pesquisa (Processo nº 06/59144-8), aos Professores Oto Araújo Vale, Sandra Maria Aluísio e Maria Cristina Parreira pela leitura e valiosas contribuições ao trabalho; ao Professor Osvaldo Novais de Oliveira Jr., do Instituto de Física da Universidade de São Paulo, pela consultoria técnica na área, sem a qual um trabalho desta natureza não poderia ser realizado; e, por fim, aos Professores António Teixeira e Patrícia Cunha França pelas leituras finais e sugestões que enriqueceram este artigo.

Referências

Almeida, Gladis Maria de Barcellos. 2000. Teoria Comunicativa da Terminologia: uma aplicação. Araraquara (Tese de doutorado).

Almeida, Gladis Maria de Barcellos; Aluísio, Sandra Maria; Teline, Maria Fernanda. 2003. Extração manual e automática de terminologia: comparando abordagens e critérios. In: 1o. Workshop em Tecnologia da Informação e da Linguagem Humana, 2003, São Carlos. Anais do TIL'2003.

Alsumait, Loulwah; Wang, Pu; Domeniconi, Carlota; Barbará, Daniel. Embedding semantics in LDA topic models. 2010. In: Berry, Michael W.; Kogan, Jacob. Text Mining: Application and Theory. John Wiley & Sons, Ltd., p. 183-203

Aluísio, Sandra Maria; Oliveira Jr., Osvaldo Novais; Almeida, Gladis Maria de Barcellos; Nunes, Maria das Graças Volpe; Oliveira, Leandro Henrique Mendonça de; Felippo, Ariani Di; Antiqueira, Lucas; Genoves Jr, Luiz Carlos; Caseli, Luciano; Zucolotto, Valtencir ; Santos Jr., David Sotero dos. 2006. Desenvolvimento de uma estrutura conceitual (ontologia) para a área de Nanociência e Nanotecnologia. (Relatório técnico)

Banerjee, Satanjeev; Pedersen, Ted. 2003. The Design, Implementation, and Use of the Ngram Statistics Package In: Conference On Intelligent Text Processing And Computational Linguistics, 4., 2003, Cidade do México. Proceedings…, Cidade do México, p. 370-381.

Berners-Lee, Tim; Hendler, James; Lassila, Ora. 2001. The Semantic Web. Scientific American. p. 35-43.

Biderman, Maria Tereza Camargo. 1996. Léxico e vocabulário fundamental. Alfa. São Paulo, v.40, p. 27-46.

Cabré, Maria Tereza. 1999. La terminología. Representación y comunicación. Barcelona: IULATERM.

Cabré, Maria Tereza. 2003. Theories of terminology: their description, prescription and explanation. Terminology, v.9, n.2, p.163-200.

Coleti, Joel S.; Mattos, Daniela F.; Genoves Jr., Luiz Carlos; Candido Jr., Arnaldo; Di Felippo, Ariani; Almeida, Gladis Maria de Barcellos;

Aluísio, Sandra M.; Oliveira Jr., Osvaldo Novais. 2008. A compilação de corpus em língua portuguesa na área de nanociência/nanotecnologia: problemas e soluções. In: Tagnin, Stella E. O.; Vale, Oto Araújo (Org.). Avanços da Lingüística de Corpus no Brasil. 1 ed. São Paulo: Humanitas, p. 167-191.

Dorr, Bonnie J. 1992. The use of lexical semantics in interlingual machine translation. v.7, n.3, Springer Netherlands, p. 135-193.

Dorr, Bonnie J. 1993. Machine Translation: a view from the lexicon. Cambridge: MIT Press.

Ercan, Gonenc; Cicekli, Ilyas. 2008. Lexical Cohesion Based Topic Modeling for Summarization. Lecture Notes in Computer Science. v. 4919, p. 582-592.

Fox, Edward A. 1980. Lexical relations: Enhancing effectiveness of information retrieval systems. SIGIR Forum, v.15, n.3, p. 5-36.

Graudina, Vita. 2008. OWL Ontology Transformation into Concept Map. Scientific Proceedings of Riga Technical University. 5th Series, Computer Science, Applied Computer Science, Vol. 34, 79-90.

Gruber, Tom. 2008. Ontology. In: Liu, Ling; Özsu, M. Tamer (Eds.) Encyclopedia of Database Systems, v. 1, Springer-Verlag.

Guarino, Nicola. 1998. Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Itália, 6-8 Junho 1998. Amsterdam, IOS Press, p. 3-15.

Hutchins, W. John; Somers, Harold L. An introduction to machine translation. London: Academic Press, 1992.

Jackendoff, Ray. 1983. Semantics and cognition. Cambridge: The MIT Press.

Kasama, Deni Y.; Almeida, Gladis Maria de Barcellos; Zavaglia, Claudia. 2008. A influência das novas tecnologias no léxico: processos de formação neológica no domínio da nanociência e nanotecnologia. Debate Terminológico, v. 4, p. 3.

Katz, Jerrold J.; Fodor, Jerry A. 1963. The Structure of a Semantic Theory. Language, v. 39, n. 2, p. 170-210.

Noy, Natalya F.; Sintek, Michael; Decker, Stefan; Crubézy, Monica; Fergerson, Ray W.; Musen, Mark A. 2001. Creating Semantic Web Contents with Protégé-2000. IEEE Intelligent Systems, v. 16, n. 2, p. 60-71.

Pottier, Bernard. 1985. Linguistique Générale: théorie et description. 2. ed. Paris: Éditions Klincksieck.

Pustejovsky, James. 1995. The Generative Lexicon. Cambridge: The MIT Press.

Reeve, Lawrence H.; Han, Hyoil. 2007. The Use of Domain- Specific Concepts in Biomedical Text Summarization. Information Processing and Management, v.43, n.6, p. 1765–1776.

Rino, Lúcia Helena Machado; Pardo, Thiago Alexandre Salgueiro (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. In: Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial, p. 203-245.

Zavaglia, Claudia. 2002. Análise da homonímia no português: tratamento semântico com vistas a procedimentos computacionais. Araraquara (Tese de doutorado).

3.14. Material nanoestruturado / Nanomaterial
3.14.1. Material nanoestruturado bidimensional / Nanomaterial bidimensional
3.14.1.1. Filme fino / Poço quântico
3.14.2. Material nanoestruturado unidimensional / Nanomaterial unidimensional
3.14.2.1. Fio quântico
3.14.2.1.1. Nanotubo
3.14.2.1.1.1. Nanotubo de carbono
3.14.2.1.1.1.1. Nanotubo de carbono de parede múltipla/ Nanotubo de carbono de múltiplas paredes
3.14.2.1.1.1.2. Nanotubo de carbono de parede simples / Nanotubo de carbono de parede única
3.14.2.1.2. Nanofio
3.14.2.1.3. Nanofita
3.14.2.2. Nanobastonete
3.14.3. Material nanoestruturado zero dimensional / Nanomaterial zero-dimensional
3.14.3.1. Nanofibra
3.14.3.1.1. Nanofibra de carbono
3.14.3.2. Nanopartícula
3.14.3.2.1. Nanopartícula de hidrogel
3.14.3.2.2. Nanopartícula de metal
3.14.3.2.2.1. Nanopartícula de ferrita
3.14.3.2.2.2. Nanopartícula de ferro
3.14.3.2.2.3. Nanopartícula de ouro
3.14.3.2.2.4. Nanopartícula de prata
3.14.3.2.3. Nanopartícula de ni
3.14.3.2.4. Nanopartícula de óxido
3.14.3.2.5. Nanopartícula de semicondutor
3.14.3.2.6. Nanopartícula de sílica
3.14.3.2.7. Nanopartícula polimérica
3.14.3.2.7.1. Nanocápsula / Lipossoma
3.14.3.2.7.2. Nanoesfera
3.14.3.3. Ponto quântico / Quantum dot
3.14.3.3.1. Nanocristal
3.14.4. Material nanoporoso
3.14.5. Nanocompósito
3.14.5.1. Nanocompósito cerâmico / Nanocompósito de matriz cerâmica
3.14.5.2. Nanocompósito polimérico / Nanocompósito de matriz polimérica
3.14.5.3. Nanoporo
3.14.6. Nanohélice
3.14.7. Nanoimã
3.14.8. Nanomola
3.14.9. Nanomotor
3.14.10. Nanorobô
3.14.11. Nanorotor
3.14.12. Nanossensor

Figura 7: Classe “Material nanoestruturado” e suas subclasses.

