Volume 3, Número 2- Dezembro 2011

# linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 3, Número 2 – Dezembro 2011

# LinguaMÁTICA

ISSN: 1647–0818

# Conteúdo

# Editorial

*Com esta edição completamos um ciclo de três anos de Linguamática, contabilizando seis edições regulares e uma edição especial. O número de artigos aceites e a quantidade de artigos enviados para apreciação mostram o crescente interesse nesta nossa/vossa revista que muito nos orgulha.*

*Preparamos neste momento o quarto ano de vida da Linguamática que, tal como no ano passado, irá contar com três edições. Esperamos mesmo que, antes do próximo número regular tenhamos uma edição especial disponível.*

*No sentido de continuar a melhorar a revista o próximo ano irá contar com uma nova rúbrica, talvez menos científica e mais técnica, em que convidamos os autores a escrever artigos mais pequenos apresentando ferramentas, recursos e/ou aplicações. Estamos especialmente interessados em artigos que apresentem algum destes tipos de recursos que sejam de acesso aberto e já disponíveis para uso generalizado.*

*Esta nova secção, juntamente com as que já são regulares (dossier, artigos de investigação, novas perspectivas e apresentação de projectos), tornam a Linguamática um meio de comunicação privilegiado entre a comunidade de processamento das línguas ibéricas.*

*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
**Antonio Moreno Sandoval**, Universidad Autónoma de Madrid
**António Teixeira**, Universidade de Aveiro
**Arantza Díaz de Ilarraza**, Euskal Herriko Unibertsitatea
**Belinda Maia**, Universidade do Porto
**Carmen García Mateo**, Universidade de Vigo
**Diana Santos**, Linguateca/FCCN
**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
**Juan-Manuel Torres-Moreno**, Laboratoire Informatique d'Avignon - UAPV
**Kepa Sarasola**, Euskal Herriko Unibertsitatea
**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
**Patrícia Cunha França**, Universidade do Minho
**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

# Dossier

# Analizadores Multilingües en FreeLing

Lluís Padró
Dept. Lenguajes y Sistemas Informáticos
Centro de Investigación TALP
Universitat Politècnica de Catalunya
padro@lsi.upc.edu

**Resumen**

FreeLing es una librería de código abierto para el procesamiento multilíngüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas. FreeLing ofrece a los desarrolladores de aplicaciones de Procesamiento del Lenguaje Natural funciones de análisis y anotación lingüística de textos, con la consiguiente reducción del coste de construcción de dichas aplicaciones. FreeLing es personalizable y ampliable, y está fuertemente orientado a aplicaciones del mundo real en términos de velocidad y robustez. Los desarrolladores pueden utilizar los recursos lingüísticos por defecto (diccionarios, lexicones, gramáticas, etc), ampliarlos, adaptarlos a dominios particulares, o –dado que la librería es de código abierto– desarrollar otros nuevos para idiomas específicos o necesidades especiales de las aplicaciones. Este artículo presenta los principales cambios y mejoras incluidos en la versión 3.0 de FreeLing, y resume algunos proyectos industriales relevantes en los que se ha utilizado.

## 1. Introducción

FreeLing[1] es una librería de código abierto para el procesamiento multilingüe, que proporciona una amplia gama de funcionalidades de análisis para varios idiomas.

El proyecto FreeLing se inició desde el centro TALP[2] de la UPC para avanzar hacia la disponibilidad general de recursos y herramientas básicos de Procesamiento del Lenguaje Natural (PLN). Esta disponibilidad debería posibilitar avances más rápidos en proyectos de investigación y costes más reducidos en el desarrollo de aplicaciones industriales de PLN.

El proyecto se estructura como una librería que puede ser llamada desde cualquier aplicación de usuario que requiera servicios de análisis del lenguaje. El software se distribuye como código abierto bajo una licencia *GNU General Public License*[3] y bajo licencia dual a empresas que deseen incluirlo en sus productos comerciales.

El planteamiento como un proyecto de código abierto ha sido muy fructífero durante los ocho años de vida de FreeLing (la primera versión fue lanzada en 2003). La versión 2.2 ha sido descargada más de 64.000 veces desde su lanzamiento en septiembre de 2010 por una amplia comunidad de usuarios, la cual ha ampliado el número inicial de tres idiomas (inglés, español y catalán) a nueve, además de la inclusión de la variante diacrónica del español de los siglos XII al XVI (Sánchez-Marco, Boleda, y Padró, 2011). La naturaleza de código abierto del proyecto ha hecho también posible –junto con su arquitectura modular– incorporar el código de otros proyectos similares, como el módulo de desambiguación del sentido de las palabras basado en UKB (Agirre y Soroa, 2009).

La versión actual soporta (a diferentes niveles de completitud) las siguientes lenguas: asturiano, catalán, castellano, galés, gallego, inglés, italiano, portugués, y ruso. Las funcionalidades existentes para cada idioma se resumen en la tabla 1.

La sección 2 describe los principales módulos y servicios de FreeLing. A continuación se describen las principales novedades de la versión 3.0, y la sección 4 resume algunos de los proyectos industriales en los que se ha utilizado la librería. Por último, se esbozan algunas conclusiones y líneas de trabajo futuro.

## 2. Estructuras de datos y servicios de análisis lingüístico

FreeLing está concebido como una librería sobre la cual se puedan desarrollar potentes aplicaciones de PLN, y orientado a facilitar la integración con las aplicaciones de niveles superiores de los servicios lingüísticos que ofrece.

[1] http://nlp.lsi.upc.edu/freeling
[2] http://www.talp.cat
[3] http://www.gnu.org/copyleft/gpl.html

| | as | ca | cy | en | es | gl | it | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| Tokenization | X | X | X | X | X | X | X | X | X |
| Sentence splitting | X | X | X | X | X | X | X | X | X |
| Number detection | | X | | X | X | X | X | X | X |
| Date detection | | X | | X | X | X | | X | X |
| Morphological dictionary | X | X | X | X | X | X | X | X | X |
| Affix rules | X | X | X | X | X | X | X | X | |
| Multiword detection | X | X | X | X | X | X | X | X | |
| Basic named entity detection | X | X | X | X | X | X | X | X | X |
| B-I-O named entity detection | | | | X | X | X | | | |
| Named Entity Classification | | | | X | X | | | | |
| Quantity detection | | X | | X | X | X | | X | X |
| PoS tagging | X | X | X | X | X | X | X | X | X |
| WN sense annotation | | X | | X | X | | | | |
| UKB sense disambiguation | | X | | X | X | | | | |
| Shallow parsing | X | X | | X | X | X | | X | |
| Full/dependency parsing | X | X | | X | X | X | | | |
| Coreference resolution | | | | | X | | | | |

Cuadro 1: Servicios de análisis disponibles para cada lengua.

La arquitectura de la librería se basa en un enfoque de dos capas cliente-servidor: una capa básica de servicios de análisis lingüístico (morfológico, morfosintáctico, sintáctico, ...) y una capa de aplicación que, actuando como cliente, realiza las peticiones deseadas a los analizadores y usa su respuesta según la finalidad de la aplicación.

La arquitectura interna de la librería se estructura en dos tipos de objetos: los que almacenan datos lingüísticos con los análisis obtenidos y los que realizan el procesamiento en sí.

## 2.1. Clases de almacenamiento de datos lingüísticos

Las clases básicas de la librería tienen la finalidad de contener los datos lingüísticos (palabras, etiquetas morfológicas, frases, árboles sintácticos, párrafos, ...) resultado de los análisis realizados. Cualquier aplicación cliente debe usar estas clases para poder proporcionar a los módulos de análisis los datos en el formato oportuno, y para poder recuperar el resultado de los analizadores.

Las clases de datos lingüísticos en la versión actual son las siguientes:

- `analysis`: Una tupla <lema, etiqueta, probabilidad, lista de sentidos>.
- `word`: Forma de una palabra, con una lista de posibles objetos `analysis`.
- `sentence`: Una lista de objetos `word` marcada como una frase completa. Puede contener también un árbol de constituyentes o de dependencias.
- `paragraph`: Una lista de objetos `sentence` marcada como un párrafo independiente.
- `document`: Una lista de objetos `paragraph` que forman un documento completo. Puede contener también información sobre la coreferencia entre las menciones a entidades del documento.

La figura 1 presenta un diagrama UML con las clases de datos lingüísticos.

## 2.2. Clases de procesamiento

Aparte de las clases para contener datos lingüísticos descritas anteriormente, la librería proporciona también clases para transformarlos, usualmente enriqueciendolos con información adicional. La figura 2 muestra un diagrama UML con las clases de procesamiento que se describen a continuación:

- `lang_ident`: Identificador de idioma. Recibe texto plano y devuelve una lista de pares <idioma,probabilidad>.
- `tokenizer`: Recibe texto plano y devuelve una lista de objetos `word`.
- `splitter`: Recibe una lista de objetos `word` y devuelve una lista de objetos `sentence`.
- `morfo`: Recibe una lista de objetos `sentence` y analiza morfológicamente cada `word` de cada `sentence` de la lista. Esta clase es un meta-analizador que simplemente aplica una cascada de analizadores especializados (detección de números, fechas, locuciones y multipalabras, búsqueda en formario, etc.) cada uno de los

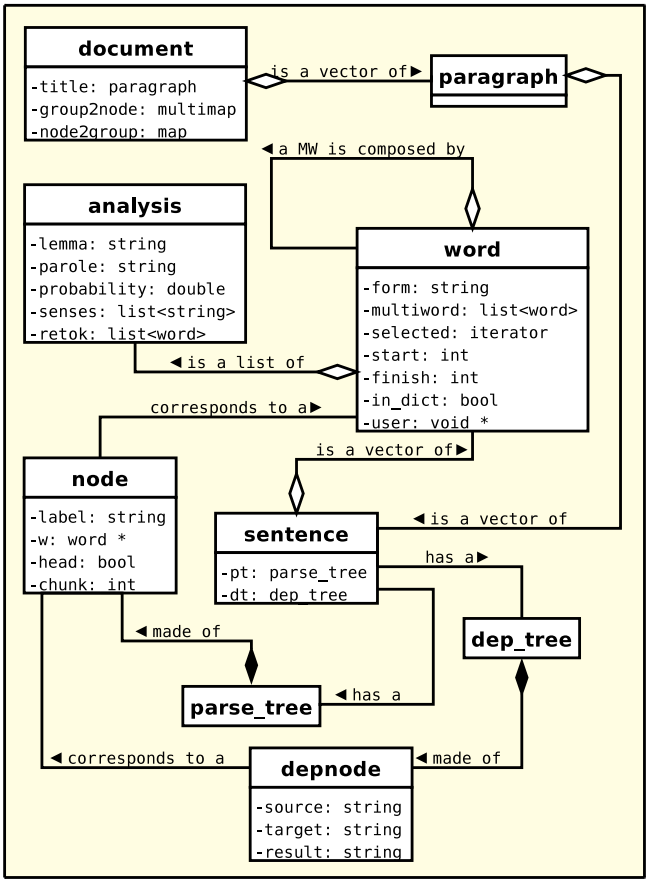



Figura 1: Clases de datos lingüísticos en FreeLing-3.0.

cuales es a su vez una clase de procesamiento que puede ser llamada independientemente si es necesario. Estas clases son:

- `user_map`: Reconocedor de expressiones regulares definidas por el usuario, que permite la asignación directa de pares lema/categoría a palabras que cumplan ciertos patrones.
- `locutions`: Reconocedor de multipalabras.
- `dictionary`: Búsqueda en formario y gestión de afijos.
- `numbers`: Reconocedor de expresiones numéricas.
- `dates`: Reconocedor de expresiones temporales (fechas/horas).
- `quantities`: Reconocedor de expresiones de proporciones, porcentajes, magnitudes físicas y monetarias.
- `punts`: Anotador de signos de puntuación.
- `probabilities`: Anotador de probabilidades léxicas y gestión de palabras desconocidas.
- `ner`: Reconocedor de nombres propios. Se proporcionan dos módulos para esta tarea: Un analizador rápido y simple basado en patrones de mayúsculas (con una precisión alrededor del 90 %), y un reconocedor basado en el sistema ganador de la *shared task* del CoNLL-2002 (Carreras, Màrquez, y Padró, 2002), sensiblemente más lento, pero con una precisión superior al 94 %.

- `tagger`: Recibe una lista de objetos `sentence` y desambigua la categoría morfosintáctica de cada palabra en las frases de la lista. Si el análisis seleccionado incorpora información de retokenización (p.e. *del* → *de+el*, *dárselo* → *dar+se+lo*) la palabra puede separarse en varias. FreeLing ofrece dos *taggers* con una precisión del estado del arte (97 %-98 %): Uno basado en modelos ocultos de markov, según se describe en (Brants, 2000) y otro basado en *relaxation labelling* (Padró, 1998) que permite la combinación de información estadística con reglas manuales.

- NE `classifier`: Recibe una lista de objetos `sentence` y clasifica cada `word` etiquetada como nombre propio que aparezca en las frases dadas. Este módulo está basado en el sistema ganador de la *shared task* del CoNLL-2002 (Carreras, Màrquez, y Padró, 2002).

- `sense annotator`: Recibe una lista de `sentence` y añade information sobre los sentidos posibles (segun WordNet) a los objetos `analysis` de cada palabra.

- `word sense disambiguator`: Recibe una lista de objetos `sentence` y ordena por relevancia en el contexto los posibles sentidos de cada palabra. El código de este módulo se incluye directamente del del proyecto del desambiguador UKB (Agirre y Soroa, 2009).

- `chunk parser`: Recibe una lista de `sentence` y enriquece cada una con un árbol de análisis. Este módulo consiste en un *chart parser*, y es una reimplementación y extensión de (Atserias y Rodríguez, 1998).

- `dependency parser`: Recibe una lista de `sentence` analizadas sintácticamente y las enriquece con un árbol de dependencias. Este módulo usa un conjunto de reglas escritas manualmente que operan en tres etapas: primero completan el árbol sintáctico superficial construido por el *chart parser*, a continuación transforman el árbol de constituyentes a dependencias, y finalmente etiquetan la función de cada dependencia. Este módulo es una extensión del que se describe en (Atserias, Comelles, y Mayor, 2005).

- `coreference solver`: Recibe un documento formado por objetos `sentence` analizados sintácticamente y lo enriquece con información de coreferencia. Este módulo se basa en el sistema propuesto por (Soon, Ng, y Lim, 2001).

## *3. Novedades en FreeLing 3.0*

La versión 3.0 presenta algunos cambios importantes que tienen como objetivo hacer que la herramienta más flexible, usable, y fácil de instalar. Estos cambios pueden agruparse en tres grandes clases: los cambios relacionados con la ampliación del soporte al multilingüismo, los cambios en los componentes de la librería basados en aprendizaje automático, y cambios relacionados con aspectos de ingeniería.

### 3.1. Ampliación del soporte al multilingüismo

La primera contribución relevante que amplia la cobertura de FreeLing con respecto a la cantidad y variedad de idiomas que puede procesar es el desarrollo de datos lingüísticos para el analizador y el desambiguador morfológico del español de los siglos XII al XVI (Sánchez-Marco, Boleda, y Padró, 2011). Este trabajo utiliza los datos por defecto para el español moderno, más las pertinentes adaptaciones y extensiones para procesar las variaciones ortográficas propias del español antiguo. Además, se ha desarrollado un corpus de entrenamiento del etiquetador, y se ha usado la herramienta resultante en un estudio lingüístico sobre la evolución del uso del verbo *haber* (Sánchez-Marco y Evert, 2011; Sánchez-Marco, 2012).

Otra modificación en la versión 3.0 de FreeLing es el soporte a la codificación de caracteres en Unicode (UTF-8). Este es un cambio importante, y una de las principales razones para el cambio de número de versión.

Las versiones anteriores de FreeLing soportaban varios idiomas, pero el desarrollador de los datos lingüísticos de cada lengua debía decidir que codificación usar, y cuidar de la consistencia de la codificación entre los datos de diversos módulos (diccionario, gramáticas, lexicones semánticos, etc.) o sus reglas o ficheros de configuración (p.e. la escritura de expresiones regulares para el tokenizador). Adicionalmente, dado que cada idioma podía utilizar una codificación diferente, no era fácil integrar un identificador de idioma capaz de manejar textos en diferentes alfabetos.

Con el soporte de las codificaciones UTF-8, la misma aplicación puede manipular textos en diferentes idiomas y alfabetos. Esta extensión ha permitido diversas mejoras funcionales:

- Un nuevo módulo para la identificación de lenguaje basado en (Padró y Padró, 2004) se ha integrado en la librería.
- Posibilidad de cambiar la configuración regional (*locale*) de la aplicación para que coincida con la del idioma del texto procesado, incluso si es distinta de la predeterminada en el sistema local.
- Las expresiones regulares utilizadas, ya sea en archivos de configuración o cableadas en el código son más expresivas y fáciles, ya que se permiten extensiones POSIX. Por ejemplo, la expresión regular usada por el tokenizador para reconocer una palabra formada por caracteres alfabéticos en español solía ser `[A-Za-záéíïóúüñÁÉÍÏÓÚÜÑ]+`, mientras que en la nueva versión se puede escribir como `[[:alpha:]]+`, que no sólo es más simple y simplifica su mantenimiento, sino que es independiente del idioma, ya que esta misma expresión puede utilizarse para reconocer palabras de caracteres alfabeticos en cualquier idioma y/o alfabeto simplemente cambiando la localización activa de la aplicación.

El uso de la codificación UTF8 deja via libre a los desarrolladores interesados en la adición de soporte para idiomas con alfabetos no latinos. Es el caso de los desarrolladores del ruso, que han conseguido un analizador morfológico muy completo y un etiquetador competitivo que no habría sido posible en las versiones anteriores. Además, no sólo se han desarrollado los datos lingüísticos para el ruso, sino también código fuente para los módulos dependientes del idioma, como los reconocedores de números o fechas.

### 3.2. Mejora de los módulos basados en aprendizaje automático

Otro cambio importante en la arquitectura FreeLing 3.0 es la organización y contenido de los módulos de aprendizaje automático: el motor de extracción de características por un lado, y los algoritmos de aprendizaje/clasificación por el otro.

En versiones anteriores, las funciones de aprendizaje automático eran proporcionadas por dos librerías externas a FreeLing: *Omlet&Fries*[4]. En la versión 3.0, el código de estas librerías está incluido en el paquete FreeLing, lo que ofrece una organización de código más clara y una reducción en el número de dependencias que simplifica el proceso de instalación.

[4] `http://nlp.lsi.upc.edu/omlet+fries`

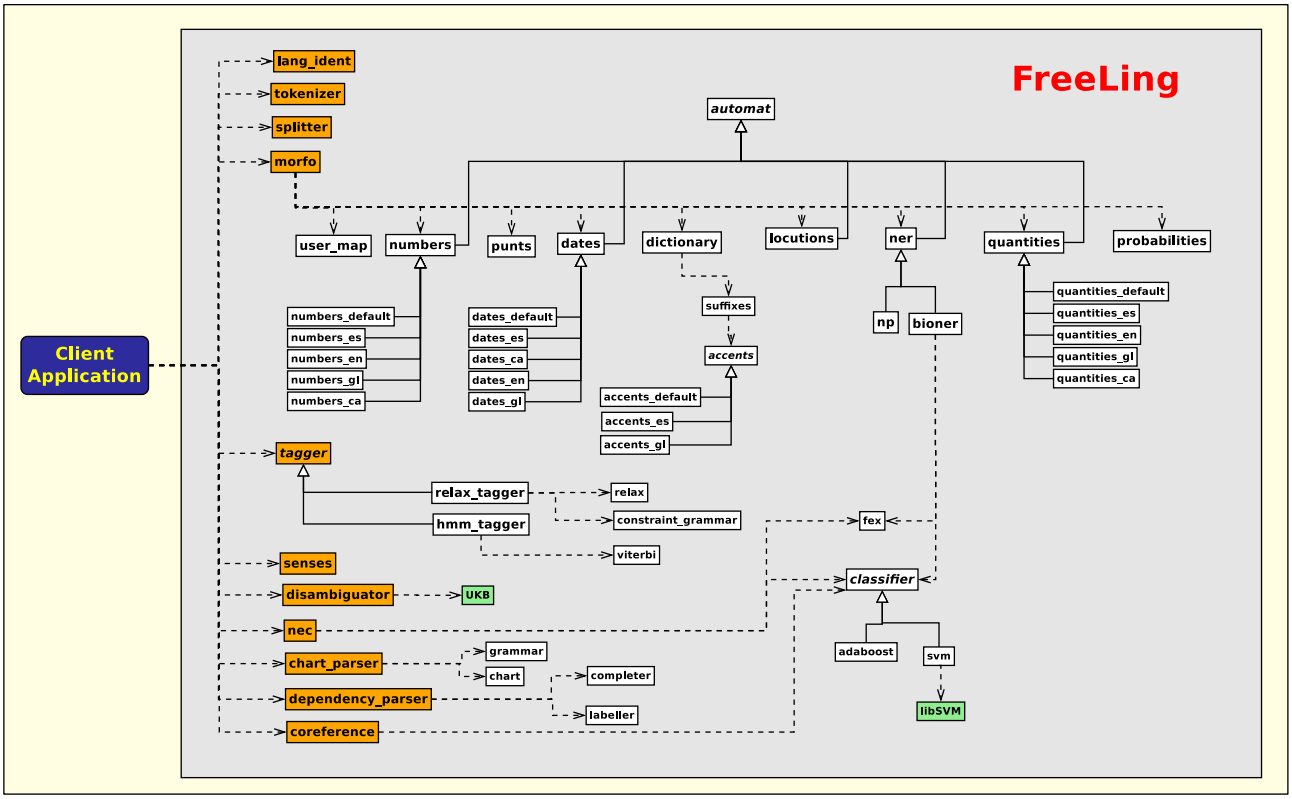



Figura 2: Clases de procesamiento en FreeLing-3.0.

El módulo de extracción de características ha sido completamente reescrito, proporcionando un formalismo de reglas de extracción más claro, y una API más flexible para aquellos desarrolladores que deseen incluir sus propias funciones de extracción de rasgos. Además, el repertorio de algoritmos de aprendizaje disponibles se ha ampliado con *Support Vector Machines* (SVM) gracias al proyecto de código abierto libSVM (Chang y Lin, 2011). El código de libSVM se ha integrado en FreeLing bajo un *wrapper* común con los clasificadores existentes. Esta estrecha integración tiene la ventaja adicional de evitar añadir un nuevo elemento a la lista de dependencias necesarias para construir e instalar FreeLing.

Por último, los modelos para reconocimiento y clasificación de entidades nombradas han sido entrenados con la nueva arquitectura. Módulos basados en aprendizaje automático para el reconocimiento y clasificación de NE se encuentran disponibles para el español, inglés, gallego y portugués. Los dos primeros ofrecen modelos AdaBoost y SVM, mientras que sólo modelos AdaBoost están disponibles para los últimos.

## 3.3. Modificaciones técnicas

El tercer tipo de cambios en FreeLing son cuestiones técnicas relacionadas con la organización de las dependencias externas, la migración a plataformas distintas de Linux, y el uso de FreeLing en modo servidor.

### 3.3.1. Dependencias externas

Un aspecto importante de estas modificaciones de ingeniería es la gestión de las dependencias de librerías externas requeridas por FreeLing. Esta es una razón principal –junto con el cambio a codificación Unicode descrito anteriormente– para la actualización del número de versión.

Se ha realizado un esfuerzo importante para reducir y simplificar la lista de dependencias, a fin de facilitar la construcción y instalación de la librería, así como simplificar su uso en productos comerciales bajo licencia dual.

Las dependencias de las versiones anteriores eran:

- BerkeleyDB - Acceso rápido a los archivos de diccionario en disco.
- PCRE - Gestión de expresiones regulares.
- libcfg+ - Gestión opciones de configuración para el programa principal `analyzer`.
- Omlet&Fries - Módulos de Aprendizaje Automático (véase más arriba).

FreeLing 3.0 ya no requiere BerkeleyDB: Los diccionarios se cargan en RAM ya sea en *prefix-trees* o en estructuras *map* de la STL. El rendimiento temporal es aproximadamente el

mismo, y el aumento en el consumo de memoria no supone ningún problema para una máquina moderna. Por otra parte, la mayor simplicidad en la instalación (menos dependencias, no necesidad de indexar los diccionarios durante la instalación), y en la gestión de los diccionarios (no es necesario reindexar después de modificar un diccionario) compensan ampliamente el ligero aumento en el tiempo de inicialización que este cambio supone.

La nueva versión FreeLing tampoco usa ya PCRE ni `libcfg+`: Tanto las funcionalidades relacionadas con las expresiones regulares como las relativas a la gestión de opciones de configuración se han transferido a las librerías *boost*[5]. Esto supone dos ventajas: las dependencias se unifican bajo un solo proveedor, y la instalación es mucho más simple ya que `liboost` es parte de todas las distribuciones Linux.

#### 3.3.2. Compilación nativa en MS-Windows

Utilizar FreeLing bajo MS-Windows solía ser una empresa difícil. Era necesario usar emuladores o compiladores cruzados como MinGW[6] o Cygwin[7], y los resultados obtenidos no siempre eran fácilmente integrables en una aplicación de MS-Windows.

Todo el código C++ en la versión 3.0 se ha adaptado para ser compilado por MS-Visual C++, y se proporcionan los archivos de proyecto para compilar la librería bajo dicho entorno, lo que simplifica enormemente la construcción y el uso de FreeLing en MS-Windows, ya que los binarios obtenidos son nativos en dicho sistema.

#### 3.3.3. Mejoras en el modo servidor

Por último, una mejora técnica de menor importancia es la capacidad multicliente que ofrece el programa de demostración `analyzer` en la nueva versión:

El modo de servidor de las versiones anteriores estaba concebida solamente como un medio para evitar la repetición de la inicialización de los módulos en el caso que se quisiera procesar un gran número de archivos pequeños. Por ello, todas las peticiones eran atendidas de forma secuencial por el mismo servidor, resultando poco adecuado o lento en el caso que se quisiera atender a muchos clientes simultáneamente.

En la nueva versión, el código sigue una arquitectura estándar de un servidor Linux: Un proceso *dispatcher* espera solicitudes de los clientes a través de un *socket*. Cuando se establece una nueva conexión con un cliente, el *dispatcher* crea un nuevo proceso *worker* que se hará cargo del cliente, mientras el *dispatcher* regresa de nuevo a esperar peticiones entrantes.

Esto hace posible el uso de FreeLing en procesamiento paralelo (por ejemplo, para procesar grandes cantidades de texto en una máquina multiprocesador, o para usarlo como un servidor en una aplicación web multiusuario) que no era posible en las versiones anteriores.

Sin embargo, el nuevo servidor no limita el número máximo de clientes conectados, ni tiene una cola de espera para las peticiones. Así pues, las aplicaciones con un cantidad potencialmente masiva de los clientes deben adaptar el código de servidor para manejar con seguridad una cola de solicitudes pendientes.

### 3.4. Otras mejoras

Otras novedades destacables incluidas en FreeLing 3.0 son las sigüientes:

- *UserMap*: Se trata de un nuevo módulo que permite al usuario definir una serie de expresiones regulares y asignar a cada una de ellas un conjunto de pares <lema,etiqueta> que se asignarán a las palabras que cumplan dicho patrón. El objetivo de este módulo es facilitar al desarrollador de aplicaciones el tratamiento específico de casos no cubiertos por otros módulos en FreeLing. Por ejemplo, una aplicación de procesamiento de *Twitter* podría requerir la anotación como nombres propios de las palabras con el patrón `@nombre`. En lugar del costoso trabajo de modificar o reentrenar el detector de nombres propios, ahora se puede simplemente añadir una regla:
  ```
  @[a-z][a-z0-9]* $$ NP00000
  ```
  que reconoce dichas palabras, asignándoles su propia forma como lema y `NP00000` como etiqueta.
- Trigramas `<Forbidden>`: El tagger basado en HMM efectúa suavizado de las probabilidades de transición no observadas, reservando cierta masa de probabilidad para casos lingüísticamente imposibles (como p.e. un determinante seguido de un verbo finito, o un *haber* auxiliar seguido de algo que no sea un participio). Estos casos se pueden explicitar en el fichero de configuración del tagger, evitando su suavizado y forzando a que tengan probabilidad cero, reduciendo así la tasa de error de la desambiguación.
- *Phonetics*: Otro servicio nuevo en FreeLing es la codificación fonética del sonido de una palabra. Este módulo usa un fichero de reglas de

[5] `http://www.boost.org`
[6] `http://www.mingw.org`
[7] `http://www.cygwin.com`

transcripción que traducen el texto a su codificación en el estándar SAMPA[8]. También es capaz de usar un diccionario fonético de transcripciones de palabras completas para las excepciones a las reglas (o para idiomas, como el inglés, con una fonética poco regular).

## 4. FreeLing en proyectos industriales

La versión de desarrollo de FreeLing 3.0 está disponible en el SVN del proyecto, y ya ha sido utilizada en varios proyectos industriales, de los cuales resumimos brevemente los más relevantes:

- Ruby Reader: Aplicación para el iPhone que ayuda a los hablantes de japonés a comprender textos en inglés. Desarrollado por CA-Mobile (`http://www.camobile.com`).
- Vi-Clone: Impresionantes asistentes virtuales para páginas web corporativas. Algunos componentes de FreeLing se están integrando en el sistema de diálogo. Vi-Clone está financiando el desarrollo del módulo de corrección ortográfica que permitirá a FreeLing procesar frases del usuario escritas en variantes no estándar. `http://www.vi-clone.com`.
- TextToSign: Traductor de texto en español al lenguaje de señas, que utiliza FreeLing para el procesamiento de texto. `http://www.textosign.es`.
- Dixio: diccionario inteligente capaz de ayudar al lector de un texto, ofreciendo definiciones contextualizadas. Desarrollado por Semantix (`http://www.semantix.com`).
- Aport News: Portal de noticias que usa FreeLing como un preprocesador para enriquecer texto en ruso. El resultado de la anotación se utiliza en la clasificación y agrupación de las noticias. `http://news.aport.ru`.

## 5. Conclusiones y trabajo futuro

Hemos presentado las principales mejoras y cambios realizados en la versión 3.0 de FreeLing, y algunos proyectos industriales en los que se ha utilizado.

Gracias a estos cambios, y a la activa comunidad en torno a este proyecto, esperamos seguir ampliando el número de idiomas soportados, ampliando las funcionalidades proporcionadas, y mejorando la usabilidad de estos analizadores en aplicaciones industriales del PLN.

[8] `http://www.phon.ucl.ac.uk/home/sampa`

Una de las líneas de trabajo que más interés despierta es la inclusión de un módulo de corrección ortográfica que –como parte de una cadena de análisis robusto– constituya una piedra angular para el desarrollo de aplicaciones orientadas a textos no estándar como chats de internet, foros, microblogs, etc.

## Agradecimientos

Este trabajo ha sido parcialmente financiado por el Gobierno Español a través de los proyectos KNOW-2 (TIN2009-14715-C04-03/04) y OpenMT-2 (TIN2009-14675-C03-01), así como por la Unión Europea a través del proyecto FAUST (FP7-ICT-2009-4). Agradecemos también a ViClone (`www.viclone.com`) por financiar parte del desarrollo de FreeLing.

## Bibliografía

Agirre, Eneko y Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. En *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Atserias, Jordi, Elisabet Comelles, y Aingeru Mayor. 2005. Txala un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, (35):455–456, September.

Atserias, Jordi y Horacio Rodríguez. 1998. Tacat: Tagged corpus analizer tool. Technical report lsi-98-2-t, Departament de LSI. Universitat Politècnica de Catalunya.

Brants, Thorsten. 2000. Tnt - a statistical part-of-speech tagger. En *Proceedings of the 6th Conference on Applied Natural Language Processing, ANLP*. ACL.

Carreras, Xavier, Lluís Màrquez, y Lluís Padró. 2002. Named entity extraction using adaboost. En *Proceedings of CoNLL Shared Task*, páginas 167–170, Taipei, Taiwan.

Chang, C.C. y C.J. Lin. 2011. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, April.

Padró, Lluís. 1998. *A Hybrid Environment for Syntax–Semantic Tagging*. Ph.D. tesis, Dep. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, February. http://www.lsi.upc.es/˜padro.

Padró, Muntsa y Lluís Padró. 2004. Comparing methods for language identification. *Proce-*

*samiento del Lenguaje Natural*, (33):155–162, September.

Soon, W.M., H. T. Ng, y D.C.Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Sánchez-Marco, Cristina. 2012. *Tracing the development of Spanish participial constructions: An empirical study of language change*. Ph.D. tesis, Universitat Pompeu Fabra, Barcelona, Spain. (forthcoming).

Sánchez-Marco, Cristina, Gemma Boleda, y Lluís Padró. 2011. Extending the tool, or how to annotate historical language varieties. En *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, páginas 1–9, Portland, OR, USA, June. Association for Computational Linguistics.

Sánchez-Marco, Cristina y Stefan Evert. 2011. Measuring semantic change: The case of spanish participial constructions. En *Proceedings of 4th Conference on Quantitative Investigations in Theoretical Linguistics (QITL-4)*, Berlin, Germany, March.

# Artigos de Investigação

# Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos

Hugo Gonçalo Oliveira
CISUC, Universidade de Coimbra
hroliv@dei.uc.pt

Leticia Antón Pérez
CISUC, Universidade de Coimbra
& Universidade de Vigo
leticiaap86@gmail.com

Hernani Costa
CISUC, Universidade de Coimbra
hpcosta@dei.uc.pt

Paulo Gomes
CISUC, Universidade de Coimbra
pgomes@dei.uc.pt

**Resumo**

Este artigo apresenta o CARTÃO, uma nova rede léxico-semântica para o português, composta por relações extraídas a partir de três dicionários electrónicos. Após analisarmos a estrutura das definições nos três, concluímos que as mesmas regras podiam ser utilizadas para extrair relações a partir de vários dicionários. Assim, aproveitamos este facto para utilizar o mesmo conjunto de gramáticas na construção desta rede. As relações que compõem o CARTÃO são apresentadas em termos de quantidade e de acordo com o dicionário de onde foram extraídas. Verifica-se que foi possível aumentar em mais de 70% o PAPEL, uma rede semelhante já existente, o que mostra os ganhos em utilizar mais do que um recurso na construção destas redes. A cobertura do CARTÃO e os resultados da validação automática de alguns tipos de relação são aqui também apresentados e discutidos.

## *1 Introdução*

Os dicionários são repositórios que reúnem palavras e expressões de uma língua, acompanhadas pelas definições dos seus possíveis sentidos, que por sua vez são descrições escolhidas e escritas por lexicógrafos, os especialistas neste campo. Apesar dos dicionários não estarem preparados para serem directamente utilizados como ferramentas de processamento de linguagem natural (PLN), não é de admirar que, desde cedo, tenham sido vistos como recursos fundamentais na construção automática (e manual) de bases de conhecimento lexical (veja-se, por exemplo Calzolari, Pecchia e Zampolli (1973), Chodorow, Byrd e Heidorn (1985) ou Richardson, Dolan e Vanderwende (1998)) – recursos computacionais onde itens lexicais (palavras e expressões) se encontram organizados de acordo com uma determinada teoria de semântica lexical (como a teoria apresentada por Cruse (1986)).

As bases de conhecimento revelaram-se essenciais na realização de várias tarefas PLN, incluindo a desambiguação do sentido das palavras (Agirre, Lacalle e Soroa, 2009), resposta automática a perguntas (Pasca e Harabagiu, 2001), sumarização (Plaza, Díaz e Gervás, 2010) ou tradução automática (Knight e Luk, 1994). Para o português, além de outros recursos lexicais de larga cobertura (veja-se Santos et al. (2010) para mais informação acerca deste tipo de recursos para português), existe já uma rede léxico-semântica pública extraída a partir de um dicionário, o PAPEL (Gonçalo Oliveira, Santos e Gomes, 2010). No âmbito deste trabalho, consideramos que rede léxico-semântica é uma base de conhecimento baseada em itens lexicais que, de acordo com os seus significados, podem estar ligados uns aos outros através de relações semânticas.

De forma a minimizar o problema da incompletude – há muito apontado à construção automática de bases de conhecimento lexical a partir de dicionários (veja-se Ide e Veronis (1995)) – neste trabalho utilizamos não um, mas três dicionários na criação de uma rede léxico-semântica. Ou seja, o nosso principal objectivo passa por enriquecer o PAPEL, extraído apenas de um dicionário, com informação obtida a partir de outros dois dicionários da língua portuguesa. O resultado é o CARTÃO, uma rede léxico-semântica de grandes dimensões extraída a partir de três dicionários da língua portuguesa.

Para o inglês, a WordNet de Princeton (Fellbaum, 1998) é a base de conhecimento lexical mais utilizada, e foi construída de forma manual. Por outro lado, a construção automática deste tipo de recursos é uma alternativa ao es-

força e tempo necessários para o seu desenvolvimento manual.

Neste contexto, para além da sua cobertura em termos de palavras e significados, umas das principais razões para se utilizarem dicionários é o facto de estes terem definições simples onde, normalmente, é utilizado um vocabulário controlado, o que faz com que muitas destas definições sejam quase previsíveis e, por isso, fáceis de processar automaticamente. Tendo em conta o nosso objectivo principal – extrair informação a partir de mais de um dicionário – comparamos a presença dos padrões mais frequentes nas definições de cada dicionário, e concluímos que o vocabulário estrutural é muito semelhante em todos eles. Assim, de forma a minimizar o esforço necessário para a criação manual de uma gramática para cada dicionário, reutilizamos as gramáticas em que se baseou a construção do PAPEL.

Neste artigo, depois de apresentarmos os três recursos explorados, mostramos os padrões mais frequentes e as suas ocorrências em cada dicionário. De seguida, descrevemos o procedimento utilizado para a extracção de relações semânticas. Tendo em conta cada dicionário, relatamos depois os resultados que formam o CARTÃO, que aumenta a última versão do PAPEL em mais de 70%, e damos a conhecer o (novo) Folheador, uma interface para interrogação, através da rede, dos conteúdos de recursos baseados em relações semânticas. Por fim, antes de concluir, a última secção deste artigo é dedicada à avaliação da cobertura e validação automática de algumas das relações semânticas extraídas. A primeira tarefa foi realizada através da comparação dos lemas abrangidos pelo CARTÃO e lemas abrangidos por dois *thesauri* electrónicos de larga cobertura para o português. Na segunda tarefa, compararam-se as relações de sinonímia com as relações presentes num dos *thesauri* anteriores, criado manualmente. Algumas das outras relações foram transformadas em frases que, por sua vez, foram procuradas num corpo jornalístico.

## 2 Trabalho relacionado

Desde cedo que os dicionários foram vistos como uma fonte de conhecimento lexical de uma língua – provavelmente a mais importante. Não é por isso surpresa que os dicionários electrónicos tenham sido também dos primeiros recursos a ser explorados na extracção automática deste tipo de conhecimento.

Os trabalhos pioneiros nesta área datam já das décadas de 1970 (Calzolari, Pecchia e Zampolli, 1973) e 1980 (Amsler, 1981), quando se começou a estudar a possibilidade de tirar partido dos dicionários na construção automática de bases de conhecimento lexical. Partindo destes primeiros estudos, surgiram também os primeiros procedimentos automáticos (Chodorow, Byrd e Heidorn, 1985), que focavam essencialmente a extracção de taxonomias. Depois de vários anos e de vários trabalhos baseados em dicionários electrónicos (e.g. Alshawi (1987), Guthrie et al. (1990), Dolan, Vanderwende e Richardson (1993), Dolan (1994)), foram apontadas algumas críticas à sua utilização na obtenção automática de informação léxico-semântica de grande cobertura (Ide e Veronis, 1995). Entre as críticas referia-se que os dicionários são inconsistentes e incompletos. No entanto, foi também mostrado que alguns dos problemas poderiam ser minimizados se fosse utilizado mais do que um dicionário. Foi também sugerido que este tipo de trabalhos poderia ser complementado com conhecimento obtido a partir de outros tipos de recurso (e.g. corpos).

Apesar das críticas em torno deste tipo de trabalho, a MindNet foi a primeira base de conhecimento lexical, completamente independente e extraída automaticamente a partir de dicionários (Richardson, Dolan e Vanderwende, 1998). Mais do que um recurso estático, a MindNet representa uma metodologia que inclui um conjunto de ferramentas para adquirir, estruturar, aceder e explorar informação semântica em texto, não só de dicionários, mas também de enciclopédias e de corpos.

Nesta altura, havia já surgido a WordNet de Princeton (Fellbaum, 1998), um recurso criado manualmente para o inglês que se revelaria como a base de conhecimento lexical mais amplamente utilizada pela comunidade de PLN. Tal como os dicionários, a WordNet, que viria a expandir-se para outras línguas[1], também é um recurso baseado em palavras e sentidos, para os quais existem definições. Por outro lado, a estrutura da WordNet encontra-se preparada a ser explorada por, ou integrada em, aplicações computacionais. As suas estruturas fundamentais são os chamados *synsets*, grupos de palavras sinónimas que representam lexicalizações de conceitos da linguagem natural. Os *synsets* podem estar ligados entre si através de relações semânticas de vários tipos (e.g. hiperonímia, parte-de).

Desde o surgimento da WordNet, começou a haver um maior interesse na aquisição automática de conhecimento que pudesse ser uti-

[1]Os projectos WordNet pelo mundo encontram-se listados na página da Global WordNet Association: `http://www.globalwordnet.org/gwa/wordnet_table.html`, de onde destacamos a WordNet.PT (Marrafa, 2002).

lizado no enriquecimento deste recurso (veja-se, por exemplo, Hearst (1998), Navigli et al. (2004), ou Toral, Muñoz e Monachini (2008)), o que levou a que a quantidade de trabalhos na extracção de conhecimento lexical a partir de dicionários diminuísse.

Ainda assim, houve na última década alguns trabalhos recentes nesta área, como Nichols, Bond e Flickinger (2005), para o japonês, ou Gonçalo Oliveira, Santos e Gomes (2010) para o português. Há ainda a referir trabalhos recentes que exploram o dicionário colaborativo Wikcionário no contexto da extracção de informação. Apesar de menos popular que a sua parente Wikipédia, o Wikcionário foi já utilizado, por exemplo, no cálculo de proximidades semânticas (Weale, Brew e Fosler-Lussier, 2009)(Zesch, Müller e Gurevych, 2008), na criação de ontologias lexicais (Wandmacher et al., 2007), ou no enriquecimento de recursos léxico-semânticos existentes (Sajous et al., 2010).

## *3 Recursos explorados e formato dos dados*

Uma razão que também contribui para que os dicionários nem sempre sejam explorados no contexto da extracção de informação é a sua disponibilidade. Os dicionários comerciais têm normalmente o seu conteúdo protegido, mesmo para fins de investigação, e nem sempre existem alternativas gratuitas.

No caso do português, o PAPEL é um recurso livre que, no entanto, resultou do processamento de um dicionário proprietário, o Dicionário PRO da Língua Portuguesa (doravante DLP) (DLP, 2005). Além da versão actual do PAPEL, o PAPEL 3.0, neste trabalho foram explorados outros dois dicionários livres, nomeadamente o Dicionário Aberto (doravante DA) (Simões e Farinha, 2011) e a versão portuguesa da iniciativa Wikcionário[2].

Nesta secção apresentamos, primeiro, os três recursos referidos anteriormente e, depois, o formato utilizado para representar os dicionários, pronto a ser processado pelos módulos que lidam com a extracção de relações.

### 3.1 Apresentação dos recursos

O DA é a versão electrónica de um dicionário de português cuja versão original data de 1913. A sua ortografia está actualmente a ser modernizada. O DA contém cerca de 128 mil entradas e está disponível no formato PDF e ainda em dois formatos textuais, onde se inclui uma versão em XML[3]. No entanto, neste trabalho foi utilizado o estado actual da segunda revisão da modernização do DA, gentilmente cedida pela sua equipa de desenvolvimento.

O Wikcionário é uma iniciativa mantida pela fundação Wikimedia que tem o objectivo de disponibilizar um conjunto de dicionários multilingues. Além de informação que geralmente se encontra em dicionários, tal como a categoria gramatical das palavras, a sua etimologia e pronunciação, ou traduções, algumas entradas do Wikcionário incluem informação acerca de relações semânticas relevantes para a entrada, como sinónimos, antónimos ou hiperónimos. No entanto, por se tratar de um projecto dependente de voluntários, este tipo de informação é escasso e incompleto para a versão portuguesa do Wikcionário (doravante Wikcionário.PT).

Os Wikcionários estão disponíveis em ficheiros XML, onde as entradas se encontram escritas em texto *wiki*. Neste trabalho foi utilizada a versão de 8 de Dezembro de 2011 do Wikcionário.PT, para a qual desenvolvemos um analisador para aceder à informação de cada entrada (Pérez, Gonçalo Oliveira e Gomes, 2011). A versão do Wikcionário.PT utilizada contém cerca de 210 mil entradas, das quais cerca de 115 mil estão identificadas como tendo pelo menos a definição de uma palavra portuguesa. Tratando-se de um dicionário multilingue, as restantes entradas referem-se apenas a palavras noutras línguas.

O PAPEL 3.0[4] é a mais recente versão de uma rede léxico-semântica pública, extraída de forma automática a partir do DLP. Contém cerca de 102 mil itens lexicais e 190 mil ligações entre eles, que simbolizam relações semânticas e que estão representadas através de triplos com a seguinte estrutura:

`arg1 RELACIONADO_COM arg2`
(e.g. `animal HIPERONIMO_DE cão`)

Um triplo indica que um sentido do item lexical no primeiro argumento (`arg1`) se relaciona com um sentido do item lexical no segundo argumento (`arg2`), através de uma relação identificada por `RELACIONADO_COM`.

### 3.2 Formato dos dados

De forma a obter informação léxico-semântica no formato descrito anteriormente a partir de outros dicionários, convertemos os seus formatos XML para um formato mais amigável, onde cada linha

[2]Ver `http://pt.wiktionary.org/`

[3]Ver `http://www.dicionario-aberto.net/`

[4]Disponível a partir do endereço `http://www.linguateca.pt/PAPEL/`

contém apenas o lema, a sua categoria gramatical e a sua definição. Veja-se o exemplo seguinte, para a palavra *coco*:

```
coco    nome    fruto gerado pelo coqueiro, muito
                usado para se fazer doces e para
                consumo de seu líquido
```

Neste formato, palavras que têm mais do que uma definição dão origem a mais de uma linha. Além disso, como o Wikcionário inclui listas de sinónimos para várias entradas, transformamos também essas listas em definições com apenas uma palavra, tal como no seguinte exemplo para a palavra *bravo*.

```
Sinónimos:    corajoso, destemido
              ⇓
bravo   adj   corajoso
bravo   adj   destemido
```

Apenas definições de categorias abertas são utilizadas, e transformadas numa notação comum: `nome` para substantivos, `verbo` para verbos, `adj` para adjectivos e `adv` para advérbios.

Intencionalmente não mantivemos informação acerca do número da acepção a que cada definição corresponde, informação que é geralmente incluída nos dicionários. Uma das razões para esta opção é, dada a ambiguidade, a impossibilidade de fazer uma correspondência clara e directa entre as ocorrências das palavras nas definições e a acepção a que dizem respeito. Além disso, a lista de acepções de uma palavra num dicionário raramente tem correspondência directa com a mesma palavra noutro dicionário, já que não existe um critério bem definido para divisão de palavras em sentidos (Dolan, 1994) (Kilgarriff, 1996) (Peters, Peters e Vossen, 1998). Os sentidos da mesma palavra podem ir desde intimamente relacionados (e.g. na polissemia ou metonímia) até totalmente não relacionados (e.g. na homonímia). Sendo assim, ao invés de desenvolver heurísticas para desambiguar as palavras nas definições (como em Navigli (2009)), e também para encontrar correspondências entre as acepções de palavras em diferentes dicionários, tratamos da mesma forma todas as ocorrências de palavras com a mesma ortografia.

Após a conversão do DA e do Wikcionário.PT, obtivemos cerca de 229 mil e 72 mil definições, respectivamente, de cada dicionário. Para além do Wikcionário ser um recurso que, apesar de estar em crescimento, ter ainda uma dimensão pequena, as suas definições são menos porque descartamos definições: (i) correspondentes apenas a palavras noutras línguas; (ii) correspondentes a palavras de categorias fechadas ou a palavras flexionadas (incluindo formas verbais); (iii) em entradas com sintaxes alternativas, não previstas pelo nosso analisador. Devido a ser criado por voluntários, nem sempre especialistas, e por não existir um padrão para o texto *wiki* das entradas, não é possível construir um analisador que preveja todas as variantes de sintaxe, nem que seja 100% livre de erros. Tal como para o Wikcionário.PT, este problema parece ser comum a outras edições do Wikcionário (veja-se por exemplo Navarro et al. (2009)).

## 4 Análise das regularidades nas definições

Uma das principais razões para os dicionários serem a primeira escolha na aquisição de relações semânticas está relacionada com a simplicidade e sistematicidade das suas definições, o que os torna fáceis de processar e de ser explorados na extracção automática de informação. Foi por isso que a extracção das relações do PAPEL se baseou num conjunto de gramáticas, desenvolvidas manualmente, e que incluíam padrões léxico-sintácticos que, no DLP, indicam frequentemente a presença de relações semânticas.

Além da identificação de regularidades nas definições, de forma a evitar a criação manual de uma gramática para cada dicionário, procuramos verificar se estas regularidades eram preservadas em diferentes dicionários. Assim, comparamos as quantidades dos padrões mais frequentes nas definições de cada dicionário. Os padrões considerados mais produtivos, ou seja, que são frequentes e apropriados para a exploração na extracção automática de relações, são apresentados na tabela 1, juntamente com a sua frequência em cada dicionário, bem como a relação semântica que geralmente indicam.

Esta análise permitiu-nos confirmar que a maior parte das regularidades são preservadas nos três dicionários. Desta forma é possível utilizar as mesmas regras para extrair relações semânticas a partir de todos eles, não havendo por isso necessidade de criar uma gramática específica para cada um. Assim, tendo em conta que já incluíam a maior parte dos padrões na tabela 1, foi-nos possível reutilizar as gramáticas do PAPEL na criação do CARTÃO. Consequentemente, o CARTÃO engloba também os mesmos tipos de relações semânticas que o PAPEL.

Entre as alterações mínimas que fizemos encontram-se, por exemplo:

- A utilização do padrão `o mesmo que` para extracção da relação de sinonímia.

| Padrão | Cat. gram. | Frequência | | | Relação |
|---|---|---|---|---|---|
| | | DLP | DA | Wikcionário | |
| *o mesmo que* | Substantivo | 0 | 10.627 | 1.107 | Sinonímia |
| *a[c]to ou efeito de* | Substantivo | 3.851 | 2.501 | 645 | Causa |
| *pessoa que* | Substantivo | 1.320 | 47 | 329 | Hiperonímia |
| *aquele que* | Substantivo | 1.148 | 3.357 | 545 | Hiperonímia |
| *conjunto de* | Substantivo | 1.004 | 316 | 298 | Membro |
| *espécie de* | Substantivo | 798 | 2.846 | 223 | Hiperonímia |
| *género/gênero de* | Substantivo | 29 | 4.148 | 48 | Hiperonímia |
| *variedade de* | Substantivo | 455 | 621 | 52 | Hiperonímia |
| *[a] parte do/da* | Substantivo | 445 | 433 | 107 | Parte |
| *qualidade de* | Substantivo | 777 | 775 | 126 | Qualidade |
| *qualidade do que é* | Substantivo | 663 | 543 | 105 | Qualidade |
| *estado de* | Substantivo | 299 | 223 | 73 | Estado |
| *natural ou habitante de/da/do* | Substantivo | 536 | 0 | 79 | Local/Origem |
| *instrumento[,] para* | Substantivo | 94 | 284 | 25 | Finalidade |
| *.. produzid[o/a] por/pel[o/a]* | Substantivo | 155 | 146 | 60 | Produtor |
| *o mesmo que* | Verbo | 0 | 166 | 97 | Sinonímia |
| *fazer* | Verbo | 1.680 | 1.294 | 364 | Causa |
| *tornar* | Verbo | 1.359 | 1.672 | 266 | Causa |
| *ter* | Verbo | 467 | 519 | 139 | Propriedade |
| *o mesmo que* | Adjectivo | 0 | 2.685 | 197 | Sinonímia |
| *relativo a/á/ao* | Adjectivo | 1.236 | 5.554 | 1.063 | Propriedade |
| *que se* | Adjectivo | 1.602 | 1.599 | 485 | Propriedade |
| *que tem* | Adjectivo | 2.698 | 4.291 | 477 | Parte/Propriedade |
| *diz-se de* | Adjectivo | 2.066 | 738 | 313 | Propriedade |
| *relativo ou pertencente* | Adjectivo | 1.647 | 9 | 61 | Membro/Propriedade |
| *habitante ou natural de* | Adjectivo | 0 | 0 | 189 | Local/Origem |
| *que não é/está* | Adjectivo | 485 | 608 | 98 | Antonímia |
| *de modo* | Advérbio | 398 | 2.261 | 109 | Maneira |
| *de maneira* | Advérbio | 49 | 9 | 36 | Maneira |
| *de forma* | Advérbio | 30 | 3 | 19 | Maneira |
| o mesmo que | Advérbio | 0 | 182 | 21 | Sinonímia |

Tabela 1: Padrões nas definições, frequentes e produtivos

- A possibilidade de trocar a ordem das palavras chave `natural` e `habitante` na extracção de relações Local/Origem;
- A consideração de algumas grafias na variante brasileira, que ocorrem no Wikcionário. Por exemplo, as palavras `ato` e `gênero`.

Além da utilização dos padrões estáticos representados na tabela 1, foram aplicadas mais duas regras que se revelaram bastante produtivas para extrair relações semânticas a partir dos três dicionários:

- Sinonímia pode ser extraída a partir de definições com apenas uma palavra, ou uma enumeração de palavras.
- A maior parte das definições de substantivos são estruturadas por *genus* e *differentia*, ou seja, iniciam-se com a apresentação de um género próximo, normalmente um hiperónimo do lema (eventualmente modificado por um adjectivo) e a diferença específica.[5]

[5]A excepção a estes casos é quando a palavra no início da definição é considerada uma "cabeça vazia" (*empty head*, ver Chodorow, Byrd e Heidorn (1985) ou Guthrie et al. (1990)). As cabeças vazias são substantivos que, apesar de iniciarem normalmente definições, não devem ser considerados como hiperónimos. Podem ser ignorados ou, preferencialmente, ser também explorados na extracção de hiperonímia (e.g. `espécie`, `variedade`) ou outras relações, como parte (e.g. `parte`) ou membro (e.g. `membro`, `conjunto`).

## *5 Aquisição automática de relações semânticas*

Como já vimos na secção anterior, as regularidades nas definições de dicionário permitem que a extracção de relações semânticas se baseie num conjunto finito de regras criado manualmente. Este procedimento opõem-se aos algoritmos vulgarmente denominados de *bootstrapping*, que são habitualmente utilizados na extracção de relações semânticas a partir de texto não estruturado (veja-se, por exemplo, o trabalho de Pantel e Pennacchiotti (2006)). O funcionamento desses algoritmos baseia-se num pequeno conjunto de relações de um determinado tipo (sementes), em que existe a máxima confiança, que é utilizado para aprender novas relações.

Em relação a algoritmos de *bootstrapping*, tendo em conta que estes também aprendem

os padrões de extracção, a nossa abordagem tem a desvantagem de requerer mais tempo na construção das gramáticas e de estas não ficarem desde logo adaptáveis a todos os tipos de texto. No entanto, já vimos que no caso dos três dicionários esta desvantagem não se revela um problema. Por outro lado, a nossa abordagem permite-nos um controlo superior sobre os padrões de extracção.

O procedimento para a criação do CARTÃO é, também ele, fortemente inspirado na construção do PAPEL, relatada em Gonçalo Oliveira, Santos e Gomes (2010), e consiste, por isso, também numa fase manual e em duas automáticas. As relações semânticas, estabelecidas entre itens lexicais nas definições e o item lexical definido, são extraídas após o processamento das entradas dos dicionários, representadas no formato descrito na secção 3.2. As instâncias de cada relação são representadas como triplos, da mesma forma que no PAPEL (ver secção 3.1). Descrevemos de seguida o procedimento (exemplificado na figura 1 para um pequeno conjunto de regras e duas definições):

1. **Criação das gramáticas de extracção:** Os padrões mais produtivos são compilados manualmente em gramáticas, especialmente criadas para a extracção de relações entre itens lexicais nas definições e os lemas definidos.
2. **A própria extracção:** As gramáticas são utilizadas em conjunto com um analisador sintáctico, o PEN (Gonçalo Oliveira e Gomes, 2008), que processa as definições do dicionário, representadas no formato introduzido na secção 3.2. Apenas definições de palavras de categoria aberta são processadas. No fim, se uma definição respeita um padrão, são extraídas instâncias de relações semânticas e representadas como triplos $p_1$ $R$ $p_2$, onde $p_1$ é um item lexical na definição, $p_2$ é o lema definido, e $R$ é o nome da relação estabelecida entre um sentido de $p_1$ e um sentido de $p_2$.
3. **Limpeza e lematização:** Após a extracção, algumas relações apresentam argumentos inválidos, incluindo sinais de pontuação ou preposições. Nesta fase, as definições vêem a categoria gramatical dos seus elementos anotada. É para isso utilizado o anotador de categoria gramatical disponibilizado pelo projecto OpenNLP[6], utilizando os módulos para a língua portuguesa[7].

   Esta anotação não é feita antes da extracção porque os modelos do anotador foram treinados em texto de corpos, e não têm a mesma precisão na anotação de definições de dicionário. Além disso, as gramáticas do PAPEL também não consideram estas anotações.

   Depois da anotação, os triplos com argumentos inválidos são descartados. Além disso, se os argumentos dos triplos se encontrarem flexionados, e por isso não definidos directamente no dicionário, são aplicadas algumas regras de lematização.

Apesar das definições do DA se encontrarem em processo de modernização de grafia, os lemas definidos foram mantidas na sua forma original. Por isso, os triplos extraídos a partir deste recurso passaram por uma quarta fase, em que argumentos com sequências que caíram em desuso são modernizados de acordo com as sugestões de Simões, Almeida e Farinha (2010). Ainda assim, de forma a minimizar as possibilidade de gerar palavras inexistentes, de todos os triplos com argumentos alterados mantivemos apenas os 9.163 em que ambos os argumentos se encontravam também no PAPEL. Os demais, que totalizam 23.226 triplos, foram descartados.

## *6 Análise quantitativa*

Nesta secção apresentamos as relações que fazem parte do PAPEL 3.0, e também aquelas que resultaram do processamento dos outros dois dicionários. A partir do DA foram extraídos cerca de 134 mil triplos e do Wikcionário.PT cerca de 57,3 mil. Estes foram depois juntos com os cerca de 190 mil triplos do PAPEL, de forma a constituir o CARTÃO, que aumenta o PAPEL 3.0 em 72%, em relação aos triplos, e em 52% relativamente ao número de lemas abrangidos. Os números apresentados confirmam que, apesar dos dicionários pretenderem cobrir toda a língua, acabam por ser incompletos. A melhor forma de conseguir um recurso mais abrangente é, portanto, juntar conhecimento obtido a partir de vários recursos.

### 6.1 As relações em números

Na tabela 2 apresentam-se os números de triplos extraídos, de acordo com o dicionário de onde são originários e com o tipo de relação, para o qual é fornecido um exemplo real. À semelhança do que acontecia no PAPEL, para cada relação existe uma relação inversa definida, que se pode obter pela troca dos argumentos e alteração do nome. Por exemplo, as relações

[6] http://incubator.apache.org/opennlp/
[7] http://opennlp.sourceforge.net/models-1.5/

1. Excerto de gramática:

```
...
RAIZ ::= HIPERONIMO_DE <&> ...
...
RAIZ ::= CABECA_VAZIA
CABECA_VAZIA ::= parte
...
RAIZ ::= ...  <&> usado <&> para <&> FAZ_SE_COM
RAIZ ::= parte <&> de <&> TEM_PARTE
RAIZ ::= ...  <&> que <&> contém <&> DET <&> PARTE_DE
...
```

2. Definições e relações extraídas:

```
candeia   nome   utensílio doméstico rústico usado para iluminação, com pavio abastecido a óleo
→ utensílio HIPERONIMO_DE candeia
→ com FAZ_SE_COM candeia
→ iluminação FAZ_SE_COM candeia
espiga   nome   parte das gramíneas que contém os grãos
→ espiga PARTE_DE gramíneas
→ grãos PARTE_DE espiga
...
```

3. Resultado da anotação, limpeza e lematização:

```
candeia   nome   utensílio#n doméstico#adj rústico#adj usado#v-pcp para#prp iluminação#n ,#punc
com#prp pavio#n abastecido#v-pcp a#prp óleo#n
→ utensílio HIPERONIMO_DE candeia
→ iluminação FAZ_SE_COM candeia
espiga   nome   parte#n de#prp as#art gramíneas#n que#pron-indp contém#v-fin os#art grãos#n
→ espiga PARTE_DE gramínea
→ grão PARTE_DE espiga
...
```

Figura 1: Exemplo do processo de aquisição de relações semânticas.

inversas de `sentimento` hiperónimo-de `afecto` e de `vírus` causador-de `doença` são respectivamente `afecto` hipónimo-de `sentimento` e `doença` resultado-de `vírus`.

Verifica-se que cerca de 40% dos triplos do CARTÃO são relações de sinonímia. As relações de hiperonímia são aproximadamente um terço. Dentro das relações restantes destacam-se os triplos referente-a, que se estabelecem entre adjectivos e outras categorias gramaticais, representando cerca de 12% dos triplos do CARTÃO.

Para dar uma ideia da contribuição de cada recurso em termos dos triplos e das suas intersecções, apresentamos a figura 2, onde é possível verificar não só a quantidade de triplos extraídos de cada recurso, mas também a quantidade extraída de dois ou dos três recursos.

A tabela 3 dá outra perspectiva na contribuição de cada dicionário para o CARTÃO. Os conjuntos de triplos extraídos de cada dicionário são comparados dois a dois através do cálculo da sua semelhança e da novidade de cada um em relação a outro, utilizando as seguintes medidas:

$$Sem(A,B) = Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

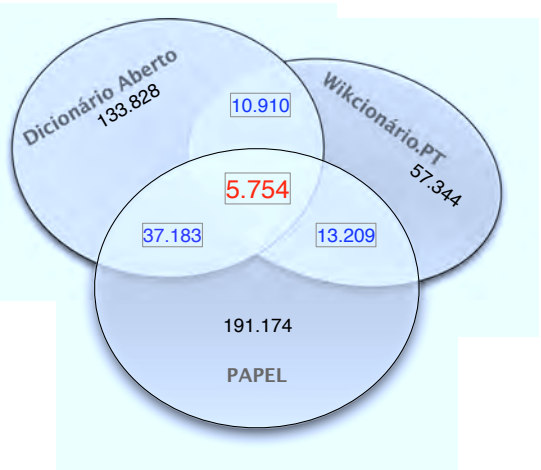



Figura 2: Intersecções dos conjuntos de triplos extraídos. A **preto** o número de triplos de cada recurso, a **azul** as intersecções dois a dois, e a **vermelho** os triplos extraídos dos três dicionários.

$$Novidade(A,B) = \frac{|A| - |A \cap B|}{|A|} \qquad (2)$$

Como seria de esperar, devido às suas diferenças de tamanho, as maiores novidades são do PAPEL e do DA em relação ao Wicionário.PT.

| Relação | Args. | Quantidade | | | | Exemplo |
|---|---|---|---|---|---|---|
| | | PAPEL | DA | Wikcionário | Únicos | |
| Sinónimo-de | n,n | 40,306 | 25.046 | 13.812 | 67.620 | *alegria,satisfação* |
| | v,v | 18.927 | 11.113 | 4.650 | 28.108 | *esticar,estender* |
| | adj,adj | 21.726 | 10.505 | 6.611 | 32.364 | *racional,filosófico* |
| | adv,adv | 1.178 | 1.199 | 277 | 2.286 | *imediatamente,já* |
| Hiperónimo-de | n,n | 62.591 | 44.777 | 17.068 | 97.924 | *sentimento,afecto* |
| Parte-de | n,n | 2.424 | 1.146 | 614 | 3.893 | *núcleo,átomo* |
| | n,adj | 3.033 | 3.414 | 520 | 5.872 | *vício,vicioso* |
| | adj,n | 43 | 45 | 16 | 104 | *sujeito,oração* |
| Membro-de | n,n | 5.679 | 928 | 1.161 | 7.328 | *aluno,escola* |
| | n,adj | 77 | 26 | 25 | 120 | *coisa,coletivo* |
| | adj,n | 968 | 80 | 138 | 1.071 | *rural,campo* |
| Contido-em | n,n | 216 | 124 | 53 | 381 | *tinta,tinteiro* |
| | n,adj | 176 | 124 | 34 | 287 | *óleo,oleoso* |
| Material-de | n,n | 335 | 513 | 146 | 888 | *folha_de_papel,caderno* |
| Causador-de | n,n | 951 | 193 | 317 | 1.423 | *vírus,doença* |
| | n,adj | 17 | 8 | 5 | 25 | *paixão,passional* |
| | adj,n | 494 | 148 | 173 | 748 | *horrível,horror* |
| | n,v | 40 | 17 | 6 | 60 | *fogo,fundir* |
| | v,n | 6.256 | 7.140 | 1.631 | 10.664 | *mover,movimento* |
| Produtor-de | n,n | 910 | 605 | 333 | 1.741 | *oliveira,azeitona* |
| | n,adj | 49 | 26 | 6 | 77 | *fermentação,fermentado* |
| | adj,n | 352 | 236 | 37 | 515 | *fonador,som* |
| Finalidade-de | n,n | 3.659 | 2.353 | 1.442 | 6.978 | *sustentação,mastro* |
| | n,adj | 56 | 40 | 9 | 88 | *habitação,habitável* |
| | v,n | 4.609 | 2.230 | 1.610 | 7.824 | *calcular,cálculo* |
| | v,adj | 236 | 204 | 27 | 374 | *comprimir,compressivo* |
| Tem-qualidade | n,n | 740 | 465 | 87 | 1.055 | *mórbido,morbidez* |
| | n,adj | 888 | 667 | 128 | 1.273 | *assíduo,assiduidade* |
| Tem-estado | n,n | 265 | 118 | 44 | 376 | *exaltação,desvairo* |
| | n,adj | 129 | 102 | 23 | 220 | *disperso,dispersão* |
| Lugar-de | n,n | 834 | 405 | 601 | 1.483 | *Equador,equatoriano* |
| Maneira-de | adv,n | 795 | 1.537 | 164 | 2.172 | *ociosamente,indolência* |
| | adv,adj | 345 | 1.624 | 135 | 1.854 | *virtualmente,virtual* |
| Maneira | adv,n | 116 | 147 | 16 | 250 | *prontamente,demora* |
| sem | adv,v | 6 | 5 | 3 | 13 | *seguido,parar* |
| Antónimo-de | n,n | 388 | 410 | 59 | 684 | *direito,torto* |
| Referente-a | adj,n | 6.287 | 5.024 | 1.793 | 10.652 | *daltónico,daltonismo* |
| | adj,v | 17.718 | 11,076 | 3.569 | 27.902 | *musculoso,ter_músculo* |
| **Total** | | 191.174 | 133.828 | 57.344 | 326.798 | |

Tabela 2: Quantidades e exemplos das relações extraídas.

Ainda assim, verifica-se que todos os recursos apresentam novidades elevadas (sempre superiores a 70%) em relação a cada um dos outros.

| A \ B | PAPEL | | DA | | Wikc.PT | |
|---|---|---|---|---|---|---|
| | Sem | Nov | Sem | Nov | Sem | Nov |
| **PAPEL** | | | 0,13 | 0,81 | 0,06 | 0,93 |
| **DA** | 0,13 | 0,72 | | | 0,06 | 0,92 |
| **Wikc.PT** | 0,06 | 0,77 | 0,19 | 0,81 | | |

Tabela 3: Semelhança (Sem) e novidade (Nov) dos recursos dois a dois, em termos de triplos

## 6.2 Lemas abrangidos

Fazemos também uma análise da cobertura em termos de lemas abrangidos pelo CARTÃO. A tabela 4 mostra os números de lemas diferentes nos argumentos dos triplos extraídos a partir de cada dicionário, distribuídos de acordo com a sua categoria gramatical. A maior parte dos lemas são substantivos. Depois, para o PAPEL, as categorias mais representadas são os verbos e os adjectivos, por esta ordem. Por outro lado, foram extraídos do DA e do Wikcionário.PT mais adjectivos do que verbos. O PAPEL é o recurso que fornece mais lemas ao CARTÃO, mas o DA não fica muito atrás. Os triplos extraídos do DA englobam mesmo mais substantivos e mais do dobro dos advérbios que o PAPEL.

| Cat. gram. | PAPEL | DA | Wikc.PT | Total |
|---|---|---|---|---|
| Substantivos | 55.769 | 59.879 | 23.007 | 89.895 |
| Verbos | 22.440 | 16.672 | 6.932 | 32.572 |
| Adjectivos | 22.381 | 18.563 | 7.113 | 29.964 |
| Advérbios | 1.376 | 3.073 | 473 | 3.443 |
| **Total** | 101.966 | 98.187 | 37.525 | 155,187 |

Tabela 4: Lemas únicos em relações semânticas

A figura 3 apresenta a contribuição e sobreposição de cada recurso em termos de lemas abrangidos. Além disso, da mesma forma que

calculámos a semelhança e novidade dos conjuntos de triplos extraídos, na tabela 5 apresentamos os mesmos valores, desta vez comparando os lemas envolvidos nos triplos de cada recurso. Tratando-se de lemas, as semelhanças são superiores e as novidades inferiores aos mesmos valores para os triplos. Ainda assim, as novidades são sempre superiores a 35%, o que mostra que para além de diferentes dicionários descrevem diferentes relações semânticas, de cada dicionário foi obtido mais de um terço de novo vocabulário relativamente aos outros.

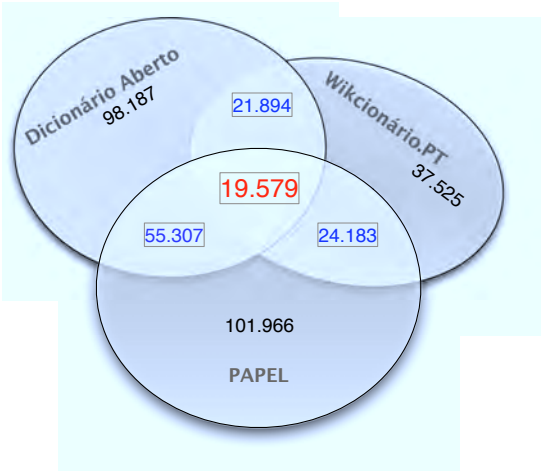



Figura 3: Intersecções dos itens lexicais extraídos. A **preto** o número de itens de cada recurso, a **azul** as intersecções dois a dois, e a **vermelho** os itens obtidos dos três recursos.

| A \ B | PAPEL | | DA | | Wikt.PT | |
|---|---|---|---|---|---|---|
| | Sem | Nov | Sem | Nov | Sem | Nov |
| **PAPEL** | | | 0,38 | 0,46 | 0,21 | 0,76 |
| **DA** | 0,38 | 0,44 | | | 0,19 | 0,78 |
| **Wikc.PT** | 0,21 | 0,36 | 0,19 | 0,42 | | |

Tabela 5: Semelhança (Sem) e novidade (Nov) dos recursos dois a dois, relativamente a lemas incluídos.

### 6.3 O (novo) Folheador

Em colaboração com a Linguateca[8] foi desenvolvido um interface na rede[9] onde é possível interrogar e navegar pelos conteúdos de recursos baseados em relações semânticas, representadas como triplos, ou seja, da mesma forma que no CARTÃO. A interface, apesar de ter sido criada de raíz, foi inspirada numa interface já existente, e utilizada para navegar no PAPEL, o Folheador. Como a ideia é que esta nova interface venha a substituir o (antigo) Folheador, a nova interface terá o mesmo nome.

A figura 4 mostra o novo Folheador e os primeiros resultados devolvidos para a palavra 'santo'. Tal como no antigo Folheador, é possível procurar por um item lexical para consultar todas as relações onde ele entra. No entanto, agora é possível seleccionar o tipo de relações a procurar, antes de fazer a pesquisa, ou procurar por todas as relações entre dois itens lexicais determinados. O novo Folheador está também feito de forma a permitir a navegação sobre triplos extraídos a partir de vários recursos. Para cada triplo encontrado, são apresentados identificadores dos recursos de que fazem parte ou de onde foram extraídos. O número de recursos de onde o triplo foi obtido pode ser, só por si, utilizado como um indicador de confiança.

Outra novidade importante do novo Folheador é a sua ligação a serviços de pesquisa em corpos, que oferecem uma nova dimensão a relações semânticas representadas como triplos e, consequentemente, ao CARTÃO. Actualmente é possível fazer uma ligação à interface do projecto AC/DC (Santos e Bick, 2000)(Santos, 2011), que permite consultar todas as frases de um conjunto de corpos portugueses onde dois itens lexicais, relacionados num triplo, co-ocorrem. Para alguns tipos de relação existe também uma ligação ao serviço VARRA (Freitas et al., 2010), que usa indirectamente o AC/DC para procurar ocorrências do próprio triplo nos corpos. Para tal, o VARRA transforma o triplo seleccionado num conjunto de frases em que os itens relacionados estão ligados por um padrão que normalmente indica a relação. Com base nestas duas ligações, estão actualmente a ser calculados graus de confiança para cada triplo.

Mais informação sobre o novo Folheador pode ser encontrada em Costa (2011).

## *7 Cobertura por outros recursos e validação automática*

Esta secção é dedicada à avaliação da cobertura do CARTÃO em relação a outros recursos, e também à sua validação automática, baseada na interrogação de um corpo de notícias. A nossa opção pela validação automática deve-se não só ao facto da avaliação manual deste tipo de conhecimento ser tediosa e requerer muito tempo, mas também devido a esta última ser, geralmente, um tipo de avaliação algo subjectivo e difícil de reproduzir. Ainda assim, não descartamos, no futuro, vir a realizar uma avaliação manual de uma amostra significativa do CARTÃO. Além disso, é sempre possível utilizar o serviço

[8] `http://www.linguateca.pt/`

[9] Disponível a partir de `http://linguateca.pt/Folheador/`

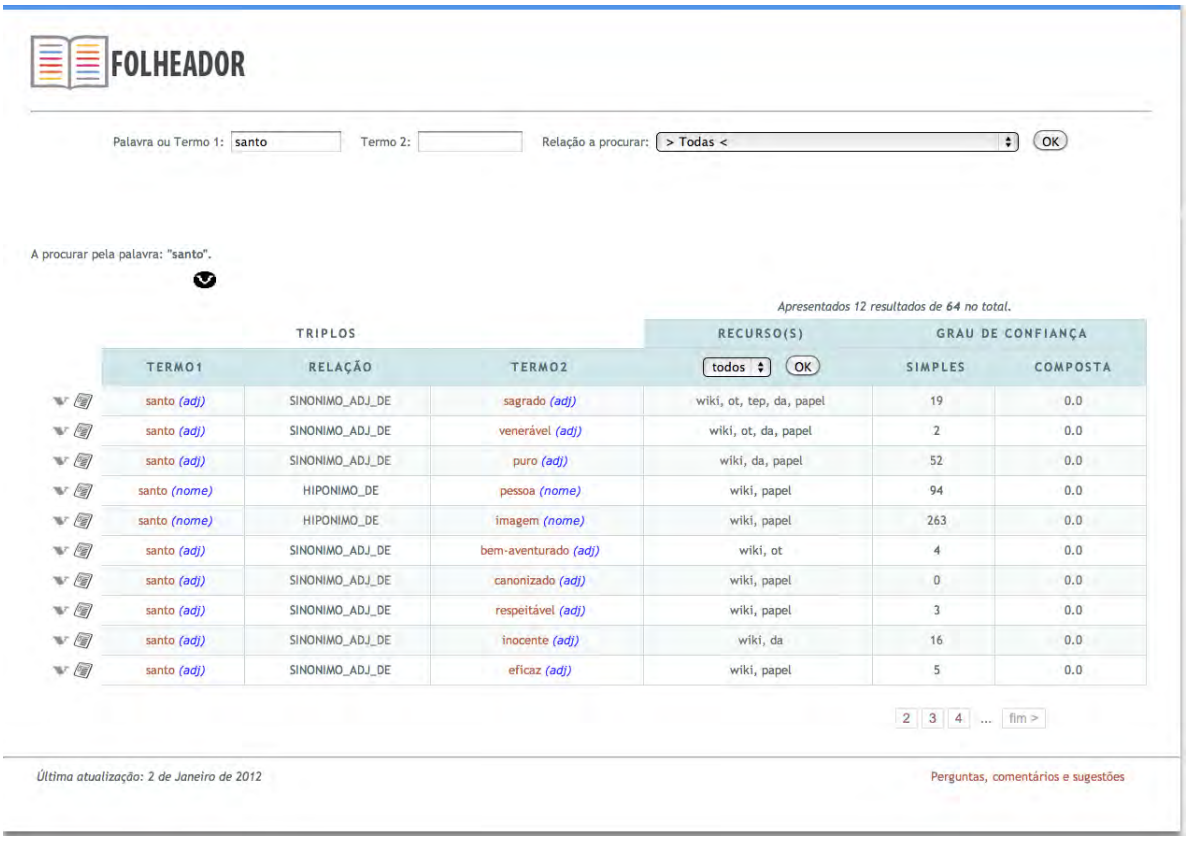



Figura 4: Primeiros resultados da pesquisa por 'santo' no Folheador.

VARRA[10] (Freitas et al., 2010) para validar triplos, também manualmente, com base no contexto em que as palavras relacionadas ocorrem.

## 7.1 Cobertura por *thesauri* criados manualmente

A cobertura do CARTÃO foi medida em relação a dois recursos lexicais livres para o português, criados de forma manual, nomeadamente o TeP 2.0 (Maziero et al., 2008) e o OpenThesaurus.PT[11] (OT). Estes recursos são ambos *thesauri*, organisados em *synsets*, tal como a WordNet, ainda que não possuam relações entre synsets[12]. O TeP foi criado para o português do Brasil e contém 43.666 itens lexicais, organizadas em 18.795 *synsets*. O OT é uma iniciativa colaborativa, cerca de quatro vezes mais pequena que o TeP, que contém 13.258 itens lexicais organizados em 4.102 *synsets*.

A tabela 6 apresenta a cobertura dos lemas do CARTÃO por ambos os *thesauri*. Entre cerca de 21% (substantivos no DA) e 60% (adjectivos no Wikcionário.PT) dos lemas estão abrangidos pelo TeP. Por outro lado, devido à sua dimensão, para o OT estes números ficam entre os 3% (advérbios no DA) e os 31% (adjectivos no Wikcionário.PT). Considerando apenas o TeP, existe uma maior proporção de adjectivos e advérbios cobertos, comparando com o mesmo número para os substantivos. A baixa proporção de advérbios do DA cobertos e a elevada proporção de substantivos do Wikcionário.PT cobertos são as excepções.

Os triplos do Wikcionário.PT têm a maior proporção de lemas cobertos para todas as categorias, o que se pode explicar pela natureza colaborativa deste recurso. O Wikcionário.PT ainda está em crescimento, e é criado por voluntários, normalmente não peritos, enquanto que o DLP e o DA são dicionários comerciais, criados por lexicógrafos. Assim, enquanto o DLP e o DA, para além de terem vocabulário mais comum, incluírem também definições mais formais e menos convencionais, o Wikcionário tende a utilizar vocabulário mais convencional. Há ainda a destacar que o Wikcionário.PT contém bastantes definições escritas na variante brasileira do português, que é a variante alvo do TeP, contribuindo isto também para a maior proporção de lemas cobertos do primeiro recurso pelo segundo.

A tabela 7 mostra a cobertura de cada triplo de sinonímia pelo TeP – se o TeP tiver um *synset*

[10] Ver `http://www.linguateca.pt/VARRA/`

[11] `http://openthesaurus.caixamagica.pt/`

[12] Na verdade, o TeP contém ligações que representam relações antonímia, mas que não foram utilizadas neste trabalho.

| Cat. gram. | TeP | | | | | | OT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PAPEL | | DA | | Wikc.PT | | PAPEL | | DA | | Wikc.PT | |
| Substantivos | 13.137 | 23,6% | 12.701 | 21,2% | 8.079 | 35,1% | 5.736 | 10,3% | 5.532 | 9,2% | 4.440 | 19,3% |
| Verbos | 6.029 | 26,9% | 5.835 | 35,0% | 3.138 | 45,3% | 2.731 | 12,2% | 2.644 | 15,9% | 1.977 | 28,5% |
| Adjectivos | 9.104 | 40,7% | 8.264 | 44,5% | 4.265 | 60,0% | 3.249 | 14,5% | 2.846 | 15,3% | 2.256 | 31,7% |
| Advérbios | 574 | 41,7% | 683 | 22,2% | 264 | 55,8% | 94 | 6,8% | 94 | 3,1% | 79 | 16,7% |

Tabela 6: Cobertura de lemas por thesauri criados manualmente.

que contenha ambos os argumentos de um triplo de sinonímia, consideramos que o triplo é abrangido pelo TeP. A proporção de triplos cobertos é apresentada para todos os triplos de sinonímia do recurso em questão (Total), bem como considerando apenas os triplos em que ambos os argumentos do triplo existem no TeP (ArgsNoTeP). Decidimos omitir os mesmos dados para o OT por se tratar de um recurso demasiado pequeno. A cobertura da sinonímia de acordo com a categoria gramatical é consistente para os três recursos – mais elevada para sinonímia entre verbos, seguida pela sinonímia entre adjectivos. À semelhança da cobertura dos lemas, a proporção de triplos de sinonímia cobertos pelo TeP é também maior para o Wikcionário.PT.

Para além de darem uma ideia acerca da cobertura do CARTÃO, estes números mostram que os *thesauri* disponíveis e criados manualmente podem ser uma fonte adicional de relações de sinonímia. Além do mais, por se encontrarem em recursos criados manualmente, a confiança na qualidade destas relações é elevada.

## 7.2 Validação com base na interrogação de um corpo

O procedimento de validação que vamos apresentar de seguida é inspirado num procedimento já utilizado para validar as relações semânticas do PAPEL (ver Gonçalo Oliveira, Santos e Gomes (2010)). Baseia-se num conjunto de padrões discriminadores, indicadores de relações semânticas em texto, e procura por ocorrências desses padrões a ligar os argumentos das relações semânticas extraídas.

Contudo, os resultados apresentados não devem ser confundidos com a precisão das relações extraídas, tendo em conta que:

- Um corpo é um recurso com conhecimento limitado;
- Há imensas formas de exprimir uma relação semântica em texto, o que torna impossível a codificação de todos os padrões e variações possíveis;
- Alguns tipos de relação são específicos dos dicionários, e não é expectável que estejam explícitas em texto de corpos. Isto acontece, por exemplo, para relações entre substantivos e verbos, que implicam a nominalização do verbo, tal como em `aumentar` causador-de `aumento`.
- Alguns estudos (Dorow, 2006) mostram que palavras sinónimas não co-ocorrem frequentemente em corpos, especialmente na mesma frase[13]. Esta ideia vai ao encontro do pressuposto de um sentido por discurso (Gale, Church e Yarowsky, 1992), dado que, principalmente em textos especializados, o autor tenderá a utilizar sempre a mesma palavra para se referir ao mesmo conceito. Também por isso realizamos previamente a validação das relações de sinonímia.

Apesar disto, estes resultados dão-nos uma ideia da utilização das relações extraídas em texto não estruturado. Mais do que isso, se for utilizado o mesmo conjunto de padrões e o mesmo corpo, os resultados são um indicador que pode ser utilizado na comparação de recursos baseados em relações semânticas, e que pode informar acerca da aplicabilidade das suas relações.

Dadas as limitações referidas, apenas foram validados quatro tipos de relações, todos eles entre substantivos. Nesta validação utilizamos o corpo jornalístico CETEMPúblico (Rocha e Santos, 2000) (Santos e Rocha, 2001), onde procuramos por relações de hiperonímia, parte-de, membro-de e finalidade-de, extraídas a partir dos três dicionários utilizados. Apesar de versões anteriores do PAPEL terem já sido validadas através de um procedimento semelhante, repetimos essa validação seguindo os mesmos critérios para os três recursos, de forma a tornar possível uma comparação directa dos resultados. A lista de padrões discriminadores utilizada foi construída com base nos padrões léxico-sintácticos utilizados nas validações anteriores do PAPEL e ainda dos padrões do serviço VARRA.

A tabela 8 apresenta os resultados da validação automática dos triplos extraídos dos três dicionários. A tabela mostra o número, e respectiva proporção, de todos os triplos dos quatro tipos validados cujos argumentos co-ocorrem

[13] Palavras sinónimas tenderão antes a ocorrer em contextos semelhantes.

| Cat. gram. | PAPEL | | | DA | | | Wikc.PT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cobertos | Total | ArgsNoTeP | Cobertos | Total | ArgsNoTeP | Cobertos | Total | ArgsNoTeP |
| Substantivos | 11.920 | 30,0% | 56,2% | 6.821 | 27,2% | 41,4% | 4.126 | 29,9% | 50,4% |
| Verbos | 10.063 | 53,1% | 83,5% | 5.927 | 53,3% | 76,2% | 2.532 | 54,3% | 78,5% |
| Adjectivos | 8.506 | 39,2% | 69,7% | 4.891 | 46,6% | 66,9% | 2.903 | 43,9% | 71,8% |
| Advérbios | 267 | 22,7% | 38,1% | 208 | 17,3% | 27,6% | 131 | 32,9% | 47,3% |

Tabela 7: Cobertura da sinonímia pelo TeP

em pelo menos uma frase do corpo (ArgsCooc). Mostra-se ainda o número de triplos suportados pelo corpo e a proporção dos triplos cujos argumentos co-ocorrem a que esse número corresponde (Suportados).

Constata-se que a proporção de triplos cujos argumentos co-ocorrem nunca é mais de 37,5% (hiperonímia no Wikcionário.PT), nem menor de 17,5% (hiperonímia no DA). Curiosamente os valores máximo e mínimo obtêm-se para o mesmo tipo de relação, mas recurso diferente. Tanto para o PAPEL, como para o Wikcionário.PT, a proporção de relações membro-de cujos argumentos co-ocorrem no CETEMPúblico é inferior à mesma proporção para as demais relações, nos mesmos recursos. Isto poderá depois contribuir para que seja a relação com mais triplos suportados.

Nos três recursos, a proporção de triplos suportados no corpo é sempre mais elevada para a relação membro-de e mais baixa para finalidade. Acreditamos que a baixa proporção de relações de finalidade suportadas se deve ao facto desta relação não estar tão bem definida semanticamente como as outras três. Além disso, existirão mais padrões discriminadores para esta relação, e os padrões utilizados serão menos frequentes e com mais variações.

Dos triplos de hiperonímia e parte-de do PAPEL e do DA cujos argumentos co-ocorrem no CETEMPúblico, cerca 30% são suportados. Mais uma vez, devido ao seu tamanho e natureza colaborativa, as proporções mais elevadas são obtidas pelo Wikcionário.PT, De forma a dar uma ideia mais clara daquilo em que consistiu a avaliação, a tabela 9 mostra exemplos de frases que suportam triplos do CARTÃO. Nas mesmas frases, os padrões discriminadores encontram-se a negrito.