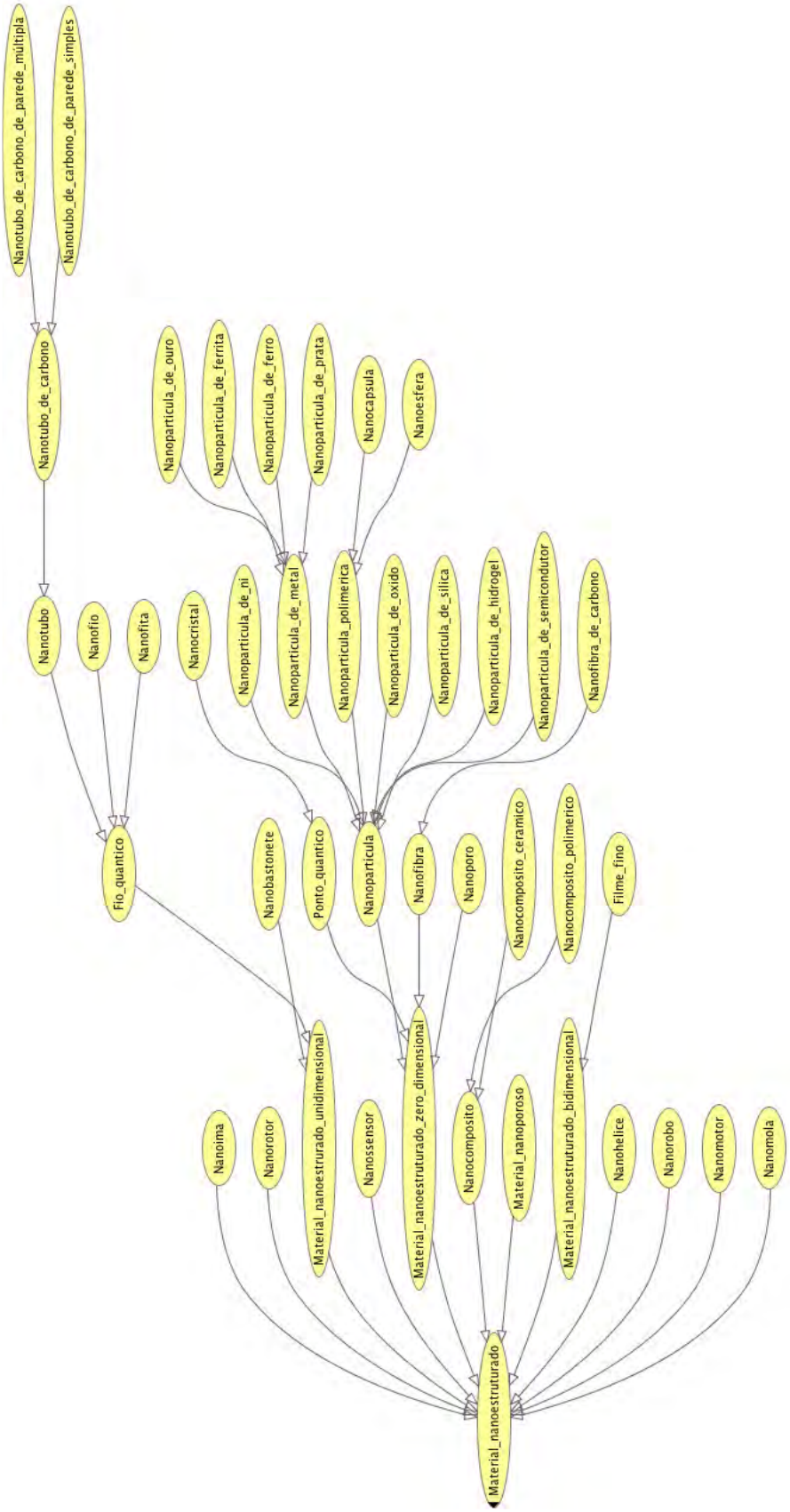

Figura 8: Classe “Material nanoestruturado” e suas subclasses.

Anexo: Concordâncias

Reproduzimos aqui, a título de exemplo, algumas das concordâncias geradas pelo Unitex segundo os critérios de busca descritos na seção 4.4.

```
eja realmente na superfície do material é a aplicação de um "coating" (recobrimento) do aditivo recobri
] {S}O objetivo desta etapa do processo é a aplicação de um filme uniforme de fotorresiste sobre o subs
létrica. {S}A constante dielétrica (k') é a propriedade determinante da capacitância do circuito, sendo
James (2005) a resistência ao glifosato é a propriedade mais frequente nestes cultivos, presente em 72%
izadas no projeto.{S} A fonte principal é a Teoria do Controle Ótimo, que aborda entre outros fatores,
izadas no projeto.{S} A fonte principal é a Teoria do Controle Ótimo, que aborda entre outros fatores,
droga, que teve sua fórmula patenteada, é material de estudo do doutorando Raul Ribeiro, orientado pela
ndutores.{S} A hidroxiapatita sintética é material inorgânico composto por fosfato de cálcio que tem si
a dimensão atómica.{S} O sistema de MBE é um equipamento sofisticado.{S} Equipamentos mais versáteis po
rredura por tunelamento (STM).{S} O STM é um equipamento sofisticado e de uso dedicado, permitindo um o
rredura por tunelamento (STM).{S} O STM é um equipamento sofisticado e de uso dedicado, permitindo um o
pa 2: {S}? O precipitador eletrostático é um equipamento apto para a remoção de nanopartículas, obtendo
}B.3.{S} Perfilômetro {S}O perfilômetro é um equipamento de medida mecânica de perfis ou topologia de f
ubstrato de silício. {S}Metalização:{S} É um método de deposição de um filme de metal que pode ser feit
am que o método de redução carbotérmica é um método viável para o crescimento de nanoestruturas unidime
lhos demonstrando que a moagem mecânica é um método eficiente de obtenção de espinélios LiMn2O (KOSOVA
rabalhos demonstram que moagem mecânica é um método eficiente para controle das característica morfológ
 de CVD [17]. {S}- A Implantação Iônica é um método de modificação superficial no qual um feixe de íons
ipais componentes.{S} Em síntese, a PCA é um método que tem por finalidade básica a redução de dados a
m a carne morta em tecido vivo.{S} Esse é um método para vencer a morte e promover a ressurreição dos s
. "O que mais nos entusiasma é que este é um método modular de montagem que irá nos permitir juntar pra
e o método de redução carbotérmica, que é um método na qual os óxidos são misturados com carbono para p
conhecido por redução carbotérmica, que é um método de simples utilização, mas que não tem sido muito e
Image @2005 AIST. {S}A tecnologia LIBWE é um método de uma etapa para a microfabricação de placa de vid
o por Lift-off [01, 14, 22] {S}Lift-off é um método simples que é muito utilizado na definição de linha
```

Quadro 2: Expressões que apontam para subdomínios.

```
esquisa, assim como a de pesquisadores, é feita por uma quantificação mais abrangente, o número de arti
an a decomposição da radiação espalhada é feita por meio de grades d difração, enquanto que no espalham
 substrato. {S}A desidratação da lâmina é feita por evaporação, pelo aquecimento do substrato em uma es
dores, enquanto que a avaliação de água é feita por análise química em laboratório e são bastante demor
dores, enquanto que a avaliação de água é feita por análise química em laboratório e são bastante demor
 potável. {S}A regeneração do nanofilme é feita por aquecimento do material. {S}Também é possível se ob
ente, a avaliação do sabor dos produtos é feita por pessoas especialmente treinadas, que analisam senso
gravação, leitura e desgravação do bits é feita por agulhas do tipo usado em microscópios de varredura
ssim como de todos os programas do PPA) é feito por meio do sistema de informações gerenciais do Minist
500.{S} O controle da pressão na câmara é feito por um sensor Pirani Balzers modelo TPR250, os fluxos d
ntrada/saída especificados.{S} O ajuste é feito por ciclos em que a cada entrada apresentada à rede os
ão, menos de 15% do gasto privado total é feito por empresas com menos de 250/300 empregados.{S} O mesm
forme. {S}Por isso, as medidas de cores são feitas por meio de métodos espectrais.{S} Neste caso, o equ
m tempo e custo elevados, e as análises são feitas por amostragem ao invés de medidas em tempo real. {S
rios com alguns nanômetros de diâmetro, são feitos por feixes ("jatos") de elétrons, obtidos de um micr
```

Quadro 3: Relação “é/são” <fazer> “por” denota relação *Agentiva*.

```
m redox/eletrodo/vidro/camada espelhada é usada em espelhos eletrocrômicos automotivos (industr
icas). {S}Hoje em dia, essa mesma idéia é usada em computadores de alto desempenho, com micropr
. {S}Nanotecnologia {S}A nanotecnologia é usada em cosméticos para trazer vantagens sensoriais
m redox/eletrodo/vidro/camada espelhada é usada em espelhos eletrocrômicos automotivos (industr
icas). {S}Hoje em dia, essa mesma idéia é usada em computadores de alto desempenho, com micropr
. {S}Nanotecnologia {S}A nanotecnologia é usada em cosméticos para trazer vantagens sensoriais
e a baixa pressão, hidrogênio molecular é usado em abundância na alimentação do gás para gerar
itrofenil-ß-D galactopiranosídeo (ONPG) é usado para detectar a enzima ß-D-galactosidase, a qua
metil-umbeliferil- ß glicuronídeo (MUG) é usado para detectar a enzima ß-glicuronidase, a qual
m chamado de BOE (Buffered Oxide Etch), é usado para corroer {S}SiO2 (óxido de silício) e SiNx
, Anritsu MS2601B. {S}A terminação (9d) é utilizada para a observação da presença da linha Bril
 fibra, (FBG - Fiber Bragg {S}Grating), é utilizada para refletir os campos ópticos chegando ao
4.3.2, extraída da Lei de Lambert-Beer, é utilizada para correlacionar a intensidade (I), a abs
ologia será baseada no esquema que hoje é utilizado em computação quântica com ressonância magn
of microparticles - LAM", Juang (1994), é utilizado para fabricação de nanopartículas em pequen
lador acusto-óptico, Intra-Action ME40, é utilizado para induzir um desvio conhecido na frequên
stão interconectados, o modelo de M. S. é utilizado para predizer a taxa de 1 densificação (ma
onente é denominada coprecipitação, que é utilizada para a obtenção de óxidos mistos, pois, per
```

Quadro 4: Busca por relações *Télicas* utilizando os verbos “utilizar” e “usar”.

Módulo de acentuación para o galego en Freeling

Miguel Anxo Solla Portela
Universidade de Vigo
miguelsolla@uvigo.es

Resumo

Descrición do módulo de acentuación para a lingua galega que se desenvolveu para a súa inclusión en vindeiras versións da biblioteca de ferramentas de análise lingüística Freeling.

1. Introdución

A biblioteca de ferramentas de análise lingüística Freeling ofrece amplas posibilidades no desenvolvemento de aplicacións lingüísticas para un conxunto de linguas cada vez máis extenso, polo momento: inglés, español, catalán, galego, italiano, portugués, asturiano e galés. Ata a versión actual, Freeling vén empregando no recoñecemento de formas resultantes da segmentación dunha afixación as mesmas funcións de restauración ou de supresión do acento gráfico para o galego que para o español. Esta característica limita as posibilidades do recoñecemento morfolóxico de numerosas raíces en lingua galega, xa que as regras de afixación en Freeling contemplan tanto a afixación léxica mediante prefixos e sufixos coma a segmentación de formas verbais con pronomes enclíticos, que é a posición habitual ou non-marcada do pronome persoal en lingua galega. A frecuencia deste tipo de secuencias xunto cun tratamento inadecuado da restauración da acentuación gráfica nas formas verbais segmentadas producen anomalías na análise. No entanto, Freeling é un proxecto de código aberto cunha atinada arquitectura modular para as especificidades de cada lingua, que permite o desenvolvemento de código para o tratamento do acento gráfico en cada lingua sen que interfira coas necesidades das demais. Co fin de evitar as interferencias do tratamento da acentuación gráfica de raíces para o castelán, desenvolveuse un novo módulo para o procesamento da acentuación gráfica que reúne un conxunto de funcións específicas que actúan sobre estas formas tras a segmentación da afixación consonte as regras de acentuación da lingua galega e remodeláronse as regras de afixación dos datos lingüísticos do galego para obter a forma illada no dicionario da aplicación.