## *8 Notas finais*

Neste artigo apresentamos o CARTÃO, uma rede léxico-semântica de grandes dimensões, extraída automaticamente a partir de três dicionários da língua portuguesa. Após analisarmos a estrutura das definições nos dicionários utilizados, verificamos que podíamos tirar partido das mesmas regras, baseadas em padrões léxico-sintácticos, para extrair relações semânticas a partir dos três.

Além de utilizar as mesmas gramáticas, a construção do CARTÃO é inspirada na construção do PAPEL, uma rede léxico-semântica pública, também extraída de um dicionário, e incluída no CARTÃO. A versão do PAPEL utilizada neste trabalho é o PAPEL 3.0, a sua versão mais recente, desenvolvida em paralelo com o resto do trabalho aqui apresentado. Analisando os resultados de extracção, mostramos a contribuição de cada dicionário para o CARTÃO, e verificamos que este recurso aumenta o PAPEL 3.0 em mais de 70%. Tendo em conta que os dicionários pretendem abranger toda a língua, é um aumento significativo, e confirma que há vantagens na utilização de mais de um dicionário neste tipo de trabalho.

A cobertura do CARTÃO foi avaliada através da comparação dos lemas que este inclui com os lemas em *thesauri* portugueses de larga cobertura, livres e criados manualmente. Além disso, algumas relações entre substantivos foram validadas automaticamente com base na sua ocorrência num corpo jornalístico.

Entre outras tarefas, no futuro pretendemos, por exemplo, refinar algumas relações (e.g. finalidade-de) e rever o contributo de alguns dos padrões utilizados. De forma a termos uma informação mais clara sobre a qualidade dos vários tipos de relação, estamos a ponderar a realização de uma avaliação manual de uma parte do CARTÃO. Esta avaliação poderá mesmo ser integrada no projecto VARRA, que contribuirá ainda na atribuição de graus de confiança a cada triplo, e na investigação das formas em que palavras relacionadas co-ocorrem no texto de corpos. Além do trabalho aqui descrito, utilizando procedimentos semelhantes, o CARTÃO pode ainda ser enriquecido com relações léxico-semânticas obtidas a partir de outros recursos, incluindo não apenas dicionários, mas também *thesauri*, ou mesmo a Wikipédia (veja-se Herbelot e Copestake (2006) para o inglês, ou Gonçalo Oliveira, Costa e Gomes (2010), para o português).

Desde o início do projecto PAPEL que a opção foi construir um recurso lexical em que os lemas não estivessem divididos em sentidos. Esta opção é justificada, inicialmente, pela ar-

| Relação | PAPEL | | | | DA | | | | Wikcionário.PT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ArgsCooc | | Suportados | | ArgsCooc | | Suportados | | ArgsCooc | | Suportados | |
| Hiperonímia | 13.724 | 21,9% | 4.098 | 29,7% | 7.846 | 17,5% | 2.255 | 28,7% | 6.405 | 37,5% | 2.086 | 32,6% |
| Parte-de | 573 | 23,6% | 186 | 32,5% | 247 | 21,6% | 81 | 32,8% | 226 | 36,8% | 94 | 41,6% |
| Membro-de | 1.089 | 19,2% | 464 | 42,6% | 303 | 32,7% | 109 | 36,0% | 317 | 27,3% | 147 | 46,4% |
| Finalidade | 1.017 | 27,8% | 164 | 16,1% | 473 | 20,1% | 65 | 13,7% | 498 | 34,5% | 75 | 15,1% |

Tabela 8: Validação automática do CARTÃO

| Relação | Suporte |
|---|---|
| *língua* hiperónimo-de *alemão* | *As iniciativas deste gabinete passam geralmente pela promoção de conferências, exposições, workshops e aulas de* ***línguas, como o inglês, alemão*** *ou japonês.* |
| *ciência* hiperónimo-de *paleontologia* | *A* ***paleontologia é uma ciência*** *que depende do que se descobre.* |
| *rua* parte-de *quarteirão* | *De resto, o* ***quarteirão formado pelas ruas*** *de São João e de Mouzinho da Silveira está, por esse motivo, assente em estacas de madeira...* |
| *mão* parte-de *corpo* | *As* ***mãos são a parte do corpo*** *mais atingida (29,7%).* |
| *pessoa* membro-de *comissão* | *A* ***comissão é constituída por pessoas*** *que ficaram marcadas pela presença de Dona Amélia: ...* |
| *lobo* membro-de *alcateia* | *Mech e os seus colegas constataram que alguns dos cheiros contidos nas marcas de urina servem para os* ***lobos de uma alcateia*** *saberem por onde andou o lobo que deixou as marcas ...* |
| *transporte* finalidade-de *embarcação* | *... onde foi descoberto o resto do casco de uma* ***embarcação presumivelmente utilizada no transporte*** *de peças de cerâmica ...* |
| *espectáculo* finalidade-de *anfiteatro* | *Sobre a hipótese da construção de «stands» de artesanato e de um* ***anfiteatro para espectáculos****, a edilidade portuense diz ainda não estar nada decidido.* |

Tabela 9: Frases exemplo, que suportam relações semânticas.

tificialidade da divisão em sentidos (Kilgarriff, 1996), que é muitas vezes diferente de lexicógrafo para lexicógrafo. Outra razão é, na prática, a inexistência de sinónimos. Existem sim quase-sinónimos, que suscitam questões interessantes, e que também devem ser exploradas. Além disso, em linguagem natural, o estudo da vagueza é tão ou mais importante que o estudo da ambiguidade, ou seja, muitos dos casos mais interessantes e frequentes são casos de vagueza, o que favorece a opção em não se separar os lemas, muitas vezes de forma arbitrária, em sentidos.

Por outro lado, reconhecemos que, ainda que artificial, em muitas tarefas PLN seja útil a existência de um recurso em que as palavras estejam separadas nos seus sentidos mais típicos, tal como acontece numa *wordnet*. Assim, no projecto Onto.PT (Gonçalo Oliveira e Gomes, 2010) (Gonçalo Oliveira e Gomes, 2011c), onde se integra o desenvolvimento do CARTÃO, pretendemos manter o mesmo recurso livre e estruturado de duas formas alternativas – rede baseada em lemas, como o CARTÃO, e ontologia semelhante à *wordnet*, o Onto.PT propriamente dito – para que os investigadores utilizem aquela que lhes for mais útil.

Muito resumidamente, para a construção automática do Onto.PT é necessário passar por três fases. A primeira, passa pela extracção de relações semânticas a partir de recursos textuais, o que dá origem exactamente ao CARTÃO. De seguida, o objectivo é identificar automaticamente *synsets* no meio das relações de sinonímia (Gonçalo Oliveira e Gomes, 2011a). Por fim, procuram-se integrar, também automaticamente, os restantes triplos, através da associação dos lemas dos seus argumentos a *synsets* adequados (Gonçalo Oliveira e Gomes, 2011b). O resultado é uma ontologia lexical para o português, com uma estrutura semelhante a uma *wordnet*, ou seja, uma rede constituída por relações semânticas entre *synsets*.

As relações que compõem o PAPEL podem ser livremente descarregadas a partir do sítio da Linguateca, enquanto que as relações extraídas a partir do DA e do Wikcionário.PT podem ser descarregadas a partir do sítio do projecto Onto.PT, respectivamente:

- `http://www.linguateca.pt/PAPEL/`
- `http://ontopt.dei.uc.pt/`

## *Agradecimentos*

Gostaríamos de agradecer à Diana Santos pela sugestão do nome para o recurso, pela orientação no desenvolvimento do Folheador e pela argumentação da utilidade de recursos lexicais em que não é feita a distinção entre sentidos.

Agradecemos à Linguateca por ter financiado o desenvolvimento do novo Folheador, através da

contratação do Hernani Costa.

Agradecemos também à Cláudia Freitas por, em conjunto com a Diana Santos, ter participado activamente na discussão opções para o PAPEL 3.0, e ao Alberto Simões por nos ter cedido a mais recente revisão do DA modernizado.

Por fim, agradecemos aos revisores pelos seus valiosos comentários, que contribuíram para melhorar este este trabalho.

Hugo Gonçalo Oliveira é apoiado pela bolsa de doutoramento da FCT SFRH/BD/44955/2008, co-financiada pelo FSE.

## *Referências*

Agirre, Eneko, Oier Lopez De Lacalle, e Aitor Soroa. 2009. Knowledge-based WSD on specific domains: performing better than generic supervised WSD. Em *Proceedings of 21st International Joint Conference on Artifical Intelligence (IJCAI)*, pp. 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alshawi, Hiyan. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics*, 13(3-4):195–202.

Amsler, Robert A. 1981. A taxonomy for english nouns and verbs. Em *Proceedings of 19th annual meeting on Association for Computational Linguistics*, pp. 133–138, Morristown, NJ, USA. ACL Press.

Calzolari, Nicoletta, Laura Pecchia, e Antonio Zampolli. 1973. Working on the italian machine dictionary: a semantic approach. Em *Proceedings of 5th Conference on Computational Linguistics*, pp. 49–52, Morristown, NJ, USA. ACL Press.

Chodorow, Martin S., Roy J. Byrd, e George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. Em *Proceedings of 23rd annual meeting on Association for Computational Linguistics*, pp. 299–304, Morristown, NJ, USA. ACL Press.

Costa, Hernani. 2011. O desenho do novo folheador. Relatório técnico, Linguateca, Dezembro, 2011. `http://www.linguateca.pt/Equipa/Hernani/HernaniCostarelFolheador.pdf`.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

2005. *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto.

Dolan, William, Lucy Vanderwende, e Stephen D. Richardson. 1993. Automatically deriving structured knowledge bases from online dictionaries. Em *Proceedings of the 1st Conference of the Pacific Association for Computational Linguistics*, PACLING'93, pp. 5–14.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. Em *Proceedings of 15th International Conference on Computational Linguistics*, COLING'94, pp. 712–716, Morristown, NJ, USA. ACL Press.

Dorow, Beate. 2006. *A Graph Model for Words and their Meanings*. Tese de doutoramento, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, e Violeta Quental. 2010. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. Em *Livro do IX Encontro de Linguística de Corpus*, ELC 2010.

Gale, William A., Kenneth W. Church, e David Yarowsky. 1992. One sense per discourse. Em *Proceedings of the HLT'91 workshop on Speech and Natural Language*, pp. 233–237, Morristown, NJ, USA. ACL Press.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2008. Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário. Relatório técnico, CISUC, November, 2008. PAPEL Tech Report 3, `http://linguateca.dei.uc.pt/papel/GoncaloOliveiraetal2008relPAPEL3.pdf`.

Gonçalo Oliveira, Hugo, Hernani Costa, e Paulo Gomes. 2010. Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. Em *Actas do II Simpósio de Informática (INFORUM 2010)*, pp. 537–548. Universidade do Minho.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. Em *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*. IOS Press.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011a. Automatic discovery of fuzzy synsets from dictionary definitions. Em *Proceedings of 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1801–1806, Barcelona, Spain. IJCAI/AAAI.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011b. Ontologising relational triples into a portuguese thesaurus. Em *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, pp. 803–817, Lisbon, Portugal, October, 2011. APPIA.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011c. Onto.PT: Construção automática de uma ontologia lexical para o português. Em Ana R. Luís, editor, *Estudos de Linguística*, volume 1. Imprensa da Universidade de Coimbra, Coimbra. No prelo.

Gonçalo Oliveira, Hugo, Diana Santos, e Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, May, 2010.

Guthrie, L., B. Slator, Y. Wilks, e R. Bruce. 1990. Is there content in empty heads? Em *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3 of *COLING'90*, pp. 138–143, Helsinki, Finland.

Hearst, Marti A. 1998. Automated Discovery of WordNet Relations. Em *(Fellbaum, 1998)*. pp. 131–151.

Herbelot, Aurelie e Ann Copestake. 2006. Acquiring ontological relationships from wikipedia using RMRS. Em *Proceedings of ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*.

Ide, N. e J. Veronis. 1995. Knowledge extraction from machine-readable dictionaries: An evaluation. Em *Machine Translation and the Lexicon, LNAI*. Springer.

Kilgarriff, A. 1996. Word senses are not bona fide objects: implications for cognitive science, formal semantics, nlp. Em *Proceedings of 5th International Conference on the Cognitive Science of Natural Language Processing*, pp. 193–200.

Knight, Kevin e Steve K. Luk. 1994. Building a large-scale knowledge base for machine translation. Em *Proceedings of 12th national conference on Artificial intelligence (vol. 1)*, AAAI '94, pp. 773–778, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Marrafa, Palmira. 2002. Portuguese Wordnet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, e Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390–392.

Navarro, Emmanuel, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Tzu Y. Kuo, Pierre Magistry, e Chu R. Huang. 2009. Wiktionary and NLP: Improving synonymy networks. Em *Proceedings of Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Suntec, Singapore. ACL Press.

Navigli, Roberto. 2009. Using cycles and quasi-cycles to disambiguate dictionary glosses. Em *Proceedings of the 12th Conference on European chapter of the Association for Computational Linguistics*, EACL'09, pp. 594–602, Athens, Greece.

Navigli, Roberto, Paola Velardi, Alessandro Cucchiarelli, e Francesca Neri. 2004. Extending and enriching Wordnet with OntoLearn. Em *Proceedings of 2nd Global WordNet Conference (GWC)*, pp. 279–284, Brno, Czech Republic. Masaryk University.

Nichols, Eric, Francis Bond, e Dan Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. Em *Proceedings of 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1111–1116. Professional Book Center.

Pantel, Patrick e Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. Em *Proceedings of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, pp. 113–120, Sydney, Australia. ACL Press.

Pasca, Marius e Sanda M. Harabagiu. 2001. The informative role of WordNet in open-domain question answering. Em *Proc. NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pp. 138–143, Pittsburgh, USA.

Pérez, Leticia Anton, Hugo Gonçalo Oliveira, e Paulo Gomes. 2011. Extracting lexical-semantic knowledge from the portuguese wiktionary. Em *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, pp. 703–717, Lisbon, Portugal, October, 2011. APPIA.

Peters, Wim, Ivonne Peters, e Piek Vossen. 1998. Automatic Sense Clustering in EuroWordNet.

Em *Proceedings of 1st International Conference on Language Resources and Evaluation*, LREC'98, pp. 409–416, Granada, May, 1998.

Plaza, Laura, Alberto Díaz, e Pablo Gervás. 2010. Automatic summarization of news using wordnet concept graphs. *International Journal on Computer Science and Information System (IADIS)*, V:45–57.

Richardson, Stephen D., William B. Dolan, e Lucy Vanderwende. 1998. Mindnet: Acquiring and structuring semantic information from text. Em *Proceedings of 17th International Conference on Computational Linguistics*, COLING'98, pp. 1098–1102.

Rocha, Paulo Alexandre e Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pp. 131–140, São Paulo, 19-22 de Novembro, 2000. ICMC/USP.

Sajous, Franck, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, e Yannick Chudy. 2010. Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. Em *Advances in Natural Language Processing, 7th International Conference on NLP (IceTAL)*, volume 6233 of *LNCS*, pp. 332–344, Reykjavik, Iceland. Springer.

Santos, Diana. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3(2):113–128, Junho, 2011. Volume edited by J.B.Johannessen, Language variation infrastructure.

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, e Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. Em A. M. Brito, F. Silva, J. Veloso, e A. Fiéis, editores, *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 681–700.

Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em *Proc. of the 2nd International Conf. on Language Resources and Evaluation (LREC)*, pp. 205–210.

Santos, Diana e Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. Em *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 442–449. ACL Press, 9-11 July, 2001.

Simões, Alberto e Rita Farinha. 2011. Dicionário Aberto: Um novo recurso para PLN. *Viceversa*, (16):159–171, December, 2011.

Simões, Alberto, José João Almeida, e Rita Farinha. 2010. Processing and extracting data from dicionário aberto. Em *Proceedings of International Conference on Language Resources and Evaluation*, LREC 2010, Malta.

Toral, Antonio, Rafael Muñoz, e Monica Monachini. 2008. Named entity wordnet. Em *Proc. International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco. ELRA.

Wandmacher, Tonio, Ekaterina Ovchinnikova, Ulf Krumnack, e Henrik Dittmann. 2007. Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. Em *3rd Australasian Ontology Workshop (AOW)*, volume 85 of *CRPIT*, pp. 61–69, Gold Coast, Australia. ACS.

Weale, Timothy, Chris Brew, e Eric Fosler-Lussier. 2009. Using the wiktionary graph structure for synonym detection. Em *Proceedings of 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pp. 28–31, Stroudsburg, PA, USA. ACL Press.

Zesch, Torsten, Christof Müller, e Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. Em *Proceedings of 23rd National Conference on Artificial Intelligence (AAAI)*, pp. 861–866. AAAI Press.

# Conversão de Grafemas para Fonemas em Português Europeu – Abordagem Híbrida com Modelos Probabilísticos e Regras Fonológicas

Arlindo Veiga
Instituto de Telecomunicações,
Polo de Coimbra
DEEC, Universidade de Coimbra
aveiga@co.it.pt

Sara Candeias
Instituto de Telecomunicações,
Polo de Coimbra
saracandeias@co.it.pt

Fernando Perdigão
Instituto de Telecomunicações,
Polo de Coimbra
DEEC, Universidade de Coimbra
fp@co.it.pt

**Resumo**

A conversão de grafema para fonema diz respeito à tarefa de encontrar a pronúncia de um vocábulo dado na sua forma escrita, a qual tem uma forte componente de aplicação em sistemas de reconhecimento e de síntese de fala. Uma nova abordagem na conversão de grafema para fonema é proposta, aplicando um modelo híbrido para o qual concorrem regras fonológicas e decisões estatísticas. Os resultados mostram que a incorporação de regras fonológicas em algoritmos de informação estatística melhora acentuadamente o desempenho do conversor. Para este trabalho, foi construído um dicionário de pronunciação com mais de 40000 vocábulos derivados do corpus CETEMPúblico. Os dicionários fonológicos de pronunciação para o português europeu, bem como outros recursos produzidos durante este trabalho, estão disponibilizados publicamente. O sistema que aqui se descreve foi aplicado à língua portuguesa escrita, sem e com o Acordo Ortográfico de 1990, e, ainda que aplicado ao português na sua vertente europeia, observa características que permitem a sua aplicação a outras línguas românicas.

## *1. Introdução*

Comummente designado por G2P[1], o mapeamento entre grafemas e fonemas tem por objetivo converter um texto escrito numa sequência de símbolos que representam os sons da fala de uma determinada língua, de uma forma inequívoca. Numa atualidade motivada para o uso de várias aplicações tecnológicas ativadas pela fala, a conversão de G2P tem sido alvo de estudos e de desenvolvimento. Apesar de ser já sólida e madura a investigação na área, em português europeu (PE), o problema do G2P ainda não se encontra totalmente resolvido, como se pode comprovar quer pelas taxas de erro publicadas nos artigos da área, quer pelos erros de conversão que persistem nos atuais sistemas existentes no mercado. Por outro lado, ainda que a maior causa para os problemas encontrados tenha sido já diagnosticada, e que envolve questões de natureza essencialmente morfológica e sintática (cf., a título de exemplo, (Braga *e* Marques, 2007)), as soluções até ao momento apresentadas, muitas vezes acompanhadas por um dicionário extenso de exceções, não estão claramente publicadas nem se encontram acessíveis para possível melhoria e extensão. O presente estudo levou à disponibilização, em (SPL, 2011), de *i*) dicionários de pronunciação; *ii*) modelos de sequências de conjuntos grafema-fonema e de *iii*) programas que permitem converter grafemas em fonemas.

No sentido de encontrar uma solução para os problemas da conversão G2P para o PE, são várias as abordagens que têm vido a ser propostas, de entre as quais destacamos as seguintes: *i*) por regras linguísticas, expostas em (Braga *et al.*, 2006), (Oliveira *et al*., 1992) e (Teixeira, 2004); *ii*) por regras inferidas a partir dos dados (Teixeira *et al*., 2006); *iii*) por máquinas de estados finitos (Caseiro *e* Trancoso, 2002), (Oliveira *et al.*, 2004); *iv*) por máxima entropia (Barros *e* Weiss, 2006); *v*) baseadas em redes neuronais (Trancoso *et al.*, 1994); e *vi*) por CARTs - Classification and Regression Trees (Oliveira *et al.*, 2001). Uma das técnicas a referenciar, e que tem sido aplicada essencialmente em línguas que não apresentam uma clara correspondência entre grafema e fone(ma), como é o caso do inglês, é a abordagem por modelos probabilísticos, apresentado por (Demberg *et al.*, 2007) e por (Bisani *e* Ney, 2008). Contrastando com as abordagens baseadas em regras, as quais são suportadas por um conhecimento linguístico da língua, que se pretende exaustivo, a abordagem estatística baseia-se no pressuposto de que a pronunciação de um vocábulo é possível de ser prevista, por analogia, a partir de exemplos de sequências suficientes de grafonemas – unidades identificativas da associação entre grafema e respetivo fonema (cf. (Bisani *e* Ney, 2008)). Uma das vantagens apontada pelos modelos probabilísticos é de não implicar uma verificação constante da interdependência das

[1] Do inglês *Grapheme to Phoneme*; por vezes também designado L2S: *Letter to Sound*.

regras, em especial quando surge uma sequência de grafemas que sai fora das regularidades até então admitidas. Por outro lado, tem-se verificado que a conversão G2P proveniente de modelos probabilísticos não capta um contexto suficientemente amplo de forma a impedir que a estrutura fonológica da língua seja violada. A língua portuguesa na sua vertente europeia, assim como as línguas românicas na sua generalidade, apresenta uma razoável regularidade fonética e fonológica bem como uma ortografia de base fonológica. Estas características explicam o sucesso da aplicação de regras linguísticas tais como a marcação de sílaba tónica (descrito em (Candeias *e* Perdigão, 2008), (Braga *et al.*, 2006), (Almeida *e* Simões, 2001) e (Teixeira *e* Freitas, 1998), como exemplos para o PE).

O estudo que aqui se apresenta propõe uma abordagem híbrida ao problema da conversão G2P, onde se utiliza um modelo probabilístico, no qual são incorporadas regras fonológicas.

Este artigo encontra-se estruturado da seguinte forma: a Secção 2 descreve o modelo probabilístico de sequências conjuntas e a Secção 3 descreve a criação do modelo híbrido usando regras fonológicas. Na Secção 4 são apresentados os resultados da contribuição de cada componente do módulo. A Secção 5 apresenta as conclusões. Estudos futuros são igualmente referidos. Parte deste trabalho foi apresentado parcialmente em (Veiga *et al.*, 2011).

## *2. Modelo probabilístico de sequências conjuntas*

A tarefa de conversão de grafemas para fonemas pode ser formulada na determinação da sequência ótima de fonemas dada a sequência de grafemas, usando uma abordagem probabilística. Definindo $G = G_1^N = \{g_1, g_2, \cdots, g_N\}$ como sendo uma sequência de $N$ grafemas e $F = F_1^M = \{f_1, f_2, \cdots, f_M\}$ como uma sequência de $M$ fonemas, a determinação da sequência ótima de fonemas, $F^*$, é descrita da seguinte forma:

$$F^* = \arg\max_F P(F \mid G). \qquad (1)$$

Não sendo fácil determinar $F^*$ calculando diretamente a probabilidade *a posteriori* $P(F \mid G)$ para todas as sequências $F$ possíveis, podemos usar o teorema de Bayes e rescrever o problema como:

$$F^* = \arg\max_F \frac{P(G \mid F)}{P(G)} P(F). \qquad (2)$$

O fator $1/P(G)$ pode ser eliminado uma vez que é comum para todas as sequências $F$. Assim, o seu valor não influencia a escolha de $F^*$, pelo que o problema pode ser simplificado de acordo com a seguinte equação:

$$F^* = \arg\max_F P(G \mid F) P(F). \qquad (3)$$

A estimação de $P(F)$ é feita usualmente recorrendo aos modelos "$n$-grama". Quanto à determinação de $P(G|F)$, as abordagens que utilizam modelos de Markov simplificam o problema assumindo a independência entre os grafemas que constituem uma sequência. Neste caso, o cálculo de $P(G|F)$ pode ser decomposto da seguinte forma:

$$P(G \mid F) = \prod_{n=1}^{N} P(g_n \mid F). \qquad (4)$$

Esta simplificação parte de princípio que a dependência entre fonemas é suficiente para modelar o problema e que os contextos de fonemas replicam os contextos de grafemas (cf. (Taylor, 2005), (Demberg, 2006) e (Jiampojamarn *e* Kondrak, 2009)).

Existem outras abordagens que propõem a utilização de modelos de probabilidade conjunta, $P(F,G)$, para determinar a sequência ótima de fonemas usando diretamente a expressão da probabilidade conjunta em (1) no lugar da probabilidade condicional (em (Bisani *e* Ney, 2002) e (Galescu *e* Allen, 2001)). Estas abordagens possibilitam a modelação da dependência entre grafemas, a dependência entre fonemas e a dependência entre grafemas e fonemas.

Qualquer abordagem estatística adotada na tarefa de conversão de grafemas em fonemas requer a existência de um dicionário fonológico, necessário para estimar as probabilidades dos padrões encontrados, e, a maioria das abordagens, requer ainda um algoritmo que permita o alinhamento entre grafemas e fonemas. Em comparação com a que usa modelos de Markov, a abordagem que reclama de um algoritmo de alinhamento apresenta um melhor desempenho, pelo que é a seguida neste trabalho.

### 2.1 Alinhamento entre grafemas e fonemas

Muitos grafemas do português têm uma correspondência unívoca com os fonemas, situação na qual a conversão G2P é direta. É o caso de muitas das consoantes, como o <p> e o <t>, em <**p**or**t**uguês> que são diretamente convertidos nos

fonemas /p/ e /t/, respetivamente. Contudo, existem grafemas em que a correspondência com os fonemas é dependente de vários fatores, nomeadamente o contexto grafemático (caso dos grafemas <r> e <u> em <por**r**tug**u**ês>) e o estatuto morfológico[2], algumas das vezes com interdependência sintática (caso dos grafemas <e> e <o>, os quais, dependendo da sua condição morfológica, podem ser convertidos nos fonemas /e/[3] ou /E/ e /o/ ou /O/, respetivamente: <s**e**lo> (nome) →/s**e**lu/, <s**e**lo> (verbo) → /s**E**lu/; <**o**lho> (nome) → /**o**Lu/, <**o**lho> (verbo) → /**O**Lu/). Existem situações em que um único grafema pode originar vários fonemas, assim como existem situações em que vários grafemas podem originar um único fonema (como <g> e <j> → /Z/, conversão esta igualmente dependente de contexto). Todas as abordagens estatísticas deparam-se com este problema, sendo necessário, durante o processo de treino, segmentar e alinhar as duas sequências com igual número de segmentos. A solução nem sempre é trivial ou única e depende da forma como os algoritmos de alinhamento associam os grafemas aos fonemas de um dado vocábulo.

De acordo com (Jiampojamarn *et al.*, 2007), os alinhadores podem ser classificados em dois tipos:

1) **"um-para-um"**

Neste tipo de alinhador cada grafema é associado a apenas um fonema, originando segmentos com apenas um símbolo. Ainda assim, é necessário utilizar um símbolo nulo ('_') para lidar com casos em que um grafema pode originar vários fonemas (inserção de fonemas) ou casos em que vários grafemas originam apenas um fonema (apagamento de fonemas). As inserções de fonemas podem ser evitadas, no caso do PE, uma vez que ocorrem em pouquíssimos contextos, facilmente identificados, como é o caso do iode que ocorre em algumas das estruturas, tais como em <extra> /6iStr6/. Este tipo de alinhador é de fácil implementação (por exemplo, através do algoritmo de Levenshtein (Gusfield, 1997)), mas necessita do conhecimento prévio do mapeamento entre grafemas e fonemas. Na literatura da área, denomina-se "01-01" quando inserções e apagamentos de fonemas são permitidos e "1-01" quando apenas apagamentos de fonemas são permitidos. No presente estudo, o alinhador usado é o de "um-para-um", na vertente de "1-01".

2) **"muitos-para-muitos"**

Neste tipo os segmentos podem ser compostos por vários símbolos, o que possibilita a associação de vários grafemas a vários fonemas. Este alinhador é mais genérico, pode ser utilizado sem nenhum conhecimento prévio do mapeamento entre grafemas e fonemas e lida com os casos de inserções e de apagamentos de fonemas sem necessidade de recorrer a símbolos especiais. No entanto, os modelos resultantes são mais difíceis de estimar e o desempenho é geralmente inferior ao dos modelos com alinhamento "um-para-um" (Bisani *e* Ney, 2008). Este tipo de associação é também conhecido como alinhamento "*m*–*n*".

## 2.2 Grafonemas

Depois de efetuado o alinhamento entre grafemas e fonemas, as sequências de grafemas e de fonemas apresentam o mesmo número de segmentos. É proposta na literatura uma nova entidade, composta pela associação de um segmento de grafemas a um segmento de fonemas, denominada de *grafonema* (Bisani *e* Ney, 2002). Mostra-se um exemplo com o vocábulo <compõem> no qual os grafonemas estão entre parênteses retos. Neste exemplo considerou-se, tanto quanto possível, um alinhamento de "um-para-um", mas onde se admitem casos de alinhamento de "2-para-1" e de "1-para-2".

$$\begin{matrix} \textit{Grafemas} \\ \textit{Fonemas} \end{matrix}\begin{bmatrix} \text{c} \\ \text{k} \end{bmatrix}\begin{bmatrix} \text{om} \\ \text{o\textasciitilde} \end{bmatrix}\begin{bmatrix} \text{p} \\ \text{p} \end{bmatrix}\begin{bmatrix} \tilde{\text{o}} \\ \text{o\textasciitilde i\textasciitilde} \end{bmatrix}\begin{bmatrix} \text{e} \\ \text{6\textasciitilde} \end{bmatrix}\begin{bmatrix} \text{m} \\ \text{i\textasciitilde} \end{bmatrix}.$$

Uma sequência de *K* grafonemas é anotada como $Q(F,G) = \{q_1, q_2, \cdots, q_K\}$ e o problema de conversão de grafemas para fonemas pode agora ser escrito como:

$$F^* = \arg\max_F P\big(Q(F,G)\big). \qquad (5)$$

Dada uma sequência de *K* grafonemas, $Q(F,G)$, e não admitindo independência entre símbolos, a probabilidade da sequência, $P\big(Q(F,G)\big)$, pode ser calculada como:

$$P\big(Q(F,G)\big) = P(q_1)P(q_2 \mid q_1)P(q_3 \mid q_1 q_2)\cdots P(q_K \mid q_1 q_2 \cdots q_{K-1}) \qquad (6)$$

No modelo estatístico é frequente limitar-se o contexto (ou história) dos grafonemas utilizando os chamados modelos "*n*-grama", que correspondem a

[2] Em inglês, PoS – *part-of-speech*.
[3] Alfabeto SAMPA; cf. Tabela 1.

sequências limitadas a um comprimento até $n$ símbolos. Deste modo, a equação (6) pode ser aproximada a:

$$P\big(Q(F,G)\big) \approx \prod_{i=1}^{K} P(q_i \mid q_{i-n+1} \cdots q_{i-1}) \,. \qquad (7)$$

## 2.3 Estimação do modelo

Os modelos "$n$-grama" são utilizados para estimar a probabilidade de um símbolo, neste caso grafonema, conhecendo os $n-1$ símbolos anteriores (história). A estimação da probabilidade de um "$n$-grama" é baseada em contagens de ocorrências num dado conjunto de treino. Definindo a frequência de um "$n$-grama" por $C(\cdot)$, a sua probabilidade pode ser estimada através de:

$$P(q_i \mid q_{i-n+1} \cdots q_{i-1}) = \frac{C(q_{i-n+1} \cdots q_i)}{C(q_{i-n+1} \cdots q_{i-1})} \,, \qquad (8)$$

onde

$$C(q_{i-n+1} \cdots q_{i-1}) = \sum_j C(q_{i-n+1} \cdots q_{i-1} q_j) \,. \qquad (9)$$

Ainda que esta probabilidade seja de cálculo simples, e acarreta o problema de atribuir probabilidade zero aos "$n$-grama" que não estão presentes no dicionário de treino. Além disso, podem existir "$n$-grama" que estão presentes no dicionário mas em número sem significado estatístico. Para evitar estes constrangimentos, é preciso precaver a existências de sequências que nunca foram encontradas no dicionário de treino (usando os chamados "descontos"), ou que são pouco frequentes (a "suavização")[4]. Assim, uma pequena massa de probabilidade é retirada dos "$n$-grama" mais frequentes e é reservada para os "$n$-grama" ausentes ou pouco frequentes no dicionário de treino.

Existem vários algoritmos propostos para resolver o problema da redistribuição da massa de probabilidade. São exemplos os algoritmos de desconto (de Good-Turing (Good, 1953), de Witten-Bell (Witten *e* Bell, 1991), de Kneser-Ney (Kneser *e* Ney, 1995)), de desconto absoluto de Ney (Ney *et al.*, 1994) e de suavização de Katz (Katz, 1987).

A estimação da probabilidade de "$n$-grama" com frequência inferior a um dado limiar é feita à custa de $(n-1)$-gramas (*backoff*).

O algoritmo implementado neste trabalho é igual ao utilizado em (Demberg *et al.*, 2007). Faz a "suavização" por interpolação e utiliza a versão modificada do algoritmo de Kneser-Ney (Chen *e* Goodman, 1998). O valor de $n$ varia de 2 a 8 conforme será descrito na secção 4.

# *3. Criação do modelo híbrido*

Nesta secção descreve-se um modelo híbrido em que se opera uma transformação da sequência de grafemas, introduzindo novos símbolos com significado fonológico preciso, proporcionando desta forma uma integração de regras fonológicas no modelo estatístico. O modelo estatístico não é alterado; apenas passam a existir mais símbolos em que a associação entre grafema e fonema é mais precisa.

## 3.1 Vocabulário

Originando o sistema de conversão um dicionário de pronunciação, foi necessário, numa primeira fase, ter disponível como base de trabalho uma listagem de vocábulos atuais e representativos do PE. O material utilizado para esse fim foi o corpus CETEMPúblico (Santos *e* Rocha, 2001), o qual contem 180 milhões de palavras[5], advindas de uma coleção de extratos do jornal Público de entre os anos 1991 e 1998.

O processo de criação dessa listagem consistiu em tomar todas as cadeias de carateres anotadas como palavras, obedecendo simultaneamente aos seguintes critérios: *i*) começar com um grafema do alfabeto português (a-z, A-Z, á-ú, Á-Ú); *ii*) não conter dígitos; *iii*) não apresentar todos os grafemas em maiúscula (caso de siglas); *iv*) não conter o caráter '.' (caso de URLs); *v*) terminar com um grafema do alfabeto português ou com '-'; *vi*) o lema correspondente não conter o caráter '=' (caso de nomes compostos).

A partir do resultado obtido, formou-se uma lista de cerca de 50k vocábulos (excluindo nomes próprios), os quais correspondem a uma contagem de ocorrências no corpus de mais do que 70 vezes. Sendo arbitrária, a consideração desta medida para a configuração do vocabulário de base deveu-se ao facto de anular a possibilidade de se estarem a incluir erros tipográficos e de se obter uma primeira listagem representativa do PE extensível até cerca de 50k vocábulos. Por fim, foram retirados quer vocábulos estrangeiros quer estrangeirismos, usando, em primeiro lugar, critérios automáticos e, seguidamente, uma confirmação manual. A pesquisa automática excluiu todos os vocábulos que apresentavam grafemas ou sequências grafemáticas que não

[4] Em inglês, *discount* e *smoothing*, respectivamente.

[5] Por palavras entendem-se, aqui, todos os átomos do corpus que contêm, pelo menos, um grafema ou dígito.

fazem parte do sistema do PE, tais como <k>, <w> e <y>; <sh> e <pp>; e <b>, <d> ou <p> em posição final de vocábulo. Alguns destes dados serão depois base de constituição de um dicionário de pronunciação de estrangeirismos (cf. descrito em 3.2). Como resultado final deste processo, consistiu-se uma lista de cerca de 40k vocábulos, os quais correspondem ao vocabulário de referência tomado para este trabalho, referenciado infra por "voc_CETEMP_40k". Na medida em que as palavras que constituem o CETEMPúblico apresentam uma grafia de acordo com as normas anteriores ao Acordo Ortográfico de 1990 (AO), houve necessidade de se constituir uma listagem adicional com vocábulos grafados de acordo com o AO. Com esse fim, usou-se a ferramenta Lince (Lince) para converter os vocábulos na nova grafia. Dos 41586 vocábulos utilizados no vocabulário pré-AO, 915 sofreram alterações de grafia, nomeadamente a eliminação das consoantes mudas (<c> e <p>), a eliminação dos hífenes e alteração de acentuação gráfica. De acordo com a possibilidade de coexistirem duas grafias, este novo vocabulário apresenta pares de vocábulos ditos 'parónimos', tais como <conceptual> e <consetual> ou <desconectar> e <desconetar>. O vocabulário pós-AO é constituído por 41598 vocábulos, sendo referenciado como "voc_CETEMP_40k_ao".
Nas secções seguintes não é feita a distinção entre estes dois vocabulários, a não ser quando pertinente, como é o caso da secção dos resultados.

## 3.2 Transcrição fonológica

A transcrição fonológica do vocabulário de referência foi feita por um processo iterativo. Em primeiro lugar, foi feito um modelo estatístico, conforme descrito em 2.2, tendo por base o dicionário de pronunciações da base de dados SpeechDat (SpeechDat) com cerca de 15k vocábulos. Para a constituição do dicionário foram retirados os estrangeirismos e foram feitas algumas correções de pronunciação. Este dicionário foi também convertido para notação SAMPA (Wells, 1997), convencionando-se que os símbolos representativos das glides [j] e [w] fossem notados como as vogais correspondentes, por razões de uniformização que ultrapassam o âmbito deste artigo. Com o mesmo princípio de uniformização, não se distinguiu a lateral velarizada da lateral, ainda que sistemas reconhecidos de anotação para o português, como o usado na SpeechDat, admitam a presença de [l~] ([5] em X-SAMPA) e de [l]. Admitiu-se, igualmente, a necessidade de inclusão do iode, como foi já observado em 2.1, nomeadamente para nos aproximarmos, o mais possível, do PE padronizado. Neste ponto, requer-se esclarecer que os símbolos SAMPA adotados (cf. Tabela 1) resultaram de uma ponderação cuidada sobre representatividade do PE falado. A observação atenta dos alfabetos fonéticos SAMPA para o Português (SAMPA-PT) e X-SAMPA, dá conta de alguma indefinição de regularidade, exemplificada na atribuição de mais do que um símbolo para o mesmo som. Na verdade, o símbolo [r] no SAMPA-PT parece ter como correspondente no X-SAMPA o símbolo [4] (IPA: [ɾ]), simbolizando o [r] no X-SAMPA a vibrante alveolar múltipla (IPA: [r]).
Relativamente à transcrição, uma dilucidação acerca da opção pela transcrição fonológica, e não fonética, é ainda devida. No que concerne o binómio fonética/fonologia, é comummente aceite pela comunidade linguística, que a fonética diz respeito às propriedades físicas e articulatórias de todos os sons que ocorrem na produção linguística, cabendo à fonologia o estudo da função de cada som pronunciado numa dada língua, a qual permite ao falante distinguir significados. É igualmente aceite que qualquer opção metodológica no que à análise da fala diz respeito, liga, inevitavelmente, as duas faces do binómio, uma vez que lida tanto com a relação que existe entre as unidades e a sua pertinência na língua falada (i.e., os fonemas) como com a realidade física que resulta na pronunciação dessas mesmas unidades (i.e., os fones e alofones) (cf. definições dos termos em (Crystal, 2001)). Tem sido frequente a alternância, muitas vezes não claramente justificada, entre os termos fone e fonema nos vários estudos efetuados no âmbito do G2P (cf., a título de exemplo, que a unidade fone é a adotada em (Caseiro *e* Trancoso, 2002) enquanto que (Barros *e* Weiss, 2006) apresenta o fonema como o resultado da conversão do grafema). Neste estudo, consideramos trabalhar ao nível do fonema, uma vez que o procedimento de conversão adotado admite valências do contexto mais ou menos alargado no âmbito da unidade acentual (vulgo palavra), considerando a unidade para a qual o grafema é convertido como uma escolha significativa por entre todas as outras unidades que o sistema de língua coloca ao dispor. Assim, aceitamos a unidade fonológica, ou fonema, como uma classe à qual pode corresponder um fone ou um feixe de realizações alofónicas disponíveis no PE (acolhendo-se, assim, a possível inserção de pronúncias alternativas). A transcrição fonológica resultante corresponde ao PE que admitimos como padronizado e não representa qualquer arquifonema ou neutralização de oposições. A transcrição é registada entre barras oblíquas, fazendo uso do alfabeto SAMPA, conforme descrito supra.
De forma informal, verificou-se que o resultado da aplicação do modelo estatístico ao vocabulário

CETEMPúblico ("voc_CETEMP_40k") era já bastante preciso, apresentando, pontualmente, algumas incorreções.

Tabela 1 – Símbolos SAMPA e símbolos unicaráter (uc) associados a grafemas possíveis com vocábulos exemplificativos.

| SAMPA | uc | Grafemas possíveis | Exemplos |
|---|---|---|---|
| 6 | | a, e, â | c**a**ma, s**e**nha, c**â**mara |
| a | | a, á, à | p**á**, p**a**la |
| @ | | e | d**e** |
| e | | e, ê | v**ê**, d**e**do |
| E | | e, é, | p**é**, p**e**le |
| i | | i, í, e | v**i**, a**í**, r**e**al |
| o | | o, ô, ou | **o**co, av**ô**, l**ou**co |
| O | | o, ó | p**ó**, p**o**te |
| u | | u, ú, o | t**u**, ti**o**, at**o**, ba**ú** |
| 6~ | ã | ã, an, ân, em, am, e, âm, é | br**an**co, **ân**cora, c**am**po, t**em**, l**âm**pada, al**ém** |
| e~ | ë | ên, en, em | p**en**te, ag**ên**cia, **em**pate |
| i~ | ï | i, in, im, ím, ín, m | mu**i**to, tr**in**car, s**im**, **ím**pio, **ín**timo, hom**em** |
| o~ | õ | õ, ôn, ôm, on, om | p**õe**, c**ôn**sul, c**ôm**puto, p**on**te, p**om**bo |
| u~ | ü | u, ún, un, um, úm | m**ui**to, an**ún**cio, at**um,** c**úm**plice |
| b | | b | **b**e**b**er |
| d | | d | **d**a**d**o |
| g | | g, gu | **g**ato, **gu**elra |
| p | | p | **p**ato |
| t | | t | **t**oca |
| k | | qu, c | a**qu**ela, **c**asa |
| f | | f | **f**é |
| s | | s, ç, x, c[eiéí], ss | **s**ol, ca**ç**a, trou**x**e, **c**éu, **c**ima, a**ss**im |
| S | | ch, s, z, x | **ch**ave, pá**s**, pa**z**, **x**á |
| v | | v | **v**ida |
| z | | z, s, x | ca**s**a, **z**ebra, e**x**emplo |
| Z | | j, g, s, z, x | **j**á, **g**ira, de**s**viar, e**x-**bar |
| l | | l | **l**âmpada |
| L | | lh | ve**lh**o |
| r | | r | ca**r**o |
| R | | r, rr | ca**rr**o, **r**ato |
| m | | m | **m**ão |
| n | | n | **n**ada |
| J | | nh | se**nh**a |

Seguiu-se então um processo moroso de confirmação e correção manual das transcrições obtidas automaticamente. O passo seguinte consistiu em comparar as transcrições do dicionário com as transcrições geradas por um sintetizador de fala comercial. Esta comparação permitiu-nos confiar no nosso resultado já que, maioritariamente, as transcrições coincidiram. As transcrições que diferiram foram analisadas individualmente e corrigidas quando necessário, no sentido da representatividade do PE. Deste processo resultou o dicionário de transcrição fonológica que referenciamos como "dic_CETEMP_40k". Com este dicionário foi feito um novo modelo estatístico. O teste do modelo com o próprio dicionário de treino permitiu ainda corrigir alguns erros subjacentes, bem como uniformizar algumas transcrições. Por exemplo, os vocábulos iniciados por <ex-> são transcritos como /6iS/ (observando-se a inserção do iode) assim como em <extra> /"6iStr6/, mas não em <extenso> /@St"e~su/, não se transcrevendo a sequência <ex> como /6iS/ em certos contextos de atonicidade.

Tabela 2 – Símbolos unicaráter (uc) convencionados a partir de sequências SAMPA, exemplificados com vocábulos. Os primeiros 13 símbolos representam fonemas vocálicos em posição tónica; os restantes 7 representam sequências específicas de fonemas.

| SAMPA | uc | Exemplos |
|---|---|---|
| "6 | â | c**a**ma → /k**"6**m6/ → /k**â**m6/ |
| "a | á | c**a**sa → /k**"a**z6/ → /k**á**z6/ |
| "e | ê | t**e**ma → /t**"e**m6/ → /t**ê**m6/ |
| "E | É | s**e**te → /s**"E**t@/ → /s**É**t@/ |
| "i | í | t**i**ro → /t**"i**ru/ → /t**í**ru/ |
| "o | ô | **o**vo → /**"o**vu/ → /**ô**vu/ |
| "O | Ó | l**o**go → /l**"O**gu/ → /l**Ó**gu/ |
| "u | ú | **u**va → /**"u**v6/ → /**ú**v6/ |
| "6~ | Ã | c**am**po → /k**"6~**pu/ → /k**Ã**pu/ |
| "e~ | Ë | c**en**tro → /s**"e~**tru/ →/s**Ë**tru/ |
| "i~ | Ï | c**in**co →/s**"i~**ku/ →/s**Ï**ku/ |
| "o~ | Õ | c**on**to → /k**"o~**tu/ → /k**Õ**tu/ |
| "u~ | Ü | ass**un**to → /6s**"u~**tu/ → /as**Ü**tu/ |
| 6i | æ | **e**xtrair → /**6i**Str6"ir/ → /**æ**Str6ír/ |
| "6i | Æ | **e**xtra → /**"6i**Str6/ → /**Æ**Str6/ |
| 6~i~6~ | Ê | t**êm** → /t"**6~i~6~**i~/ → /t**Êï**/ |
| o~i~ | ¤ | p**õe**m → /p"**o~i~**6~i~/→ /p**¤ã**ï/ |
| ks | K | a**x**ila → /a**ks**"il6/ → /a**K**íl6/ |
| ai | Å | c**ae**m → /k"**ai**6~i~/ → /k**Å**ãï/ |
| Oi | ® | constr**oe**m→/ko~Str**Oi**6~i~/→ /kÖStr**®**ãï/ |

Ao longo do desenvolvimento deste trabalho, o dicionário sofreu um processo constante de revisão e de correção. Apesar de admitirmos a presença de alguns erros de transcrição, estamos confiantes na sua precisão, pelo que acreditamos que o dicionário "dic_CETEMP_40k" constitui uma base de trabalho interessante para estudos na língua portuguesa, em especial na área da fonética e da fonologia. Acrescente-se que do processo que acompanhou este estudo, resultou igualmente um

Tabela 3.1 – Possíveis fonemas associados a cada grafema para alinhamento – caso de vogais. A coluna "G" indica os grafemas e "F" os fonemas possíveis. A coluna "Alt" indica os fonemas alternativos, segundo a Tabela 2. A coluna à direita mostra exemplos de transcrição sem e com marcação de acentuação.

| G | F | Alt | Exemplos |
|---|---|---|---|
| a | 6 | â | c**a**ma → /k**6**m6/ → /k**â**m6/ |
| | a | á | m**a**la → /m**a**l6/ → /m**á**l6/ |
| | ã | Ã | c**a**nto → /k**ã**tu/ → /k**Ã**tu/ |
| á,à | a | á | **à**s → /**a**S/ → /**á**S/ |
| | | | p**á**s → /p**a**S/ → /p**á**S/ |
| ã | ã | Ã | an**ã**o → /6n**ã**ü/ → /6n**Ã**ü/ |
| â | ã | Ã | atl**â**ntico→/6tl**ã**tiku/→ /6tl**Ã**tiku/ |
| | 6 | â | c**â**mara → /k**6**m6r6/ → /k**â**m6r6/ |
| e | 6 | â | des**e**nho →/d@z**6**Ju/→ /d@z**â**Ju/ |
| | @,e | ê | aqu**e**le → /6k**e**l@/ → /6k**ê**l@/ |
| | E | É | p**e**le → /p**E**l@/ → /p**É**l@/ |
| | i | | ar**e**al → /6r**i**al/ → /6r**i**ál/ |
| | ë | Ë | v**e**nto → /vëtu/ → /vËtu/ |
| | ï | | visõ**e**s → /vizõïS/ → /vizÕïS/ |
| | 6i | æ | **e**xtrair → /**6i**Str6ir/ → /**æ**Str6ir/ |
| | 6i | Æ | **e**xtra → /**6i**Str6/ → /**Æ**Str6/ |
| | ã | Ã | t**e**ns → /tãïS/ → /tÃïS/ |
| | | | vot**e**m → /vOtãï/ → /vÓtãï/ |
| é | E | É | caf**é** → /k6f**E**/ → /k6f**É**/ |
| | ã | Ã | cont**é**m → /kõnt**ã**ï/ → /kõt**Ã**ï/ |
| ê | 6 | â | am**ê**ijoa→/6m**6**iZu6/→/6m**â**iZu6/ |
| | e | ê | voc**ê** → /vOs**e**/ → /vOs**ê**/ |
| | ë | Ë | p**ê**ndulo → /p**ë**dulo/ → /p**Ë**dulo/ |
| | 6i | Æ | **ê**xito → /**6i**zitu/ → /**Æ**zitu/ |
| | ãïã | Ê | t**ê**m → /tãïã**ï**/ → /tÊï/ |
| i,í | i | í | c**i**ma → /s**i**m6/ → /s**í**m6/ |
| | | | sa**í** → /s6**i**/ → /s6**í**/ |
| | ï | Ï | c**i**nco → /sïku/ → /s**Ï**ku/ |
| | | | l**í**mpida → /l**ï**pid6/ → /l**Ï**pid6/ |
| o | o | ô | c**o**rpo → /k**o**rpu/ → /k**ô**rpu/ |
| | O | Ó | c**o**po → /k**O**pu/ → /k**Ó**pu/ |
| | õ | Õ | c**o**nto → /k**õ**tu/ → /k**Õ**tu/ |
| | ü | | visã**o** → /vizã**ü**/ → /vizÃ**ü**/ |
| | u | ú | p**o**rque → /p**u**rk@/ → /p**ú**rk@/ |
| ó | O | Ó | c**ó**pia → /k**O**pi6/ → /k**Ó**pi6/ |
| ô | o | ô | av**ô** → /6v**o**/ → /6v**ô**/ |
| | õ | Õ | c**ô**nsul → /k**õ**sul/ → /k**Õ**sul/ |
| õ | õ | Õ | p**õ**e → /p**õ**ï/ → /p**Õ**ï/ |
| | õï | ¤ | p**õ**em → /p**õ**ïã**ï**/ → /p**¤**ãï/ |
| u,ú | u | Ú | ap**u**ra → /6p**u**r6/ → /6p**ú**r6/ |
| | | | t**ú**nel → /t**u**nEl/ → /t**ú**nEl/ |
| | ü | | **u**nto → /**ü**tu/ → /**Ü**tu/ |
| | | Ü | an**ú**ncio → /6n**ü**siu/ → /6n**Ü**siu/ |
| | | _ | q**u**ente → /këthough/ → /k_Ët@/ |

dicionário de pronunciação de cerca de 1300 estrangeirismos. Este dicionário de estrangeirismos será incorporado no sistema final de conversão G2P, como tabela de exceções.

Tabela 3.2 – Continuação da Tabela 3.1 – caso de consoantes.

| G | F | Alt | Exemplos |
|---|---|---|---|
| c | k | | **c**apa → /**k**ap6/ → /**k**áp6/ |
| | s | | **c**edo → /**s**edu/ → /**s**êdu/ |
| | ks | K | fi**cc**ional →/fi**ks**iunal/ → /fi**K**iunál/ |
| | | _ | a**c**tuar → /6tuar/ → /6_tuár/ (pré-AO) |
| g | g | | **g**atu → /**g**atu/ → /**g**átu/ |
| | Z | | **g**irafa → /**Z**iraf6/ → /**Z**iráf6/ |
| l | l | | **l**ado → /**l**adu/ → /**l**ádu/ |
| | L | | ma**l**ha → /ma**L**6/ → /má**L**_6/ |
| m | m | | **m**al → /**m**al/ → /**m**ál/ |
| | ü | | apela**m** → /6pElã**ü**/ → /6pÉlã**ü**/ |
| | ï | | vote**m** → /vOtã**ï**/ → /vÓtã**ï**/ |
| | | _ | ca**m**po → /kãpu/ → /kÃ_pu/ |
| n | n | | ca**n**a → /k6**n**6/ → /k**â**n6/ |
| | ï | | te**n**s → /tã**ï**S/ → /tÃ**ï**S/ |
| | J | | ma**n**ha →/m6**J**6/→/mâ**J**6/→/mâ**J**_6/ |
| | | _ | ca**n**to → /kãtu/ → /kÃ_tu/ |
| p | p | | **p**ar → /**p**ar/ → /**p**ár/ |
| | | _ | ó**p**tica → /Otik6/ → /Ó_tik6/ (pré-AO) |
| r | r | | ca**r**o → /ka**r**u/ → /ká**r**u/ |
| | R | _ | ca**rr**o → /ka**R**u/→ /ká**R**u/→ /ká**R**_u/ |
| s | s | _ | ma**s**sa → /ma**s**6/→ /má**s**6/→ /má**s**_6/ |
| | S | | à**s** → /a**S**/ → /á**S**/ |
| | z | | ca**s**a → /ka**z**6/ → /ká**z**6/ |
| | Z | | abi**s**mo → /6bi**Z**mu/ → /6bí**Z**mu/ |
| x | s | | má**x**imo → /ma**s**imu/ → /má**s**imu/ |
| | S | | **x**adrez → /**S**6dreS/ → /**S**6drêS/ |
| | z | | e**x**ame → /e**z**6m@/ → /e**z**âm@/ |
| | Z | | ex-diretor→/6i**Z**dirEtor/→/Æ**Z**dirEtôr/ |
| | ks | K | fi**x**o → /fi**ks**u/ → /fí**K**u/ |
| z | S | | arro**z** → /6Ro**S**/ → /6R_ô**S**/ |
| | z | | a**z**ar → /6**z**ar/ → /6**z**ár/ |
| | Z | | feli**z**mente→/f@li**Z**mët@/→ /f@li**Z**mËt@/ |

## 3.3 Alinhador de grafemas com fonemas

Um passo importante na criação do modelo estatístico, no qual cada grafema dá origem a zero ou a um fonema (vertente "1-01"; cf. 2.1, 1)), consiste num passo inicial de alinhamento entre grafemas e fonemas. A opção pelo modelo de "1-01" foi desde logo tomada pela verificação de que em apenas 7 casos um grafema pode dar origem a mais do que um fonema. Esses casos, e respetivos contextos de ocorrência, estão indicados na Tabela 2, nas últimas 7 linhas.

O problema aportado pelo alinhamento segundo a vertente "1-01" foi resolvido definindo símbolos que correspondessem a mais do que um fonema

(por exemplo, definimos o símbolo /Ê/ para representar /6~i~6~/ em "têm";cf. Tabela 2). Com esta solução, cada grafema pode dar origem a um fonema ou a zero fonemas (sendo este último o caso do <n> em <canto> → /k"6~_tu/ ou do <h> inicial em <homem> → /"_Om6~i~/, onde '_' representa o símbolo nulo). A cada fonema foi, então, associado um único símbolo, do conjunto de carateres ISO Latino (ISO-8859-1), tal como se apresentam nas Tabela 1 e Tabela 2.

O alinhamento entre grafemas e fonemas é então obtido usando o conhecido algoritmo de alinhamento entre cadeias de carateres (*edit distance* ou algoritmo de Levenshtein). Para tal, foi necessário definir uma distância entre cada grafema e cada fonema. Esta distância, ou custo de associação, foi definida através da equação

$$d(g,p) = -\log_2\left(P(p \mid g)\right), \qquad (10)$$

onde a probabilidade condicional do fonema *p* dado o grafema *g*, $P(p \mid g)$, é estimada a partir de um dicionário de transcrições alinhado. Definiu-se também um valor máximo para essa distância, $d_{\max}$, para os casos onde não existe qualquer associação entre grafema e fonema. Existe um apagamento de um grafema sempre que esse grafema não dá origem a um fonema e, para que isso aconteça, o apagamento tem de ter um custo menor que $d_{\max}$, admitindo ser preferível apagar um grafema a fazer uma associação errada.

As alternativas de transcrição para cada grafema estão documentadas na Tabela 3.1 e Tabela 3.2 (focando casos de vogais e de consoantes, respetivamente), tendo em conta o alinhamento individual entre grafemas e fonemas. Exemplos para cada alternativa estão também nelas apresentados.

Observando as Tabelas 3.1 e 3.2, pode verificar-se que os grafemas suscetíveis de serem apagados são <u|c|m|n|p|r|s>. Os grafemas <r> e <s> apenas podem sofrer apagamento quando o alinhamento é feito sem o uso da convenção de transformação de dígrafos, explicitados de seguida, em 3.4.1, e indicados na Tabela 4. Os grafemas <c> e <p> são apagados apenas quando se converte o vocabulário grafado na forma prévia à aplicação do AO.

## 3.4 Regras fonológicas

Apresentando o PE uma certa regularidade fonética e fonológica e uma ortografia de base fonológica, adicionamos ao módulo de G2P constrições linguísticas do PE pertinentes à tarefa de transcrever o grafema em fonema. Assim, foram propostos algoritmos baseados em regras fonológicas para a acentuação vocálica, reconhecendo o núcleo de sílaba tónica de cada vocábulo, e para a identificação da correspondência exata entre um grafema e respetivo fonema, de acordo com o contexto.

As regras resultam na definição de símbolos grafemáticos que as exprimem e que são introduzidos no modelo estatístico. Foram, assim, criados símbolos para dígrafos, vogais tónicas e grafemas em certos contextos fonológicos.

Tabela 4 – Grafemas especiais ("G") para dígrafos, associados a fonemas possíveis ("F") e a exemplos convencionados. Os primeiros 7 símbolos representam consoantes; os restantes 11 representam sequências específicas de vogal ditongada e de vogais nasais.

| G | Dígrafos | F | Exemplo |
|---|---|---|---|
| C | cc | s, ks | fi**cc**ional → fi**C**ional |
| Ç | cç | s, ks | fi**cç**ão → fi**Ç**ão |
| R | rr | R | ca**rr**o → ca**R**o |
| § | ss | S | ma**ss**a →ma**§**a |
| L | lh | L | mo**lh**o → mo**L**o |
| J | nh | J | u**nh**a → u**J**a |
| S | ch | S | **ch**ave → **S**ave |
| º | ou | o | d**ou**rada → dºrada |
| Ä | an, am | ã, Ã | c**an**to →cÄto, c**am**po →cÄpo |
| Ë | en, em | ë, Ë | s**en**te →sËte, s**em**pre → sËpre |
| Ï | in, im | ï, Ï | l**im**bo →lÏbo |
| Ö | on, om | õ, Õ | c**on**to→cÖto, |
| Ü | un, um | ü, Ü | ass**un**to → a§Üto |
| Â | ân, âm | Ã | p**ân**tano →pÂtano, |
| Ê | ên, êm | Ë | **ên**fase → Êfase |
| Í | ín, ím | Ï | **ín**dio→Índio, l**ím**pido→lÍpido |
| Ô | ôn, ôm | Õ | c**ôn**sul→cÔsul |
| Ú | ún, úm | Ü | den**ún**cia→denÚcia, c**úm**plice→cÚplice |

### 3.4.1 Dígrafos

Um dígrafo ocorre quando dois grafemas são pronunciados apenas por um único som. Na Tabela 4 apresentam-se símbolos para representar esses dígrafos.

A nossa proposta altera a representação dessas sequências de dois grafemas de forma a permitir uma associação ótima entre o símbolo grafado e o símbolo sonoro. Neste estudo foram consideradas como dígrafos sequências consonânticas e sequências vocálicas. No âmbito das sequências vocálicas, consideramos a sequência oral <ou>, a qual, seguindo a pronúncia padronizada do PE, corresponde ao fonema singular /o/, e as sequências

nasais <a|e|i|o|u> +<m|n>, admitidas em contexto silábico de <*V*+$C_{nasal}$> (cf. Tabela 4). Na medida em que o modelo implementado recebe igualmente informação sobre a vogal tónica (cf. 3.4.2), às vogais que no âmbito dos dígrafos apresentam acento gráfico foi-lhes igualmente atribuído um único símbolo (uni caráter).

### 3.4.2 Marcação de tonicidade

Seguindo os pressupostos teóricos discutidos em (Mateus e d'Andrade, 2000), admitimos tratar-se de uma tarefa de maior importância a marcação das vogais acentuadas, núcleos de sílaba, no âmbito de um vocábulo enquanto unidade acentual. A informação sobre a vogal tónica ($V_{tónica}$) tem sido reconhecida em trabalhos prévios de conversão de G2P, quer para a implementação de regras de transmutação do grafema em fone(ma), quer para a modulação de índices prosódicos (em especial se a informação for alargada à sílaba tónica). Sendo o contexto do "*n*-grama" fixo, curto e sem informação silábica, o conhecimento da $V_{tónica}$ traduziu-se num melhoramento ao modelo estatístico, uma vez que permitiu definir grafonemas de forma unívoca. Assim como em (Andrade *e* Viana, 1985), a nossa proposta considerou ser pertinente a marcação da $V_{tónica}$ (identificada com o símbolo SAMPA ' " ') e não da respetiva unidade silábica.

O processo de identificação de $V_{tónica}$ foi conseguido de uma forma que, tanto quanto nos é dado a perceber, não é usual noutros trabalhos. Atendendo ao contexto vocálico grafemático, de cada vocábulo, se alguma vogal (<*V*>) recebe um acento gráfico, essa <*V*> é identificada como $V_{tónica}$ (cf. Tabela 5, regra 1). Caso não apresente graficamente qualquer marca de tonicidade, analisamos a penúltima <*V*> nos vocábulos terminados em <a>, <o>, <e> ou <m> e nas correspondentes determinações de plural (cf. regra 2, Tabela 5). Excluindo os casos de presença de sequência ditongada, os quais são analisados à parte (regras 5 e 6, Tabela 5), essa <*V*> passa a ter a indicação de tonicidade (cf. regra 6, Tabela 5). Os restantes vocábulos (sem grafemas acentuados), recebem indicação de $V_{tónica}$ em posição oxítona (regras 3 e 4, Tabela 5). A aplicação das regras descritas são suficientes para não marcar tonicidade nos vocábulos com uma única <*V*> não acentuada graficamente, como é o caso: *i*) das preposições <com>, <de>, , <sem>, <sob> e das contrações <do(s)>, <no(s)>; *ii*) dos pronomes pessoais oblíquos <me>, <te>, <se>, <nos>, <vos>, <lhe(s)>, <o(s)> e <a(s)>, <lo(s)>, <no(s)>, <vo(s)> e das contrações <mo(s)>, <to(s)>, <lho(s)>; *iii*) do pronome relativo <que>; das conjunções <e>, <nem>, <que>, <se>; as quais se agregam frequentemente a um grupo de força acentual no âmbito do sintagma prosódico.

Tabela 5 - Regras para acentuação de vogais (<*V*>)

| | **Regra** | **Exemplo** |
|---|---|---|
| **1** | **Se** vocábulo apresenta alguma <*V*> acentuada graficamente, e**ntão** <*V*> → <$V_{tónica}$> [6]. | aux**í**lio, an**á**lise, avaliaç**ã**o, **à**s, t**ú**nel |
| **2** | **Se** vocábulo não apresenta acento gráfico e termina em <a>, <e> ou <o>, seguido ou não de <m\|n\|s>, e**ntão** <*V*> anterior a <a>, <e> ou <o> → <$V_{tónica}$>. | c**a**rta, d**a**nça, d**a**nçam, cont**e**nte(s), h**o**mem, h**o**mens, est**u**do(s) |
| **3** | **Se** vocábulo não apresenta acento gráfico e termina em <l>, <r>, <x> ou <z>, e**ntão** <*V*> anterior → <$V_{tónica}$>. | cant**a**r, emit**i**r, dev**e**r, can**a**l, pap**e**l, fun**i**l, cet**i**m, telef**a**x, dupl**e**x, cab**a**z |
| **4** | **Se** vocábulo não apresenta acento gráfico e termina em <i> ou <u>, seguidas ou não de <m\|n\|s>, e**ntão** <*V*> <i> ou <u> → <$V_{tónica}$>. | delf**i**m, bot**i**ns, par**i**s, alg**u**m, com**u**ns, jes**u**s |
| **5** | **Se** em 2, 3 e 4, a <*V*> <i> ou <u> é precedida de outra <*V*>, e**ntão** essa outra <*V*> → <$V_{tónica}$>. | p**a**i(s), r**e**i(s), l**e**u, m**a**u(s), decid**i**u, c**a**ixa(s),ad**e**us, p**e**ixe, p**a**uta(s) |
| **6** | **Se** em 5 a <*V*> <i> ou <u> é seguida de <ch>, <nh>, <m + *C*\|#> ou <n + *C*>, e**ntão** <*V*> <i> ou <u> → <$V_{tónica}$>. | sandu**i**che, vento**i**nha, ra**i**nha, amendo**i**m, co**i**mbra |

Um problema levanta-se quando nos confrontamos com vocábulos derivados morfologicamente, como é o caso dos advérbios de modo cuja terminação é <mente>, em especial quando a forma adjetival da qual derivam é marcada por um acento gráfico (exemplos: <rápido> → <rapidamente>, <dócil> → <docilmente>). O processo para a marcação da $V_{tónica}$ nos advérbios de modo terminados em <mente> passa pela seguinte solução: implementou-se um algoritmo que pesquisa os vocábulos com esse perfil e os divide em duas

[6] São exceções a esta regra, palavras como órfão(s), órfã(s), órgão(s), sótão(s), ímã(s), as quais, embora apresentem mais do que um acento gráfico, apenas têm uma sílaba tónica (em posição paroxítona).

partes (<*RAIZ*+mente>. A <*RAIZ*> passa por um módulo de tratamento específico, o qual apresenta uma lista de sequências grafemáticas em posição final, segundo padrões específicos, já com a determinação específica de $V_{tónica}$. Este método resolveu todos os casos presentes no vocabulário de referência, embora se admita que possam surgir casos remanescentes não resolvidos.

### 3.4.3 Regras para contextos frequentes

A descodificação da transmutação de grafema em fonema sem ambiguidade foi também auxiliada pela indicação de regras simples que atendem ao contexto grafemático. A título de exemplo, a determinação da sequência grafemática <al+*C*> resulta na notação de <a> em /a/ (em <**a**lmoçar> → /**a**lmus"ar/); a definição de <*V*+s+*V*> resulta na notação de <s> em /z/ (em <ca**s**a> → /k"a**z**6/). Foram ainda definidas outras regras para o <s> e para os grafemas <r>, <z>, <c>, <g> e <x>, inseridos em contextos mais restritos. Considerando um contexto mais alargado, na sequência grafemática <muit>, as <$V_{orais}$> <u> e <i> passam a /$V_{nasais}$/.

## *4. Resultados*

Todas as experiências foram baseadas no dicionário de pronunciação de 41586 vocábulos da língua portuguesa, descrito na subsecção 3.1. Aplicando diferentes formas de pré-processamento ao dicionário base, foram criados vários outros dicionários, nomeadamente:

1- Dicionário alinhado: nele apresenta-se a correspondência de "um-para-um" entre grafemas e fonemas. Todos os fonemas são representados com um único símbolo, incluindo as vogais com marcação de tónica (cf. Tabela 2 e Tabela 3). É introduzido um fonema especial (símbolo '_') para indicar o apagamento de um grafema, embora não existam inserções de fonemas (cf. subsecção 3.3).

2- Dicionário com símbolos para dígrafos: vocábulos em que os dígrafos são convertidos nos símbolos representados na Tabela 4. Tendo sido o objetivo da conversão de dígrafos num único símbolo facilitar a correspondência "um-para-um" entre grafemas e fonemas, este dicionário é alinhado.

3- Dicionário com acentuação: presença da marcação da vogal tónica em cada pronunciação.

4- Dicionário com acentuação e com símbolos para dígrafos: composição dos dois anteriores, usando alinhamento "um-para-um" entre grafemas e fonemas.

No total são tomados 5 dicionários, os 4 descritos mais o dicionário base. Estes dicionários estão disponibilizados com os seguintes nomes:
- dic_CETEMP_40k;
- dic_CETEMP_40k_alinhado;
- dic_CETEMP_40k_acentuado;
- dic_CETEMP_40k_alinhado_digrafos;
- dic_CETEMP_40k_acentuado_alinhado_digraf.

Para testar o modelo estatístico, cada um destes dicionários foi particionado em 5 dicionários de treino e 5 dicionários de teste, de forma rotativa. O dicionário inicial foi dividido em 5 partes, cada uma com 20% dos vocábulos (8317), escolhidos de forma aleatória. Os vocábulos foram mutuamente exclusivos em cada uma das 5 partes. Cada uma das partes deu origem a um dicionário de teste e os restantes 4 partes (33269 vocábulos) a um dicionário de treino. A rotação das partes deu origem a 5 ciclos de treino e teste dos modelos estatísticos para validação cruzada. Os resultados indicados correspondem à média dos 5 resultados parciais.

O desempenho do sistema de conversão de grafemas para fonemas é expresso em duas taxas médias de erros de conversão verificados nos dicionários de teste: taxa média de erro de fonemas (PER – "phoneme error rate") e taxa média de erro de vocábulos (WER – "word error rate"). A Tabela 6 sumariza os resultados obtidos usando "*n*-grama" entre 2 e 8 e utilizando o dicionário alinhado (dic_CETEMP_40k_alinhado), enquanto a Tabela 7 sumariza os resultados obtidos usando regras fonológicas.

Os gráficos da Figura 1 e da Figura 2 ilustram o contributo de cada etapa de pré-processamento fonológico no desempenho do sistema de conversão, apresentando as percentagens da taxa de erro de conversão de vocábulos. Como se pode observar, a marcação da vogal tónica é o processamento que mais contribui para o melhoramento do desempenho do sistema.

Tabela 6 – Resultados com modelo base (sem regras fonológicas)

| *n*-grama | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| WER (%) | 35.1 | 15.5 | 7.90 | 5.96 | 6.08 | 6.51 | 7.01 |
| PER(%) | 4.69 | 1.86 | 0.95 | 0.72 | 0.74 | 0.79 | 0.86 |

Tabela 7 – Resultados com modelo base (com todas as regras fonológicas)

| *n*-grama | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| WER (%) | 9.73 | 4.70 | 2.60 | 2.31 | 2.42 | 2.58 | 2.82 |
| PER (%) | 1.27 | 0.60 | 0.33 | 0.30 | 0.31 | 0.33 | 0.36 |

É de notar que, ao contrário do que se poderia supor, a utilização de "*n*-grama" com grandes contextos (*n* maior que 5) não aumenta o desempenho, verificando-se, ao invés, um ligeiro aumento das taxas de erros. Isto pode ser explicado pela falta de amostras suficientes para estimar convenientemente "*n*-grama" com grandes contextos. Verifica-se, pois, que o valor ideal de *n* fica dependente da dimensão da base de dados usada para treino.

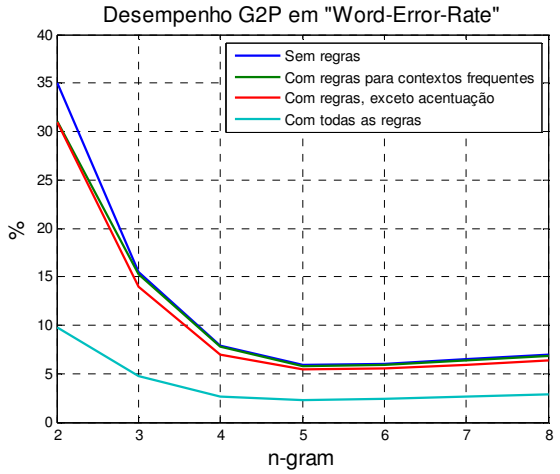



Figura 1 – Taxas de erro de palavras em função do comprimento do "*n*-grama" e da inclusão de regras fonológicas.

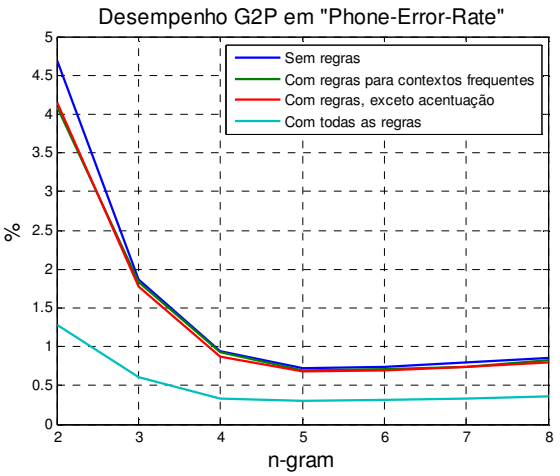



Figura 2 – Taxas de erro de fonemas em função do comprimento do "*n*-grama" e da inclusão de regras fonológicas.

O desempenho do sistema para o caso do dicionário pós-AO está representado na Figura 3 e na Figura 4. Observando os dados, é possível afirmar que o desempenho com este novo dicionário é ligeiramente inferior ao desempenho dos modelos anteriores ao AO (pré-AO), em todas as combinações de pré-processamento. Analisando os erros dos modelos com os melhores desempenhos (5-grama com marcação da vogal tónica e demais regras) observa-se que a soma dos erros dos 5 conjuntos de validação cruzada dos modelos pré-AO totaliza 958, enquanto nos modelos pós-AO totaliza 1022. Deve ainda ser referido que os dois modelos partilham 644 erros (resultantes de vocábulos que foram mal convertidas em ambos os modelos, especialmente no que diz respeito à conversão de <e> em /Ele/ e de <o> em /Olo/), sendo 313 e 378 os erros em vocábulos diferentes provocados pelos modelos pré- e pós-AO, respetivamente.

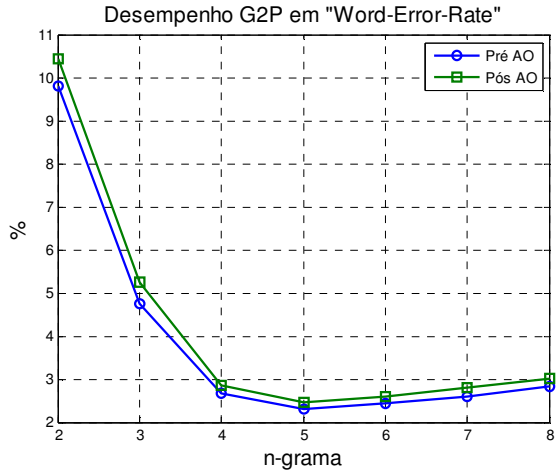



Figura 3 – Comparação do desempenho dos modelos com os dicionários pré- e pós-AO em termos de erros de vocábulos.

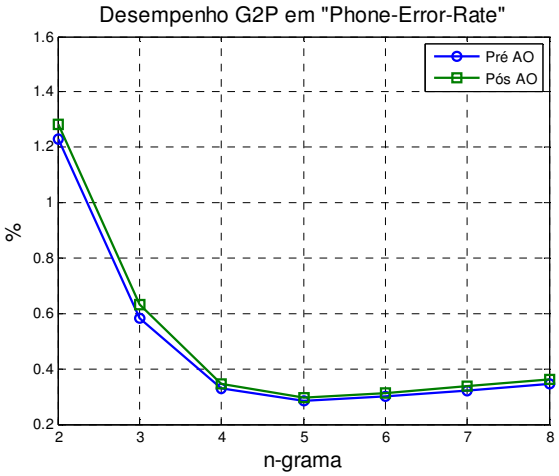



Figura 4 – Comparação do desempenho dos modelos com os dicionários pré- e pós-AO em termos de erros de fonemas.

Mais de 80% dos erros são provocados pela confusão gerada pelo grau de abertura dos fonemas que devem corresponder aos grafemas <o>, <e> e <a>, os quais podem ser pronunciados como /o/ ou /O/, /e/ ou /E/, /6/ ou /a/. Nos modelos pós-AO, não existem erros provocados pelas consoantes mudas <c> e <p>, tendo-se verificado um acréscimo expressivo de erros provocados pela conversão do grafema <a>. Tal facto pode justificar o decréscimo do desempenho dos modelos pós-AO evidenciando

o papel das consoantes mudas, indicativo, muitas das vezes, da abertura da vogal a elas antecedente e auxiliador, em muitos dos casos, da desambiguação entre pronúncia aberta e pronúncia fechada. A supressão da acentuação gráfica em alguns vocábulos, como é exemplo <boia> (<bóia>, pré-AO), contribui também para aumentar a ambiguidade na aprendizagem e na determinação da pronunciação. A supressão do hífen usado no processo de prefixação veio igualmente dificultar a determinação da vogal tónica, sendo este um pré-processamento que, como vimos, contribui muito para o aumento do desempenho dos modelos.

Em termos de implementação, o conversor de grafema para fonema utiliza o dicionário como uma tabela de exceções que é carregado para uma tabela "*hash*" ("*hash table*"). A conversão com o modelo estatístico só é evocada para vocábulos que não constam no dicionário. Esta implementação requer mais recursos de memória mas, por outro lado, é muito mais célere e mais precisa. Uma tabela "*hash*" com 100k entradas para o dicionário base de 40k vocábulos necessita de cerca de 2MB de memória, apresentando cerca de 7500 colisões.

O ritmo de conversão é de cerca de 1M vocábulos por segundo, contra 20k vocábulos por segundo usando a conversão com um modelo estatístico de 2-grama (num PC "*quad core*" a 2,8GHz).

## *5. Conclusões e trabalho futuro*

Neste trabalho pretendemos mostrar uma nova abordagem na tarefa de converter grafemas em fonemas em português europeu. Propomos um modelo de base estatística, imbuído de regras fonológicas. Sequências de grafonemas foram modeladas através de um algoritmo de alinhamento entre grafemas e fonemas, nas quais foram também consideradas informações advenientes do contexto fonológico da língua portuguesa, tais como a digrafia, a acentuação tónica e a vizinhança fonético-fonológica. Todas estas informações foram testadas individualmente, tendo-se verificado que a inclusão de informação sobre a tonicidade da vogal foi decisiva para o aumento do desempenho do conversor. Contrariamente, a inclusão de informação sobre dígrafos não trouxe benefícios acentuados.

Os modelos de "*n*-grama" foram treinados e testados usando a grafia pré-AO e pós-AO, tendo-se verificado um ligeiro, mas consistente, decréscimo de desempenho dos modelos pós-AO.

Decorrente da tarefa de conversão, foi gerado um dicionário de pronunciação com mais de 40 mil vocábulos oriundos do corpus CETEMPúblico, do qual derivaram outros dicionários, com informação de alinhamento, de acentuação e de dígrafos.

Os diferentes dicionários, bem como os modelos de "*n*-grama" estão livremente disponíveis em (SPL, 2011). O dicionário de estrangeirismos e o dicionário de múltipla pronunciação de homógrafos serão incluídos no sistema, a breve prazo. A pronunciação de adjetivos, de verbos e de nomes flexionados encontra-se em estudo, também com o objetivo de vir a integrar o sistema.

***Agradecimentos***

Os autores agradecem o contributo dos revisores deste artigo, pelas sugestões e comentários apresentados. Agradecem igualmente ao Instituto de Telecomunicações e à FCT as bolsas de doutoramento (Arlindo Veiga) e de pós-doutoramento (Sara Candeias, SFRH/BPD/ 36584/2007). Este trabalho recebeu ainda o apoio financeiro do projeto FCT - PTDC/CLE-LIN/11 2411/2009.

## *Referências*

Almeida, J. J.; Simões, A. 2001. Text to Speech – "A Rewriting System Approach". *Procesamiento del Lenguaje Natural*, 27, pp. 247–255.

Andrade, E.; Viana, M. C. 1985. Corso I - Um Conversor de Texto Ortográfico em Código Fonético para o Português. *Technical report*, CLUL-INIC, Lisboa.

Barros, M. J.; Weiss, C. 2006. Maximum Entropy Motivated Grapheme-To-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech, *IV Jornadas en Tecnologías del Habla*, pp. 177-182. Zaragoza, España.

Bisani, M.; Ney, H. 2008. Joint-Sequence Models for Grapheme-To-Phoneme Conversion, *Speech Communication*, vol. 50 (5), pp. 434–451.

Bisani, M.; Ney, H. 2002. Investigations on Joint-Multigram Models for Grapheme-to-Phoneme Conversion, *Proc. 7th International Conference on Spoken Language Processing (ICSLP'02)*, pp. 105–108. Denver, USA.

Braga, D.; Coelho, L.; Resende Jr., F. 2006. A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese, *VI Int. Telecommunications Symposium*, pp. 328-333. Fortaleza-CE, Brazil.

Braga, D.; Marques, M. A. 2007. Desambiguação de Homógrafos para Sistemas de Conversão Texto-Fala em Português, *Diacrítica*, 21.1 Série Ciências da Linguagem, pp. 25-50. Braga, CEHUM/Universidade do Minho.

Candeias, S.; Perdigão, F. 2008. Conversor de Grafemas para Fones Baseado em Regras para Português, Costa, L.; Santos, D.; Cardoso, N. (Eds.). *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, cap. 14. Linguateca. Lisboa.

Caseiro, D. A.; Trancoso, I. 2002. Grapheme-to-Phone Using Finite-State Transducers, *Pro. 2002 IEEE Workshop on Speech Synthesis*, USA.

Chen, S.; Goodman, J. 1998. An Empirical Study of Smoothing Techniques for Language Modeling, *Tech. Report TR–10–98*. Center for Research in Comp.Tech., Harvard Univ.

Crystal, D. 2001. *A Dictionary of Linguistics and Phonetics*. Blackwell, Oxford.

Demberg, V.; Schmid, H.; Möhler, G. 2007. Phonological Constraints and Morphological Preprocessing for Grapheme-to-phoneme Conversion, *Proc. 45th Annual Meeting of the Association for Computational Linguistics* (ACL-07), pp. 96-103. Prague, Czech Republic.

Demberg, V. 2006. *Letter-to-Phoneme Conversion for a German Text-to-Speech System*. PhD Thesis. Stuttgart University, Germany,

Galescu, L.; Allen, J. 2001. Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model, *Proc. 4th ISCA Workshop on Speech Synthesis*, Scotland.

Good, I. 1953. The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, vol. 40 (3,4), 237-264.

Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Jiampojamarn, S.; Kondrak, G.; Sherif, T. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion, *HLT-NAACL*, pp. 372-379. Rochester, New York.

Jiampojamarn, S.; Kondrak, G. 2009. Online Discriminative Training for Grapheme-to-Phoneme Conversion, *Proc. INTERSPEECH*, pp. 1303-1306, Brighton, UK.

Katz, S. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, *IEEE Trans. Acoustics, Speech and Signal Processing*, 35(3), 400-401.

Kneser, R.; Ney, H. 1995. Improved Backing-Off for M-gram Language Modeling, *Proc. IEEE ICASSP*, vol. 1, pp. 181–184.

Lince. Lince - Conversor para a Nova Ortografia, http://www.portaldalinguaportuguesa.org/lince.php

Mateus, M. H.; d'Andrade, E. 2000. *The Phonology of Portuguese*. Oxford University Press.

Ney, Hermann; Essen, Ute; Kneser, Reinhard. 1994. On Structuring Probabilistic Dependences in Stochastic Language Modelling, *Computer Speech and Language*, vol. 8 (1), pp. 1-38.

Oliveira, C.; Moutinho, L.; Teixeira, A. 2004. Um Novo Sistema de Conversão Grafema-Fone para PE Baseado em Transdutores, *Actas II Congresso Int. Fonética e Fonologia*, Brasil.

Oliveira, L.; Viana, M. C.; Mata, A. I.; Trancoso, I. 2001. *Progress Report of Project Dixi+: A Portuguese Text-to-Speech Synthesizer for Alternative and Augmentative Communication*. Technical Report, FCT.

Oliveira, L.; Viana, M. C.; Trancoso; I. 1992. A Rule-Based Text-to-Speech System for Portuguese, *Proc. ICASSP'92*, San Francisco, USA.

Santos, D.; Rocha, P. 2001. Evaluating CETEMPúblico, a Free Resource for Portuguese, *Proc. 39th Annual Meeting of the Association for Computational Linguistics*, pp.442-449. Toulouse, France.

SpeechDat. Databases for the Creation of Voice Driven Teleservices, http://www.speechdat.org/SpeechDat.html

SPL, 2011. Material disponibilizado no âmbito deste artigo, http://lsi.co.it.pt/spl/resources.htm

Taylor, P. 2005. Hidden Markov Models for Grapheme to Phoneme Conversion, *Proc. INTERSPEECH*, pp. 1973-1976, Lisbon, Portugal.

Teixeira, A.; Oliveira, C., Moutinho, L., 2006. On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme-Phone Conversion, *Proc. PROPOR'2006*, pp. 212-215.

Teixeira, J. P. 2004. *A Prosody Model to TTS Systems*. PhD Thesis, Faculdade de Engenharia da Universidade do Porto.

Teixeira, J. P.; Freitas, D. 1998. MULTIVOX-Conversor Texto-Fala para Português, *Proc. PROPOR'98*, Porto Alegre, Brasil

Trancoso, I.; Viana, M. C.; Silva, F.; Marques, G.; Oliveira, L. 1994. Rule-based vs. Neural Network Based Approaches to Letter-to-Phone Conversion for Portuguese Common and Proper Names", *Proc. ICSLP'94*, Yokohama, Japan. pp. 1767-1770.

Veiga, A.; Candeias, S.; Perdigão, F. 2011. Generating a Pronunciation Dictionary for European Portuguese Using a Joint-Sequence Model with Embedded Stress Assignment, *Proc. Brazilian Symposium in Information and Human Language Technology - STIL*, Cuiabá, Brazil, pp. 144 – 153.

Wells, J. C. 1997. SAMPA Computer Readable Phonetic Alphabet. Gibbon, D., Moore, R. and Winski, R. (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*, part IV. Mouton de Gruyter, Berlin and New York.

Witten, I.; Bell, T. 1991. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression, *IEEE Trans. Information Theory*, vol. 37 (4), pp. 1085-1094.

# Novas Perspectivas

# Criação e Acesso a Informação Semântica Aplicada ao Governo Eletrónico

Mário Rodrigues
Universidade de Aveiro
mjfr@ua.pt

Gonçalo Paiva Dias
Universidade de Aveiro
gpd@ua.pt

António Teixeira
Universidade de Aveiro
ajst@ua.pt

**Resumo**

Os cidadãos, empresas ou serviços públicos - os clientes - que procuram informações no contexto do Governo Eletrónico visam obter respostas objetivas às suas questões. Para isso é necessário que os sistemas de pesquisa consigam manipular a informação de modo a que seja disponibilizada de uma forma eficaz e adequada às necessidades de cada cliente. Uma vez que grande parte dos documentos do governo estão escritos em formatos não estruturados e em linguagem natural, é necessário desenvolver métodos para obter e estruturar este tipo de informação. A alternativa seria indexar pelo seu texto a grande quantidade de documentos existente, uma solução desadequada no contexto do Governo Eletrónico, uma vez que assim seriam retornados frequentemente muitos resultados a cada pesquisa.