2. O tratamento da acentuación gráfica das formas afixadas en Freeling

As regras de afixación para cada lingua da biblioteca atópanse no ficheiro afixos.dat dos datos lingüísticos correspondentes.

Freeling diferencia as regras de segmentación de elementos que anteceden na secuencia á forma que debe buscar no dicionario (*prefixes*) das regras para secuencias nas que debe segmentar un elemento ao final da secuencia (*sufixes*). Neste último grupo inclúense tanto as regras de sufixación léxica coma as de segmentación de formas verbais e pronomes enclíticos, pero os parámetros que permite establecer a aplicación en cada regra rexen comportamentos moi diferentes:

```
mente *  ^AQ0CS  RG 1  0 0  L  1  -
lle   *  ^V      *  0  1 0  L  1  $$+lle:$$+PP
```

A regra para o sufixo *-mente* vai segmentar esta terminación, activar a función de acentuación para bases de sufixos léxicos (5ª columna) que crea un candidato sen ningún acento gráfico e un candidato con acento en cada unha das vogais que conteña a raíz e, se atopa unha base adxectival, etiquetará como adverbio o derivado deadxectival; mentres que a regra para o sufixo *-lle* vai segmentar a raíz e procesala coas funcións de acentuación gráfica para as formas verbais segmentadas (6ª columna) e, se atopa no dicionario unha forma verbal, etiquetará os dous segmentos da secuencia, a forma verbal flexionada e o pronome persoal.

Porén, as funcións de restauración ou supresión do acento gráfico para o español non resultan adecuadas en raíces verbais galegas, xa que as regras de acentuación gráfica do español difiren das do galego na acentuación diacrítica, na silabación de certos encontros vocálicos (español *atribuimos* / galego *atribuímos, atribuíu, atribuïamos*) e na consideración das secuencias polisilábicas que rematan en ditongo decrecente ou en ditongo decrecente seguido de *-n* ou *-s* (español *comeréis, fuereis* / galego *comerei, amábeis*). Ademais, os encontros cos pronomes enclíticos presentan particularidades propias: tres alomorfos para o pronome persoal acusativo de terceira persoa en distribución complementaria segundo a terminación verbal (*la, las, lo* ou *los* tras as formas que rematan en *-r* ou *-s*: *comerala* ~ *comerás* + *a*; *na, nas, no* ou *nos* tras as formas que rematan en ditongo decrecente: *comereinas* ~ *comerei* + *as*; e *a, as, o*

ou *os* nos casos restantes: *cómoa* ~ *como* + *a*) e perda do *-s* final das formas flexionadas en primeira persoa do plural cando levan enclítico un pronome persoal que tamén sexa de primeira persoa de plural (*achegámonos* ~ *achegamos* + *nos*).

Co fin de procesar estas características, optouse polo desenvolvemento dun novo módulo de acentuación para a lingua galega, que mantén o tratamento da acentuación gráfica para a afixación léxica que operaba co módulo do español, pero que establece un novo comportamento para engadirlles, manterlles ou suprimirlles a acentuación gráfica ás formas verbais segmentadas. Asemade, revisáronse as regras de afixación para que operasen consonte o novo módulo e retomouse un ficheiro moi similar ao sufixos.dat da versión 2.0 de Freeling que resulta máis acorde, pois as versións posteriores intentaron suplir as limitacións na análise co módulo de acentuación para o español ampliando o número de regras para cada pronome mediante a inclusión da parte final da terminación verbal, mais esta ampliación non é necesaria cun comportamento máis axeitado do módulo de acentuación específico. Nesta regresión do ficheiro afixos.dat os cambios respecto do sufixos.dat da versión 2.0 son pouco significativos:

- Adaptouse o nome do ficheiro e o número de columnas ao das versións actuais da biblioteca.
- Aumentáronse as regras para o encontro do alomorfo do pronome persoal átono acusativo cando vai enclítico a unha forma verbal que termina en ditongo decrecente (ditongos *-ei* e *-íu*).
- Incluíronse novas regras para o encontro de formas verbais en 1ª persoa de plural cun pronome –ou unha secuencia de pronomes que o leve en pirmeira posición– de 1ª persoa de plural (cruz*ámonos* ~ cruz*amos* + *nos*).
- Engadíronse algunhas regras para secuencias de dous pronomes enclíticos (*cheme*, *chelle*...).
- Incluíuse unha regra regexp para tratar de evitar a segmentación da primeira forma do pronome persoal acusativo de terceira persoa nas secuencias nas que a presenza lle corresponda a algún dos seus alomorfos. Deste xeito, impídese, por exemplo, a separación de *comeras* en *comer* + *as*, pois este encontro sempre se resolve en galego co segundo alomorfo do pronome de acusativo: *comelas*):

```
Comeras comer VMIM2S0 1
Comelas comer+o VMN0000+PP3FPA00 0.330674
comer+o VMN03S0+PP3FPA00 0.330674
comer+o VMN01S0+PP3FPA00 0.330674
comer+o VMSF3S0+PP3FPA00 0.00398928
comer+o VMSF1S0+PP3FPA00 0.00398928
```

Porén, esta limitación non resulta suficiente en todos os casos, particularmente co terceiro alomorfo do pronome acusativo *-na, -nas, -no* ou *-nos* que aparece tras ditongo decrecente, xa que a restrición da segmentación imponse tras a etiquetación de Freeling sobre as formas da flexión verbal. As formas flexivas que rematan en *-r* ou *-s* son constantes na flexión de todos os verbos, pero algunhas formas flexivas que rematan en ditongo decrecente non presentan unha recorrencia tan constante: a terceira persoa de singular do pretérito de indicativo de todos o verbos regulares remata en ditongo decrecente (*andou, varreu, partiu*), mais moitos irregulares adoptan outras terminacións (*dixo, houbo*); na primeira persoa deste mesmo tempo, os verbos regulares da primeira conxugación rematan en ditongo decrecente (*andei*) fronte aos da segunda e da terceira conxugación (*varrín, partín*); algúns verbos irregulares teñen formas que rematan en ditongo decrecente na primeira e na terceira persoa do singular do presente de indicativo (*dou, vou, hei, sei, fai, vai*)... Como non se atopou unha restrición satisfactoria para evitar que a regra da primeira forma do acusativo se aplique tras un ditongo decrecente, nalgunhas secuencias que presentan esta ditongación seguida de *-a, -as, -o* ou *-os* prodúcese unha segmentación incorrecta*:*

```
Vaia ir VMSP3S0 0.477778 ir VMSP1S0 0.477778 ir VMM03S0 0.0111111 vaia I 0.0111111 ir+o VMIP3S0+PP3FSA00 0.0111111 ir+o VMM02S0+PP3FSA00 0.0111111
Faios faio NCMP000 0.758333 faios NP00000 0.210714 facer+o VMIP3S0+PP3MPA00 0.0297619 facer+o VMM02S0+PP3MPA00 0.00119048
```

Nos dous exemplos, as dúas últimas análises parten dunha segmentación incorrecta, pois a forma *vaia* non admite unha segmentación en *vai* + *a*, xa que esta secuencia se resolve co terceiro alomorfo do pronome, *vaina* e, do mesmo modo, *faino* é o resultado de *fai* + *o*:

```
Vaina ir+o VMIP3S0+PP3FSA00 0.997312 ir+o VMM02S0+PP3FSA00 0.00268817
Faino facer+o VMIP3S0+PP3MSA00 0.997312 facer+o VMM02S0+PP3MSA00 0.00268817
```

- Engadiuse tamén outra regra regexp que evita a segmentación do pronome enclítico *se* (ou de secuencias de pronomes que comecen por *se*) con formas verbais flexionadas en primeira ou en segunda persoa.

3. Comportamento da acentuación gráfica no novo módulo

Cada regra de segmentación dun afixo establece o tratamento da acentuación gráfica que requiren os candidatos a formas de dicionario. A candidatura vai recibir un procesamento moi diferente cando provén da segmentación de pronomes persoais enclíticos de cando provén dunha afixación léxica.

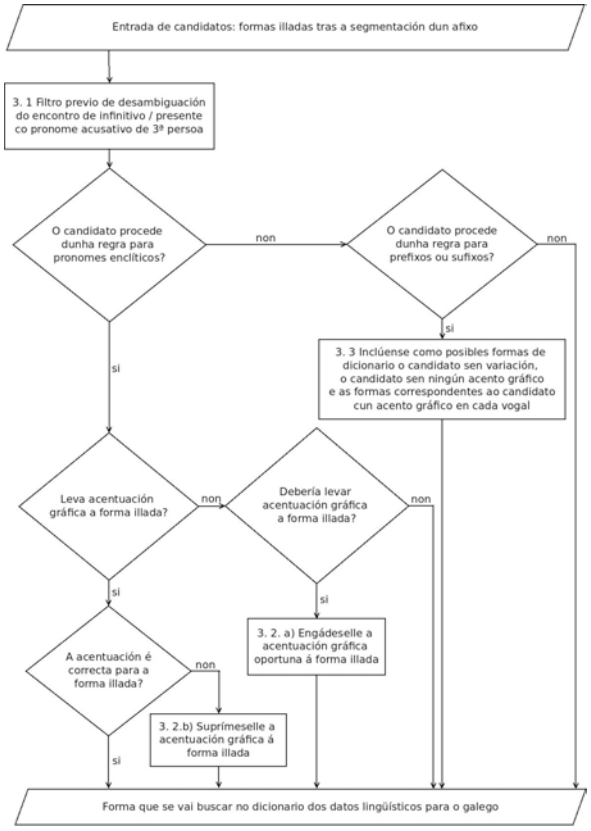

Ilustración 1. Diagrama de fluxo do módulo

3.1 O filtro de desambiguación

Incluíuse unha función de desambiguación de formas resultantes de segmentar o alomorfo do pronome persoal átono de terceira persoa *-la, -las, -lo, -los*. A regra para este enclítico establece que se lle debe engadir un *-s* ou un *-r* á raíz verbal para recuperar a forma verbal no dicionario da aplicación. A función de desambiguación compara ambas as dúas candidaturas e examina se o acento que tiña co enclítico ten función fonolóxica antes de que se apliquen as demais funcións que determinarán se lle corresponde ou non levar acento gráfico á forma verbal resultante; isto é, diferencia exemplos como *comela* (~ *comer* + *a*) de *cómela* (~ *comes* + *a*) e mantén a dobre posibilidade de análise con determinados presentes de indicativo polisilábicos oxítonos (*prevelos ~ prevés + os / prever + os*). O motivo de establecer este filtro con anterioridade ao tratamento da acentuación gráfica é que, cando a ambas as dúas formas illadas lles corresponde eliminar a acentuación gráfica, a desambiguación a posteriori xa non sería posible, pois calquera das dúas constitúe unha entrada no dicionario.