Este artigo apresenta um primeiro protótipo de uma aplicação que gera informação semântica a partir de textos escritos em Português. A informação semântica gerada corresponde a um domínio de conhecimento definido por um operador humano através de uma interface gráfica, de modo a que o sistema seja adaptável às diferentes áreas de atuação do Governo Eletrónico. O conteúdo é acessível através de uma interface em linguagem natural e através de uma interface de pesquisa que aceita entradas SPARQL. Deste modo é possível aos clientes aceder diretamente ou integrar este sistema com os seus próprios sistemas de informação. A aplicação está organizada em três grandes módulos: Representação do Conhecimento que permite definir domínio de conhecimento e sua semântica e criar exemplos semente, nos textos, de conceitos do domínio de conhecimento; Processamento de Linguagem Natural que permite obter estruturas sintáticas associadas às frases em linguagem natural; e Extração e Integração Semântica que utiliza os exemplos semente para treinar classificadores estatísticos a identificar nas estruturas sintáticas os conceitos do domínio de conhecimento, que utiliza os classificadores treinados para detetar esses conceitos em estruturas sintáticas de novas frases, e que contém as interfaces para pessoas e máquinas.

Neste artigo apresentamos igualmente exemplos ilustrativos da utilização do sistema e os resultados de uma primeira avaliação de desempenho. O sistema funciona para o Português e foi construído reutilizando software do estado da arte, maioritariamente desenvolvido visando o Inglês. A sua modularidade permite alterar a língua base do sistema, de Português para outra, alterando o módulo de Processamento de Linguagem Natural e sem ser necessário alterar os restantes módulos da aplicação.

## *1 Introdução*

O Governo Eletrónico (e-gov) é uma expressão utilizada para descrever a utilização das Tecnologias da Informação e da Comunicação (TIC) no âmbito do governo e da administração pública. Refere-se a vários conceitos alternativos ou complementares, incluindo o uso das TIC para tornar mais fácil, mais rápido e mais barato o acesso a informação e a serviços aos clientes do governo: cidadãos, empresas, e outros organismos governamentais (Layne e Lee, 2001).

Os órgãos de governo e da administração pública produzem grandes quantidades de informação sob a forma de leis, regulamentos, editais, atas, etc. Estes documentos são normalmente escritos em linguagem natural (Português, Castelhano, etc.) em texto livre, sem uma estrutura em meta linguagem que indique qual o significado das diferentes partes do documento. Mesmo que estes documentos sejam armazenados em computadores, o seu formato de texto livre dificulta a manipulação automática da informação neles contida de modo a ir de encontro às necessidades específicas dos clientes do governo. Frequente são retornados muitos resultados às pesquisas efetuadas com vista a que a informação relevante esteja no conjunto de resposta. Por exemplo, se a procura for "processo de obra Maria", normalmente são devolvidos todos os documentos que contenham (pelo menos) uma das palavras, ordenados pela maior semelhança com a procura, e não apenas aquele(s) que contenha(m) informação acerca de processos de obra aplicados por cidadãs chamadas Maria.

Este comportamento é adequado em sistemas de informação genéricos como os motores de pesquisa da Internet. O mesmo já não acontece quando o contexto é o e-gov. Quando cidadãos, empresas ou serviços públicos procuram informações no contexto do e-gov querem obter respostas às suas questões e não uma lista de documentos acerca de tópicos relacionados. Por outro lado, como o governo tem servir a totalidade da população, incluindo cidadãos com poucos conhecimentos de TIC ou acerca dos processos do governo, as respostas devem ser curtas, claras e concisas de modo a evitar dificuldades na sua localização ou interpretação. Além disso, as respostas devem ser textos criados com base nos documentos oficiais.

É por isso importante desenvolver aplicações e tecnologias que permitam um acesso fácil à informação disponibilizada pelos órgãos de governação e administração. Isso implica a utilização de tecnologias que permitam perceber e manipular o conteúdo de documentos escritos em linguagem natural. O e-gov beneficiaria da existência de sistemas capazes de organizar e integrar diversas fontes de informação e capazes de compreender documentos escritos em linguagens naturais tais como o Português (Rodrigues, Paiva Dias e Teixeira, 2010a).

Em virtude disto temos vindo a desenvolver um sistema que utiliza tecnologias de Processamento de Linguagem Natural (PLN) para interpretar o conteúdo dos documentos, e tecnologias de Representação do Conhecimento (RC) para organizar e manipular o conteúdo obtido. O sistema permite definir os tipos de conteúdo que serão procurados e armazenados, permite a integração de informação relevante de outras fontes, e permite o acesso à informação de diversas formas incluindo perguntas em linguagem natural, por referência geográfica ou através de normas da web semântica.

Focámos a aplicação na disponibilização de informação municipal apesar do sistema poder ser utilizado com diversos tipos de informação. A importância dos municípios reside no fato de serem muitas vezes o ponto mais próximo de serviço para os cidadãos e empresas. São também interessantes devido a integrarem, numa única organização, decisão política e execução administrativa (Paiva Dias, 2006).

Neste artigo apresentamos um sistema capaz de gerar e disponibilizar informação semântica a partir de documentos não estruturados escritos em linguagem natural. A próxima subsecção apresenta trabalho relacionado. A secção 2 descreve detalhadamente a concepção e desenvolvimento do sistema. A secção 3 apresenta os exemplos de utilização e a avaliação de desempenho. O artigo termina, na secção 4, com as respetivas conclusões.

### 1.1 Trabalho Relacionado

A atividade de investigação em e-gov tem sido geralmente centrada na resolução de problemas como a integração e interoperabilidade de serviços, que são problemas muito importantes e devem continuar a ser estudados. Em tais projetos é geralmente considerado que a informação está no sistema, quer tenha sido colocada manualmente ou utilizando bases de dados existentes, como por exemplo o OneStopGov (Chatzidimitriou e Koumpis, 2008) e o Acess-eGov (Sroga, 2008). Tanto quanto sabemos, até hoje nenhum projeto foi dedicado ao problema da aquisição automática de informações a partir de documentos do governo em linguagem natural, quer para Português quer para outras línguas.

Relativamente à extração de informação, vários projetos foram dedicados à tarefa de extração de informação escalável e independente do domínio. DBPedia (Bizer et al., 2009) é uma base de conhecimento criada pela extração de informação das caixas de informações da Wikipedia e utilizou a estrutura da Wikipedia para inferir a semântica. Uma abordagem semelhante foi seguida para criar a base de conhecimento Yago (Suchanek, Kasneci e Weikum, 2007). Além da Wikipedia, o YAGO também utiliza um conjunto de regras para melhorar a precisão da extracção de informação e a WordNet para desambiguar os significados das palavras. Estas bases de conhecimento foram criados sem qualquer processamento de linguagem natural.

O sistema Kylin (Wu, Hoffmann e Weld, 2008) usa informações das caixas de informação da Wikipedia para treinar classificadores estatísticos que mais tarde são usados para extrair informações a partir de textos de linguagem natural. Os textos são analisados pelas suas etiquetas morfo-sintáticas e caraterísticas de superfície (posição das palavras na frase, a capitalização, a presença de dígitos ou carateres especiais, etc.) Não usa informação sintática.

Outros sistemas do estado da arte não utilizam a Wikipedia como fonte de conhecimento. O TextRunner (Banko et al., 2007) pretende extrair todas as instâncias de todas as relações significativas a partir de páginas web. Constrói a sua ontologia a partir do corpus sem controlar se as relações ontológicas estão bem definidas e sem desambiguar as entidades. O KnowItAll (Etzioni et al., 2004) utiliza exemplos especificados manu-

almente que expressam um conjunto de relações, por exemplo `amigo(João,Pedro)`. Esses exemplos são utilizados para obter padrões textuais que podem expressar as relações, por exemplo "o João é amigo do Pedro". Os padrões textuais são usados para treinar um conjunto de informações pré-definidas.

O sistema Leila (Suchanek, Ifrim e Weikum, 2006) aperfeiçoou o método do KnowItAll usando tanto exemplos e contra-exemplos como sementes, a fim de gerar padrões mais robustos, e usando análise sintática para gerar padrões de extração de informações. A maior robustez dos padrões conjugada com uma análise sintática que permite capturar informações em frases mais complexas foram as pricipais razões para que esta abordagem fosse adoptada no nosso sistema.

Relativamente a interfaces em linguagem natural escrita o que tem sido estudado é, essencialmente, como mapear frases em linguagem natural para os esquemas de armazenamento de informação. Um dos sistemas relevantes é o NaLIX, uma interface para bases de dados XML que aceita frases arbitrárias em Inglês. Esta interface traduz as pesquisas em expressões XQuery e, por exemplo, é possível consultar uma base de dados acerca de filmes com frases do tipo "find the title of publications with more that 5 authors" que traduz para: encontra os títulos de obras com mais de 5 autores (Li, Yang e Jagadish, 2005).

O Panto é outra interface em linguagem natural escrita que aceita consultas genéricas em linguagem natural, produzindo como saída consultas *Simple Protocol and RDF Query Language* (SPARQL), que é atualmente a linguagem padrão para acesso de dados da Web semântica. Foi concebido para ser aplicável a qualquer ontologia não pressupõe nada acerca do domínio do conhecimento. Os seus autores argumentam que obtém bons resultados e que ajuda a fazer a ponte entre a lógica da web semântica e os utilizadores (Wang et al., 2007).

O ESTER é um sistema modular que conjuga pesquisas de texto completo e pesquisas em ontologia. Responde a consultas SPARQL básicas reduzindo-as a um pequeno número de duas operações básicas: pesquisa de prefixo e junção. Suporta uma mistura de consultas semânticas com consultas de texto normais e sugere ao utilizador possíveis interpretações semânticas da consulta (Bast et al., 2007).

## *2 Sistema Desenvolvido*

O sistema desenvolvido está organizado conforme o modelo conceptual apresentado na Figura 1. O modelo separa claramente o domínio da linguagem natural do domínio da representação do conhecimento e está organizado em três componentes:

- Representação do Conhecimento - componente que contém ferramentas para definir a semântica do sistema - através de uma ontologia representada em *Web Ontology Language Description Logic* (OWL-DL) - e para permitir a operadores humanos adicionar exemplos de correspondência entre essa semântica e elementos presentes nos textos;
- Processamento de Linguagem Natural - componente baseado em tecnologias da área de PLN que inclui tecnologias de processamento de informação para obter estruturas sintáticas que representam as frases encontradas nos textos em linguagem natural. Conforme os exemplos definidos na RC, algumas destas estruturas serão associadas a elementos da ontologia;
- Extração e Integração Semântica - componente que aprende as associações entre as estruturas sintáticas e a ontologia e aplica-as a novos textos para obter novas informações. Este componente pode complementar a informação contida nos textos com fontes estruturadas de informação, como por exemplo coordenadas geográficas dos locais via Google Maps API e organização política do território via Geo-Net-PT01 (Chaves, Silva e Martins, 2005). Inclui ainda interfaces de acesso aos dados.

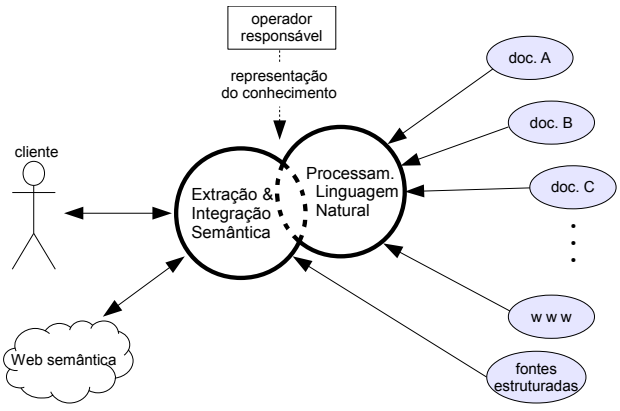



Figura 1: O modelo conceptual. A informação semântica é extraída das estruturas provenientes do PLN conforme definido pela RC definida pelo operador responsável pelo sistema. As setas unidirecionais representam aquisição do conteúdo e as bidirecionais representam as interfaces.

O resultado é informação semântica que pode ser consultada e acedida em vez, ou em complemento, dos documentos originais (ver Figura 1).

O sistema foi construído reutilizando software de código aberto - algum adaptado para trabalhar com Português - para tirar vantagem do estado da arte em termos de abordagens e ferramentas existentes. Foi desenvolvido software especifico para integrar o software reutilizado num sistema coerente. A arquitetura da instanciação do modelo conceptual está representada na Figura 2 e será descrita em mais detalhe nas subsecções que se seguem.

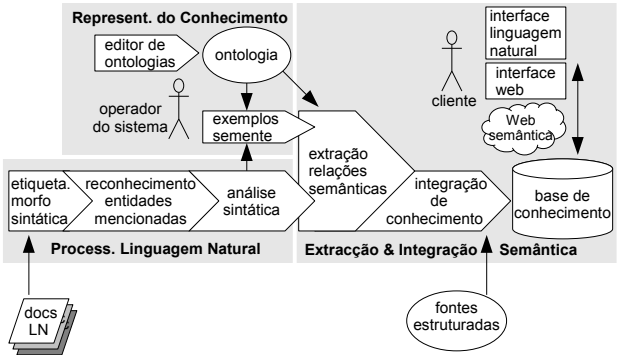



Figura 2: Instanciação do modelo conceptual. Os três grandes módulos são delimitados pelo sombreado. A Representação do Conhecimento define a semântica do sistema e fornece exemplos semente dos conceitos nos textos. O módulo de Processamento de Linguagem Natural enriquece o texto com etiquetas morfo-sintáticas, entidades mencionadas e estruturas sintáticas. O módulo de Extração e Integração Semântica treina modelos de extracção com base nos exemplos semente, aplica-os em todos os textos, integra outras fontes de informação e disponibiliza a informação aos clientes.

## 2.1 Representação do Conhecimento

O primeiro passo para construir uma representação do conhecimento é definir uma estrutura que represente conceitos de um domínio e respetivas relações. Para isso utilizámos ontologias que formalmente são definidas como "*a formal, explicit specification of a shared conceptualisation*", o que traduz para: especificação explícita e formal de uma conceptualização partilhada (Gruber, 1993). Ser "explícita" implica que todos os conceitos usados e respetivas restrições têm de estar definidos explicitamente e ser "formal" refere-se a ter de ser legível para máquinas. Uma "conceptualização" é um modelo abstrato de que representa um domínio, identificando conceitos e relações relevantes a essa parte do mundo. Ser "partilhada" é importante porque uma ontologia deverá servir para partilhar conhecimento e por isso deve ser aceite por um grupo ao invés de ficar restrita a um indivíduo.

As ontologias permitem representar um conjunto de conceitos pertencentes a um domínio, bem como as relações existentes entre esses conceitos. O fato de ser uma especificação formal bem definida permitiu o desenvolvimento de ferramentas de *software* que inferem novos fatos através de implicações lógicas acerca dos dados já conhecidos. Na nossa aplicação as ontologias são criadas e/ou editadas usando o Protégé (versão 4). Para o domínio do e-gov, a ontologia criada inclui as ontologias Friend-of-a-Friend (FOAF) (Brickley e Miller, 2010), Dublin Core (Weibel et al., 2007), World Geodetic System revisão de 1984 (National Imagery and Mapping Agency, 2000), e GeoNames versão integral (GeoNames, 2010). Inclui também classes especificamente criadas para lidar com assuntos relativos aos municípios. Foi criada uma classe denominada *Assunto_executivo* que é subclasse da classe de nível superior *Thing* e que possui sete subclasses (ver Figura 3). As subclasses de *Assunto_executivo* e respetiva descrição encontram-se na Tabela 1.

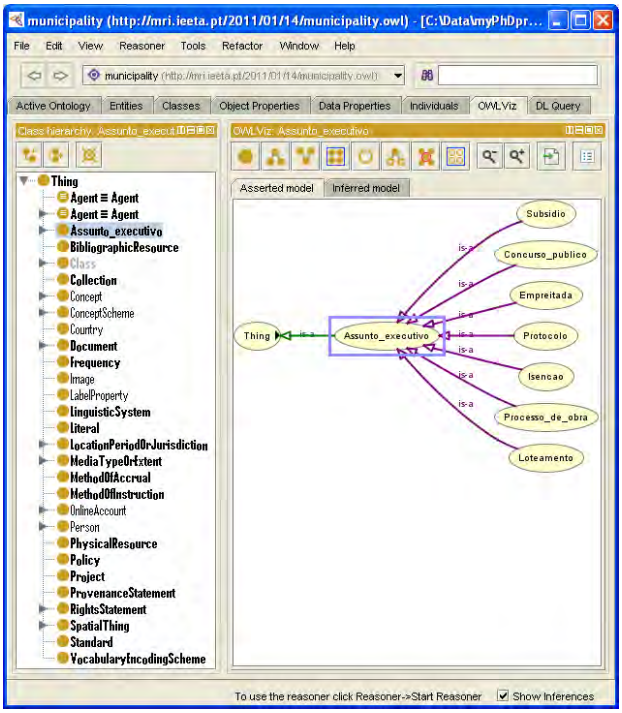



Figura 3: A interface de criação da ontologia. No painel da esquerda está a lista de todas as subclasses de *Thing*. No painel direita encontra-se uma representação gráfica da classe *Assunto_executivo* e respetivas subclasses.

Cada assunto executivo pode conter seis propriedades para estabelecer relações com outras classes da ontologia, como por exemplo com as classes de *Person* (Pessoa) das ontologias importadas FOAF e Dublin Core. As propriedades estão enumeradas e descritas na Tabela 2.

| Classe | Descrição |
|---|---|
| Loteamento | Pedido de permissão para lotear ou alterar loteamentos de terrenos. |
| Empreitada | Relativo a processos de construção em execução. |
| Processo_de_obra | Anúncios relativos a processos de construção genéricos: início de trabalhos, alterações em orçamentos, expropriações, etc. |
| Isenção | Pedidos de isenções de taxas e outros pagamentos municipais. |
| Protocolo | Protocolos assinados com outras instituições. |
| Concurso_publico | Anúncios de concursos públicos relativos a aquisição de equipamento, construções, contratação, etc. |
| Subsidio | Subsídios pedidos e/ou concedidos pela autarquia. |

Tabela 1: Subclasses da classe *Assunto_executivo* e respetiva descrição.

| Propriedade | Descrição |
|---|---|
| deliberação | Resultado do pedido. |
| identificador | Identificador unívoco dado pelos serviços municipais. |
| montante | Qualquer quantia de dinheiro envolvida no processo. |
| motivo | O motivo do processo. |
| local | O local da construção ou do loteamento, morada da entidade que assinou o protocolo ou que pediu isenção ou subsídio. |
| submetidoPor | Entidade ou entidades que estão envolvidas no processo, excluindo o município. |

Tabela 2: Tipos de relações associadas à classe *Assunto_executivo*.

### 2.1.1 Exemplos Semente

Após a definição do domínio do conhecimento é necessário encontrar exemplos dos conceitos nos textos que o sistema deverá processar. Estes exemplos serão utilizados para treinar algoritmos de aprendizagem automática de modo a que o sistema detete esses conceitos em todos os documentos a processar.

A associação entre as amostras de texto e classes da ontologia e as relações são feitas usando o anotador AKTive Media (Chakravarthy, Ciravegna e Lanfranchi, 2006). No arranque da interface de anotação é necessário escolher ou criar uma sessão de anotação e escolher os textos a anotar e a ontologia que define o domínio do conhecimento. Após este passo é possível iniciar o processo de anotação ou então pedir ao sistema para pré-anotar partes do texto.

A pré-anotação foi uma funcionalidade desenvolvida para facilitar o processo de anotação quando existe uma grande quantidade de textos a anotar. Serve para pré-anotar no texto as classes da ontologia mas não as relações da ontologia. O seu comportamento é definido por um ficheiro de configuração que contém, em cada linha, uma entrada com uma expressão regular a localizar seguida da classe ou classes da ontologia a associar a essa palavra (ver Figura 4).

As palavras pré-anotadas ficam destacadas por um fundo colorido em que a cor está associada com a classe da ontologia (ver Figura 5). No fim do processo de anotação todas as pré-anotações que não foram validadas ou completadas pelo utilizador serão descartadas. Deste modo os exemplos semente são todas e apenas as anotações validadas pelo utilizador.


Figura 4: Ficheiro de configuração da pré-anotação. Cada linha contém a expressão regular a detetar no texto seguida da(s) classe(s) da ontologia a associar ao texto abrangido por essa expressão regular.

O procedimento para anotar uma frase é o seguinte (ver Figura 5):

1. Selecionar a classe da ontologia no painel superior esquerdo. Ao escolher a classe da ontologia surgem, no painel debaixo da caixa de procura, as relações possíveis para essa classe;
2. Selecionar a(s) palavra(s) a associar a essa classe. As palavras ficam por cima de um fundo colorido cuja cor está associada à classe escolhida;
3. Escolher a relação da ontologia a associar ao texto selecionado e marcar no texto o objeto dessa relação. A relação surgirá no painel inferior esquerdo;
4. Repetir o passo 3 até todas as relações estarem marcadas;
5. Voltar ao passo 1 até todo o texto relevante estar marcado.

A Figura 5 mostra a anotação de um subsídio cuja motivação é "execução do Plano Anual e a Escola Artística", foi submetido pela "ARCEL" e o montante envolvido é "8.640,00€".

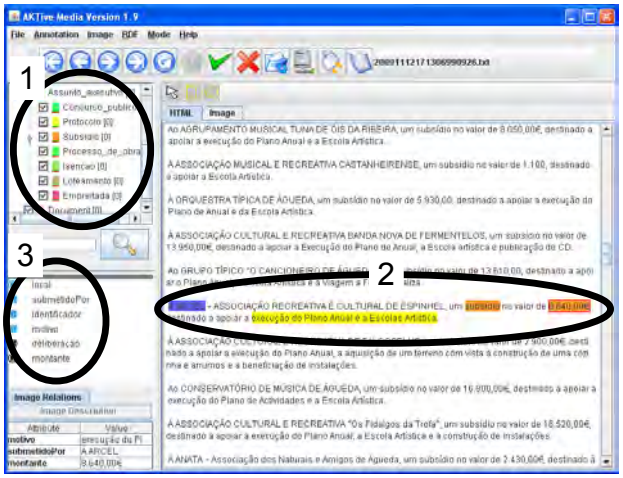



Figura 5: A interface de anotação. Os números correspondem à sequência de passos descritos no procedimento.

O resultado desta etapa é a ontologia e o conjunto de exemplos nos textos de entidades das classes e relações da ontologia.

## 2.2 Processamento de Linguagem Natural

Esta parte do sistema inclui ferramentas para obter e extrair o conteúdo de documentos da Web e/ou do sistema de ficheiros local. O sistema permite definir a fonte dos dados, sendo o conteúdo dos ficheiros processado automaticamente num encadeamento de operações sem intervenção dos utilizadores. A sequência de operações é igual à encontrada num vasto conjunto de sistemas de PLN (um bom exemplo é (Ferreira et al., 2009)): etiquetagem morfo-sintática, reconhecimento e classificação de entidades mencionadas e análise sintática.

O primeiro passo, a etiquetagem morfo-sintática (em Inglês *Part of Speech (POS) tagging*), tem por objetivo associar os diversos elementos do texto com classes morfo-sintáticas tais como substantivo, adjetivo, etc (Mihalcea, 2010). No sistema implementado a etiquetagem é realizada pelo TreeTagger que anota o texto com etiquetas morfo-sintáticas e com lemas e tem sido usado com sucesso para marcar várias linguagens naturais, incluindo Português (Schmid, 1994). O TreeTagger foi treinado com o Bosque v7.3, uma versão especificamente escolhidas por ser a única no formato aceite também pelo analisador sintático (descrito adiante). O Bosque é um subconjunto da Floresta (Freitas, Rocha e Bick, 2008) revisto por linguistas. O léxico utilizado foi enriquecido com o LABEL-LEX-sw (Ranchhod, Mota e Baptista, 1999).

De seguida é efetuado o Reconhecimento de Entidades Mencionadas (REM) e respetiva classificação que tem por objetivo detetar e classificar elementos atómicos no texto em categorias pré-definidas tais como nomes de pessoas, organizações, locais, etc (Santos e Cardoso, 2007). Além das classes de REM e seu significado serem diferentes das de POS *tagging*, uma diferença fundamental é que o processo de REM implica frequentemente o agrupamento de palavras numa única entidade. O REM do sistema é feito com o Rembrandt (Cardoso, 2008). O Rembrandt é um sistema de REM desenvolvido para Português que utiliza a estrutura e conteúdo da Wikipédia como uma fonte de conhecimento para classificar todos os tipos de entidades mencionadas no texto. Rembrandt tenta classificar cada entidade mencionada de acordo com as diretivas do segundo HAREM (Mota e Santos, 2008).

O terceiro passo, a análise sintática, é o processo de determinar a estrutura gramatical de uma sequência de palavras segundo uma determinada gramática formal. A análise sintática transforma um texto numa estrutura de dados. Este passo é efetuado por um analisador, em Inglês *parser*, de dependências chamado MaltParser (Hall et al., 2007). O MaltParser já foi utilizado com sucesso para analisar várias línguas o Inglês, Francês, Grego, Sueco e Turco. Foi treinado para Português com o Bosque v7.3 que existe no formato aceite por esta ferramenta, o formato CoNLL-X.

O funcionamento geral deste módulo está esquematizado na Figura 6.

## 2.3 Extração e Integração de Informação Semântica

Esta parte do sistema tem dois modos de operação: modo de treino e modo de execução.

No modo de treino, o sistema aprende a associar as estruturas sintáticas das frases às classes e relações da ontologia. Esta aprendizagem é baseada em exemplos anotados manualmente.

No modo de execução, o sistema aplica as associações aprendidas às estruturas sintáticas de todas as frases dos documentos a processar para extrair classes e relações semânticas do texto. O procedimento de ambos os modos é explicado seguidamente.

### 2.3.1 Treino de Modelos para Extração

Resumidamente, o processo desenrola-se da seguinte forma:

1. Processar todos os documentos de treino com o módulo de PLN para se obter estruturas sintáticas de todas as frases de treino;

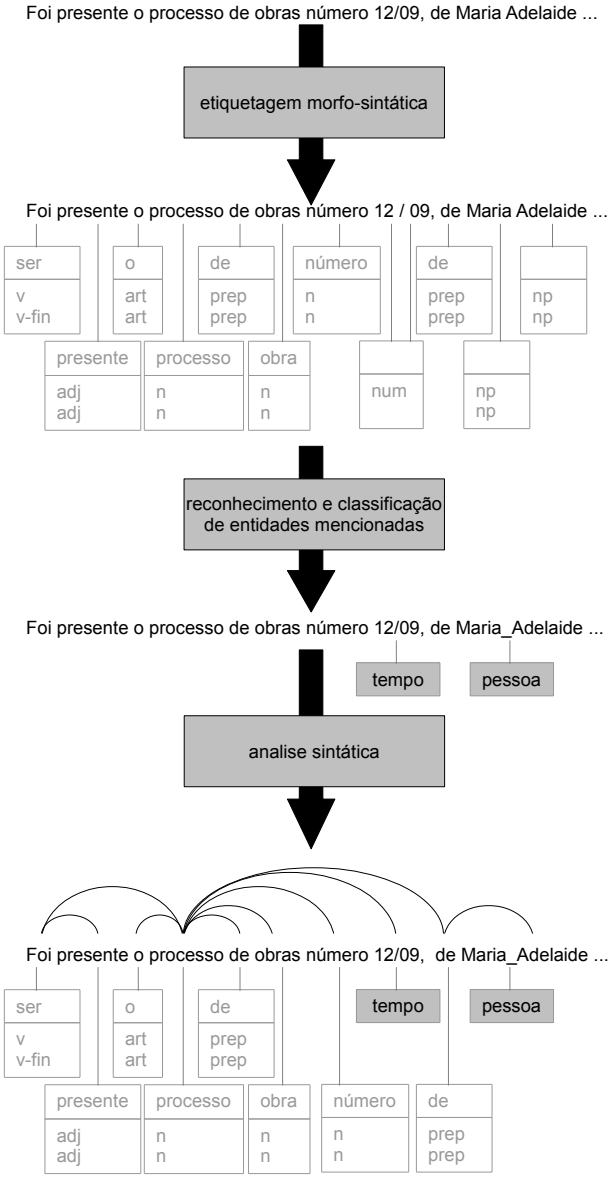



Figura 6: Sequência de passos do Processamento de Linguagem Natural e respetivos resultados intermédios. A entrada do módulo são frases sem estrutura definida e a saída é uma estrutura sintática enriquecida com etiquetas morfo-sintáticas e entidades mencionadas.

2. Para todas as estruturas sintáticas do conjunto de treino e para todas as relações (semânticas) anotadas: se as palavras da relação estiverem na estrutura sintática fazer os passos 3 e 4;
3. Guardar o caminho da árvore sintática entre as palavras envolvidas na relação. Este caminho é considerado um exemplo de elemento da relação ontológica e é composto por: sequência de ligações necessárias e lemas e etiquetas morfo-sintáticas das palavras que estão no caminho;
4. Guardar o contexto das palavras envolvidas na relação. O contexto é considerado um exemplo de elemento da classe ontológica e é composto por: lema, etiqueta morfo-sintática e tipos ligações sintáticas que a palavra possui;

No final deste processo, os caminhos encontrados para cada relação da ontologia são agrupados e utilizados para gerar um classificador estatístico por relação. Também os contextos encontrados para cada classe da ontologia são agrupados e utilizados para gerar um classificador estatístico por classe. Os classificadores estatísticos utilizados são baseados no algoritmo *k-nearest neighbor* (k-NN) e são semelhantes aos utilizados no LEILA (Rodrigues, Paiva Dias e Teixeira, 2011).

O último passo é melhorar a precisão dos classificadores. Este passo assume que todas as relações existentes no conjunto de treino foram marcadas. Assim, os classificadores começam por avaliar todas as estruturas sintáticas do conjunto de treino. Todas as estruturas sintáticas que são avaliadas como representando classes e relações da ontologia e que não as representam, ou seja não são exemplos anotados, passam a ser contra-exemplos para o classificador que gerou essa avaliação errada. Após da recolha de todos os contra-exemplos, todos os classificadores são novamente treinados agora utilizando os exemplos e os contra-exemplos.

### 2.3.2 Aplicação dos Modelos

À semelhança do treino, a execução da extração de informação inicia-se com o módulo de PLN a processar todos os documentos de modo a se obterem estruturas sintáticas para todas as frases dos documentos. Seguidamente, os classificadores estatísticos gerados na fase de treino avaliam se as estruturas sintáticas representam alguma classe ou relação da ontologia. Caso a avaliação do classificador seja mais elevada que o limiar de aceitação, essa informação é recolhida para uma base de conhecimento temporária.

Após a extração de informação segue a integração de informação. O motor de inferência semântico aplica as regras ontológicas a todos os dados e verifica se não existem implicações impossíveis, ou seja verifica se a nova informação é coerente com a ontologia e com informação já presente no sistema. O motor de inferência utilizado é o Pellet (Sirin et al., 2007) que suporta integralmente o formalismo OWL-DL. Toda a informação coerente passa da base de conhecimento temporária para a base de conhecimento do sistema. A informação incoerente não é adicionada e gera um aviso no registo do sistema para se averiguar a causa da incoerência. A base de conheci-

mento é armazenada e gerida pelo Virtuoso Universal Server[1]. Este servidor tem, entre outras características, um motor de base de dados nativo para *Resource Description Language* (RDF), suporta pesquisas SPARQL e, como indicador do seu desempenho, é o servidor da DBpedia[2] que contém atualmente 3,64 milhões de fatos dos quais 1,83 milhões estão classificados numa ontologia consistente (416.000 pessoas, 526.000 lugares, 169.000 organizações, etc.).

Nesta fase também se procura informação em falta, de acordo com a ontologia, em fontes externas estruturadas de informação. É necessário que a informação proveniente destas fontes seja estruturada de modo a se poder definir uma semântica apropriada para elas, uma vez que nesta fase do processamento a informação entra diretamente na base de conhecimento, não passando pelos classificadores estatísticos responsáveis por detetar informação semântica relevante.

Por agora existem dois tipos de informação adicionados caso estejam em falta na base de conhecimento: as coordenadas *Global Positioning System* (GPS) de entidades que deverão ter uma localização fixa e a organização política dos espaços.

As entidades que estão definidas na ontologia como tento uma localização fixa são, por exemplo, cidades, ruas, sedes de organizações e alguns eventos. Nestes casos, caso não existam na base de conhecimento, as coordenadas GPS destes locais são consultadas via Google Maps API.

Sendo esta aplicação um sistema de pesquisa de informação para a área do e-gov é relevante saber quais os locais políticos relativos à informação (rua $\subset$ freguesia $\subset$ cidade $\subset$ concelho...). Assim, além das coordenadas GPS também é adicionada a organização política dos espaços que é obtida utilizando uma ontologia geográfica de Portugal com cerca de 418 mil entradas chamado Geo-Net-PT01 (Chaves, Silva e Martins, 2005).

Estes dois tipos de informação adicionados permitem o sistema exibir informações espacialmente num mapa e procurar e relacionar informações em função da sua localização (Rodrigues, Paiva Dias e Teixeira, 2010b).

## 2.4 Interfaces de Acesso à Informação

Foram implementadas duas formas de aceder à informação gerida pelo sistema. Uma destina-se a ser utilizada de um modo fácil e intuitivo por pessoas e corresponde ao acesso via interface de linguagem natural. A outra destina-se a ser utilizada por sistemas que queiram aceder à informação semântica contida na base de conhecimento e corresponde ao acesso via interface para máquinas. Ambas as interfaces são explicadas seguidamente.

### 2.4.1 Interface para Utilizadores Humanos

A interface para humanos suporta linguagem natural escrita e permite a interacção usando Português. É uma interface flexível o suficiente para permitir a pesquisa por palavras chave, tal como os motores de pesquisa da Web, ou através da formulação de perguntas em Português. Esta flexibilidade é importante uma vez que o e-gov tem de servir a totalidade da população independentemente do seu nível de proficiência nas TIC. Assim, utilizadores habituados a pesquisar informação na Web podem pesquisar de um modo que já lhes é familiar ou então podem formular as perguntas às quais querem obter respostas.

A interface utilizada é baseada no NLP-Reduce (Kaufmann e Bernstein, 2007), uma interface em linguagem natural para a web semântica, em Inglês e independente do domínio. A escolha do NLP-Reduce foi motivada por esta independência de domínio, o que o torna adaptável aos vários assuntos do e-gov, e por ser facilmente adaptável ao Português uma vez que não contém componentes específicos para processar Inglês. A sua abordagem evita deliberadamente quaisquer tecnologias semântica ou linguística complexas e não interpreta ou tenta compreender as perguntas efetuadas. Consiste em associar as palavras (e seus sinónimos) contidas na pergunta às expressões utilizadas para descrever classes, relações e indivíduos presentes na base de conhecimento. Deste modo, se a ontologia estiver descrita Português, uma parte considerável do sistema fica automaticamente em Português. Apenas foram necessárias pequenas adaptações para Português na formulação das perguntas como por exemplo palavras muito frequentes e com pouco significado (*stopwords*) e os pronomes interrogativos (qual, quem, etc.).

A interface constrói automaticamente um léxico usando as palavras contidas em todos os fatos explícitos ou inferidos da base de conhecimento. Ao léxico são igualmente adicionados os sinónimos das palavras já presentes nele. A procura de sinónimos é efetuada através da ontologia lexical PAPEL (Oliveira, Santos e Gomes, 2010), criado pela Linguateca a partir do Dicionário PRO da Língua Portuguesa da Porto Editora. Também são adicionadas ao léxico os lemas das palavras nele presentes de modo a aumentar a abrangência lexical.

[1] http://virtuoso.openlinksw.com/
[2] http://dbpedia.org/

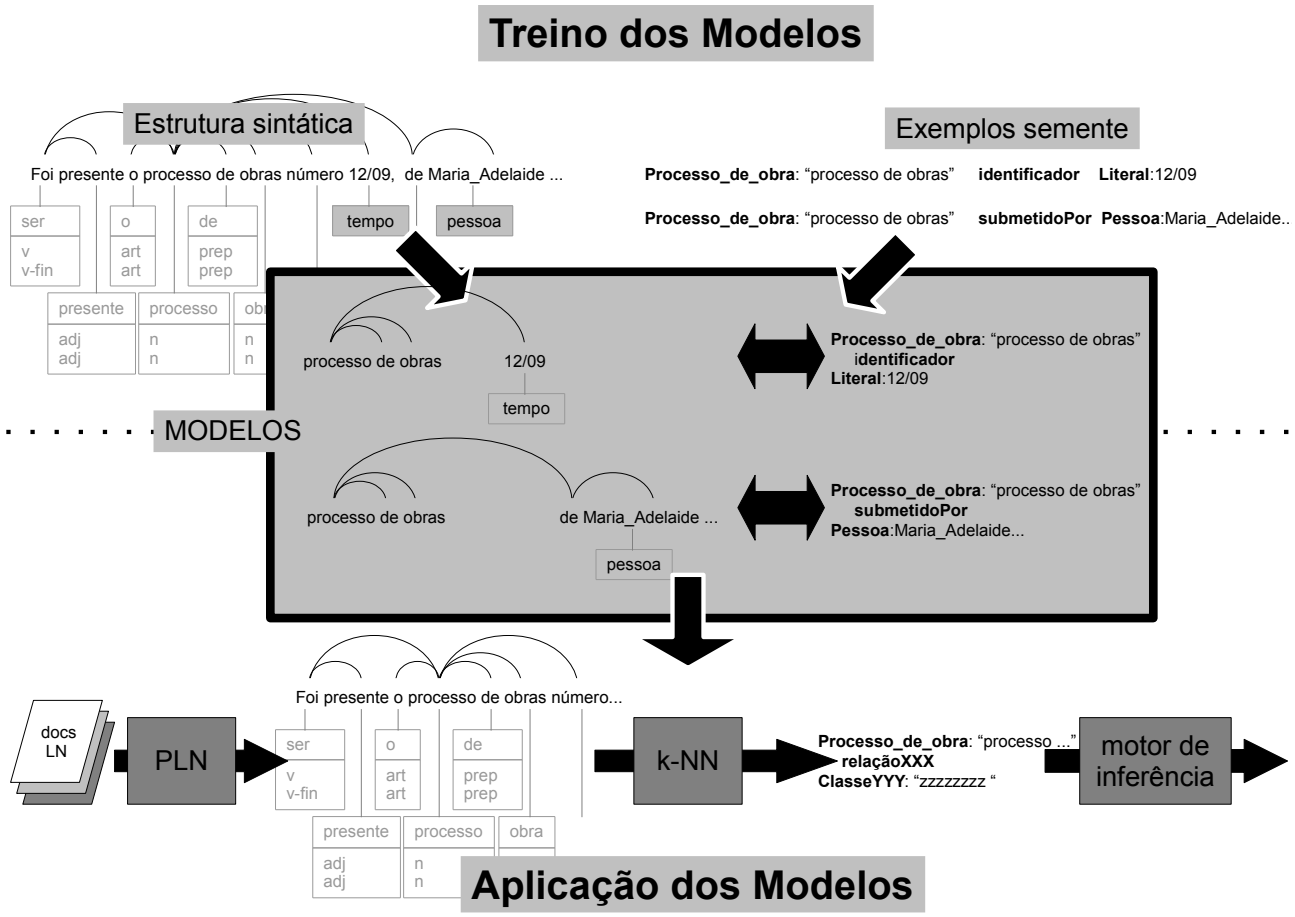



Figura 7: Treino dos modelos e respetiva aplicação. O treino começa por guardar o caminho da árvore sintática que liga as palavras envolvidas no exemplo semente. Após a recolha de todos os exemplos (e contra-exemplos) os caminhos são usados para treinar um classificador estatístico baseado no algoritmo k-NN. Durante a execução, os classificadores treinados são aplicados avaliam todas as estruturas sintáticas de modo a verificar se estas representam uma relação ontológica.

O processamento das entradas dos utilizadores inicia-se com a remoção de sinais de pontuação e de *stopwords*, tais como artigos, preposições, algumas conjunções. Seguidamente, com base nos lemas das palavras sobrantes, é construído uma pesquisa SPARQL que será submetida a um motor de pesquisas SPARQL. Considere-se o exemplo da pergunta "Qual a deliberação do processo de obra submetido por Maria?". A construção da pesquisa SPARQL é efetuada do seguinte modo:

1. São procurados os fatos em que pelo menos um dos lemas da pesquisa faz parte da etiqueta de uma propriedade de objeto. Considerando o exemplo, os fatos contendo as propriedades <deliberacao> e <submetidoPor> serão retornados. As propriedades de objeto encontradas são ordenadas de acordo com o ajustamento entre a sua etiqueta e as palavras da pesquisa, por exemplo a etiqueta <submetidoPor> obtém melhor classificação com as palavras "submetido por" que uma etiqueta que fosse <submetido>;
2. São procuradas no léxico elementos que podem ser conjugados com as propriedades encontradas no passo 1, usando os restantes lemas da pesquisa e tomando em consideração os seus domínio e contra-domínio. No nosso exemplo são procuradas os elementos que contenham "qual", "processo", "obra" e "maria". Como a classe <Processo_de_obra> contém a palavra "processo" e "obra" e é o domínio de ambas as propriedades obtidas no passo 1, este passa a ser o elemento de ligação entre elas;
3. São procuradas no léxico as propriedades relativas a dados cujos valores correspondem aos restantes lemas da pesquisa. Estas propriedades são combinadas com as identificadas anteriormente, tendo em conta os domínios e contra-domínios de todas as propriedades envolvidas e ordenados conforme o seu ajustamento às palavras sobrantes. Das palavras sobrantes do nosso exemplo, "qual"

e "maria", a palavra "maria" existe como valor da propriedade <Nome>. Como o domínio de <Nome> é a classe <Pessoa> que por sua vez é contra-domínio da relação <submetidoPor>, a propriedade <Nome> é adicionada à pesquisa.

4. Por último é gerada a pesquisa SPARQL com a junção de propriedades que obtiveram a classificação mais alta nos passos 1 e 3. Adicionalmente são removidos os duplicados semanticamente equivalentes e é efetuada a pesquisa com o SPARQL gerado.

### 2.4.2 Interface para Máquinas

A interface para máquinas aceita como entrada pesquisas em SPARQL e devolve um RDF contendo o conjunto de resultados e respetiva marcação semântica. Esta interface pode ser utilizada pela interface em linguagem natural, depois de gerar a pesquisa SPARQL, ou por sistemas externos que pretendam aceder à informação semântica. O seu objetivo é possibilitar a interoperabilidade entre este e outros sistemas.

A interoperabilidade é importante para o conceito de e-gov como uma plataforma. Este conceito é uma visão para o futuro em que um dos papéis principais dos sistemas de e-gov é o fornecimento de informação usando formatos abertos e livres e interpretáveis por máquinas (Frissen et al., 2007; United Nations, 2010). A ideia é que se a informação estiver disponível, existirá maior transparência na definição de políticas públicas e permitirá que entidades extra-governamentais utilizem essa informação combinando-a de formas inovadoras e úteis para as populações.

A interoperabilidade também tem um papel central na Web semântica, um conceito introduzido em (Berners-Lee, Hendler e Lassila, 2001). A Web semântica é uma extensão da Web atual que visa atribuir um significado aos conteúdos de modo que seja perceptível por pessoas e por computadores simultaneamente. Uma vez que a ontologia é tornada pública e o seu modo de acesso é uma norma aberta, qualquer entidade externa tem conhecimento do tipo de dados contidos na base de conhecimento, do seu significado semântico e de qual o protocolo de acesso. Uma forma de explorar esta funcionalidade é mostrada na Secção 3.