3.2 Formas verbais de secuencias con pronomes persoais enclíticos

O tratamento que recibe o acento gráfico cando unha regra activa o módulo de acentuación para formas verbais resultantes da segmentación de pronomes enclíticos é, a grandes trazos, o seguinte:

a) Cando da segmentación se obtén unha forma verbal que non ten acento gráfico, compróbase que non se trate dunha forma polisílabica, que termine nas vogais *-a, -e, -o,* e *-i* cando non forma parte dun ditongo decrecente nin en ningunha das terminacións anteriores e mais un *-n* ou un *-s* final, que levase enclítico un pronome persoal monosílabo. Se se trata dun destes casos, incorpórase o acento gráfico da forma verbal oxítona para validar a forma no dicionario (*prevese ~ prevé + se*, *darasme ~ darás + me*) e, se non, valídase sen o acento gráfico (*deille ~ dei + lle, enviounos ~ enviou + nos*).

b) Cando da segmentación resulta unha raíz con acento gráfico, compróbase que o acento sexa correcto:

- Acentos diacríticos (*dálle ~ dá + lle*).
- Segunda persoa do singular ou terceira persoa, singular ou plural, do futuro de indicativo de todos os verbos e segunda ou terceira persoa de singular ou terceira de plural de formas oxítonas de certos verbos en presente de indicativo, que levasen enclítica unha secuencia polisílaba de pronomes persoais (*fáranche*~*me ~ farán + che + me, estánvola ~ están + vos + a*).
- Primeira ou segunda persoa do plural do pretérito de subxuntivo (*cantásemoslles ~ cantásemos + lles*).
- Acentuación dunha vogal pechada que marca un hiato (*sabíao ~ sabía + o, saínlles ~ saín + lles, constituíuna ~ constituíu + a*).

Se non se trata de ningún dos casos anteriores, elimínase o acento gráfico da forma resultante (*quíxeno ~ quixen + o, perseguíndoas ~ perseguindo + as, cáelles ~ cae + lles, caéralle ~ caera + lle, tróuxoma ~ trouxo + me + a, atéivolas ~ atei + vos + as, cantáballe ~ cantaba + lle*).

3.3 Afixación léxica

As regras dun prefixo ou dun sufixo seguen contando, coma no caso do español, cunha posibilidade diferente no módulo de acentuación, unha función específica coa que a forma candidata vaise reconstruír en varias: unha forma sen ningún acento gráfico e esa mesma forma con acento gráfico en cada unha das vogais que conteña. Cada unha destas formas vaise procurar no dicionario. Deste xeito, aínda que *calidamente* non figura no dicionario, Freeling identifica que se trata dunha

derivación de *cálido* grazas á regra do sufixo *-mente* que segmenta e reconstrúe a forma candidata, e activa este tratamento da acentuación no ficheiro afixos.dat dos datos lingüísticos para o galego.

4. O código do módulo

O código do módulo de acentuación incorporouse ao repositorio de subversion da versión en desenvolvemento de Freeling e pódese obter mediante a instrución `svn checkout http://devel.cpl.upc.edu/freeling/svn/latest/freeling`.

Ademais das modificacións que xa se viron para o ficheiro dos datos lingüísticos coas regras de afixación, o ficheiro accents.cc modificouse para que a análise en lingua galega deixe de utilizar o módulo de acentuación para o español e pase a utilizar o módulo novo. No ficheiro accents_modules.h decláranse as clases e as funcións que se definen no ficheiro accents_modules.cc, no que figuran diferentes funcionalidades de adecuación da acentuación gráfica para as linguas que as precisan.

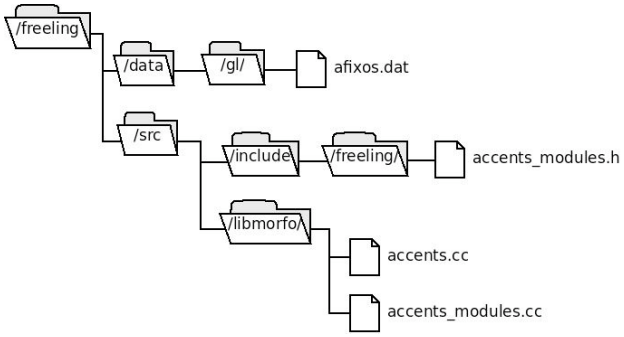

Ilustración 2. Rutas dos ficheiros que se modificaron na versión en desenvolvemento.

5. Avaliación dos resultados

Para a avaliación dos resultados analizouse o ficheiro corpus_economia_prensa_oracions.txt do *Corpus xiada*, versión 2.4, que distribúe o Centro Ramón Piñeiro para a Investigación en Humanidades baixo os termos da licenza Lesser General Public License for Linguistic Resources. O corpus analizouse primeiro coa versión estable de Freeling 2.2 e despois coa versión en desenvolvemento. O ficheiro contén, consonte o cómputo do editor de textos, 205.370 palabras.

Nesta análise sobre o mesmo corpus, a cantidade de secuencias lingüísticas que a versión en desenvolvemento segmenta como formas verbais con pronomes enclíticos (3.213) practicamente duplica as secuencias que atopa a versión 2.2 de Freeling (1.636). A continuación figuran algúns exemplos deste comportamento:

Resultados con Freeling 2.2	**Resultados coa versión en desenvolvemento**
faise faise NCFS000 0.894226 faise VMSI3S0 0.0769279 faise VMSP3S0 0.0288462	faise facer+se VMIP3S0+PP3CN000 1
reclamándollo reclamándollo NCMS000 1	reclamándollo reclamar+lle+o VMG0000+PP3CSD00+PP3MSA00 1
adoptárense adoptárense NP00000 1	adoptárense adoptar+se VMN03P0+PP3CN000 1
encontrámonos encontrámonos NCMP000 0.962947 encontrámonos AQ0MP0 0.0370529	encontrámonos encontrar+nos VMIP1P0+PP1CP000 0.5 encontrar+nos VMIS1P0+PP1CP000 0.5
vaise vaise NCFS000 0.894226 vaise VMSI3S0 0.0769279 vaise VMSP3S0 0.0288462	vaise ir+se VMIP3S0+PP3CN000 1
Báixansenos báixansenos NP00000 1	Báixansenos baixar+se+nos VMIP3P0+PP3CN000+PP1CP000 1
déixenos déixenos RG 0.893127 déixenos AQ0MP0 0.0733362 déixenos NCMP000 0.0335372	déixenos deixar+o VMSP3P0+PP3MPA00 0.333269 deixar+nos VMSP3S0+PP1CP000 0.333269 deixar+nos VMSP1S0+PP1CP000 0.333269 deixar+o VMM03P0+PP3MPA00 9.62927e-05 deixar+nos VMM03S0+PP1CP000 9.62927e-05
mantela mantela NCFS000 1	mantela manter+o VMIP2S0+PP3FSA00 0.6 mantela NCFS000 0.1 manter+o VMN0000+PP3FSA00 0.1 manter+o VMN03S0+PP3FSA00 0.1 manter+o VMN01S0+PP3FSA00 0.1
subilo subilo NCMS000 0.470691 subilo NP00000 0.382288 subilo AQ0MS0 0.14702	subilo subir+o VMN0000+PP3MSA00 0.330674 subir+o VMN03S0+PP3MSA00 0.330674 subir+o VMN01S0+PP3MSA00 0.330674 subir+o VMSF3S0+PP3MSA00 0.00398928 subir+o VMSF1S0+PP3MSA00 0.00398928
faino faino AQ0MS0 0.409727 faino NCMS000 0.360778 faino NP00000 0.213106 faino VMIP1S0 0.0163881	faino facer+o VMIP3S0+PP3MSA00 0.997312 facer+o VMM02S0+PP3MSA00 0.00268817
Dío dío NP00000 1	Dío dicir+o VMM02S0+PP3MSA00 1
podérense podérense NP00000 1	podérense poder+se VMN03P0+PP3CN000 1
foise foise NCFS000 0.894226 foise VMSI3S0 0.0769279 foise VMSP3S0 0.0288462	foise ir+se VMIS3S0+PP3CN000 1
Dise dise NP00000 1	Dise dicir+se VMIP3S0+PP3CN000 1

Para ilustrar os resultados da aplicación do novo módulo, contrastouse o número de análises destas secuencias coa cantidade de veces en que a etiquetación era certeira.

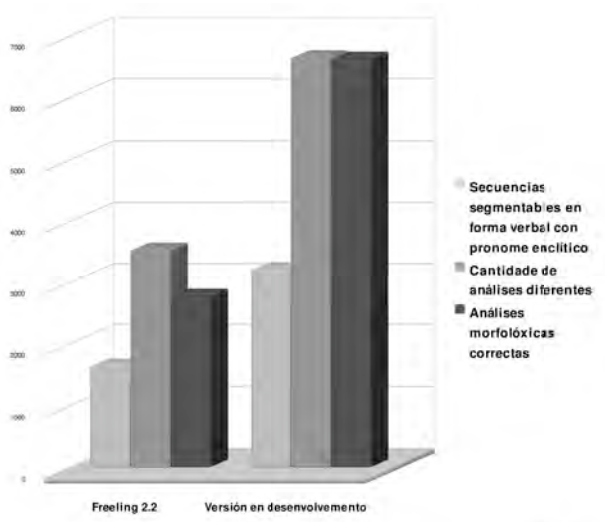

Ilustración 3. Gráfico comparativo das secuencias de forma verbal con pronomes enclíticos nas etiquetacións do *Corpus xiada*.

Consideráronse análises incorrectas as que contiñan descricións morfolóxicas da secuencia inadecuadas e as que etiquetaron o pronome persoal *se* en énclise cunha forma verbal flexionada en 1ª ou 2ª persoa. Deste xeito, constatouse que a marxe de erro da versión 2.2 de Freeling era bastante superior (21,41%, 756 etiquetacións erróneas nas 3.531 análises das 1.636 secuencias segmentadas) respecto da versión en desenvolvemento (0,30%, 20 análises erradas nas 6.647 etiquetacións de formas verbais con pronomes enclíticos nas 3.213 secuencias que se detectaron).

As etiquetacións que se obtiveron coa versión en desenvolvemento resultaron máis axeitadas ca as da versión estable e ofrecen, en xeral, unha descrición máis precisa de secuencias homógrafas:

Resultados con Freeling 2.2	**Resultados coa versión en desenvolvemento**
mantela mantela NCFS000 1	mantela manter+o VMIP2S0+PP3FSA00 0.6 mantela NCFS000 0.1 manter+o VMN0000+PP3FSA00 0.1 manter+o VMN03S0+PP3FSA00 0.1 manter+o VMN01S0+PP3FSA00 0.1
ilusionante ilusionar+te VMIP3P0+PP2CSA00 1	ilusionante ilusionante AQ0CS0 0.5 ilusionante AQ0MS0 0.5
convertela converter+o VMIP2S0+PP3FSA00 0.880952 converter+o VMN0000+PP3FSA00 0.0238095 converter+o VMN03S0+PP3FSA00 0.0238095 converter+o VMN01S0+PP3FSA00 0.0238095 converter+o VMSF3S0+PP3FSA00 0.0238095 converter+o VMSF1S0+PP3FSA00 0.0238095	convertela converter+o VMN0000+PP3FSA00 0.330674 converter+o VMN03S0+PP3FSA00 0.330674 converter+o VMN01S0+PP3FSA00 0.330674 converter+o VMSF3S0+PP3FSA00 0.00398928 converter+o VMSF1S0+PP3FSA00 0.00398928
serrano serrar+o VMIP3P0+PP3MSA00 1	serrano serrano AQ0MS0 1
préstamos préstamo NCMP000 1	préstamos préstamo NCMP000 0.857639 prestar+me+o VMIP3S0+PP1CS000+PP3MPA00 0.140556 prestar+me+o VMM02S0+PP1CS000+PP3MPA00 0.00180556
importe importe NCMS000 0.988095 importar VMM03S0 0.00396825 importar VMSP3S0 0.00396825 importar VMSP1S0 0.00396825	importe importe NCMS000 0.986395 importar VMM03S0 0.00226757 importar VMSP3S0 0.00226757 importar VMSP1S0 0.00226757 impor+te VMN0000+PP2CSA00 0.00226757 impor+te VMN03S0+PP2CSA00 0.00226757 impor+te VMN01S0+PP2CSA00 0.00226757
verse versar VMSP3S0 0.49988 versar VMSP1S0 0.49988 versar VMM03S0 0.000240674	verse ver+se VMN0000+PP3CN000 0.391657 ver+se VMN03S0+PP3CN000 0.391657 versar VMSP3S0 0.107597 versar VMSP1S0 0.107597 versar VMM03S0 0.00149195
dálle dar+lle VMM02S0+PP3CSD00 1	dálle dar+lle VMIP3S0+PP3CSD00 0.997312 dar+lle VMM02S0+PP3CSD00 0.00268817
querela querela NCFS000 1	querela querela NCFS000 0.321672 querer+o VMN0000+PP3FSA00 0.226109 querer+o VMN03S0+PP3FSA00 0.226109 querer+o VMN01S0+PP3FSA00 0.226109
días día NCMP000 1	días día NCMP000 0.991968 dicir+o VMIP3S0+PP3FPA00 0.00401606 dicir+o VMM02S0+PP3FPA00 0.00401606
vaise vaise NCFS000 0.894226 vaise VMSI3S0 0.0769279 vaise VMSP3S0 0.0288462	vaise ir+se VMIP3S0+PP3CN000 1
Tense tensar VMSP3S0 0.49988 tensar VMSP1S0 0.49988 tensar VMM03S0 0.000240674	Tense ter+se VMIP3S0+PP3CN000 0.86083 tensar VMSP3S0 0.069332 tensar VMSP1S0 0.069332 tensar VMM03S0 0.000506073
quedarmos quedar VMN01P0 0.75 quedar VMSF1P0 0.25	quedarmos quedar VMN01P0 0.228571 quedar+me+o VMN0000+PP1CS000+PP3MPA00 0.228571 quedar+me+o VMN03S0+PP1CS000+PP3MPA00 0.228571 quedar+me+o VMN01S0+PP1CS000+PP3MPA00 0.228571 quedar VMSF1P0 0.0285714 quedar+me+o VMSF3S0+PP1CS000+PP3MPA00 0.0285714 quedar+me+o VMSF1S0+PP1CS000+PP3MPA00 0.0285714
afaste afastar VMSP3S0 0.444444 afastar VMSP1S0 0.444444 afastar VMM03S0 0.111111	afaste afastar VMSP3S0 0.416667 afastar VMSP1S0 0.416667 afastar VMM03S0 0.0833333 afacer+te VMIP2S0+PP2CSA00 0.0833333

Con todo, na versión en desenvolvemento aínda se xeran análises que parten de segmentacións incorrectas nalgúns casos. A forma *vaia* aparece 7 veces no corpus e produce 14 análises que parten dunha segmentación errónea, pois, como xa se viu, as dúas últimas análises do exemplo correspóndenlle á secuencia *vainas*:

```
vaia ir VMSP3S0 0.477778 ir VMSP1S0 0.477778 ir VMM03S0
0.0111111 vaia I 0.0111111 ir+o VMIP3S0+PP3FSA00 0.0111111 ir+o
VMM02S0+PP3FSA00 0.0111111
```

As restantes análises erróneas non aparecen con tanta frecuencia no corpus:

```
explícaselles explicar+lle VMSI3S0+PP3CPD00 0.492234
explicar+lle VMSI1S0+PP3CPD00 0.492234 explicar+se+lle
VMIP3S0+PP3CN000+PP3CPD00 0.0155317

débeselle deber+lle VMSI3S0+PP3CSD00 0.492234 deber+lle
VMSI1S0+PP3CSD00 0.492234 deber+se+lle VMIP3S0+PP3CN000+PP3CSD00
0.0155317

predios predio NCMP000 0.857639 predicir+o VMIP3S0+PP3MPA00
0.140556 predicir+o VMM02S0+PP3MPA00 0.00180556

dias dicir+o VMIP3S0+PP3FPA00 0.997312 dicir+o VMM02S0+PP3FPA00
0.00268817
```

As dúas primeiras análises das secuencias *explícaselles* e *débeselle* correspóndense, en realidade, coa análise atinada das secuencias *explicáselles* e *debéselle* respectivamente:

```
Explicáselles explicar+lle VMSI3S0+PP3CPD00 0.5 explicar+lle
VMSI1S0+PP3CPD00 0.5

Debéselles deber+lle VMSI3S0+PP3CPD00 0.5 deber+lle
VMSI1S0+PP3CPD00 0.5
```

A segmentación da secuencia *predios* nunha forma verbal cun pronome enclítico tamén é errónea, xa que esta análise correspóndelle á secuencia *predíos*:

```
Predíos predicir+o VMIP3S0+PP3MPA00 0.997312 predicir+o
VMM02S0+PP3MPA00 0.00268817
```

No caso da secuencia **dia,* trátase dun erro ortográfico da forma *día*:

```
Día día NCMS000 0.758333 día NP00000 0.210714 dicir+o
VMIP3S0+PP3FSA00 0.0297619 dicir+o VMM02S0+PP3FSA00 0.00119048
```

6. Conclusións e traballo futuro

A etiquetación morfolóxica das secuencias de formas verbais con pronomes enclíticos mellorou sensiblemente co desenvolvemento do novo módulo de acentuación para a lingua galega. Unha identificación máis adecuada dos núcleos verbais debera mellorar tamén outras análises da biblioteca, como a análise de dependencias, así como o funcionamento xeral doutras aplicacións que utilicen Freeling.

O módulo que se desenvolveu estase adaptando á formalización das futuras versións de Freeling coa pretensión de acadar os mesmos resultados que na versión de desenvolvemento do Freeling 2.2 que se vén de describir, e tamén coa intención de tratar de deseñar estratexias que eviten as etiquetacións erróneas que se detectaron ata o momento.

Referencias

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, e Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, páxinas 48-55.

Real Academia Galega / Instituto da Lingua Galega, *Normas ortográficas e morfolóxicas do idioma galego*, 18ª edición, 2003.

Freeling user manual, 2.2, September 2010 `http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf`.

Technical reference manual, 2.2, `http://nlp.lsi.upc.edu/freeling/doc/refman`.

Centro Ramón Piñeiro para a Investigación en Humanidades, *Etiquetador/Lematizador do Galego Actual (XIADA),* versión 2.4, *`http://corpus.cirp.es/xiada`*, [Consultado o: 20/10/2010].

Apresentação de Projectos

P-PAL: Uma base lexical com índices psicolinguísticos do Português Europeu

Ana Paula Soares[1], Montserrat Comesaña[1], Álvaro Iriarte[2], José João de Almeida[3], Alberto Simões[3], Ana Costa[4], Patrícia Cunha França[4], João Machado[4]

[1]Escola de Psicologia, Universidade do Minho
[2]Instituto de Letras e Ciências Humanas, Universidade do Minho
[3]Departamento de Informática, Universidade do Minho
[4]Centro de Investigação em Psicologia, Universidade do Minho

{asoares,mvila}@psi.uminho.pt, alvaro@ilch.uminho.pt
{jj,ambs}@di.uminho.pt, {ana.costa,patfranca,joaoffm}@psi.uminho.pt

Resumo

Neste trabalho apresentamos o projecto Procura-PALvras (P-PAL) cujo principal objectivo é desenvolver uma ferramenta electrónica que disponibilize informação sobre índices psicolinguísticos objectivos e subjectivos de palavras do Português Europeu (PE). O P-PAL será disponibilizado gratuitamente à comunidade científica num formato amigável a partir de um sítio na Internet a construir para o efeito. Ao utilizar o P-PAL, o investigador poderá fazer uma utilização personalizada do programa ao seleccionar, da ampla variedade de análises oferecidas, os índices que se adequam aos propósitos da sua investigação e numa dupla funcionalidade de utilização: pedir ao programa para analisar listas de palavras previamente constituídas nos índices considerados relevantes para a investigação ou para obter listas de palavras que obedeçam aos parâmetros definidos. O P-PAL assume-se assim como uma ferramenta fundamental à promoção e internacionalização da investigação em Portugal.