## *3 Exemplos de Utilização*

As experiências relatadas nesta secção foram concebidas para extrair informações sobre os assuntos municipais públicos mais frequentes e mais procurados por cidadãos e empresas. Para isso foram selecionados três temas em atas municipais públicas: os subsídios concedidos, as licenças de construção solicitadas, e protocolos assinados com outras instituições.

Um *crawler* web obteve todos os documentos disponíveis nos portais da Internet de sete municípios portugueses. Foram selecionados dois conjuntos aleatórios e disjuntos de 50 documentos cada. O documentos selecionados estavam no formato pdf. Um conjunto foi anotado manualmente por uma pessoa e as anotações foram utilizadas para treinar o sistema de de classificação. O outro conjunto foi utilizado em tempo de execução para ter conhecimento extraído pelo sistema.

Os utilizadores podem obter informação utilizando a interface de linguagem natural. A Figura 8 apresenta uma captura de ecrã contendo a resposta à pergunta "Qual a deliberação do processo de obra submetido por Maria?". Na janela por baixo da pergunta verifica-se que o SPARQL gerado é (URL's e variáveis SPARQL abreviados para ficar mais conciso):

```
select distinct * WHERE {
 ?Proc <#SubmetidoPor> ?Pess .
 ?Proc <#Deliberacao> ?Delib .
 ?Pess <#Nome> ?Pess_Nome .
 FILTER(REGEX(?Pess_Nome,'maria','i')).
 ?Proc <#type> <#Processo_de_obra> .
 ?Pess <#type> <#Pessoa>
}
```

A resposta à interrogação SPARQL gerada contém apenas duas entradas na base de conhecimento. A vantagem de ter uma base de conhecimento semântica fica patente neste exemplo uma vez que o sistema associa as palavras "processo de obra" à classe da ontologia "Processo_de_obra", associa a palavra "Maria" a uma pessoa, e procura obter o valor da propriedade "deliberacao". Assim apenas são verificadas as informações que o sistema capturou como relacionando processos de obra com pessoas chamadas Maria e não todas as frases que incluem (algumas das) palavras "processo", "obra" e "Maria". Outra vantagem é ser possível mostrar imediatamente apenas as informações consideradas relevantes, tais como o resultado da deliberação e o nome completo da pessoa, sem mostrar todos os outros dados conhecidos. Contudo é possível obter mais dados uma vez que também são devolvidas as referências da base de conhecimento correspondentes aos processos de obra retornados.

A funcionalidade do acesso para aplicações externas em SPARQL é demonstrada com uma página web (Figura 9) onde são mostrados num mapa os locais que estão envolvidos nos

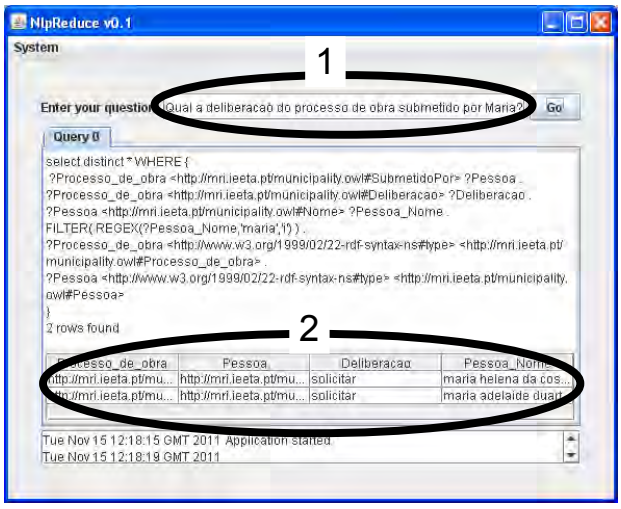



Figura 8: A interface de linguagem natural escrita. A entrada da pergunta é efectuada em 1 e a resposta dada em 2. Neste caso, solicitar significa que a Câmara Municipal solicitou mais documentação. Entre 1 e 2 pode-se ver a pesquisa SPARQL gerada.

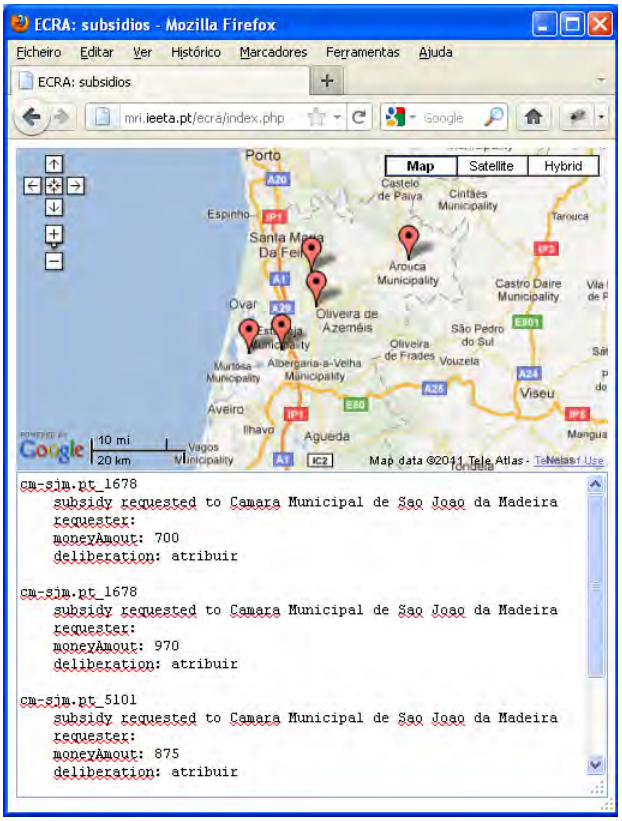



Figura 9: A interface Web. No mapa estão marcados os locais para os quais existe informação. Ao pressionar um local são mostradas as informações relativas ao mesmo na parte inferior da página.

subsídios existentes na base de conhecimento do sistema. A página web desenvolvida faz uma pergunta SPARQL onde questiona várias informações como a latitude e longitude das entidades que concederam os subsídios, o montante de dinheiro pedido e se foi atribuído e a quem. Depois de obter a resposta em RDF, a página web exibe a informação num mapa, usando para isso as coordenadas de latitude e longitude obtidas. Ao selecionar uma localização são mostradas todas as informações relativas a essa localização.

### 3.1 Avaliação de Desempenho

Recentemente foi efetuada uma avaliação de desempenho do sistema e os resultados obtidos foram apresentados na EPIA2011 - 15th Portuguese Conference on Artificial Intelligence (Rodrigues, Paiva Dias e Teixeira, 2011). A avaliação implicou que uma pessoa verificasse que fatos relevantes estavam contidos nos documentos do conjunto de teste. O conjunto detetado pela pessoa passou a ser a "verdade" e serviu de base de comparação para verificar que fatos foram encontrados ou não pelo sistema, e quais os que foram incorretamente extraídos. Os fatos foram considerados detetados se o sistema extraiu o tipo de fato (subsídio, processo de obra, protocolo) mesmo que estivessem em falta alguns dados como os pretendentes e as quantias envolvidas. Os resultados estão sumarizados na Tabela 3.

Existiam um total de 32 subsídios no conjunto de teste, dos quais o sistema detetou 14 e não houve nenhum falso positivo, isto é, todos os subsídio detetados eram realmente subsídios. Relativamente a processos de obra o sistema detetou 67 de um total de 68. Contudo foram também extraídos como processos de obra 4 informações que não o eram: 4 falsos positivos, Quanto aos protocolos, o sistema detetou 8 dos 41 existentes e houve um falso positivo. A baixa cobertura na deteção de protocolos (0.20) está em grande parte associada à existência de enumerações. Uma vez que estas não existiam no conjunto de treino, o sistema apenas detetou a primeira instituição em enumerações do tipo "... protocolos ... com as seguintes instituições:" seguida da listagem de instituições, uma por linha. Esta falha causou a baixa cobertura uma vez que por cada instituição listada, à exceção da primeira, foi considerado um protocolo não identificado.

O desempenho global do sistema relativamente à extração de informação semântica (precisão 0.95; cobertura 0.63) está em linha com o estado da arte para Inglês: DBpedia (precisão 0.86 a 0.99; cobertura 0.41 to 0.77), Kylin (precisão 0.74 a 0.97; cobertura 0.61 a 0.96), e YAGO/NAGA (precisão 0.91 a 0.99; cobertura

| | município | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | d | e | f | g | precisão | cobertura | $F_1$ |
| subsídio | 0(2) | 3(3) | 4(11) | 1(1) | 1(1) | 3(14) | 0(0) | 1.00 | 0.44 | 0.61 |
| processo de obra | 3(4)[1] | 13(13) | 47(47) | 0(0) | 0(0) | 4(4) | 0(0) | 0.94 | 0.99 | 0.97 |
| protocolo | 3(4) | 3(3) | 0(3) | 0(0) | 7(24) | 2(7)[2] | 0(0) | 0.89 | 0.20 | 0.32 |
| total | | | | | | | | 0.95 | 0.63 | 0.76 |

Tabela 3: Quantidade de fatos detetados pelo sistema. Os resultados apresentados para o conjunto de documentos de cada município são: a quantidade total de fatos corretamente detetados e, entre parêntesis, o número total de fatos encontrados pela pessoa, nesses mesmos documentos. Adicionalmente existem em ([1]) 4 processo de obra incorretamente extraídos e em ([2]) 1 protocolo incorretamente extraído.

não reportada).

## *4 Conclusões*

Este artigo apresenta pela primeira vez o sistema completo com a nova interface em linguagem natural escrita. O artigo descreve ainda a criação do domínio de conhecimento e dos exemplos semente com um maior nível de detalhe em relação a publicações anteriores, permitindo assim ter-se uma percepção mais aprofundada dos procedimentos a efetuar para utilizar o sistema em casos concretos. A descrição efetuada contempla todos os módulos do sistema, proporcionando-se deste modo uma visão global do mesmo.

A aplicação desenvolvida adiciona informação semântica ao conteúdo existente em documentos escritos numa linguagem natural, o Português, e disponibiliza essa informação via uma interface em linguagem natural ou via protocolos abertos de acesso a dados. As suas principais caraterísticas são: aceitar diversos domínios do conhecimento desde que definido por uma ontologia, obter informações acerca desse domínio em textos escritos em linguagem natural, e disponibilizar a informação via interfaces apropriadas para pessoas e para máquinas.

A preparação da aplicação a um novo domínio implica um conjunto reduzido de tarefas que incluem a definição desse domínio e o fornecimento de alguns exemplos do conteúdo desse domínio nos textos a processar. É igualmente possível alterar a língua base do sistema de Português para outra alterando o módulo de PLN, sem ser necessário alterar os restantes módulos da aplicação.

Este tipo de aplicações são importantes para o e-gov porque o seu próprio sucesso depende, em grande medida, da facilidade de obtenção de informação e utilização dos seus serviços. Contudo, o desenvolvimento deste tipo de aplicações para o e-gov e para Português é uma tarefa que ainda apresenta desafios. Um deles é a adaptação a esta área específica, uma vez que o e-gov contém documentos que abarcam diversos assuntos e que, frequentemente, contêm frases de difícil interpretação devido à sua extensão e ao estilo de escrita. Outro desafio é o desenvolvimento de sistemas de extração e disponibilização de informação semântica para Português que, apesar da maturidade de vários recursos e ferramentas disponíveis, ainda não são comuns trabalhos acerca da sua integração e utilização em aplicações concretas.

Para concluir, o sistema funciona para o Português e foi construído reutilizando software do estado da arte maioritariamente desenvolvido visando o Inglês. Isto mostra que é possível - e deve ser tentado - integrar ferramentas de software de alto desempenho mesmo que inicialmente tenham sido concebidas para outras línguas naturais.

## *Agradecimentos*

Os autores gostariam de agradecer ao Ciro Martins pela cuidada anotação dos exemplos semente nos documentos de treino do sistema.

## *Referências*

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, e Oren Etzioni. 2007. Open information extraction from the Web. Em *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India.

Bast, Holger, Alexandru Chitea, Fabian Suchanek, e Ingmar Weber. 2007. ESTER: Efficient Search on Text, Entities, and Relations. Em *Proc. 30th ACM SIGIR*, pp. 679–686.

Berners-Lee, Tim, James Hendler, e Ora Lassila. 2001. The Semantic Web. *Scientific American*, 284(5):34–43.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, e Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the WWW*, 7(3):154–165.

Brickley, Dan e Libby Miller. 2010. FOAF Vocabulary Specification. Publicado online em 9 de Agosto May 24th, 2010 at `http://xmlns.com/foaf/spec/`.

Cardoso, Nuno. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Chakravarthy, A., F. Ciravegna, e V. Lanfranchi. 2006. Cross-media document annotation and enrichment. Em *Proc. 1st Semantic Web Authoring and Annotation Workshop (SAAW2006)*.

Chatzidimitriou, M. e A. Koumpis. 2008. Marketing One-stop e-Government Solutions: the European OneStopGov Project. *IAENG International Journal of Computer Science*, 35(1):74–79.

Chaves, M.S., M.J. Silva, e B. Martins. 2005. A Geographic Knowledge Base for Semantic Web Applications. Em *Proc. of Simpósio Brasileiro de Banco de Dados*.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, e A. Yates. 2004. Web-scale information extraction in KnowItAll:(preliminary results). Em *Proceedings of the 13th international conference on World Wide Web*, pp. 100–110. ACM.

Ferreira, Liliana, César Telmo Oliveira, António Teixeira, e João Paulo Silva Cunha. 2009. Extracção de Informação de Relatórios Médicos. *Linguamática*, 1(1).

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2008. Floresta Sintá (c) tica: Bigger, Thicker and Easier. *Computational Processing of the Portuguese Language*.

Frissen, Valerie, Jeremy Millard, Noor Huijboom, Jonas Svava Iversen, Linda Kool, Bas Kotterink, Marc van Lieshout, Mildo van Staden, e Patrick van der Duin. 2007. The Future of eGovernment: An exploration of ICT-driven models of eGovernment for the EU in 2020.

GeoNames. 2010. GeoNames Geographical Database. `http://www.geonames.org/export`.

Gruber, Thomas R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5:199–220.

Hall, Johan, Jens Nilsson, Joakim Nivre, Gülşen Eryiğit, Beáta Megyesi, Mattias Nilsson, e Markus Saers. 2007. Single Malt or Blended? A Study in Multilingual Parser Optimization. Em *Proc. of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*.

Kaufmann, Esther e Abraham Bernstein. 2007. How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users? Em Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, e Philippe Cudré-Mauroux, editores, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pp. 281–294. Springer.

Layne, Karen e Jungwoo Lee. 2001. Developing fully functional E-government: A four stage model. *Government Information Quarterly*, 18(2):122–136.

Li, Yunyao, Huahai Yang, e H. V. Jagadish. 2005. NaLIX: an interactive natural language interface for querying XML. Em *Proc. of the ACM SIGMOD international conference on Management of data*, pp. 902.

Mihalcea, R. 2010. Performance Analysis of a Part of Speech Tagging Task. *Computational Linguistics and Intelligent Text Processing*, pp. 299–321.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

National Imagery and Mapping Agency. 2000. Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems. `http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html`.

Oliveira, Hugo Gonçalo, Diana Santos, e Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93.

Paiva Dias, Gonçalo. 2006. *Arquitectura de suporte á integração de serviços no governo electrónico*. Tese de doutoramento, Universidade de Aveiro.

Ranchhod, Elisabete, Cristina Mota, e Jorge Baptista. 1999. A Computational Lexicon of Portuguese for Automatic Text Parsing. Em *Proc. of SIGLEX99: Standardizing Lexical Resources - ACL*.

Rodrigues, Mário, Gonçalo Paiva Dias, e António Teixeira. 2010a. Human Language Technologies for e-Gov. Em *Proc. of the 6th International Conference on Web Information Systems and Technologies*, pp. 400–403, Valencia, Spain.

Rodrigues, Mário, Gonçalo Paiva Dias, e António Teixeira. 2010b. Knowledge Extraction from Minutes of Portuguese Municipalities Meetings. Em *Proc. of the FALA 2010 - VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*.

Rodrigues, Mário, Gonçalo Paiva Dias, e António Teixeira. 2011. Ontology Driven Knowledge Extraction System with Application in e-Government. Em *Proc. of the 15th Portuguese Conference on Artificial Intelligence*, pp. 760–774, Lisboa, Portugal.

Santos, Diana e Nuno Cardoso, editores. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Em *Proc. of International Conference on New Methods in Language Processing*, volume 12.

Sirin, E., B. Parsia, B.C. Grau, A. Kalyanpur, e Y. Katz. 2007. Pellet: A Practical OWL-DL Reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53.

Sroga, Magdalena. 2008. Access-eGov-Personal Assistant of Public Services. Em *Proc. of the International Multiconference on Computer Science and Information Technology*, pp. 421–427.

Suchanek, Fabian M., Gjergji Kasneci, e Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. Em *WWW '07*, pp. 697–706, New York, NY, USA. ACM.

Suchanek, F.M., G. Ifrim, e G. Weikum. 2006. LEILA: Learning to Extract Information by Linguistic Analysis. Em *Proc. of the ACL Workshop OLP*.

United Nations. 2010. United Nations E-Government Survey 2010 - Leveraging e-government at a time of financial and economic crisis.

Wang, C., M. Xiong, Q. Zhou, e Y. Yu. 2007. Panto: A portable natural language interface to ontologies. *LNCS*, 4519:473.

Weibel, S., J. Kunze, C. Lagoze, e M. Wolf. 2007. Dublin Core Metadata for Resource Discovery. RFC 5013 (Informational). `http://www.ietf.org/rfc/rfc5013.txt`.

Wu, Fei, Raphael Hoffmann, e Daniel S. Weld. 2008. Information extraction from Wikipedia: moving down the long tail. Em *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 731–739, New York, NY, USA. ACM.

# Estudio sobre el impacto de los componentes de un sistema de recuperación de información geográfica y temporal

Fernando S. Peregrino
Universidad de Alicante
fsperegrino@dlsi.ua.es

David Tomás Díaz
Universidad de Alicante
dtomas@dlsi.ua.es

Fernando Llopis Pascual
Universidad de Alicante
llopis@dlsi.ua.es

**Resumen**

La inmensa mayoría de los motores de búsqueda comerciales se centran principalmente en la recuperación de información textual, tratando de igual forma cualquier otro tipo de información contenida en el texto. Dicho tratamiento hace que cuando se añade alguna otra dimensión, como pueden ser la geográfica o la temporal, los citados buscadores obtienen unos pobres resultados. El presente trabajo pretende centrarse en los sistemas de recuperación de información geográfica y temporal abordando toda la problemática relacionada con éstos. Para ello, se ha desarrollado un sistema completo para el tratamiento de la dimensión geográfica y temporal en el texto y su aplicación a la recuperación de información, basando el citado sistema en múltiples motores de búsqueda y en técnicas de búsqueda de respuestas. Este sistema se ha evaluado en la tarea *GeoTime* del *workshop NTCIR*, lo que ha permitido comparar el sistema con otras aproximaciones actuales al tema.

**Palabras clave:** recuperación de información geográfica, etiquetado geográfico, información espacial, información temporal.

## 1. Introducción

En la sociedad actual, prácticamente se puede acceder a toda la información en formato digital. Dicha información se encuentra en constante crecimiento, por lo que se hacen imprescindibles herramientas que sean capaces de obtener los documentos deseados de una forma eficaz, rápida y sencilla.

La recuperación de la información (*IR: Information Retrieval*) es la ciencia de la búsqueda de información en documentos electrónicos dando como resultado un conjunto de estos mismos documentos ordenados según la relevancia que tengan con la consulta formulada.

Según un estudio realizado por (Zhang, Rey, y Jones, 2006), el 12,7 % sobre 4 millones de consultas de ejemplo contenía un topónimo, de lo que se desprende que la geografía también se ve involucrada en *IR*. Consultas del tipo *"Catedrales en Europa"*, o *"Dónde murió Osama bin Laden"*, hacen necesaria la intervención de dicha materia.

La recuperación de información geográfica (*GIR: Geographical Information Retrieval*) es una especialización de *IR* con metadatos geográficos asociados. Los sistemas de *IR*, generalmente, ven los documentos como una colección o "bolsa de palabras". Por el contrario, los sistemas *GIR* necesitan información semántica, es decir, necesitan de un lugar o rasgo geográfico asociado a un documento. Debido a esto, es común en los sistemas *GIR* que se separe el análisis y la indexación de texto de la indexación geográfica.

En la recuperación de información sobre textos sin estructurar se dificulta aún más la obtención del topónimo al que se hace referencia. La estructura típica de una consulta que requiere de *GIR* es <qué_se_busca> + <relación> + <localización>. Si nos centramos en un ejemplo concreto, dada la siguiente consulta: "*Estaciones de esquí en España*", deberíamos de limitar los resultados devueltos a aquellas estaciones ubicadas dentro del ámbito geográfico de la consulta ("*España*").

Los sistemas *GIR* son un campo de investigación en auge en los últimos años debido a la falta de buenos resultados cuando se realiza una búsqueda centrada en una ubicación específica. Son diversas las competiciones que se han organizado alrededor de este tipo de sistemas. El *CLEF*[1] (*Cross Language Evaluation Forum*) agregó una rama geográfica, *GeoCLEF*[2]. A raíz de esta rama geográfica, nacieron otro tipos de tareas como el *GikiP*, la cual fue una tarea piloto en el *GeoCLEF* 2008 pasando en 2009 a llamarse *GikiCLEF* y ser una tarea propia dentro del *CLEF*. Ésta tarea consistía en encontrar entradas

[1] http://www.clef-campaign.org/
[2] http://ir.shef.ac.uk/geoclef/

o documentos en la *Wikipedia* que contestaran a una serie de consultas que requerían de algún tipo de razonamiento geográfico. El *NTCIR*[3] (*NII Test Collection for IR Systems*) creó la tarea *GeoTime*[4]. Esta tarea combina *GIR* con búsqueda basada en el tiempo para encontrar eventos específicos en una colección de textos. También se han creado *workshops* específicos en la materia como el *Geographic Information Retrieval*[5].

Este artículo pone su foco de atención en los sistemas *GIR*, realizando un estudio exhaustivo de la situación actual en dicha materia. Además del presente estudio, se ha desarrollado un sistema *GIR* modular con el fin de discutir las dificultades expuestas en éste, evaluado cómo afectan los distintos componentes que intervienen en el proceso sobre el rendimiento final del sistema. Para dicha empresa, se ha evaluado el sistema en la tarea *GeoTime* del *NTCIR*, para lo que se ha tenido que incorporar al sistema un módulo para el tratamiento de la información temporal.

Este trabajo está estructurado según sigue. Primero, se introduce el estado de la cuestión que recopila los trabajos relacionados más relevantes hasta la fecha, así como las principales tendencias en este campo. A continuación, se da paso a la descripción detallada del sistema implementado deteniéndonos en cada uno de sus módulos y componentes. Se prosigue con los experimentos realizados en dicho sistema, así como la evaluación obtenida de los mismos y un análisis exhaustivo de los resultados. Para finalizar, se muestran las conclusiones y trabajo futuro para extender y mejorar el sistema.

## *2. Trabajo relacionado*

En los últimos años ha habido un incremento en la investigación dedicada a la recuperación de información geográfica dado su gran interés mercantil. Los grandes motores de búsqueda web comerciales (*Google*, *Yahoo!* y *Bing*) han desarrollado herramientas para poder afrontar dicha problemática, sin embargo, dichas herramientas tienen un amplio margen de mejora.

Si hay un proyecto que es referencia obligatoria para todo aquel que se quiera iniciar en la materia, y que aún hoy en día sigue marcando la pauta a seguir por el resto de investigadores en *GIR*, ese es el proyecto *SPIRIT*(Jones et al., 2007). En este proyecto se crearon herramientas software y técnicas que pueden ser usadas para crear motores de búsqueda y sitios web que muestren inteligencia en el reconocimiento de terminología geográfica. Con el fin de demostrar y evaluar los resultados del proyecto, se construyó un prototipo de motor de búsqueda *GIR*, el cual está siendo usado como plataforma para probar y evaluar nuevas técnicas en recuperación de información geográfica. Este proyecto aborda plenamente todos los frentes abiertos en la investigación en *GIR*.

En (Wang et al., 2005) clasificaron las áreas de investigación en esta materia en tres grandes grupos: identificación y desambiguación de topónimos (etiquetado geográfico), desarrollo de herramientas informáticas para el manejo de la información geográfica, y explotación de las diversas fuentes de recursos geográficos. En (Jones y Purves, 2008) podemos ver una disección más exhaustiva de los principales asuntos que podemos abordar en los sistemas *GIR*:

1. Detección de referencias geográficas.
2. Desambiguación de topónimos.
3. Terminología geográfica vaga.
4. Indexación espacial y geográfica.
5. Ranking por relevancia geográfica.
6. Iterfaces de usuario.
7. Métodos de evaluación de los sistemas *GIR*.

A continuación, se van a ver las distintas aproximaciones que se han llevado a cabo en cada uno de estos apartados.

### 2.1. Detección de referencias geográficas

La detección de referencias geográficas o *geotagging* en el Procesamiento del Lenguaje Natural (*PLN*) es una extensión de los reconocedores de entidades nombradas (*NER: Named Entity Recognition*), y versa sobre el análisis de los textos con el fin de identificar la presencia inequívoca de topónimos. Dicha problemática dista mucho de ser algo trivial, dado que en numerosas ocasiones podemos encontrarnos con que los nombres de personas, organizaciones, equipos deportivos, etc., son compartidos por los topónimos. Podemos ver como (Li et al., 2006) afronta la tarea con un mecanismo para la resolución probabilística de topónimos, dando dicho método una mejora de la efectividad sobre el subconjunto de *topics* del *GeoCLEF*.

Otra dificultad con la que nos topamos es la metonimia, en situaciones tales como "*no aceptaremos órdenes de Madrid*". En (Leveling y Hartrumpf, 2007) se muestra un enfoque para solucionar el problema de la metonimia a través del análisis de rasgos superficiales.

[3] `http://research.nii.ac.jp/ntcir/index-en.html`
[4] `http://metadata.berkeley.edu/NTCIR-GeoTime/`
[5] `http://www.geo.uzh.ch/~rsp/gir10/`

## 2.2. Desambiguación de topónimos

Una vez obtenido de manera inequívoca el nombre del lugar hay que desambiguarlo, ya que, muchos de los topónimos existentes son compartidos por varios lugares (p. ej. Granada, Springfield, etc.). Para dicha tarea se han afrontado diversas estrategias tales como la identificación por medio del resto de topónimos del texto, es decir, obteniendo el ámbito del que se está hablando para desambiguar cada uno de los lugares (Wang et al., 2005). Otra de las estrategias seguidas en esta materia, al hilo de la anterior, es el esclarecimiento del lugar nombrado jugando con las entidades geográficas de orden superior o inferior mencionadas en el texto (Silva et al., 2006).

## 2.3. Terminología geográfica difusa

Otra problemática adicional es la de las expresiones geográficas difusas, es decir, aquellas que describen lugares imprecisos que no podemos encontrar en ninguna base de datos geográfica (*gazetteer*). Ampliando el ejemplo visto en la introducción, serían expresiones del tipo "*Estaciones de esquí en el norte de España*", donde la entidad geográfica en la que deberíamos buscar ("*norte de España*") resulta imprecisa.

Los trabajos previos realizados en el campo de la definición automática de regiones geográficas difusas han seguido dos líneas fundamentales a la hora de obtener la información necesaria para definir dichas regiones. La primera aproximación se centra en la obtención de información a partir de la consulta directa a un conjunto de usuarios reales, que son los encargados de delimitar la región a estudio. La segunda aproximación está enfocada a la obtención de información a partir de fuentes de información no estructurada para la identificación de estas regiones.

Dentro del primer grupo se encuentra el trabajo de (Montello et al., 2003). En este trabajo se propone una aproximación probabilística basada en frecuencias, donde la inclusión de una determinada localización dentro de una región difusa viene condicionada por el número de usuarios que considera su pertenencia a la misma. Estos valores sirven para generar una curva de nivel que delimita la región difusa y proporciona una probabilidad de pertenencia a las localizaciones que se hallan en su interior.

El objetivo de la segunda aproximación es recuperar suficiente información de la web para poder definir espacialmente las regiones difusas estudiadas. En esta línea, en (Clough, 2005) se creó un sistema donde las coordenadas de las localidades contenidas en dichas regiones se empleaban para la definición de polígonos representativos de dicha región. Una extensión de este trabajo se puede encontrar en (Jones et al., 2008), donde utilizan esta aproximación para identificar tanto regiones difusas como precisas.

## 2.4. Indexación espacial y textual

Existen una serie de técnicas para la indexación textual de documentos. Dichas técnicas, tal y como podemos ver en (Baeza-Yates y Ribeiro-Neto, 1999), suelen basarse en la creación de índices inversos, es decir, se añaden al índice todas las palabras encontradas en el corpus, indicando en qué documentos aparece cada palabra y con qué frecuencia, para su posterior recuperación mediante la intersección con la consulta del usuario. En cuanto a la indexación espacial, los sistemas de información geográfica (*GIS: Geographical Information System*) son los que han tratado con más éxito dicho asunto. La dificultad subyace en conseguir mezclar ambos índices con acierto (Cardoso y Santos, 2007). La técnica más común para afrontar dicha mezcla ha sido la obtención de "huellas" (*footprints*) espaciales de los documentos, para posteriormente poder indexar esas huellas por documentos. El problema es que un alto número de huellas por documento (según (Vaid et al., 2005) hay unas 21 por documento web) puede hacer intratable el problema, por lo que han habido muchos trabajos orientados a obtener un mínimo número de huellas representativas por documento (Wang et al., 2005)(Silva et al., 2006).

En (Vaid et al., 2005) podemos ver tres estilos diferentes de conseguir llevar a cabo dicha tarea con celeridad y mejorando los resultados de un sistema de *IR* genérico. En dicho trabajo se comparó la indexación textual corriente (*PT: Pure Text*), contra tres tipos de indexación textual y espacial: indexación texto-espacial (*TS: Text-primary spatio-textual indexing*), indexación espacio-textual (*ST: Space-primary spatio-textual indexing*) e indexación separada (*T: Separate text and spatial indexes*), dando como resultado una gran mejora en el *recall*, aunque las indexación *TS* y *ST* supusieron un considerable aumento en el espacio requerido para su indexación.

## 2.5. Ranking por relevancia geográfica

La clasificación de documentos relevantes o ranking, determina la manera en la que debe de puntuar un documento según su idoneidad para con la consulta lanzada para su posterior devolución al usuario. La manera más usual con la que suelen afrontar la clasificación de documentos los

principales motores de búsqueda es mediante la intersección de los términos de la consulta con los de los documentos, dando un mayor valor a aquellos términos que ocurran en un menor número de documentos, a los términos que con más frecuencia aparezcan en un documento, y a la proporción de apariciones de un término en un documento dada la longitud de éste último (Robertson, Walker, y Hancock-Beaulieu, 1998).

En el caso de los sistemas *GIR* no hay que devolver una simple intersección de términos, si no que hay que saber tratar semánticamente los topónimos referenciados en la consulta de tal forma que se capturen las huellas existentes en ésta y se comparen con las de los documentos del corpus, teniendo en cuenta que dichas huellas pueden pertenecer a un ámbito geográfico inferior, superior, intersectante, etc. (Van Kreveld et al., 2005).

### 2.6. Iterfaces de usuario

El desarrollo de una interfaz de usuario eficaz para los sistemas *GIR* es un tema que se ha tratado con escaso éxito dada su dificultad. La mayoría de consultas geográficas, como ya se ha indicado anteriormente, son del modo <qué_se_busca> + <relación> + <localización>, por lo que parece sencillo crear una interfaz con tres campos, pero habría que saber tratar cualquier tipo de "relación", y también habría que tener en cuenta que no todas las consultas son del tipo descrito anteriormente (p. ej. "*Dónde murió Osama bin Ladem*"). Otra aproximación que se ha hecho en (Jones et al., 2007) es permitir al usuario trazar una zona en un mapa en la que se quiere obtener los resultados e introducir la consulta en sí.

Por otro lado, nos encontramos con la problemática de cómo devolver los resultados. ¿Se debe adjuntar un mapa sobre los sitios de los que habla cada documento? ¿A qué escala? ¿Un mapa por documento o un único mapa para todos los documentos relevantes?

### 2.7. Métodos de evaluación de los sistemas *GIR*

La evaluación de los resultados devueltos por un sistema *GIR*, al igual que en los sistemas de *IR*, es una tarea costosa dado que hay que ir examinando manualmente, uno a uno, todos los documentos devueltos por un sistema para comprobar la relevancia de cada una de las preguntas que se le ha lanzado. En tareas como la del *GeoCLEF* o *NTCIR GeoTime*, dichas evaluaciones se han resuelto mediante valores binarios, es decir, diciendo si un documento es relevante o no para la pregunta realizada, aunque también es cierto que en el *NTCIR GeoTime*, dicho juicio se hizo añadiendo una mayor escala de valores, según lo relevante que sea. En *GIR* hay que tener en cuenta que se añade una nueva dimensión a evaluar, la geográfica, por lo que un documento puede ser relevante en cierto grado geográficamente hablando, y en otro grado en cuanto a contenido.

Por ejemplo, en la tarea *GeoTime* del *NTCIR 2011*, se emplearon tres métricas distintas para la evaluación de los documentos: *Average Precision* (*AP*), *Q-measure* (*Q*), y *normalised Discounted Cumulative Gain* (*nDCG*) (Mitamura et al., 2008). La evaluación se efectuó recogiendo los $n$ primeros documentos que cada uno de los sistemas de los participantes entendió que eran relevantes para cada una de las consultas efectuadas. Dichos documentos se evaluaban de la siguiente manera:

- Relevante. Si contestaba a la pregunta dónde y cuándo.
- Parcialmente relevante (dónde). Si contestaba a la pregunta dónde.
- Parcialmente relevante (cuándo). Si contestaba a la pregunta cuándo.
- Parcialmente relevante (otro). Si contestaba de alguna manera a la pregunta.
- Irrelevante. Si no contestaba a la pregunta.

## *3. Descripción del sistema*

Los sistemas *GIR*, comúnmente, se pueden dividir en los siguientes módulos: etiquetado geográfico (obtención y desambiguación de los topónimos para su posterior procesamiento), indexación geográfica y de texto, almacenamiento de datos, clasificación geográfica por relevancia (con respecto a una consulta geográfica) y la navegación en los resultados (normalmente con un mapa como interfaz).

Para la creación del presente sistema *GIR*, se ha optado por una implementación modular con el fin de poder así añadir nuevos componentes en un futuro y crear mejoras sobrelos ya existentes.

En la figura 1 se muestran en la parte izquierda y central los elementos que intervienen en la fase inicial de indexación y preprocesamiento del corpus. En la parte central y derecha se muestran los que intervienen en las fases que tienen que ver con las consultas, es decir, los de procesado y ejecución de las mismas. También se puede observar que las líneas continuas que unen los distintos componentes del sistema están relacionadas con todas las acciones de preprocesado, las líneas discontinuas indican las acciones llevadas

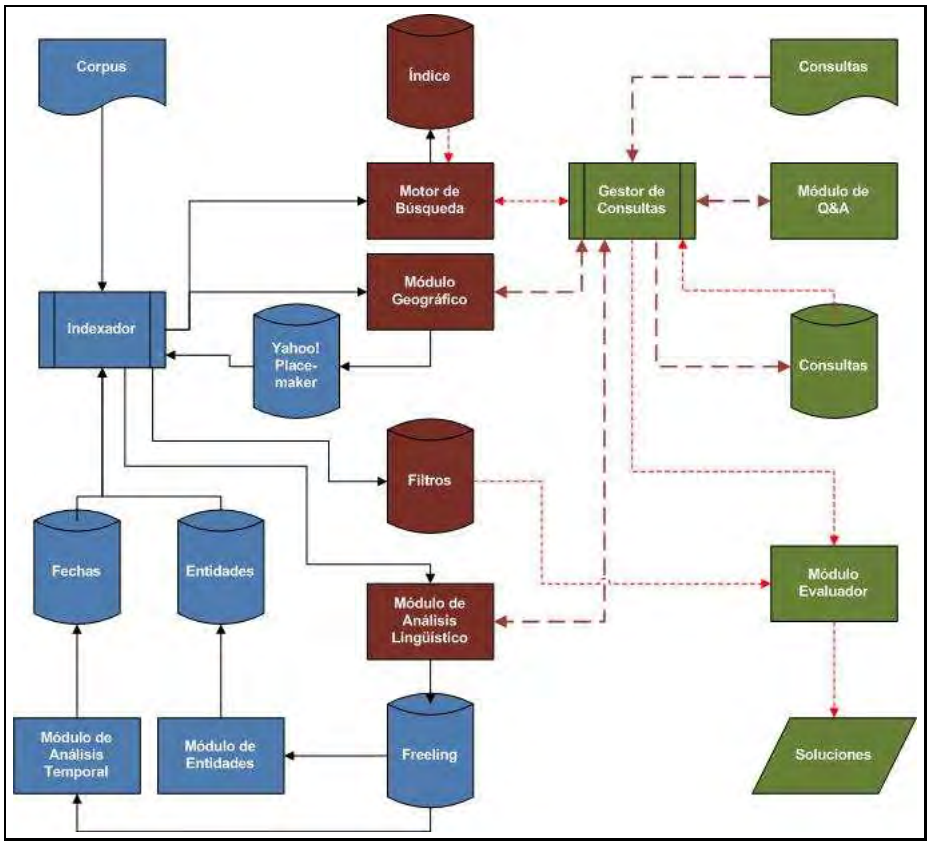



Figura 1: Esquema de funcionamiento general del sistema *GIR* implementado.

a cabo en tiempo de ejecución, siendo las líneas gruesas discontinuas las que indican el procesado de la consulta, mientras que las líneas discontinuas representan la búsqueda de la consulta en sí (sección 3.3).

Dado que se ha decidido evaluar el sistema construido en la tarea *GeoTime* del *NTCIR*, la cual tiene una componente temporal (tal y como se explicará más en profundidad en la sección 4), ha sido necesaria la creación de un módulo que sea capaz de tratar la información temporal.

## 3.1. Módulos del sistema

A continuación se describen en detalle cada uno de los módulos desarrollados en el sistema.

### 3.1.1. Motor de Búsqueda.

Como motor de búsqueda para este sistema, originalmente se optó por utilizar *Lucene*[6], un motor de búsqueda de alto rendimiento equipado con librerías de búsqueda escritas completamente en Java. Se trata de una tecnología adecuada para casi cualquier aplicación que requiera búsqueda de texto completo, especialmente en distintas plataformas. *Apache Lucene* es un proyecto de código abierto disponible para su descarga gratuita.

Viendo que la cobertura obtenida por *Lucene* apenas llegaba al 50 %, como se podrá ver con más detenimiento en la sección 4, y basándonos en el trabajo realizado por (Perea-Ortega, 2010), se decidió dotar al sistema de un motor de búsqueda adicional, *Terrier*[7]. Así pues, para obtener una puntuación definitiva normalizada por cada uno de los documentos devueltos por ambos motores de búsqueda, se procedió según sigue para cada una de las consultas lanzadas:

1. Se obtiene el valor máximo de *Lucene* de entre todos los documentos devueltos por éste.
2. Se dividen todas las puntuaciones de todos los documentos devueltos por *Lucene* entre el valor máximos indicado en el punto anterior.
3. Análogamente, se repiten los dos pasos anteriores para *Terrier*.
4. En caso de que haya documentos devueltos tanto por *Lucene* como por *Terrier*, se suman las puntuaciones de ambos.
5. Finalmente, se divide entre dos la puntuación de cada uno de los documentos devueltos por los motores de búsqueda citados an-

[6] `http://lucene.apache.org/`

[7] `http://terrier.org/`

teriormente, obteniendo así un valor normalizado entre 0 y 1.

El módulo del motor de búsqueda tiene principalmente dos funcionalidades: la indexación de todo el corpus y la recuperación de una serie de documentos dada una consulta.

Al motor de búsqueda se le han añadido una serie de características para mejorar su rendimiento, tales como un lematizador y la eliminación de palabras de parada (*stopwords*). Para la indexación del corpus, se obvian todas las *stopwords* y se indexan el lema del resto de palabras de cada uno de los documentos.

Para la ordenación de los resultados según su relevancia, se ha utilizado la función de pesado *BM25* (Robertson, Walker, y Hancock-Beaulieu, 1998) basada en los modelos probabilísticos de *IR*.

Por otro lado, también se ha utilizado el modelo de expansión de consulta *Bose-Einstein* (*Bo1*).

Finalmente, se ha establecido que el motor de búsqueda pueda recuperar hasta 1.000 documentos relevantes.

#### 3.1.2. Módulo de Análisis Lingüístico.

Este módulo se encarga de:

- Lematizar la consulta.
- Eliminar las *stopwords*.
- Obtener los topónimos de las consultas mediante una base de datos de nombres de lugar (*GeoNames*[8]).
- Obtener las fechas de la consulta mediante un analizador lingüístico (*FreeLing*[9]).
- Obtener las restricciones geográficas y temporales de las consultas (p. ej. si en la búsqueda es suficiente con encontrar el nombre del país o hay que buscar el de la ciudad también, si con el año y el mes es suficiente o hay que encontrar la fecha completa también, etc.). Para ello se analizan las partes narrativas de las consultas en busca de palabras clave como "país", "estado", "ciudad", "día", "fecha exacta", etc.
- Encontrar otro tipo de entidades (p. ej. nombres de persona, nombres de empresas, etc.) mediante un analizador sintáctico (*FreeLing*).
- Obtener otros términos lingüísticos comunes para una posible expansión de la consulta.

[8] `http://www.geonames.org/`

[9] `http://nlp.lsi.upc.edu/freeling/`

#### 3.1.3. Módulo de Q&A.

Una de las novedades que se ha introducido en este sistema *GIR* ha sido un módulo de búsqueda de respuesta (*Q&A: Question Aswering*), mediante el cual se pretenden obtener los términos geográficos y temporal que más se adecúan a la respuesta de la pregunta para su posterior intersección con los artículos del corpus. Lo que se hace en este módulo es lanzar la consulta a la interfaz de búsqueda de *Yahoo!* (*Yahoo! Search BOSS*[10]) y recoger los resúmenes de los 1000 primeros resultados devueltos. De estos resúmenes se extraen todas las fechas y lugares que se encuentren, se cuenta el número total de ocurrencias de cada uno de ellos y se normaliza, quedándose el sistema con los 10 más relevantes para cada una de las siguientes 4 categorías: fecha completa, mes y año, año, y topónimos.

#### 3.1.4. Módulo de Análisis Temporal.

Como ya se ha indicado con anterioridad, este módulo ha sido introducido con el fin de poder evaluar el sistema en la tarea *GeoTime* del *NTCIR*, pero no forma parte del principal propósito de esta investigación. La implementación se ha apoyado en el analizador lingüístico *FreeLing*. Concretamente, se ha aprovechado que *FreeLing* dispone de un módulo de detección y normalización de fechas para obtener todas las referencias a fechas que haya en cada uno de los documentos del corpus, incluida la fecha de cada documento (la de publicación en el periódico, ya que los documentos fueron extraídos de noticias periodísticas). De esta forma, en el tiempo de preproceso se crea un nuevo fichero por cada artículo existente en el corpus, en el que están registradas todas las fechas que *FreeLing* ha detectado en el artículo. Posteriormente, en tiempo de ejecución, se busca intersectar la fecha que pueda haber en la consulta con la de los documentos devueltos por el motor de búsqueda, dando mayor peso si coincide la fecha completa (día, mes y año) a si lo hace parcialmente (mes y año, o solamente el año).