1 Introdução

A importância da existência de bases lexicais informatizadas que apoiem de forma efectiva a investigação nas áreas da Psicologia Cognitiva, das Neurociências, da Linguística ou do Processamento de Linguagem Natural (PLN) é, na actualidade, um dado inquestionável. Com efeito, constituindo a palavra a matéria-prima a partir da qual grande parte da investigação nessas áreas se realiza, e constituindo as palavras, em si mesmas, um estímulo complexo, que reúnem um conjunto de propriedades ou atributos cujo controlo e/ou manipulação se revelam fundamentais ao desenvolvimento profícuo de estudos nesses domínios, a investigação nacional e internacional já não se compadece mais com a inexistência deste tipo de ferramentas.

Refira-se, a título de exemplo, a sua utilidade, nas áreas mais experimentais da Psicolinguística ou das Neurociências, onde o seu apoio à selecção de estímulos (palavras) se revela essencial. Entre as características que se desejam ver devidamente manipuladas ou controladas, encontram-se tanto propriedades mais objectivas, que podem ser determinadas directamente pela análise da própria palavra (p. ex. extensão da palavra em letras ou sílabas, divisão silábica), a análise da palavra em contexto (categoria sintáctica ou informação semântica) ou derivadas da análise da relação dessa palavra com as restantes existentes no léxico (p. ex. frequência de uso da palavra na escrita e/ou na fala, similaridade ortográfica ou fonológica com outras palavras, frequência de bigrama, etc.), como propriedades de natureza mais subjectiva que implicam a recolha de medidas que reflectem as experiências pessoais dos indivíduos com o uso da própria língua (p. ex. idade-de-aquisição, imaginabilidade, familiaridade, concreteza, emocionalidade).

A manipulação sistemática destes atributos na investigação tem contribuído de forma decisiva não só para a compreensão da arquitectura e processamento linguístico humano, como para a compreensão do funcionamento de outros sistemas cognitivos como a memória, a atenção, a representação mental de conceitos ou a compreensão de determinados processos desenvolvimentais (p. ex. aquisição da fala, leitura) tanto em populações "normais" como em populações com trajectórias atípicas de desenvolvimento. Refira-se também que os contributos associados a este

tipo de ferramentas não se limitam ao seu uso como instrumento de apoio à investigação, mas também como um meio para obter um conhecimento mais aprofundado das características da própria língua. Com efeito, não apenas a criação mas também a disponibilização pública deste tipo de recursos é importante e urgente, especialmente quando se compara com os recursos existentes para outras línguas. Assim, na linguística descritiva, será uma ferramenta útil para a análise e descrição fonológica, morfossintáctica e semântica do PE, particularmente na análise quantitativa. Poderá vir a ser também um recurso muito importante para a Linguística aplicada (por exemplo, para a Lexicografia e a Terminologia do PE), fornecendo informação sobre o uso real de palavras e acepções, bem como a sua frequência, etc., assim como para a análise estilística (não apenas do ponto de vista literário, mas também pedagógico, forense, sóciolinguístico, cultural, etc.), nomeadamente graças ao trabalho de etiquetação realizado (com índices objectivos e subjectivos). Em suma, permitirá realizar estudos com base em informação descritiva, estatística e classificativa que anteriormente não estava disponível, designadamente numa única plataforma.

Para o PLN esta base de dados poderá ser utilizada em diversas vertentes, desde a simples correcção ortográfica (tendo em conta vizinhança ortográfica e fonética, por exemplo), à síntese de voz (dada a inclusão de transcrição fonética) e à análise semântica, dado o interesse do P-Pal em integrar relações semânticas.

Em Portugal, o reconhecimento da necessidade deste tipo de bases é relativamente recente. Assim, e embora tais bases se encontrem disponíveis em línguas como o inglês (p. ex. MRC (Coltheart, 1981); N-Watch (Davis, 2005); E-Lexicon (Balota et al., 2007)), o francês (p ex. BRULEX (Content, Mousty e Radeau, 1990); LEXIQUE (New et al., 2001; New et al., 2004); French Lexicon Project(Ferrand et al., 2010)), o holandês e o alemão (p. ex. CELEX (Baayen, Piepenbrock e Gulikers, 1995; Baayen, Piepenbrock e van Rijn, 1993)), o grego (p. ex. GreekLex (Ktori, van Heuven e PitchFord, 2008)), ou o espanhol (p. ex. LEXESP (Sebastián-Gallés et al., 2000); BuscaPalabras (Davis e Perea, 2005)), elas são praticamente inexistentes para o português. Até aos anos 90, o indicador psicolinguístico mais citado pelos investigadores nacionais era o de frequência de uso das palavras num trabalho designado Português Fundamental (Nascimento, Marques e da Cruz, 1987) e baseado num corpus oral de pequenas dimensões (700.000 palavras). Embora nos últimos anos se tenha reconhecido essa limitação e se tenham desenvolvido esforços no sentido de construir bases lexicais que contivessem outros indicadores linguísticos importantes, a verdade é que elas apresentam um número muito reduzido de informações. Para além da informação ortográfica disponível em todas elas (e que configura as suas entradas lexicais), cada uma contém apenas informação relativa ou à transcrição fonética ou à caracterização morfossintáctica das palavras (Nascimento, Rodrigues e Gonçalves, 1996).

Procurando ultrapassar tais dificuldades surgiu a PORLEX (Gomes e Castro, 2003). A PORLEX é uma base lexical que reúne informações de tipo ortográfico, fonológico, fonético, gramatical e de vizinhança para um total de 29.238 palavras e que constitui um instrumento útil à investigação cognitiva em geral e à da psicolinguística em particular. Contudo, as limitações que apresenta ao nível do valor de frequência lexical que disponibiliza (importado do trabalho Português Fundamental que, para além de se revelar desactualizado, apenas é disponibilizado para cerca de 5% das suas entradas lexicais) impedem um uso mais alargado dessa ferramenta na investigação nacional. Ora, na actualidade, o PE conta já com novos léxicos de frequências extraídos de corpora de grandes dimensões (p. ex. CORLEX (Nascimento, Pereira e Saramago, 2000)) e de vários corpora como o CETEMPúblico, o ECI-EE, o FrasesPP, os Clássicos da Porto Editora, o Natura/Minho, o Vercial, o Avante e o DiaCLAV disponíveis na rede, no sítio da Linguateca[1] (Costa, Santos e Cardoso, 2008). Não obstante, embora disponibilizem informação de frequência de uso mais actualizada, diversificada e representativa, não disponibilizam outras informações sobre outras propriedades lexicais das palavras, como a PORLEX. Urge assim desenvolver novas aplicações que incorporem todas estas informações numa única ferramenta.

No que se refere aos índices psicolinguísticos subjectivos, alguns autores, reconhecendo também essa lacuna nas bases nacionais e a sua relevância na investigação cognitiva e neurocognitiva mais actual, desenvolveram estudos que procuraram avaliar a familiaridade (Garcia-Marques, 2003; Marques, 2004), a valência (Garcia-Marques, 2003), a imaginabilidade e a concreteza (Marques, 2005), e a idade de aquisição (Cameirão e Vicente, 2010; Marques et al., 2007) de palavras portuguesas. Contudo, apesar da relevância desses trabalhos a verdade é que eles incidiram sobre um número bastante

[1] `http://www.linguateca.pt/ACDC/`

restrito de palavras (p. ex. 459 para o índice de familiaridade e 249 para o índice de imaginabilidade (Marques, 2004; Marques, 2005)) e, mesmo para aqueles que avaliaram os mesmos índices, a adopção de procedimentos de avaliação distintos (veja-se, por exemplo, a forma como a variável familiaridade é avaliada nos estudos de Garcia-Marques (2003) e Marques (2004); ou a idade de aquisição nos estudos Cameirão e Vicente (2010) e Marques (2005)) impede a sua utilização conjunta.

Por último, o suporte informático em que se apresentam (Microsoft Excel), embora garanta alguma flexibilidade de pesquisa, a verdade é que pode dificultar a selecção de estímulos quando, como na maioria das vezes acontece, o investigador pretende controlar um conjunto diversificado de parâmetros relativos às palavras ao mesmo tempo. Além disso, dado que as informações das palavras se encontram em suportes distintos, o investigador terá sempre de recorrer a distintas aplicações informáticas para seleccionar os estímulos apropriados, correndo sempre o risco de, nas diferentes aplicações, não encontrar as mesmas entradas lexicais. Assim, e independentemente do paradigma experimental adoptado ou da área de investigação considerada, os investigadores portugueses deparam-se na actualidade com sérias dificuldades no planeamento e condução dos estudos que utilizem estímulos verbais, e, em geral, na análise e descrição linguística do PE baseadas em corpora. Com o presente projecto pretendemos colmatar essa necessidade desenvolvendo uma aplicação informática multi-plataforma designada Procura-PALavras (P-PAL) que, com comodidade e rapidez, permita calcular, em simultâneo, um conjunto de índices psicolinguísticos objectivos e subjectivos para palavras do PE, num formato amigável e disponibilizado gratuitamente à comunidade científica a partir de um sítio em linha a construir para o efeito.

2 Procura-PALavras (P-PAL)

O P-PAL será a versão adaptada para o PE do *software* inglês N-Watch (Davis, 2005) já adaptado para o espanhol como BuscaPalabras (Davis e Perea, 2005) e Basco como E-Hitz (Perea et al., 2006) considerando as características particulares do sistema do PE contemporâneo. Permitirá, para além da computação do valor de frequência por milhão e logarítmico (base 10) de todos os lemas e formas que constituirão as suas entradas lexicais (indexadas a partir da compilação, tratamento e análise de vários corpora recentes), a realização de um conjunto diversificado de análises relativas quer às dimensões morfológicas e morfo-sintácticas (p. ex. classe gramatical, número de morfemas, frequências por tipo, ocorrência, forma e lemas por classe, género e número); quer às dimensões ortográficas (p. ex. número de letras, estrutura consoante-vogal, ponto de unicidade, homógrafos e diversas medidas de frequências por tipo e ocorrência de bi e trigramas e de vizinhanças); fonológicas (p. ex. pronúncia da palavra, número de fonemas, vogais neutras, homófonos e diversas medidas de frequências tipo e ocorrência de bifone e de vizinhanças); silábicas (p. ex. silabificação ortográfica e fonológica da palavra, número de sílabas, estrutura silábica, padrão de acento e diversas medidas de frequências tipo e ocorrência de vizinhanças silábicas ortográfica e fonológica); e semânticas (p. ex. número de acepções da palavra, co-ocorrências e distância semântica) de palavras do PE. Permitirá ainda obter índices para pseudo-palavras (que, a par das palavras, constituem estímulos de ampla utilização nos diferentes paradigmas da investigação experimental), e para os índices subjectivos de imaginabilidade, concreteza, familiaridade, valência, activação e controlabilidade, ainda não disponíveis entre nós ou, como vimos, disponíveis para um léxico bastante restrito.