#### 3.1.5. Módulo Geográfico.

En este módulo es donde se hará todo el tratamiento geográfico. Para el desarrollo de dicho módulo, el sistema se ha basado en *Yahoo! Placemaker*[11].

*Yahoo! Placemaker* es un servicio web de *geoparsing*[12] de libre disposición. Es útil para desarrolladores que quieren hacer aplicaciones basa-

[10] `http://developer.yahoo.com/search/boss/`

[11] `http://developer.yahoo.com/geo/placemaker/`

[12] Detección y desambiguación de nombres de lugar asignándole un identificador único.

das en localización espacial mediante la identificación de topónimos existentes en textos no estructurados (p. ej. *feeds*, páginas web, noticias, etc.), de los que es capaz de devolver metadatos geográficos asociados a dichos textos. La aplicación identifica los topónimos en el texto, los desambigua, y devuelve identificadores de lugar únicos para cada uno de los lugares que tiene en su base de datos. También aporta otro tipo de información como, cuántas veces aparece el lugar en el texto, en qué lugar del texto se encontró, etc.

En el caso concreto de este sistema *GIR*, ha utilizado *Yahoo! Placemaker* para la obtención de topónimos, la desambiguación de los mismos, y la obtención de entidades administrativas de orden superior e inferior.

*Yahoo! Placemaker* devuelve un documento *XML* por cada texto que se le pase. Finalmente, el módulo almacena toda la información geográfica pertinente del artículo en un documento *XML*.

Este módulo tiene otra función a la hora de analizar las consultas. Concretamente, lo que hace es recoger los topónimos existentes en los ficheros *XML* de las consultas y los transforma al identificador inequívoco de *Yahoo!* (*WOEID: Where On Earth IDentifier*) para un procesamiento más ágil en la fase de búsqueda.

### 3.1.6. Módulo para la Detección de Entidades.

Este módulo se encarga de guardar un documento por cada uno de los artículos del corpus. En dicho documento estarán todas las entidades reconocidas por *FreeLing* en el artículo original para, posteriormente, crear un nuevo documento *XML* que sirva de filtro para las consultas introducidas.

### 3.1.7. Módulo de Reordenación.

Este módulo entra en acción únicamente en tiempo de ejecución, es decir, en el momento de realizar la búsqueda de una consulta concreta. El objetivo de este módulo es intersectar los documentos *XML* de las consultas y los documentos *XML* de los artículos de corpus que han sido devueltos como solución por el motor de búsqueda. Como resultado de dicha intersección, el módulo evaluador reorganizará el ranking de documentos devuelto por el motor de búsqueda. Dicha reorganización viene dada por dos esquemas de pesado. En ambos esquemas de pesado se puede observar la importancia del peso de *Lucene*, pero se diferencian en:

- Esquema de pesado A. Permite evaluar la importancia del módulo de Q&A de la parte descriptiva (parte de la consulta en sí misma) y la parte narrativa (parte que detalla qué es lo que se busca concretamente, p. ej. ciudad, estado, fecha exacta, etc.) de la consulta:

$$\alpha\cdot(\beta\cdot\omega_{desc}+(1-\beta)\cdot\omega_{narr})+(1-\alpha)\cdot\omega_{QA} \quad (1)$$

  Donde:

  - $\alpha$ = Peso que se le da a la parte descriptiva y narrativa de la consulta.
  - $\beta$ = Peso que se le da a la parte descriptiva de la consulta.
  - $\omega_{desc}$ = Valor normalizado de la parte descriptiva de la consulta.
  - $\omega_{narr}$ = Valor normalizado de la parte narrativa de la consulta.
  - $\omega_{QA}$ = Valor normalizado del módulo de Q&A.
- Esquema de pesado B. Permite evaluar la importancia de la parte geográfica, la temporal y la de entidades:

$$\alpha\cdot(\beta\cdot\omega_{geo}+(1-\beta)\cdot\omega_{temp})+(1-\alpha)\cdot\omega_{ent} \quad (2)$$

  Donde:

  - $\alpha$ = Peso que se le da a la parte geográfica y temporal de la consulta con respecto a la de entidades.
  - $\beta$ = Peso que se le da a la parte geográfica de la consulta con respecto a la temporal.
  - $\omega_{geo}$ = Valor normalizado de la parte geográfica.
  - $\omega_{temp}$ = Valor normalizado de la parte temporal.
  - $\omega_{ent}$ = Valor normalizado de la parte de entidades.

En ambos esquemas de pesado, una vez obtenido el resultado de las ecuaciones 1 y 2, se tiene que unir con el resultado obtenido por el motor de búsqueda, utilizándose para ambos esquemas la siguiente fórmula:

$$\lambda \cdot L + (1 - \lambda) \cdot E \quad (3)$$

Donde:

- $\lambda$ = Peso que se le da al motor de búsqueda (*Lucene+Terrier*).
- $L$ = Valor normalizado de los resultados devueltos por el módulo de motor de búsqueda.
- $E$ = Esquema de pesado elegido (ver ecuación 1 y 2).

### 3.2. Esquemas de Almacenamiento

Con el fin de agilizar y de hacer más eficiente el proceso a la hora de realizar consultas, se han guardado dos grupos de documentos *XML*: los documentos de filtrado por cada uno de los artículos del corpus y los documentos de análisis de cada una de las consultas.

#### 3.2.1. Corpus.

Estos documentos *XML* se dividen en tres partes: geográfica, temporal y de entidades (figura 2).

```
▼<document id="NYT_ENG_20040502.0019">
 ▼<geo>
  ▼<entity>
     <documentType>ancestor</documentType>
     <woeid>23424977</woeid>
     <type>Country</type>
     <name>United States</name>
   </entity>
  ▼<entity>
     <documentType>geographicScope</documentType>
     <woeid>24701772</woeid>
     <type>Zone</type>
     <name>212 New York, NY, US</name>
   </entity>
  ▼<entity>
     <documentType>place</documentType>
     <woeid>2388929</woeid>
     <type>Town</type>
     <name>Dallas, TX, US</name>
   </entity>
  ▼<entity>
     <documentType>ancestor</documentType>
     <woeid>2347591</woeid>
     <type>State</type>
     <name>New York</name>
   </entity>
  ▼<entity>
     <documentType>administrativeScope</documentType>
     <woeid>2459115</woeid>
     <type>Town</type>
     <name>New York, NY, US</name>
   </entity>
 </geo>
 ▼<temp>
    <dateDoc>[??:02/05/2004:??.??:??]</dateDoc>
    <date>[X:??/??/??:??.??:??]</date>
    <date>[??:??/3/??:??.??:??]</date>
    <date>[??:??/3/??:??.??:??]</date>
    <date>[G:??/??/??:??.??:??]</date>
    <date>[G:??/??/??:??.??:??]</date>
 </temp>
 ▼<names>
    <name/>
    <name>Jenkens jenkens</name>
    <name>Gilchrist gilchrist</name>
    <name>IRS irs</name>
    <name>IRS irs</name>
    <name>Jenkens jenkens</name>
    <name>Gilchrist gilchrist</name>
    <name>New_York new_york</name>
    <name>David_Deary david_deary</name>
 </names>
</document>
```

Figura 2: Ejemplo de fichero *XML* de filtro creado a partir de un documento del corpus.

**Geográfica:** en este apartado se guarda la parte geográfica relevante del artículo original. Esta información geográfica es extraída mediante *Yahoo! Placemaker*. Concretamente se guardan los siguientes datos por cada uno de los topónimos localizados en el texto: tipo (indica si el topónimo que describe es el ámbito genérico del texto, una entidad administrativa superior, un lugar encontrado en el texto, etc.), *WOEID*, tipo de topónimo (indica si el lugar encontrado es un país, una ciudad, una entidad vaga, etc.), y nombre.

**Temporal:** en este grupo se guardan todas las fechas (en formato normalizado) encontradas en el texto original. Al menos tendrá una entrada, la fecha del artículo, y detrás de ésta vendrán el restos de fechas que se hayan localizado en el texto.

**Entidades:** esta sección es la que recoge todas las entidades no geográficas nombradas en el artículo original.

#### 3.2.2. Consultas.

Se han creado las siguientes secciones por cada una de las consultas enviadas (figura 3):

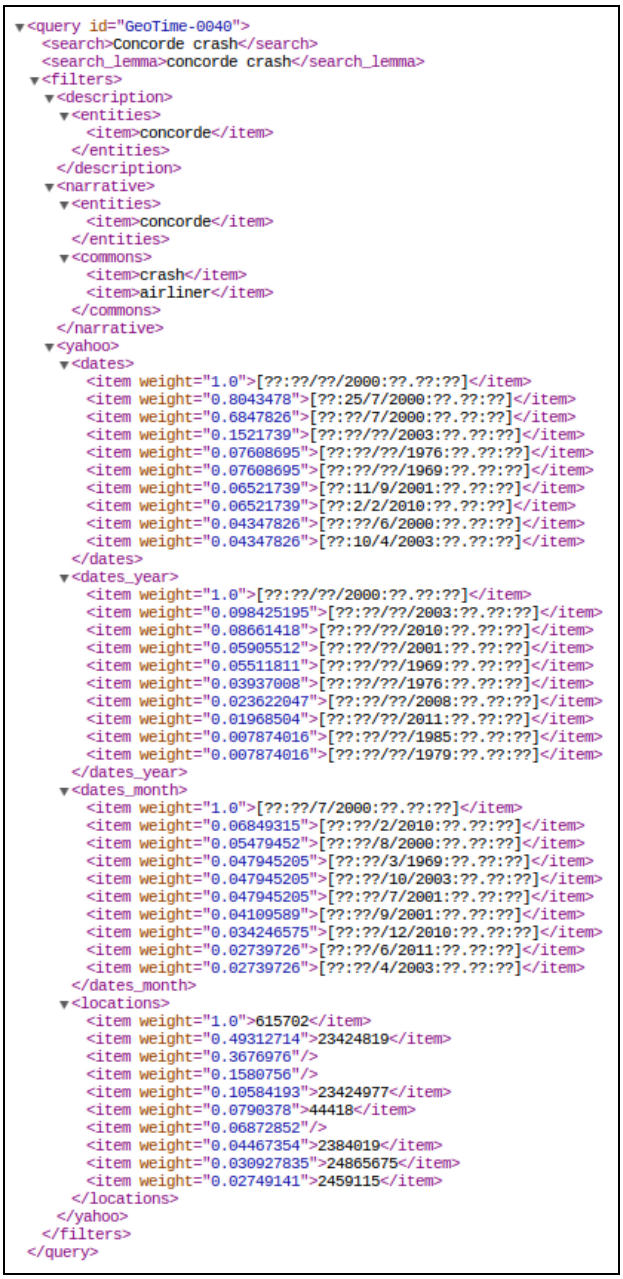
- Términos de búsqueda: Todos los términos de búsqueda, sin *stopwords*.
- Términos de búsqueda lematizados: lo mismo del apartado anterior pero lematizado.
- Filtros:
  - Parte descriptiva: fechas, topónimos y entidades encontradas en la parte descriptiva de la consulta.
  - Parte narrativa: análogamente, contendrá los mismos tres apartados descritos en la parte descriptiva más restricciones de topónimos y temporales.
  - Extensión de consulta: contendrá entradas expandidas de los términos más representativos de la consulta para una posible extensión de la misma.
  - Q&A: parte que contendrá los datos extraídos de la búsqueda de respuestas mediante *Yahoo!*, como fechas (completas o incompletas), fechas con mes y año, fechas con año, y topónimos. Tendrá los 10 valores más significativos normalizados por cada uno de los datos mencionados anteriormente.

### 3.3. Funcionamiento del sistema

El funcionamiento del sistema *GIR* se divide en tres fases: una inicial que se encarga de indexar y preprocesar todo el corpus, la segunda que procesa las consultas, y una final que es la encargada de ejecutar dichas consultas.

#### 3.3.1. Preprocesado del corpus.

En esta fase se indexa el corpus lematizado con *Lucene* y *Terrier*, se obtienen las entidades geográficas con *Yahoo! Placemaker*, y se obtienen las entidades nombradas y temporales con



```
<query id="GeoTime-0040">
  <search>Concorde crash</search>
  <search_lemma>concorde crash</search_lemma>
  <filters>
    <description>
      <entities>
        <item>concorde</item>
      </entities>
    </description>
    <narrative>
      <entities>
        <item>concorde</item>
      </entities>
      <commons>
        <item>crash</item>
        <item>airliner</item>
      </commons>
    </narrative>
    <yahoo>
      <dates>
        <item weight="1.0">[??:??/??/2000:??.??:??]</item>
        <item weight="0.8043478">[??:25/7/2000:??.??:??]</item>
        <item weight="0.6847826">[??:??/7/2000:??.??:??]</item>
        <item weight="0.1521739">[??:??/??/2003:??.??:??]</item>
        <item weight="0.07608695">[??:??/??/1976:??.??:??]</item>
        <item weight="0.07608695">[??:??/??/1969:??.??:??]</item>
        <item weight="0.06521739">[??:11/9/2001:??.??:??]</item>
        <item weight="0.06521739">[??:2/2/2010:??.??:??]</item>
        <item weight="0.04347826">[??:??/6/2000:??.??:??]</item>
        <item weight="0.04347826">[??:10/4/2003:??.??:??]</item>
      </dates>
      <dates_year>
        <item weight="1.0">[??:??/??/2000:??.??:??]</item>
        <item weight="0.098425195">[??:??/??/2003:??.??:??]</item>
        <item weight="0.08661418">[??:??/??/2010:??.??:??]</item>
        <item weight="0.05905512">[??:??/??/2001:??.??:??]</item>
        <item weight="0.05511811">[??:??/??/1969:??.??:??]</item>
        <item weight="0.03937008">[??:??/??/1976:??.??:??]</item>
        <item weight="0.023622047">[??:??/??/2008:??.??:??]</item>
        <item weight="0.01968504">[??:??/??/2011:??.??:??]</item>
        <item weight="0.007874016">[??:??/??/1985:??.??:??]</item>
        <item weight="0.007874016">[??:??/??/1979:??.??:??]</item>
      </dates_year>
      <dates_month>
        <item weight="1.0">[??:??/7/2000:??.??:??]</item>
        <item weight="0.06849315">[??:??/2/2010:??.??:??]</item>
        <item weight="0.05479452">[??:??/8/2000:??.??:??]</item>
        <item weight="0.047945205">[??:??/3/1969:??.??:??]</item>
        <item weight="0.047945205">[??:??/10/2003:??.??:??]</item>
        <item weight="0.047945205">[??:??/7/2001:??.??:??]</item>
        <item weight="0.04109589">[??:??/9/2001:??.??:??]</item>
        <item weight="0.034246575">[??:??/12/2010:??.??:??]</item>
        <item weight="0.02739726">[??:??/6/2011:??.??:??]</item>
        <item weight="0.02739726">[??:??/4/2003:??.??:??]</item>
      </dates_month>
      <locations>
        <item weight="1.0">615702</item>
        <item weight="0.49312714">23424819</item>
        <item weight="0.3676976"/>
        <item weight="0.1580756"/>
        <item weight="0.10584193">23424977</item>
        <item weight="0.0790378">44418</item>
        <item weight="0.06872852"/>
        <item weight="0.04467354">2384019</item>
        <item weight="0.030927835">24865675</item>
        <item weight="0.02749141">2459115</item>
      </locations>
    </yahoo>
  </filters>
</query>
```


Figura 3: Ejemplo de fichero *XML* creado a partir de una consulta.

*FreeLing*. Con toda esta información se crea un fichero *XML* por cada documento del corpus, que se empleará a la hora de valorar la relevancia de los documentos con respecto a la consulta en la fase de ejecución (figura 2).

#### 3.3.2. Procesado de las consultas.

En esta segunda fase, las consultas son enviadas al módulo de análisis lingüístico para obtener la información descrita anteriormente en dicho módulo. Una vez finalizado el trabajo en el módulo de análisis lingüístico, el sistema envía la consulta al módulo de Q&A. Seguidamente, con todos los resultados obtenidos de los dos módulos anteriores, el sistema envía cada una de las referencias geográficas encontradas al módulo geográfico, el cual transformará cada una de estas referencias en el identificador unívoco de *Yahoo! Placemaker* (*WOEID*). Por último, el sistema almacena todos estos datos creando un nuevo documento XML por cada una de las consultas leídas (figura 3).

#### 3.3.3. Ejecución.

En esta tercera y última fase, el sistema obtiene los archivos *XML* de las consultas y, junto con los *XML* de los documentos relevantes recuperados se las envía al módulo evaluador que ejecutará la tarea descrita en dicho módulo. Una vez finalizada la reordenación de los documentos relevantes para cada una de las consultas, el módulo evaluador guardará los resultados en un fichero de soluciones.

## *4. Experimentos y evaluación*

En esta sección se describirán por un lado las métricas utilizadas para la evaluación del sistema así como el entorno donde se realizó dicha evaluación, y por otro lado el impacto de cada uno de los componentes del sistema.

### 4.1. Métricas y entorno de evaluación

A continuación se puede ver dónde se ha evaluado el sistema descrito en este trabajo y se razonará la elección de una métrica de evaluación de los resultados.

#### 4.1.1. Entorno de evaluación

Por un lado, en una primera fase, se ha evaluado el sistema aquí descrito con las colecciones de documentos y preguntas de la tarea *GeoTime* del *NTCIR 2010*. Esta tarea combina *GIR* con búsqueda basada en el tiempo para encontrar eventos específicos en una colección de documentos. El conjunto de evaluación lo forman un corpus de documentos (artículos del *New York Times 2002-2005*) y un corpus de 25 preguntas, las cuales incluyen una parte descriptiva (p. ej. "*When and where did Astrid Lindgren die?*") y una parte narrativa ("*The user wants to know when and in what city the children's author Astrid Lindgren died.*"). Al tratarse de un *workshop* oriental, los organizadores trataron de organizar las tareas haciendo referencia a cualquier tipo de lengua asiática, aceptando como única lengua occidental el inglés, ya que es la lengua común para este tipo de campo de investigación. Debido a que no dominábamos ninguna lengua oriental, se optó por trabajar únicamente en inglés.

En una segunda fase, se evaluó el sistema en la tarea *GeoTime* del *NTCIR 2011*. Para esta nueva edición del *NTCIR 2011 GeoTime* se emplearon otras 25 preguntas con las mismas características

que las de la edición del año anterior. En cuanto al corpus empleado, continuaron empleándose los artículos del *New York Times 2002-2005* y se añadieron artículos de 3 corpora más, *Mainichi Daily 1998-2001*, *Korea Times 1998-2001* y *Xinhua English 1998-2001* y, al igual que con la tarea del año anterior, se optó por trabajar únicamente con corpora y consultas en inglés.

### 4.1.2. Métricas de evaluación

Para el análisis de los resultados expuestos en esta memoria se ha optado por la utilización de la métrica *nDCG*[13] (*normalized Discounted Cumulative Gain*) (Järvelin y Kekäläinen, 2002) utilizada en el *NTCIR 2010* y en el *NTCIR 2011* para poder comparar los sistemas participantes entre sí. Se ha optado por esta métrica de entre las tres empleadas en el *NTCIR* (sección 2.7), ya que ésta es una de las que son capaces de hacer evaluaciones graduales, es decir, que no son simplemente decisiones binarias entre válido y no válido, como hace la métrica *AP*, por ejemplo, si no que es capaz de evaluar un documento parcialmente relevante, como se hizo en las dos ediciones del *workshop* mencionado. Estas evaluaciones podían ser: completamente relevante, relevante geográficamente, relevante temporalmente, relevante de algún modo, e irrelevante. En el *NTCIR-GeoTime* esta métrica fue utilizada en tres bases diferentes: 10, 100 y 1.000. Se ha escogido la base 1.000 en los experimentos aquí mostrados (la misma base que el número de documentos recuperados por consulta), lo que significa que no se está teniendo en cuenta la posición de los documentos recuperados, si no que se está considerando si los documentos recuperados son de algún modo relevantes o no (ganancia acumulativa). Esto se ha hecho debido a que nos estamos centrando más en obtener el mayor porcentaje posible de documentos relevantes, más que en el orden correcto, ya que esto último se realizará en un trabajo futuro donde se emplearán los módulos ya existentes para obtener una puntuación por documento más precisa de los documentos recuperados inicialmente por el módulo del motor de búsqueda.

## 4.2. Impacto de los componentes

Los experimentos fueron elaboradas según los esquemas de pesados descritos en el módulo de reordenación (sección 3.1.7 en página 7). Para ambos esquemas fueron asignando distintos valores a las variables $\lambda$, $\alpha$ y $\beta$, realizado un recorrido sistemático sobre estos valores en los intervalos [0,1] en incrementos de 0,1.

### 4.2.1. Motor de búsqueda

Como ya se ha comentado en la sección 4.1.1, en una primera fase de la experimentación las consultas y el corpus utilizados fueron los del *NTCIR 2010*. Para esta fase, el motor de búsqueda utilizado fue únicamente *Lucene*, obteniendo como resultado los mostrados en la figura 4 para los mejores valores de los parámetros $\alpha$ y $\beta$ de los esquemas de pesado vistos en la sección 3.1.7.

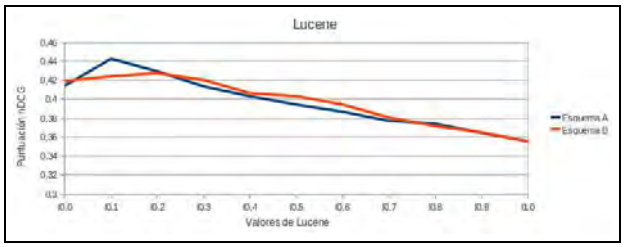



Figura 4: Gráfica de resultados de los esquemas de pesado A y B, desde la no utilización del ranking de *Lucene* (valor 0), hasta la utilización única y completa de éste (valor 1).

En la segunda fase de experimentación, la que tuvo lugar con los corpora del *NTCIR 2011*, se observó que la cobertura alcanzada por *Lucene* apenas superaba el 50 %, por lo que se llevo a cabo un experimento para comprobar qué hubiera sucedido si se hubiese alcanzado una mayor cobertura, cuyos resultados se pueden observar en la tabla 1. Estos resultados han sido clasificados en tres grupos:

1. Consultas que obtienen una cobertura entre el 0 % y el 100 %, es decir, todas las consultas.
2. Consultas que obtienen una cobertura entre el 50 % y el 100 % (12 consultas).
3. Consultas que obtienen una cobertura entre el 75 % y el 100 % (10 consultas).

En cada uno de estos tres grupos se pueden observar el tanto por cien de la cobertura de documentos relevantes recuperado por cada consulta, y la puntuación *nDCG-1000* alcanzada para la mencionada consulta. Finalmente, se obtiene la cobertura y puntuación media para las consultas que entran en cada uno de los tres grupos. Como ya se ha mencionado anteriormente, el objetivo de este experimento era comprobar lo que sucedería si se hubiese obtenido una mayor cobertura por parte del módulo del motor de búsqueda,

[13] *nDCG* mide la media normalizada de utilidad, o ganancia, de un documento basado en la posición en el ranking final de resultados. La ganancia es acumulativa desde lo más alto de la lista de resultados hasta el final, con la ganancia de cada resultado descontada al nivel que le sucede.

Tabla 1: Cobertura y puntuación *nDCG-1000* alcanzada utilizando únicamente *Lucene* por cada una de las consultas del *NTCIR 2011*.

| | 0 % - 100 % | | 50 % - 100 % | | 75 % - 100 % | |
|---|---|---|---|---|---|---|
| Topic | Cobertura | nDCG | Cobertura | nDCG | Cobertura | nDCG |
| GeoTime-0026 | 93,2945 % | 0,7730 | 93,2945 % | 0,7730 | 93,2945 % | 0,7730 |
| GeoTime-0027 | 85,7143 % | 0,2576 | 85,7143 % | 0,2576 | 85,7143 % | 0,2576 |
| GeoTime-0028 | 85,4839 % | 0,5846 | 85,4839 % | 0,5846 | 85,4839 % | 0,5846 |
| GeoTime-0029 | 43,3566 % | 0,2806 | - | - | - | - |
| GeoTime-0030 | 66,6667 % | 0,3467 | 66,6667 % | 0,3467 | - | - |
| GeoTime-0031 | 36,6667 % | 0,2905 | - | - | - | - |
| GeoTime-0032 | 35,0877 % | 0,3367 | - | - | - | - |
| GeoTime-0033 | 74,4186 % | 0,5660 | 74,4186 % | 0,5660 | - | - |
| GeoTime-0034 | 86,3636 % | 0,4655 | 86,3636 % | 0,4655 | 86,3636 % | 0,4655 |
| GeoTime-0035 | 28,5714 % | 0,1031 | - | - | - | - |
| GeoTime-0036 | 31,9149 % | 0,2849 | - | - | - | - |
| GeoTime-0037 | 0,0000 % | 0,0000 | - | - | - | - |
| GeoTime-0038 | 1,6908 % | 0,0317 | - | - | - | - |
| GeoTime-0039 | 84,1202 % | 0,6174 | 84,1202 % | 0,6174 | 84,1202 % | 0,6174 |
| GeoTime-0040 | 82,0755 % | 0,7887 | 82,0755 % | 0,7887 | 82,0755 % | 0,7887 |
| GeoTime-0041 | 98,9362 % | 0,7117 | 98,9362 % | 0,7117 | 98,9362 % | 0,7117 |
| GeoTime-0042 | 1,2739 % | 0,0145 | - | - | - | - |
| GeoTime-0043 | 91,4894 % | 0,5294 | 91,4894 % | 0,5294 | 91,4894 % | 0,5294 |
| GeoTime-0044 | 28,5714 % | 0,1920 | - | - | - | - |
| GeoTime-0045 | 75,0000 % | 0,6110 | 75,0000 % | 0,6110 | 75,0000 % | 0,6110 |
| GeoTime-0046 | 92,3077 % | 0,7454 | 92,3077 % | 0,7454 | 92,3077 % | 0,7454 |
| GeoTime-0047 | 6,6667 % | 0,0174 | - | - | - | - |
| GeoTime-0048 | 47,9167 % | 0,4963 | - | - | - | - |
| GeoTime-0049 | 60,0000 % | 0,6509 | 60,0000 % | 0,6509 | - | - |
| GeoTime-0050 | 57,1429 % | 0,2031 | 57,1429 % | 0,2031 | - | - |
| **Cobertura media** | **55,7892 %** | | **80,9295 %** | | **87,4785 %** | |
| **Puntuación media** | **0,3959** | | **0,5607** | | **0,6081** | |

pudiéndose apreciar la sustancial mejora obtenida en las dos últimas filas de la tabla 1 (pasando de una puntuación de 0,3959 a 0,5607 o 0,6081, según la cobertura mínima requerida).

Observando dichos resultados, y basándonos en el trabajo realizado por (Perea-Ortega, 2010), se incorporó un motor de búsqueda adicional, *Terrier*. Mediante la utilización de ambos motores de búsqueda, se pasó del 55,7892 % de cobertura al 87,0165 %. Dicha cobertura hizo que la puntuación obtenida pasara de 0,3959 a 0,5921 (*nDCG-1000*) utilizando únicamente los motores de búsqueda.

### 4.2.2. Módulo de análisis lingüístico

En lo que al módulo de análisis lingüístico respecta, como ya se ha mencionado previamente en la sección 3.1.2, es el encargado de procesar sintácticamente el contenido de cada uno de los artículos del corpus, por lo que para evaluar su funcionamiento se optó por ver cuan importante era la parte descriptiva de las consultas respecto a la narrativa. Para ello se utilizó el esquema de pesado $A$ visto en la ecuación 1, asignándole el

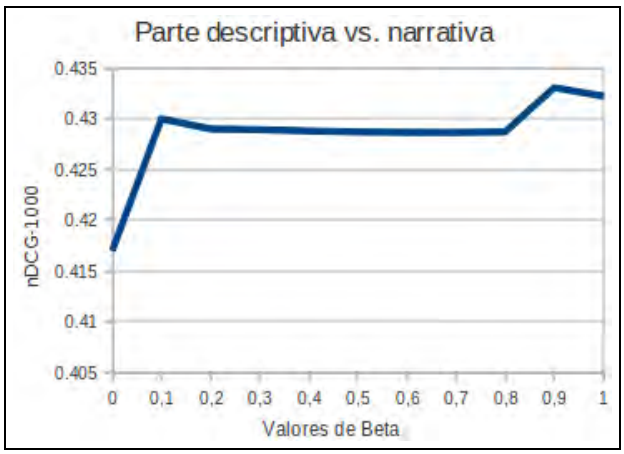



Figura 5: Importancia de la parte narrativa de la consulta contra la descriptiva sobre las consultas del *NTCIR 2011*.

valor que mejor resultados obtuvo a la variable $\alpha$ e incrementando gradualmente los valores de $\beta$ desde 0 hasta 1, tal y como se puede apreciar en la figura 5. Esta gráfica nos muestra cómo es sustancialmente más relevante para nuestro sistema la parte descriptiva de la consulta frente a la narrativa, lo que conduce a pensar que este módulo debe ser mejorado para extraer así mejor las características más importantes descritas en la parte narrativa de la consulta, las cuales, ahora mismo no se tienen en cuenta.

#### 4.2.3. Módulo de Q&A

Sobre los resultados del *NTCIR 2011*, y utilizando únicamente el motor de búsqueda *Lucene*, se realizó un experimento para evaluar la importancia del módulo de Q&A. Para ello se utilizó el esquema de pesado A (ecuación 1) explicado en la sección 3.1.7, asignándole el valor que mejores resultados obtiene a la variable $\beta$ y variando los valores de $\alpha$ para comprobar dicha importancia. Cómo se puede ver en la figura 6, para el intervalo de valores que va de 0,2 a 0,9 los resultados se mantienen muy igualados, lo que parece decir que el módulo de Q&A tiene importancia aunque no sea crucial en el resultado final.

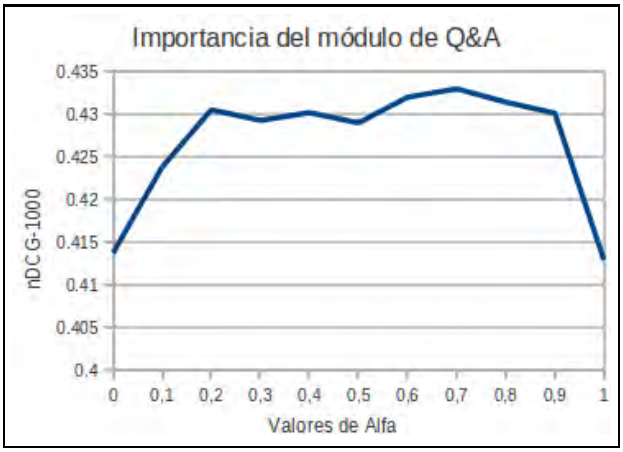



Figura 6: Importancia del módulo de Q&A sobre las consultas del *NTCIR 2011*.

Se realizó un estudio más exhaustivo sobre este módulo y se observó que los documentos *XML* creados tras el tratamiento de las consultas (figura 3), en el apartado que concierne a este módulo, en la inmensa mayoría de las ocasiones se contestaba a la parte temporal y/o geográfica de la consulta, por lo que se decidió llevar a cabo un experimento donde se expandiera la consulta lanzada al motor de búsqueda *Lucene* los 10 términos del apartado de fechas, completas o incompletas (*dates*), y los 10 términos del apartado de topónimos (*locations*), todos ellos con sus respectivos pesos. Posteriormente, se unirían los documentos devueltos por *Lucene* con los devueltos por *Terrier*, tal y cómo se explicó en la sección 3.1.1. Como resultado de este experimento se pasó de una puntuación *nDCG-1000* de 0,5921 a 0,6206.

#### 4.2.4. Módulo de análisis temporal y módulo geográfico

Se ha comprobado el peso que tiene el módulo de análisis temporal en el sistema frente al módulo geográfico. Dicha comprobación se ha realizado en el marco de las consultas del *NTCIR 2011*, utilizando para ello el esquema de pesado $B$ (sección 3.1.7) reflejado en la ecuación 2. Para este análisis, se le ha asignado el mejor valor obtenido para la variable $\alpha$ y se ha ido incrementando gradualmente el valor de la variable $\beta$ en el intervalo que va desde 0 hasta 1, correspondiendo el valor 0 a la utilización exclusiva de la parte temporal frente a la geográfica y el valor uno lo contrario. La figura 7 muestra los resultados, donde se puede observar cómo la parte geográfica adquiere una mayor importancia que la temporal.

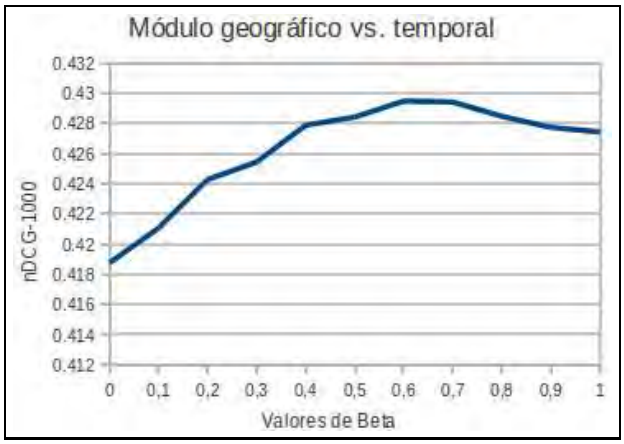



Figura 7: Importancia de los módulo geográfico y temporal sobre las consultas del *NTCIR 2011*.

#### 4.2.5. Módulo para la detección de entidades

Uno de los análisis más interesantes que cabe hacer de los resultados es el del comportamiento de las consultas individualmente para la mejor configuración del sistema.

Dentro del marco del *NTCIR 2010*, en la figura 8, se puede apreciar como hay algunas consultas para las que el sistema obtuvo un comportamiento excepcional, mejorando claramente al mejor de los sistemas de los que participaron en el *GeoTime 2010*. Estos resultados suelen darse en consultas que tienen alguna entidad (nombres de persona, de organizaciones, etc.) poco común, es decir, que aparecen rara vez en el corpus dado. De igual manera, hay algunas consultas para las cuales el sistema obtiene un pobre resultado. Estos acostumbran a darse cuando se encuentran

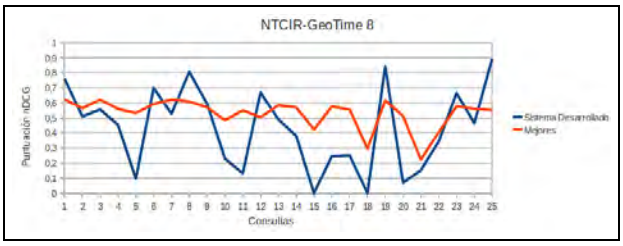



Figura 8: Gráfica de resultados por cada una de las 25 consultas del *NTCIR 2010*, comparando el sistemas expuesto con los mejores resultados para estas consultas de entre todos los participantes.

términos muy comunes en las consultas.

Debido a lo anteriormente expuesto, se decidió obtener la gráfica que mostrara la importancia que tiene el módulo de detección de entidades sobre las consultas del *NTCIR 2011*, cuyo resultado se puede observar en la figura 9. Dicha gráfica ha sido obtenida mediante la utilización del esquema de pesado *B*, plasmado en la ecuación 2, asignándole su mejor valor a la variable $\alpha$, y obteniendo valores para la variable $\beta$ dentro del intervalo 0,1. La gráfica nos muestra la gran importancia que tiene el trato de las entidades (excluyendo las geográficas y las temporales) en el corpus. Se puede apreciar como a medida que dejamos de utilizar exclusivamente las entidades detectadas (valor $\beta = 0$) para ir utilizando únicamente las entidades geográficas y temporales (valor $\beta = 1$), los resultados empeoran pese a ser unas consultas geo-temporales.

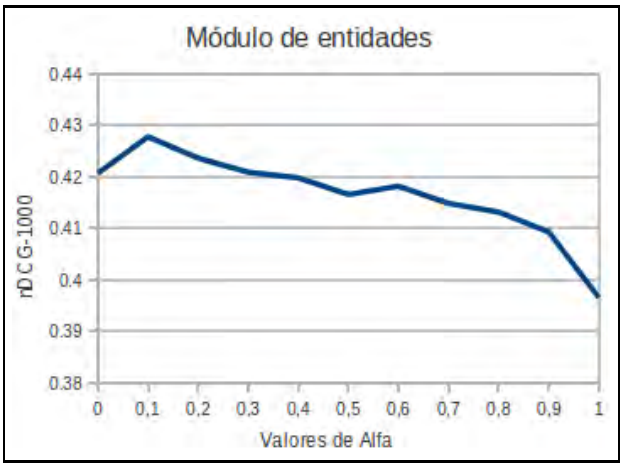



Figura 9: Importancia del módulo de detección de entidades sobre las consultas del *NTCIR 2011*.

### 4.2.6. Módulo de reordenación

La importancia de este módulo puede verse reflejada en la figura 4, donde se puede ver cómo al reordenar los resultados para las consultas del *NTCIR 2010* devueltas por *Lucene*, mejoran los resultados.

Por otro lado, al introducir las técnicas de Q&A descritas en el experimento llevado a cabo en la sección 4.2.3, y después de añadir el motor de búsqueda *Terrier*, tal y como se pudo ver en la experimentación realizada en 4.2.1, la reordenación ha sido implementada realizando una intersección de los resultados devueltos por ambos motores de búsqueda, tal y como se se explicó en la sección 3.1.1, por lo que prácticamente la reordenación la llevan a cabo dichos motores. Esta reordenación es crucial para cualquier sistema de *IR*, por lo que no se puede plantear la existencia de un sistema *GIR* sin la existencia de ésta.

## 5. *Conclusiones y trabajo futuro*

En este trabajo se ha realizado una introducción al campo de la recuperación de información geográfica, implementando un sistema *GIR* propio y evaluándolo en el contexto de una competición internacional como es el *NTCIR*.

En primer lugar, cabe destacar la gran mejora que realiza el sistema con todos sus módulos con respecto a un simple sistema de *IR*, alrededor de 10 puntos porcentuales (figura 4), por lo que, para consultas que contengan un perfil geográfico, dicho sistema supone un gran aumento en la relevancia de los resultados. Dicha mejora se ha conseguido utilizando prácticamente sólo la parte descriptiva de las consultas, ya que la parte narrativa de las mismas introducía ruido en abundancia. En un futuro habría que mejorar el módulo de análisis lingüístico para que sea capaz de extraer y/o filtrar mejor la información de la parte narrativa de las consultas.

Si nos centramos en los módulos del sistema, cabe destacar el gran funcionamiento que ha tenido el módulo de Q&A. Observando esto, es obligatorio resaltar que como uno de los posibles trabajos futuros habría que investigar técnicas más avanzadas en Q&A.

Por otro lado, como se ha podido ver en la sección 4.2.4, el módulo temporal ha tenido menor peso que el geográfico. Debido a esto se está sopesando la incorporación al sistema de un módulo temporal más completo, para lo cual se está barajando la opción del sistema *TIPSem*[14] desarrollado en el *GPLSI* de la *Universidad de Alicante*.

Centrándonos en el módulo puramente geográfico, actualmente se tienen dos frentes abiertos. Por un lado, la obtención de más metadatos de *Yahoo! Placemaker*, como pueden ser el ámbito general del que habla el documento analizado (la huella o *footprint*) para su posterior tratamiento.

El objetivo final sería el implementar un siste-

[14] `http://gplsi.dlsi.ua.es/demos/TIMEE/`

ma completo que satisfaga todos los puntos descritos en la sección 2.

## *6. Agradecimientos*

Esta investigación ha sido parcialmente financiada por el gobierno de España bajo del proyecto TEXTMESS 2.0 (TIN2009-13391-C04-01), y por la Universidad de Alicante bajo el proyecto GRE10-33.

## *Bibliografía*

Baeza-Yates, Ricardo y Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*, volumen 463. Addison Wesley.

Cardoso, Nuno y Diana Santos. 2007. To separate or not to separate : reflections about current gir practice. *English*.

Clough, Paul. 2005. Extracting metadata for spatially-aware information retrieval on the internet. En Chris Jones y Ross Purves, editores, *GIR*, páginas 25–30. ACM.

Järvelin, Kalervo y Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October.

Jones, C B, R S Purves, P D Clough, y H Joho. 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1065.

Jones, Christopher B y Ross S Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones, Christopher Bernard, A Arampatzis, P Clough, y R S Purves. 2007. The design and implementation of spirit: a spatially-aware search engine for information retrieval on the internet.

Leveling, J y S Hartrumpf. 2007. On metonymy recogntion for geographic ir. En *Proceedings of GIR2006 the 3rd Workshop on Geographical Information Retrieval*.

Li, Yi, Alistair Moffat, Nicola Stokes, y Lawrence Cavedon. 2006. Exploring probabilistic toponym resolution for geographical information retrieval. En Ross Purves y Chris Jones, editores, *GIR*. Department of Geography, University of Zurich.

Mitamura, Teruko, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin yew Lin, Ruihua Song, Chuan jie Lin, Tetsuya Sakai, y Donghong Ji Noriko K. 2008. Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access.

Montello, D R, M F Goodchild, Jonathon Gottsegen, y Peter Fohl. 2003. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition Computation*, 3(2):185–204.

Perea-Ortega, J.M. 2010. Recuperación de información geográfica basada en múltiples formulaciones y motores de búsqueda. *Procesamiento del Lenguaje Natural. N. 46 (2011). ISSN 1135-5948*, páginas 131–132.

Robertson, Stephen E, Steve Walker, y Micheline Hancock-Beaulieu, 1998. *Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive*, páginas 199–210. NIST.

Silva, Mário J, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, y N Cardoso. 2006. Adding geographic scopes to web resources. *Computers Environment and Urban Systems*, 30(4):378–399.

Vaid, S, Christopher Bernard Jones, H Joho, y M Sanderson. 2005. Spatio-textual indexing for geographical search on the web. *Advances in Spatial and Temporal Databases 9th International Symposium SSTD 2005*, 3633:218–235.

Van Kreveld, Marc, Iris Reinbacher, Avi Arampatzis, y Roelof Van Zwol. 2005. Multidimensional scattered ranking methods for geographic information retrieval*. *Geoinformatica*, 9:61–84, March.

Wang, Chuang, Xing Xie, Lee Wang, Yansheng Lu, y Wei-Ying Ma. 2005. Detecting geographic locations from web resources. En Chris Jones y Ross Purves, editores, *GIR*. ACM.

Zhang, Vivian Wei, Benjamin Rey, y Rosie Jones. 2006. Geomodification in query rewriting. En *In Proceeding of the 3 rd workshop of Geographic Information Retrieval*.

# Apresentação de Projectos

# Uma incursão pelo universo das publicações em Portugal

Diana Santos
Linguateca, FCCN & Universidade de Oslo
d.s.m.santos@ilos.uio.no

Fernando Ribeiro
Linguateca, FCCN
fernando.ribeiro@fccn.pt

**Resumo**

Neste artigo descrevemos um projeto de colaboração entre a Linguateca e o RCAAP (Repositório Científico de Acesso Aberto de Portugal) no sentido de determinar a possibilidade de melhorar a procura no meta-repositório deste último com ferramentas de processamento da língua portuguesa. Após uma breve apresentação do projeto e da sua motivação nas duas primeiras secções, na secção 3 descrevemos a quantidade de procuras a que tivemos acesso, e nas quais baseamos o estudo, assim como fazemos uma descrição do material depositado no repositório com base em oito recolhas diferentes, no que se refere ao nome dos autores. Prosseguimos descrevendo a análise e processamento dos nomes dos autores (limpeza, normalização e agrupamento), assim como a análise da população de autores nos metadados e nas procuras nas duas secções seguintes, 4 e 5. Com isso identificamos uma série de possíveis grupos de autores, e descrevemos alguns problemas encontrados. Na secção 6, a mais importante do artigo, analisamos as sessões – ou seja, sequências de procuras feitas por um mesmo utilizador a interagir no portal – para verificar se há variação, correção e alteração no nome dos autores dentro de uma sessão. As secções seguintes, 7 e 8 referem-se a assuntos relacionados com a procura em repositórios de publicações, sobre os quais se fizeram pequenas experiências piloto no âmbito do presente projeto, e que permitem ilustrar o quanto ainda estamos aquém de utilizar robustamente quer correção ortográfica quer análise de citações em ambientes realistas, mas que indicam caminhos a seguir. Acabamos a apresentação com uma discussão de possíveis formas de prosseguir, após abordar levemente trabalho relacionado na secção 9.