Ao utilizar o P-PAL o utilizador poderá assim fazer uma utilização personalizada do programa ao seleccionar, da ampla variedade de análises disponíveis aquelas que se adequam aos propósitos da sua investigação e numa dupla possibilidade de utilização: o utilizador poderá optar por pedir ao programa que avalie um conjunto de palavras previamente definidas pelo investigador num conjunto de parâmetros seleccionados do menu de análises (p. ex. frequência lexical, número de letras, estrutura consoante-vogal, vizinhos ortográficos por substituição, adição e subtracção, frequência das formas dos vizinhos de frequência alta, distância de Levenshtein) ou poderá pedir ao programa que lhe faculte as palavras que, entre as que fazem parte da base lexical, obedeçam a esses parâmetros. Cremos que esta característica da ferramenta, não disponível na versão original do N-Watch (Davis, 2005), do BuscaPalabras (Davis e Perea, 2005) ou do E-Hitz (Perea et al., 2006) oferece maior versatilidade à ferramenta. O P-PAL assume-se assim como uma ferramenta de investigação fundamental e indispensável à promoção e internacionalização da investigação em Portugal.

3 Fases de execução do projecto

O projecto P-PAL é um projecto claramente interdisciplinar onde os contributos das áreas da Psicolinguística, da Linguística e do Processamento de Linguagem Natural (PLN) se assumem como essenciais à sua execução. Embora tais contributos sejam importantes ao longo de todo o projecto, podemos distinguir três fases principais que configuram o contributo mais acentuado de alguma delas em cada momento temporal da sua implementação.

Assim, a primeira fase do projecto, já em curso (a decorrer entre Maio de 2010 e Maio 2011), envolverá essencialmente o contributo da área da Linguística e do PLN na constituição do vocabulário por defeito a incluir no P-PAL (e que consubstanciarão as suas entradas lexicais - lemas e formas) e na extracção dos seus valores de frequência lexical (absoluta, por milhão e logarítmica – base 10). Tal tarefa compreenderá a recolha, o tratamento e a análise de vários corpora recentes do PE de diversos géneros literários e dimensões com informação de frequência de uso disponível. Ainda durante este primeiro ano de execução do projecto levar-se-á a cabo a inserção semi-automática da informação linguística estrutural das entradas lexicais do P-PAL (p. ex. informação morfo-sintáctica, transcrição fonética, silabificação, padrão de acento), a verificação e correcção da base, e a selecção do conjunto de palavras sobre as quais se recolherão medidas subjectivas. Dar-se-á também inicio à construção do interface e da aplicação na rede a partir dos quais se disponibilizarão os índices à comunidade de investigadores.

A segunda fase do projecto (Maio de 2011 – Maio 2012), envolverá essencialmente o contributo das áreas do PLN na computação das métricas de frequências por tipo e ocorrência e de vizinhanças de cada um dos índices integrados no P-PAL (índices ortográficos, fonológicos, fonográficos, silábicos ortográficos e fonológicos), e da Psicolinguística na preparação dos materiais e procedimentos na recolha presencial, lápis-papel, e a recolha via aplicação na rede, dos índices subjectivos a incluir na base (familiaridade, imaginabilidade, concreteza, valência, activação e controlo).

A terceira e última fase do projecto (Maio de 2012 – Maio 2013), envolverá essencialmente o contributo das áreas do PLN na computação das métricas semânticas a incluir na base e na computação de métricas para pseudo-palavras (frequências por tipo e ocorrência de bigrama e trigrama e de vizinhanças ortográficas e fonológicas), e da Psicolinguística na conclusão da recolha e no tratamento dos índices subjectivos a incluir no P-PAL.

4 Conclusão

O Procura-PALvras (P-PAL) é um projecto interdisciplinar que cruza as áreas da Psicolinguística, da Linguística e do Processamento de Linguagem Natural (PLN) na construção de uma ferramenta electrónica que habilite os investigadores nacionais com um instrumento que funcione ora como um meio de apoio à investigação nas diferentes áreas do questionamento científico (p. ex. Psicologia Cognitiva, Neurociências, Linguística, PLN), ora como um meio para um conhecimento mais aprofundado das características da própria língua e para o apoio ao desenvolvimento de aplicações capazes de processar a linguagem natural.

Pela inovação que constitui entre nós, pela diversidade de índices que aglutina (índices de frequência lexical, índices morfológicos e morfo-sintácticos, índices ortográficos, índices fonológicos, índices fonográficos, índices silábicos ortográficos e fonológicos, índices semânticos, índices subjectivos e índices para pseudo-palavras) e pela dupla funcionalidade de análises que oferece ao utilizador (avaliar palavras em determinados parâmetros e obter palavras que obedeçam a tais parametros), consideramos estar perante uma ferramenta com um potencial inestimável à promoção e internacionalização da investigação em Portugal.

Agradecimentos

Agradecemos à FCT (Fundação para a Ciência e a Tecnologia), ao QREN (Quadro de Referência Estratégica Nacional) e ao programa COMPETE (Programa Operacional Factores de Competitividade), integrado no Fundo Europeu de Desenvolvimento Regional (FEDER), o financiamento deste projecto (PTDC/PSI-PCO/104679/2008).

Referências

Baayen, Harald R., Richard Piepenbrock, e Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Baayen, Harald R., Richard Piepenbrock, e H. van Rijn. 1993. *The CELEX Lexical Database. Release 1 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith I. Hutchison, Brett Kessler,

Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, e Rebecca Treiman1. 2007. The english lexicon project. *Behavior Research Methods*, 39:445–459. http://artsci.wustl.edu/~rtreiman/Selected_Papers/English_Lexicon_Project_userguide_in%20press.pdf.

Cameirão, Manuela L e Selene G. Vicente. 2010. Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior Research Methods*, 42(2):474–480.

Coltheart, Max. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

Content, Alain, Phillppe Mousty, e Monique Radeau. 1990. Brulex: une base de données lexicales informatisée pour le français écrit et parlé. *L'année psychologique*, 90:551–566. http://www.lexique.org/public/Brulex.pdf.

Costa, Luís, Diana Santos, e Nuno Cardoso. 2008. Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos, 11 de Setembro, 2008. http://www.linguateca.pt/LivroL10/Livro-Costaetal2008.pdf.

Davis, Colin J. 2005. N-Watch: a program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37(1):65–70. http://www.pc.rhul.ac.uk/staff/c.davis/Articles/Davis_05.pdf.

Davis, Colin J. e Manuel Perea. 2005. BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in spanish. *Behavior Research Methods*, 37(4):665–671. http://brm.psychonomic-journals.org/content/37/4/665.full.pdf.

Ferrand, Ludovic, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova, e Chistophe Pallier. 2010. The French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496. http://www.mariaaugustinova.com/site/publications_files/FERRAND-BRM-Final-2010.pdf.

Garcia-Marques, Teresa. 2003. Avaliação da familiaridade e valência de palavras concretas e abstractas em língua portuguesa. *Laboratório de Psicologia*, 1(1):21–44. http://repositorio.ispa.pt/bitstream/10400.12/124/1/LP%20%281%291%20-%2021-44.pdf.

Gomes, Inês e São Luís Castro. 2003. Porlex: A lexical database in European Portuguese. *Psychologica*, 32:31–108. http://www.fpce.up.pt/labfala/porlex_gomes&castro03.pdf.

Ktori, Maria, Walter J. B. van Heuven, e Nicola J. PitcHford. 2008. GreekLex: A lexical database of modern Greek. *Behavior Research Methods*, 40(3):773–783. http://brm.psychonomic-journals.org/content/40/3/773.full.pdf+html.

Marques, J. Frederico. 2004. Normas de familiaridade para substantivos comuns. *Laboratório de Psicologia*, 2:5–19.

Marques, J. Frederico. 2005. Normas de imagética e concreteza para substantivos comuns. *Laboratório de Psicologia*, 3:65–75.

Marques, J. Frederico, Francisca L. Fonseca, A. Sofia Morais, e Inês A. Pinto. 2007. Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables. *Behavior Research Methods*, 39(3):439–444. http://brm.psychonomic-journals.org/content/39/3/439.full.pdf.

Nascimento, Maria Fernanda Bacelar, M. Lúcia Garcia Marques, e M. Luísa Segura da Cruz. 1987. *Português Fundamental: Métodos e documentos (Vol. II, Tomo I: Inquérito de frequência)*. INIC, Centro de Linguística da Universidade de Lisboa, Lisboa.

Nascimento, Maria Fernanda Bacelar, Luísa Pereira, e João Saramago. 2000. Portuguese corpora at CLUL. Em *Second International Conference on Language Resources and Evaluation*, volume II, pp. 1603–1607, Athens.

Nascimento, Maria Fernanda Bacelar, Maria Celeste Rodrigues, e José Bettencourt Gonçalves, editores. 1996. *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, volume I: Corpora, Lisboa. Colibri.

New, Boris, Christophe Pallier, Marc Brysbaert, Ludovic Ferr, Royal Holloway, U Service, e Hospitalier Frédéric Joliot. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36:516–524.

New, Boris, Christphe Pallier, Ludovic Ferrand, e Rafael Matos. 2001. Une base de

données lexicales du Français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101:447–462. `http://www.pallier.org/papers/Lexique.2001.pdf`.

Perea, Manuel, Miriam Urkia, Colin J. Davis, A. Agirre, E. Laseka, e M. Carreiras. 2006. E-Hitz: A word-frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, 38:610–615. `http://www.uv.es/~mperea/ehitz.pdf`.

Sebastián-Gallés, Núria, Maria Antònia Martí Antonín, Manuel Francisco Carreira Valinã, e Fernando Cuetos Vega. 2000. *LEXESP: Léxico informatizado del español*. Edicions de la Universitat de Barcelona, Barcelona.

Chamada de Artigos

A revista Linguamática pretende colmatar uma lacuna na comunidade de processamento de linguagem natural para as línguas ibéricas. Deste modo, serão publicados artigos que visem o processamento de alguma destas línguas.

A Linguamática é uma revista completamente aberta. Os artigos serão publicados de forma electrónica e disponibilizados abertamente para toda a comunidade científica sob licença *Creative Commons*.

Tópicos de interesse:

- Morfologia, sintaxe e semântica computacional
- Tradução automática e ferramentas de auxílio à tradução
- Terminologia e lexicografia computacional
- Síntese e reconhecimento de fala
- Recolha de informação
- Resposta automática a perguntas
- Linguística com corpora
- Bibliotecas digitais
- Avaliação de sistemas de processamento de linguagem natural
- Ferramentas e recursos públicos ou partilháveis
- Serviços linguísticos na rede
- Ontologias e representação do conhecimento
- Métodos estatísticos aplicados à língua
- Ferramentas de apoio ao ensino das línguas

Os artigos devem ser enviados em PDF através do sistema electrónico da revista. Embora o número de páginas dos artigos seja flexível sugere-se que não excedam 20 páginas. Os artigos devem ser devidamente identificados. Do mesmo modo, os comentários dos membros do comité científico serão devidamente assinados.

Em relação à língua usada para a escrita do artigo, sugere-se o uso de português, galego, castelhano, basco ou catalão.

Os artigos devem seguir o formato gráfico da revista. Existem modelos LaTeX, Microsoft Word e OpenOffice.org na página da Linguamática.

Datas Importantes

- Envio de artigos até: 15 de Abril de 2011
- Resultados da selecção até: 15 de Maio de 2011
- Versão final até: 31 de Maio de 2011
- Publicação da revista: Junho de 2011

Qualquer questão deve ser endereçada a: `editores@linguamatica.com`

Petición de Artigos

A revista Linguamática pretende cubrir unha lagoa na comunidade de procesamento de linguaxe natural para as linguas ibéricas. Deste xeito, han ser publicados artigos que traten o procesamento de calquera destas linguas.

Linguamática é unha revista completamente aberta. Os artigos publicaranse de forma electrónica e estarán ao libre dispor de toda a comunidade científica con licenza *Creative Commons*.

Temas de interese:

- Morfoloxía, sintaxe e semántica computacional
- Tradución automática e ferramentas de axuda á tradución
- Terminoloxía e lexicografía computacional
- Síntese e recoñecemento de fala
- Extracción de información
- Resposta automática a preguntas
- Lingüística de corpus
- Bibliotecas dixitais
- Avaliación de sistemas de procesamento de linguaxe natural
- Ferramentas e recursos públicos ou cooperativos
- Servizos lingüísticos na rede
- Ontoloxías e representación do coñecemento
- Métodos estatísticos aplicados á lingua
- Ferramentas de apoio ao ensino das linguas

Os artigos deben de enviarse en PDF mediante o sistema electrónico da revista. Aínda que o número de páxinas dos artigos sexa flexíbel suxírese que non excedan as 20 páxinas. Os artigos teñen que identificarse debidamente. Do mesmo modo, os comentarios dos membros do comité científico serán debidamente asinados.

En relación á lingua usada para a escrita do artigo, suxírese o uso de portugués, galego, castelán, éuscaro ou catalán.

Os artigos teñen que seguir o formato gráfico da revista. Existen modelos LaTeX, Microsoft Word e OpenOffice.org na páxina de Linguamática.

Datas Importantes

- Envío de artigos até: 15 de abril de 2011
- Resultados da selección: 15 de maio de 2011
- Versión final: 31 de maio de 2011
- Publicación da revista: 15 de xuño de 2011

Para calquera cuestión, pode dirixirse a: `editores@linguamatica.com`

Petición de Artículos

La revista Linguamática pretende cubrir una laguna en la comunidad de procesamiento del lenguaje natural para las lenguas ibéricas. Con este fin, se publicarán artículos que traten el procesamiento de cualquiera de estas lenguas.

Linguamática es una revista completamente abierta. Los artículos se publicarán de forma electrónica y se pondrán a libre disposición de toda la comunidad científica con licencia *Creative Commons*.

Temas de interés:

- Morfología, sintaxis y semántica computacional
- Traducción automática y herramientas de ayuda a la traducción
- Terminología y lexicografía computacional
- Síntesis y reconocimiento del habla
- Extracción de información
- Respuesta automática a preguntas
- Lingüística de corpus
- Bibliotecas digitales
- Evaluación de sistemas de procesamiento del linguage natural
- Herramientas y recursos públicos o cooperativos
- Servicios lingüísticos en la red
- Ontologías y representación del conocimiento
- Métodos estadísticos aplicados a la lengua
- Herramientas de apoyo para la enseñanza de lenguas

Los artículos tienen que enviarse en PDF mediante el sistema electrónico de la revista. Aunque el número de páginas de los artículos sea flexible, se sugiere que no excedan las 20 páginas. Los artículos tienen que identificarse debidamente. Del mismo modo, los comentarios de los miembros del comité científico serán debidamente firmados.

En relación a la lengua usada para la escritura del artículo, se sugiere el uso del portugués, gallego, castellano, vasco o catalán.

Los artículos tienen que seguir el formato gráfico de la revista. Existen modelos LaTeX, Microsoft Word y OpenOffice.org en la página de Linguamática.

Fechas Importantes

- Envío de artículos hasta: 15 de abril de 2011
- Resultados de la selección: 15 de mayo de 2011
- Versión final: 31 de mayo de 2011
- Publicación de la revista: junio de 2011

Para cualquier cuestión, puede dirigirse a: `editores@linguamatica.com`

Petició d'articles

La revista Linguamática pretén cobrir una llacuna en la comunitat del processament de llenguatge natural per a les llengües ibèriques. Així, es publicaran articles que tractin el processament de qualsevol d'aquestes llengües.

Linguamática és una revista completament oberta. Els articles es publicaran de forma electrònica i es distribuiran lliurement per a tota la comunitat científica amb llicència *Creative Commons*.

Temes d'interès:

- Morfologia, sintaxi i semàntica computacional
- Traducció automàtica i eines d'ajuda a la traducció
- Terminologia i lexicografia computacional
- Síntesi i reconeixement de parla
- Extracció d'informació
- Resposta automàtica a preguntes
- Lingüística de corpus
- Biblioteques digitals
- Evaluació de sistemes de processament del llenguatge natural
- Eines i recursos lingüístics públics o cooperatius
- Serveis lingüístics en xarxa
- Ontologies i representació del coneixement
- Mètodes estadístics aplicats a la llengua
- Eines d'ajut per a l'ensenyament de llengües

Els articles s'han d'enviar en PDF mitjançant el sistema electrònic de la revista. Tot i que el nombre de pàgines dels articles sigui flexible es suggereix que no ultrapassin les 20 pàgines. Els articles s'han d'identificar degudament. Igualement, els comentaris dels membres del comitè científic seràn degudament signats.

En relació a la llengua usada per l'escriptura de l'article, es suggereix l'ús del portuguès, gallec, castellà, basc o català.

Els articles han de seguir el format gràfic de la revista. Es poden trobar models LaTeX, Microsoft Word i OpenOffice.org a la pàgina de Linguamática.

Dades Importants

- Enviament d'articles fins a: 15 d'abril de 2011
- Resultats de la selecció: 15 de maig de 2011
- Versió final: 31 de maig de 2011
- Publicació de la revista: juny de 2011

Per a qualsevol qüestió, pot adreçar-se a: `editores@linguamatica.com`

Artilulu eskaera

Iberiar penintsulako hizkuntzei dagokienean, hizkuntza naturalen prozedura komunitatean dagoen hutsunea betetzea litzateke Linguamática izeneko aldizkariaren helburu nagusiena. Helburu nagusi hau buru, aurretik aipaturiko edozein hizkuntzen prozedura landuko duten artikuluak argitaratuko dira.

Linguamática aldizkaria irekia da oso. Artikuluak elektronikoki argitaratuko dira, eta komunitate zientefikoaren eskura egongo dira honako lizentziarekin; *Creative Commons*.

Gai interesgarriak:

- Morfologia, sintaxia eta semantika konputazionala.
- Itzulpen automatikoa eta itzulpengintzarako lagungarriak diren tresnak.
- Terminologia eta lexikologia konputazionala.
- Mintzamenaren sintesia eta ikuskapena.
- Informazio ateratzea.
- Galderen erantzun automatikoa.
- Corpus-aren linguistika.
- Liburutegi digitalak.
- Hizkuntza naturalaren prozedura sistemaren ebaluaketa.
- Tresna eta baliabide publikoak edo kooperatiboak.
- Zerbitzu linguistikoak sarean.
- Ezagutzaren ontologia eta adierazpideak.
- Hizkuntzean oinarrituriko metodo estatistikoak.
- Hizkuntzen irakaskuntzarako laguntza tresnak.

Arikuluak PDF formatoan eta aldizkariaren sitema elektronikoaren bidez bidali behar dira. Orri kopurua malgua den arren, 20 orri baino gehiago ez idaztea komeni da. Artikuluak behar bezala identifikatu behar dira. Era berean, zientzi batzordeko kideen iruzkinak ere sinaturik egon beharko dira.

Artikulua idazterako garaian, erabilitako hizkuntzari dagokionean, honako kizkuntza hauek erabili daiztezke; portugesa, galiziera, gaztelania, euskara, eta katalana.

Artikuluek, aldizkariaren formato grafikoa jarraitu behar dute. "Linguamática" orrian LaTeX, Microsoft Word eta OpenOffice.org ereduak aurki ditzakegu.

Data garratzitsuak:

- Arikuluak bidali ahal izateko epea: 2011eko apirilak 15.
- Hautapenaren emaitzak: 2011eko maiatzak 15.
- Azken itzulpena: 2011eko maiatzak 31.
- Aldizkariaren argitarapena: 2011eko ekainean.

Edozein zalantza argitzeko, hona hemen helbide hau: `editores@linguamatica.com`.

http://www.linguamatica.com/