## *1 Apresentação*

O projeto RCAAP[1] tem como objetivo permitir um fácil acesso às publicações públicas em Portugal. A Linguateca[2], formalmente localizada na mesma instituição infra-estrutural portuguesa, a Fundação Científica para a Computação Nacional (FCCN), iniciou um projeto conjunto com o RCAAP para investigar se as funcionalidades de processamento computacional do português podiam melhorar a experiência dos utilizadores do RCAAP e contribuir para implementar um melhor serviço.

A Linguateca (Santos, 2009), aliás, há vários anos que tem tentado incentivar a publicação em português, e tem dado apoio à catalogação e publicitação de textos na área, através entre outras coisas do SUPeRB, um sistema de catalogação de publicações com várias funcionalidades pensadas para a língua portuguesa (Cabral, 2007; Cabral, Santos e Costa, 2008).

Neste artigo apresentamos, em relação ao projeto RCAAP, e do ponto de vista da Linguateca,

- o tipo de perguntas a que tentamos responder
- uma descrição do material do RCAAP em termos de nomes de autores
- o que pretendemos fazer em futuras iterações, se for considerado útil pelos financiadores e pelos nossos parceiros no RCAAP.

## *2 Algumas notas de motivação*

A nossa intenção primordial com uma colaboração entre a Linguateca e o RCAAP era estudar e analisar os possíveis problemas que o portal do RCAAP e a procura em geral no material por ele indexado poderia apresentar, e investigar a possível melhoria que ferramentas de processamento da língua portuguesa poderiam oferecer.

Ao contrário de afirmar peremptoriamente que faríamos isto e aquilo e que tal constituiria uma melhoria apreciável, preferimos analisar primeiro a situação e os serviços já oferecidos pelo projeto RCAAP e investigar se o nosso saber-fazer poderia de facto trazer alguma melhoria substancial.

Por nos parecer mais simples e mais rela-

[1] Repositório Científico de Acesso Aberto de Portugal, ver http://www.rcaap.pt.

[2] http://www.linguateca.pt

| Anterior (1) | presente (2) | autores1 | autores2 | comuns | perdidos | novos |
|---|---|---|---|---|---|---|
| Jun 2010 | 15 Fev 2011 | 38532 | 42011 | 30230 | 8302 | 11781 |
| 15 Fev 2011 | 1 Mar 2011 | 42011 | 43977 | 41622 | 389 | 2355 |
| 1 Mar 2011 | 12 Abr 2011 | 43977 | 52541 | 43037 | 940 | 9504 |
| 12 Abr 2011 | 19 Maio 2011 | 52541 | 54705 | 52323 | 218 | 2382 |
| 19 Maio 2011 | 19 Jul 2011 | 54705 | 57874 | 54222 | 483 | 3652 |
| 19 Jul 2011 | 1 Ago 2011 | 57874 | 58425 | 57747 | 127 | 678 |
| 1 Ago 2011 | 5 Set 2011 | 58425 | 46003 | 45557 | 12868 | 446 |
| 5 Set 2011 | 1 Out 2011 | 46003 | 48473 | 43383 | 2620 | 5090 |

Tabela 1: Evolução do número e identidade dos autores distintos no RCAAP, comparando cada recolha com a recolha anterior

cionado com a língua portuguesa (dado que um número relativamente elevado de publicações no RCAAP estão em inglês[3]), começámos por estudar o universo dos autores, ou melhor dos seus nomes, para ver se era possível atribuir com alguma certeza várias identificações a um mesmo autor, desambiguar nomes de autores e corrigi-los. Outra motivação para a concentração específica em nomes de autores foi o facto de a correta identificação e desambiguação de pessoas na rede era na altura (e continua a ser) uma área de investigação também a nível internacional, como por exemplo o WePs (Artiles, Sekine e Gonzalo, 2008; Artiles, Gonzalo e Sekine, 2009) demonstra.

Embora a Linguateca tenha considerável experiência na área do reconhecimento de entidades mencionadas em português, devido sobretudo à organização de duas edições do HAREM – a primeira (Santos e Cardoso, 2007) e a segunda (Mota e Santos, 2008), o problema que tratamos aqui não é de facto esse: não estamos a tentar identificar que cadeias de caracteres são nomes de autores – esse é um dado da interação. Estamos sim a lidar com os problemas contíguos que são o de normalizar e validar presumíveis nomes de autores e o de melhorar a procura num universo de autores. Veja-se a secção 9 para uma discussão dos vários problemas relacionados.

## *3 Dados do projeto*

O RCAAP agrega o material oriundo de várias bases de publicações, e tem um meta-repositório que as une e coordena, e que reúne todos os meses uma nova lista de metadados do conteúdo global no RCAAP.

Além disso, e como está mencionado na página do RCAAP que citamos em seguida, recolhe o conteúdo para facilitar o acesso.

> O RCAAP é portal agregador (meta-repositório) que reúne a descrição (metadados) dos documentos depositados nos vários repositórios institucionais em Portugal. O Portal RCAAP recolhe o texto integral para melhorar o resultado das pesquisas mas não guarda qualquer documento.
>
> Para além de poder pesquisar a produção científica portuguesa, pode optar por pesquisar também a produção científica brasileira que neste momento é composta por vários repositórios e revistas agregados no projecto OASIS.[4]
>
> (`http://www.rcaap.pt/help.jsp`)

Associado a esse meta-repositório existe um diário das pesquisas nele feitas (cujo desenho foi revisto por nós), a que nós temos acesso mensalmente. Não temos contudo acesso às pesquisas feitas em cada repositório, visto que isso é gerido (e guardado ou não) de forma distribuída pelos gestores dos diferentes repositórios.[5]

Numa primeira fase (cujos detalhes podem ser apreciados nos relatórios de Santos e Ribeiro (2010; Santos e Ribeiro (2011)), concentrámo-nos nos autores das publicações. Convém aliás indicar que todos estes dados se encontram disponíveis do sítio `http://www.linguateca.pt/colabRCAAP`, onde são atualizados todos os meses.

Na tabela 1 apresentamos a quantidade de dados que temos, assim como a evolução em termos de número de autores (diferentes) no RCAAP. Convém talvez explicar que, se os nomes de autores também desaparecem (quando

[3]Cerca de um terço, com dois terços em português, as outras línguas tendo um peso muitíssimo reduzido. No OASIS brasileiro (o congénere do RCAAP), para comparação, a proporção é de sete vezes mais publicações em português do que em inglês.

[4]Nota dos autores do artigo: OASIS é o OASIS.BR - Portal Brasileiro de Repositórios e Periódicos de Acesso Livre, IBICT, `http://oasisbr.ibict.br/`.

[5]Sobre a forma como o projeto RCAAP funciona a nível nacional e internacional, consulte-se a documentação produzida por este (Moreira et al., 2010; Carvalho et al., 2010), assim como o sítio do projeto.

se esperaria que apenas aumentassem ao longo do tempo), isso é devido à dinâmica da publicação de acesso aberto, em que aparentemente artigos aceites em revistas ou livros deixam de poder ser obtidos em repositórios de acesso aberto.

Estes valores provêm da análise automática que fazemos aos metadados que o RCAAP nos faculta, não podendo nós saber, por exemplo, a causa da diminuição drástica de nomes de autores – e do consequente previsível decrescimento de objetos no RCAAP – entre Agosto e Setembro do presente ano de 2011.

Mas de qualquer maneira ilustram o tamanho do universo com que temos estado a lidar, assim como a dinâmica no repositório, em que mensalmente podemos dizer que se ganham geralmente dez vezes mais autores do que se perdem, mas que este último número ainda é significativo, correspondendo por exemplo no mês de Outubro de 2011 a 5,7%.

## *4 Agrupamento de nomes de autores*

Para estudar a possibilidade de confusão entre nomes de autores diferentes, e os vários grupos que estes poderiam constituir entre si, levámos a cabo dois tipos de agrupamento diferentes.

Antes disso, procedemos a um processo extensivo de normalização e limpeza de dados, descrito pormenorizadamente em Santos e Ribeiro (2010), de que damos aqui um lamiré:

Por exemplo, nomes de autores com termos como *Universidade*, *colóquio* ou *European* são descartados; números ou datas são removidos; são adicionados pontos a iniciais sem ponto, e nomes com vírgulas são invertidos de forma a obter o nome numa forma canónica, como ilustrado na figura 1.

```
Uzan, C        -> C. Uzan
Colaço, ML     -> M. L. Colaço
Correia, P.J. -> P. J. Correia
Vacas, Joana Malta, 1977
               -> Joana Malta Vacas
Saldanha, L.[1937-1987]
               -> L. Saldanha
Falcão, Amílcar Celta
               -> Amílcar Celta Falcão
Colin, J.-P.   -> J. P. Colin
```

Figura 1: Exemplos de normalização

Depois criámos, no processo de **ambiguação**, todas as variantes possíveis de citar um dado nome em português, através da transformação em iniciais de todos ou apenas alguns constituintes (exceto o último nome), e através da omissão de partículas do nome tais como *e*, *de*, *dos*, etc., assim como a expansão ou abreviação de nomes como *Maria*, *Filho*, e *Júnior* ou *Junior* (*M.ª*, *F.*, *Jr.*) ou títulos incluídos no nome como *Pe.* (*Padre*). Nesse processo, nomes com hífens foram também tratados especialmente.

Para dar uma ideia do impacto desta fase de processamento, mencione-se que, dos 48.473 nomes de autores diferentes nos metadados do RCAAP (no mês de Outubro de 2011), após a normalização ficámos com 47.327, que foram ambiguados para 327.614, ou seja, aumentaram o seu número de um factor de 7.

Procedemos em seguida a dois tipos de agrupamento.[6]

Por um lado, transformámos todos os (nomes dos) autores na sua identificação mínima, ou seja, a inicial do primeiro nome seguido do último nome (que é infelizmente a tradição inglesa), criando assim o que chamámos **grupos de nomes máximos** no sentido de que todos os autores cobertos por esta "inicialização compulsiva" pertencem ao mesmo grupo. Obtivemos assim para o mês de Outubro de 2011, 20.970 grupos de nomes máximos, com uma média de 15,48 elementos por grupo, em que a distribuição pelo tamanho dos grupos se apresenta na tabela 2:

| Tamanho do grupo | Número de grupos |
|---|---|
| 1 | 5342 |
| 2 | 7543 |
| 3 | 374 |
| 4 (ou mais) | 7898 |

Tabela 2: Grupos de nomes máximos

Por outro lado, criámos outra forma de agrupamento a que chamámos **grupos de nomes únicos**, em que pretendemos na medida do possível separar (e como tal identificar) univocamente autores diferentes[7]. Nesse caso, o identificador do grupo é a sequência máxima de carateres que corresponde à maneira mais precisa de identificar uma pessoa, e o resto dos membros do grupo têm de ser versões compatíveis (mas mais ambíguas) dessa cadeia de carateres.

Assim, exemplificando com o nome (nome de grupo) único José João Dias de Almeida, casos como *João David de Almeida* não lhe pertencem, mas *J. D. Almeida* pertence. Pertence, aliás, a ambos os grupos de nomes únicos

[6] De facto, na realidade fizemos três, mas não temos ainda os dados referentes aos terceiro, mais radical, em que também removemos nomes e/ou iniciais.

[7] Naturalmente que estamos conscientes de, que sendo este um processo automático baseado em muito pouco conhecimento, estamos longe de chegar a esse objetivo, como será discutido mais à frente no artigo.

José João Dias de Almeida e João David de Almeida. Todos estes nomes, por outro lado, estão agrupados no mesmo grupo de nomes máximo J. Almeida.

Tal permite-nos apontar as seguintes propriedades, distintas, destes dois tipos de agrupamento: cada nome só pode pertencer a um grupo de nomes máximo, que é de certa forma o mínimo múltiplo comum de todos esses nomes; por outro lado, quanto menos comprido e complicado um nome de autor for, mais provavelmente se poderá integrar em mais grupos de nomes únicos. Assim a caracterização estatística destes dois tipos de grupos dá-nos diferentes propriedades do que convencionámos chamar universo dos autores.

Apresentamos aqui para comparação também a distribuição, em termos de tamanho, dos 38.008 grupos de nomes únicos relativos a este mesmo mês, com uma média de 9,75 nomes por grupo, na tabela 3.

| Tamanho do grupo | Número de grupos |
|---|---|
| 1 | 5.524 |
| 2 | 11.380 |
| 3 | 1.006 |
| 4 (ou mais) | 20.098 |

Tabela 3: Grupos de nomes únicos

Isto significa que em Outubro de 2011 havia 5.524 grupos de nomes únicos com apenas um elemento (ou seja, com uma designação que não é ambígua nesse universo), e que existiam 32.484 grupos com mais de um membro, possivelmente (idealmente) representando diferentes maneiras de a mesma pessoa assinar.

(Se, para efeitos de comparação, fizermos o mesmo processo aos autores brasileiros, e a todos os autores – brasileiros e portugueses – obtemos os dados das tabelas 4 e 5.)

| Tamanho | Grupos em B | Grupos em B e P |
|---|---|---|
| 1 | 1.960.294 | 2.248.567 |
| 2 | 111.543 | 12.137 |
| 3 | 27.801 | 31.508 |
| 4 (ou mais) | 45.729 | 53.601 |

Tabela 4: Grupos de nomes únicos B e P

| Tamanho | Grupos em B | Grupos em B e P |
|---|---|---|
| 1 | 9.951 | 13.956 |
| 2 | 38.928 | 44.134 |
| 3 | 2.001 | 2.341 |
| 4 (ou mais) | 88.359 | 92.007 |

Tabela 5: Grupos de nomes máximos B e P

Apresentamos também, para cada nome no universo dos autores, a sua ambiguidade potencial em termos dos agrupamentos de nomes únicos, ou seja, para cada nome (dos nomes que temos no nosso universo dos autores), a quantos grupos pode pertencer, na tabela 6.

| Número de grupos com que emparelham | Número de casos |
|---|---|
| 1 | 313.299 |
| 2 | 8016 |
| 3 | 2280 |
| 4 (ou mais) | 3830 |

Tabela 6: Ambiguidade relativa a grupos de nomes únicos

Embora possamos fazer várias considerações (e estatísticas) sobre este assunto, na realidade o nosso objetivo é ver como é que o material depositado no RCAAP corresponde (ou não) ao que os utilizadores procuram (e encontram ou não).

Por isso na próxima secção apresentamos o mesmo processo (de normalização), agora referente aos nomes encontrados nos diários da procura, antes de voltar a possíveis interpretações dos valores providenciados.

## *5 Emparelhamento simples com as procuras*

Antes de descrever este emparelhamento, apresentamos sucintamente o material de que dispomos, e que provém da análise dos diários expressamente criados para o efeito pela equipa do RCAAP (cujo formato, mais uma vez, pode ser consultado nos relatórios já mencionados).

A distribuição das consultas (desde que a elas tivemos acesso) encontra-se na figura 2, separadas entre procuras simples e avançadas (neste último caso, os utilizadores especificam campos como autor, assunto, etc).

Vemos que a maior parte dos utilizadores não usa a procura avançada, por campos, mas sim a procura simples, onde é possível também digitar autores. Por agora limitámo-nos à pesquisa avançada para obter sem esforço os autores procurados, embora uma extensão óbvia da nossa investigação seja tentar obter nomes de autores na expressão da pesquisa simples.[8]

Passemos agora ao assunto que nos interessa, nomeadamente: Quão alinhado está o universo

[8]Isto não é tão simples quanto pode parecer à primeira vista, visto que muitos apelidos em português podem ser também nomes comuns, veja-se *oliveira* ou *mota*, e que um utilizador pode procurar um autor com um nome composto "Mota Oliveira" ou um artigo com dois autores, um designado por "Mota" e outro por "Oliveira".

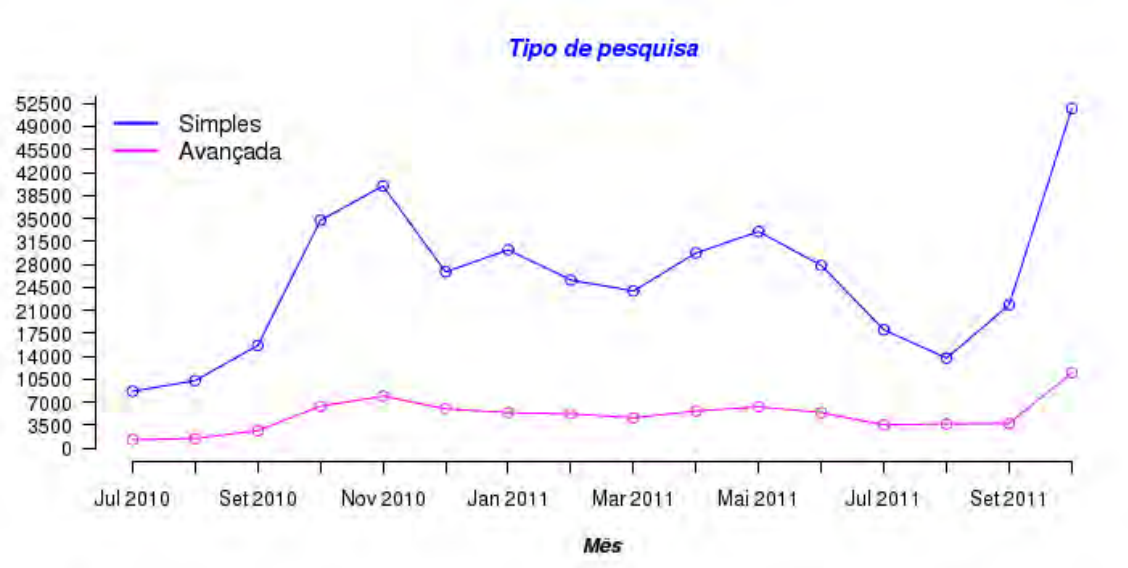



Figura 2: Número de procuras por mês

dos autores no RCAAP com o que os utilizadores procuram? Ou seja, os autores que os utilizadore procuram são os que se encontram no repositio? Para averiguar isso, começamos por caracterizar os autores procurados, verificando em que grupos se podem incluir.

Os resultados – relativos às procuras de Julho de 2010 até Outubro de 2011 inclusive – encontram-se na tabela 7.[9]

| Número de grupos | Número de casos |
|---|---|
| 1 | 3304 |
| 2 | 486 |
| 3 | 244 |
| 4 (ou mais) | 1263 |

Tabela 7: Emparelhamento com grupos de nomes únicos

A tabela seguinte (tabela 8) mostra os mesmos valores, agora em relação aos grupos de nomes máximos:

Em primeiro lugar, apontamos que os casos concretos que correspondem a apenas um grupo – os casos em que há apenas um candidato (ambiguado ou não), e que remontam a 3304 (63%) e 3487 (66%) respetivamente – coincidem na grande maioria dos casos, nomeadamente em 3266, ou seja em 98%/94% dos casos.

[9]Poderíamos empregar vários métodos de emparelhamento, como se vê, e dá a escolher, na página do serviço, mas para o presente artigo usámos o método a que chamámos **sequência exata**, e que apenas emparelha quando a sequência procurada existe sem interrupções no grupo. Ou seja, a procura de *Fernando Ribeiro* emparelha com *José Fernando Ribeiro*, mas não com *Fernando José Ribeiro*.

| Número de grupos | Número de casos |
|---|---|
| 1 | 3487 |
| 2 | 527 |
| 3 | 240 |
| 4 (ou mais) | 956 |

Tabela 8: Emparelhamento com grupos de nomes máximos

Tentemos explicar, com um exemplo concreto, as razões da semelhança entre os dados apresentados para cada tipo de grupos:

Considere-se o caso do nome fictício *Zacarias da Zelândia Martelo Rocambolesco*, que assumimos que, quer no grupo de nomes máximo Z. Rocambolesco, quer no grupo de nomes único Zacarias da Zelândia Martelo Rocambolesco, só apresenta um elemento, esse próprio nome (ou variantes inicializadas do mesmo).

Pode-se imaginar que, num universo sempre em expansão, um dia outros autores com o primeiro nome começado pela letra Z e com o mesmo apelido virão a surgir, mas sendo este um trabalho virado para a realidade concreta e para uma situação presente e não para um universo tendente para o infinito, podemos afirmar que existirão também sempre casos não ambíguos no material, de certa forma corroborando mais uma vez a lei de Zipf (Zipf, 1949).

Seja como for, e para investigar se isso se deve a que a grande maioria das procuras por autor tem um comprimento grande (ou pequeno), tabelámos o comprimento dos nomes procurados, na tabela 9.

Ao fazer esta análise, detetámos que muitas

| Palavras | Iniciais | Frequência |
|---|---|---|
| 2 | 0 | 3822 |
| 3 | 0 | 1915 |
| 1 | 0 | 1784 |
| 4 | 0 | 1343 |
| 5 | 0 | 777 |
| 1 | 1 | 434 |
| 6 | 0 | 303 |
| 1 | 2 | 223 |
| 2 | 1 | 198 |
| 7 | 0 | 114 |
| 2 | 2 | 100 |
| 1 | 3 | 44 |
| 3 | 1 | 38 |
| 4 | 1 | 32 |
| 8 | 0 | 27 |
| 3 | 2 | 25 |
| 2 | 3 | 17 |
| 9 | 0 | 10 |

Tabela 9: Forma dos autores procurados: não descartamos quaisquer palavras, tal como *de* ou *e*.

das cadeias de carateres enviadas como "autores" o não são: de facto, ao analisar manualmente os primeiros 2044 casos de autores não encontrados (ordenados por ordem alfabética), encontrámos 258 casos que não eram certamente autores, e que na sua esmagadora maioria eram palavras chave.

Para indagar por outro lado até que ponto é que as procuras por autores foram bem sucedidas, precisamos de saber o número de resultados que desencadearam, e qual o comportamento do utilizador em face dos mesmos. Para isso precisamos de estudar a interação com o sistema, olhando para as sessões.

Mas antes disso, apresentamos os dados relativos às procuras no meta-repositório: das 543.684 que temos conhecimento, removendo os duplicados correspondem a 542.545 pesquisas diferentes, às quais retirámos além disso as poucas procuras automáticas (feitas por robôs), resultando em 491.376 procuras.

Dessas, apenas 65.018 (13%) não têm nenhum resultado.

Se limitarmos a nossa atenção às procuras que preencheram o campo autor (e que no total são 28.644), as que não obtiveram nenhum resultado são 11.506, ou seja, 40%.

Mas se estes valores correspondem a satisfação do utilizador, ou se a maior parte dos resultados são lixo ou pelo menos precisam de refinamento, é algo não podemos ajuizar considerando cada procura separadamente.

## *6 Estudo das sessões*

Como matéria prima para o nosso estudo temos acesso (dados de 1 de Novembro de 2011) a 85.398 sessões diferentes, ou seja, visitas com mais de uma pesquisa); a 95.448 "utilizadores" (definidos como par IP mais browser) diferentes, representando 132.442 visitas ao RCAAP. Por **visitas** consideramos uma ou mais procuras, a que chamamos respetivamente **visitas unitárias** (46.044), ou **sessões** (com mais de uma pesquisa, portanto). Chamamos **visitas únicas** (que podem ter uma ou mais pesquisas) aos casos em que um utilizador que só comunicou uma vez com o meta-repositório do RCAAP (79.596).

Como talvez não seja preciso salientar, num sistema que não tem autenticação de utilizadores é muito difícil individualizar estes, dados os IP dinâmicos e as "proxies" das instituições, o que significa que, se conseguimos com uma certa precisão identificar sessões, o mesmo não pode ser dito de utilizadores, que repetimos, são meramente pares distintos de IP e identificador do browser.[10]

Por isso os dados mais fiáveis que temos dizem respeito às sessões, veja-se na figura 3 a distribuição das sessões em número de procuras.

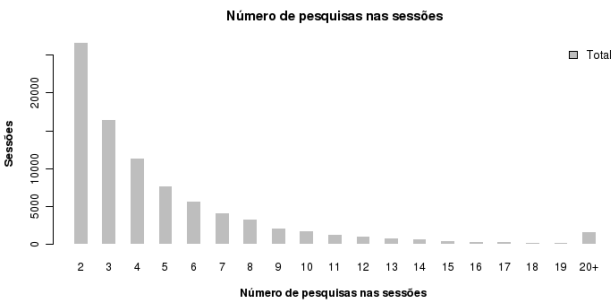



Figura 3: Número de procuras por sessão

As questões a que queremos responder, com uma análise das sessões (ou melhor, com uma análise do que os utilizadores fazem nas suas sessões), é, após consultarmos a distribuição dos resultados obtidos, conseguir pronunciar-nos sobre se o resultado é bom ou mau, ou seja, se os utilizadores conseguiram ou não obter o que pretendiam, e com quanto esforço, através de dados (indiretos) como

- Quantas vezes se refinou (até chegar a algo satisfatório, ou desistir)
- Se não se obteve nenhum resultado, isso é porque a procura foi mal feita ou porque não

[10] Para uma defesa da nossa abordagem de sessões, ver Santos e Frankenberg-Garcia (2007), por sua vez baseado em He e Göker (2000).

existia no RCAAP?

- Se se obteve resultados, mas houve mudanças à procura feita pelo utilizador, o resultado da mesma foi no sentido de aumentar, ou diminuir o número de casos encontrados?

Nas tabelas 10 e 11 contabilizamos os casos encontrados em que é possível apreender uma modificação a uma procura, respetivamente no caso de a sessão ter pesquisas em que exista o campo autor, e em todas as sessões. (A unidade são pares de consultas dentro de uma mesma sessão, e a heurística utilizada para considerar modificação em vez de nova procura não relacionada foi a existência de algo constante em duas procuras consecutivas.)

| Mudança, número de resultados | Frequência |
|---|---|
| PS para PA, diminuição | 5125 |
| PS para PA, aumento | 208 |
| PA para PS, aumento | 1763 |
| PA para PS, diminuição | 69 |
| PA, aumento | 2811 |
| PA, diminuição | 6762 |
| PA, mesmo número | 3081 |
| PS para PA, mesmo número | 75 |
| PA para PS, mesmo número | 37 |

Tabela 10: Resultado de mudanças na expressão de pesquisa em sessões incluindo o campo Autor (PS: pesquisa simples; PA: pesquisa avançada)

Por aqui vemos que na maior parte dos casos os utilizadores estão interessados, ou acabam por, obter menos resultados dos que inicialmente obtiveram, e que existe uma correlação grande entre mudar para a pesquisa avançada para restringir os resultados, e, conversamente, mudar para a pesquisa simples para obter mais resultados. Contudo, é de salientar que uma percentagem significativa de vezes os utilizadores não conseguem modificar efetivamente a pesquisa. Vemos que as generalizações descritas acima têm confirmação num universo maior, com a observação adicional de que alterações dentro da pesquisa simples tendem a ser igualmente frequentes para aumentar ou para diminuir o número de resultados, o que não seria evidente.

Convém contudo salientar que as contagens apresentadas referem-se a pares de consultas, sendo de imaginar que em alguns casos a interação à procura da pesquisa perfeita seja mais numerosa, e que o utilizador tente mais do que uma modificação, ou que apenas desista após variadas tentativas, e que dessa forma alguns dos pares não são independentes, mas sim deveriam ser contados como triplos, quádruplos, n-tuplos de modificação.

| Mudança, número de resultados | Frequência |
|---|---|
| PS para PA, diminuição | 39547 |
| PS para PA, aumento | 5313 |
| PA para PS, aumento | 8744 |
| PA para PS, diminuição | 1253 |
| PA, aumento | 33689 |
| PA, diminuição | 54415 |
| PA, mesmo número de resultados | 23815 |
| PS para PA, mesmo número | 512 |
| PA para PS, mesmo número | 1574 |
| PS, aumento | 12328 |
| PS, diminuição | 12833 |
| PS, mesmo número | 18268 |

Tabela 11: Mudanças na expressão de pesquisa em geral

Refizemos assim as contagens. No caso de sessões (de pesquisa avançada) com nome de autor, verificámos quantas vezes este foi mudado (mas não radicalmente alterado), em casos que não são apenas pares mas sim sequências de duas, três ou mais pesquisas.

Em termos das sessões correspondentes, na tabela 12 mostramos o tamanho em número de pesquisas consecutivas de uma modificação à pesquisa.

| Número de modificações | Frequência |
|---|---|
| 1 | 1375 |
| 2 | 532 |
| 3 | 247 |
| 4 | 99 |
| 5 | 61 |
| 6 | 30 |
| 7 | 18 |
| 8 (ou mais) | 46 |

Tabela 12: Tamanho da interação à volta de uma expressão de pesquisa incluindo o campo Autor

E na tabela 13 indicamos, em termos do resultado (do primeiro para o último caso), os casos em que houve aumento dos resultados, diminuição ou manteve-se o número. Aqui como nos resultados anteriores, é a redução do número

| Mudança nos resultados | Frequência |
|---|---|
| Aumento | 314 |
| Diminuição | 1447 |
| Mesmo número | 647 |

Tabela 13: Resultado da interação à volta de uma expressão de pesquisa incluindo o campo Autor

de resultados que parece ser o objetivo mais importante da reformulação de consultas.

A um nível mais geral, no entanto, o que queremos compreender, é, além de saber quantos autores são ambíguos (ou seja, passíveis de terem várias interpretações), que foi o que tentámos averiguar na secção anterior,

- quantas vezes os autores "ambíguos" têm de ser re-procurados, ou seja a sua especificação refinada na expressão da procura?
- qual é a melhor maneira de os distinguir para um utilizador? Por assunto(s), por data das publicações, pela instituição associada à publicação, pelos seus co-autores?

Para isso, tivemos de observar individualmente uma amostra de casos (pares de consultas relacionadas, em sessões com campo autor) em que houve diminuição de resultados, para ver se essa redução se deveu a uma maior desambiguação do nome do autor. A distribuição das diferentes ações após analisar 60 casos é apresentada na tabela 14. Em alguns (poucos) casos, mais do que uma destas modificações foi executada pelo utilizador. Nesses casos, contabilizámos cada mudança separadamente.

| Mudança na pesquisa | Frequência |
|---|---|
| Aumento no nome do autor | 5 |
| Adição de autor | 26 |
| Adição de autor adicional | 6 |
| Adição de tipo ou data | 12 |
| Mudança para o campo autor | 8 |
| Adição de assunto | 3 |
| Mudança no assunto | 4 |
| Diminuiçao de termos | 3 |

Tabela 14: Análise fina das alterações à pesquisa que levaram a uma diminuição de resultados

Embora o número de casos seja pequeno demais para permitir generalizações, é interessante realçar que há diferenças apreciáveis entre colocar um nome de autor na pesquisa simples ou precisá-lo (por vezes repetindo-o) como nome de autor: os oito casos analisados demonstram reduções de 594 para 31, 86 para 34, 37 para 3 e 97 para 4. Isto parece indicar que um serviço que identificasse cadeias de carateres na procura simples como possíveis autores (constando no repositório), e apresentasse primeiro essa resposta, eventualmente seguida de procura em texto livre, poderia imediatamente evitar algumas reformulações.

Por outro lado, também é interessante verificar que na maior parte das procuras analisadas os utilizadores pareciam ter grande conhecimento do que estavam à procura (sabendo nomes de co-autores, assuntos e mesmo datas). Talvez por isso apenas oito casos redundaram em resultados nulos.

Outra coisa que nos surpreendeu, foi o facto de muitas vezes os utilizadores reformularem a procura mesmo no caso de obterem poucos resultados (menos de dez). Tal leva a suspeitar que, ao contrário de minimizar o seu trabalho, os utilizadores preferem por vezes resultados sem ruído.

Como foi descrito, de uma forma um tanto provocatória, em Santos (2011), para avaliar se vale a pena implementar um sistema que permita uma identificação mais fina e rigorosa de autores (com possível necessidade de mais escolhas e cliques do utilizador), muito provavelmente apenas se conseguirão obter resultados conclusivos após a própria implementação e subsequente comparação entre duas versões.

Isto porque, quando se está em presença de um problema de engenharia (e não puramente científico) é preciso pesar os contras da insatisfação ou descontentamento que os erros de um tal sistema podem também causar. Concretamente, se em 5% dos casos (um em vinte) essa distinção não era relevante para o utilizador, o preço de ter de decidir não seria contraproducente em relação à experiência do utilizador? Algo que pelo menos na situação presente da disciplina da usabilidade, apenas é possível medir com testes com utilizadores.

Os dados de que dispomos por agora – e lembramos que este é um projeto ainda em progresso – parecem apontar para que uma muito pequena percentagem de procuras muda o autor. Das 75.651 sessões, apenas 4.148 contêm algo no campo autor, e dessas apenas cerca de 2.000 o altera.

De qualquer maneira, podemos tambem apresentar, como dados úteis para futuro estudo, a forma dos nomes dos autores procurados e que resultaram em resultados nulos, na tabela 15.

## *7 Correção de nomes de autores*

Um outro problema sobre que nos debruçámos – e que parece receber confirmação de premência dado o grande número de resultados nulos nas pesquisas feitas – corresponde ao utilizador ter digitado o nome de um autor incorretamente.

Embora a correção ortográfica seja uma das áreas mais antigas e mais utilizadas do PLN, a transformação de nomes próprios tem particularidades próprias: em primeiro lugar, porque mesmo numa língua como o português com

| Palavras | Iniciais | Frequência |
|---|---|---|
| 2 | 0 | 2269 |
| 1 | 0 | 1515 |
| 3 | 0 | 1103 |
| 4 | 0 | 616 |
| 5 | 0 | 351 |
| 1 | 1 | 217 |
| 6 | 0 | 149 |
| 1 | 2 | 119 |
| 2 | 1 | 95 |
| 7 | 0 | 66 |
| 2 | 2 | 61 |
| 3 | 1 | 37 |
| 3 | 2 | 21 |
| 4 | 1 | 19 |
| 1 | 3 | 17 |
| 8 | 0 | 13 |
| + | + | 68 |

Tabela 15: Forma dos autores procurados que deram resultados nulos

pouca variabilidade de grafia, nomes próprios provenientes de estádios anteriores da língua, e mesmo de palavras estrangeiras (apelidos), são frequentes. Em segundo lugar porque a correção de uma palavra ou grupo de palavras não pode ser feita com auxílio do contexto (como por exemplo com palavras correntes quando diferentes grafias implicam diferentes categoria gramaticais).

Finalmente, só queríamos corrigir para nomes que se encontrassem já no RCAAP. Não seria de qualquer ajuda obrigar um utilizador a corrigir *Dulcinela Alves* para *Dulce Alves* para depois informar que não havia nenhuma obra dessa autora no RCAAP.

Assim, utilizámos o Jspell (Almeida e Pinto, 1994; Almeida e Pinto, 1995; Simões e Almeida, 2002) criando um novo dicionário apenas com os nomes de autores presentes no RCAAP, e com uma sintaxe especial (visto que inicialmente os dicionários do Jspell não tinham sido desenhados para palavras com pontos e com mais do que uma palavra). Assim, *A. Bonfante* passa a *A@%Bonfante*.

Criámos tambem algumas regras padrão de confusão de letras e sons em português (tais como entre "ç" e "ss", por exemplo), que é uma funcionalidade do Jspell que permite fazer correção baseada em regras.

Também incluímos como correções possíveis nomes em que existem variantes com acento e sem acento, tal como *Raúl* e *Raul*, *Luís* e *Luís*, *Marcirio* e *Marcírio*, ou nomes que têm consoantes mudas e que é possível que uma pessoa que os procure não saiba qual a ortografia com que um dado autorgrafa o nome em questão, tal como *Christina* e *Cristina*, *David*, *Davi* e *Davide*, ou *Raquel* e *Rachel*, *Estela*, *Stella*, e *Estella*. Isto é especialmente necessário para nomes brasileiros, visto que no Brasil não existe uma lista de nomes autorizados, como em Portugal, e portanto a variação é incomparavelmente maior.

Finalmente, esta versão também remove ou adiciona acentos e cedilhas para remediar o problema de utilizadores sem teclado com diacríticos – ou sem paciência para os incluir, habituados que estão aos motores de procura internacionais os ignorarem.

De momento, dos 6600 casos de nomes de autores procurados mas não encontrados, é possível apenas corrigir 692 com a presente versão.

## *8 Experiências preliminares de medição de impacto*

Embora tal não esteja diretamente relacionado com o que nos propusemos fazer nesta primeira colaboração com o RCAAP, não há dúvida de que o ter acesso (se for possível) a uma base tão grande de publicações permite um conjunto de outros estudos de interesse sociológico, de política de investigação, e de relacionamento entre a comunidade de investigadores, estudos aliás que têm sido moda a nível internacional nos últimos tempos, cf. Tsatsaronis et al. (2011) e de que são além disso prova os serviços Arnetminer - Instant Social Graph Search[11], Microsoft Academic Research[12], Google Scholar[13] e Scholarometer[14].

Em particular, entre estas questões existe a funcionalidade, oferecida por alguns sistemas, de estudar a referência/fator de impacto de uma dada publicação, ou de um autor (Couto et al., 2009). Estas medidas são contudo pobres no sentido de que um verdadeiro impacto mede-se pelo uso e possível replicação de uma obra e não na simples menção, por vezes crítica, muitas vezes nem lida, de autores "célebres" ou que seja conveniente citar na comunidade em questão.

Mais relevante, portanto, é a capacidade de estudar a pragmática das próprias referências: ou seja, quantas meramente mencionam ou exemplificam, quantas criticam ou apontam fraquezas, e quantas finalmente usam a obra citada como ponto de partida.

[11] `http://www.arnetminer.org/`
[12] `http://academic.research.microsoft.com/`
[13] `http://scholar.google.com/`
[14] `http://scholarometer.indiana.edu/`

Tomando por exemplo a questão interessante de avaliar a contribuição de Eckhard Bick para o processamento computacional da língua portuguesa, e admitindo que o uso do PALAVRAS (que tem uma citação fixa) é algo que pode ser relativamente fácil de considerar como um indicador desse impacto, já Santos (2011) discutiu o diferente peso que deveria ser dado a formas – fictícias – de o referir, tal como

- *O nosso sistema é semelhante a Bick(2000)*
- *Outros sistemas, tal como Bick(2000)*
- *Ao contrário de Bick(2000)*
- *Usámos Bick(2000)*
- *O nosso sistema é baseado em Bick(2000)*
- *A nossa pesquisa usa o AC/DC*

Usando o GoogleScholar e a funcionalidade de ver o impacto, fizemos um pequeno teste com as referências feitas aos artigos da primeira autora (consulta feita em 22 de Outubro de 2011), que resultou em 81 artigos em forma eletrónica (de notar que a página de impacto refere também artigos e entradas que não se encontram acessíveis electronicamente, e que não foram contados artigos em que ela figurasse como autora, nem mesmo de alunos ou colaboradores seus).

Uma primeira análise manual (com a intenção, naturalmente, de desenvolver requisitos para um programa que o fizesse automaticamente no futuro) indicou

- (Uma grande maioria de) casos em que a citação aparece associada ao nome de um recurso ou de uma iniciativa (em que o HAREM (Santos e Cardoso, 2007; Mota e Santos, 2008), a propósito, é de longe o mais citado)
- 20 casos onde vem associada a uma afirmação ou descrição de trabalho (mais 3 que simplesmente remetem para um artigo para ver questões gerais, tais como *Para os diversos problemas relativos á divulgación de corpus a través da web, véxase Santos (1999)*)
- 3 casos onde é criticado ou menorizado ou contrastado negativamente, tais como *Similar comparisons had already been made (.. ; Santos and Ranchhod, 2002; ...), but in general they were not focused on free software.*
- 4 casos onde apenas aparece na lista de referências sem ser referido no texto
- 2 casos onde nova terminologia é-lhe associada (por outras palavras, é indicada como a origem de um dado termo), tal como *see also Santos 1996 and what she calls "acquisitions").*

Além de aproveitarmos a ocasião para indicar que consideramos que o impacto das diversas publicações foi maior do que a sua citação (contagem de citações) testemunha, dado o número de visitas ao sítio da Linguateca e o levantamento massiço dos recursos criados, notamos de qualquer maneira que estas referências são diferentes entre si em sentido e em peso.

É além disso extraordinário o facto de haver imensos erros nas citações analisadas – erros na ortografia de nomes (tal como *Aries* para *Aires*, *Dianna* em vez de *Diana*), mas sobretudo datas e referências completas – que muitas vezes variam o ano, ou variam a própria referência que seria apropriada. Isso só (quase) a própria autora pode distinguir, mas não deixa de ser digno de menção.

## *9 Trabalho relacionado*

Tanto quanto sabemos, Spink e Jansen (2004) foram dos primeiros a debruçar-se sobre a procura de nomes de pessoas na rede, e nessa altura consideraram que não constituía uma fatia suficientemente importante das procuras.

Quase dez anos mais tarde, o surgimento de muitos sistemas e serviços baseados precisamente num tipo especial de nomes de pessoas (académicos), como os mencionados na secção anterior, parece indicar, das duas uma, ou que houve uma mudança nos hábitos e na cultura internáutica mundial, ou que, mesmo que a nível de motores de procura genéricos esses pedidos não sejam maioritários, fazem-se certamente em serviços especiais. Além de bases de dados de teses e dissertações públicas, e de repositórios institucionais, são também rotineiramente executadas em serviços, ou sítios, como a Wikipédia, como a Linguateca aliás espera poder estudar em breve com o Págico[15] (Costa, Mota e Santos, 2012; Mota, 2012; Simões, Mota e Costa, 2012).

Assim, existem variados artigos (Han et al., 2004; Han e Zhao, 2009; Ferreira et al., 2010; Gong e Oard, 2009; Kern, Zechner e Granitzer, 2011) que se preocupam com a desambiguação dos autores na procura na rede, o que é um objetivo (decomponível em vários problemas distintos) diferente do que abordámos aqui. Assim, podemos apreciar e ver formas de solucionar, na procura na rede

[15] `http://www.linguateca.pt/Pagico/`

- muitos autores com o mesmo nome, que se queiram individualizar para o cálculo de citações, e para encontrar a pessoa certa
- muitas formas de descrever o mesmo autor que se querem juntar, pelas mesmas razões do que o ponto anterior
- o facto de os sistemas de autoridade bibliográfica eles mesmos conterem mutas incorreções
- o facto de uma procura por autores na rede vai produzir um conjunto de respostas em que muitas delas são já o resultado de sistemas agregadores (Tan, Kan e Lee, 2006)
- o facto de autores prolíficos poderem publicar em várias áreas (Sun et al., 2011)

sendo que uma das propostas mais recentes (Qian et al., 2011) advoga o uso de colaboração homem-máquina para resolver o problema.

Outros autores (Dervos et al., 2006; Sun et al., 2011; Tsatsaronis et al., 2011), pelo contrário, dedicam-se a artigos científicos, embora os métodos e as bases documentais sejam bastante diferentes. Por exemplo, Tsatsaronis et al. (2011) categorizam os autores em termos da sua produtividade e do número de co-autores com que publicam, enquanto Han et al. (2004) reparam que o nome da revista em que publicam é a melhor forma de desambiguar autores através de aprendizagem automática.

Está claro que neste aspeto um dos trabalhos pioneiros é o do CiteSeer, continuado pelo CiteSeer X[16], embora arquivos internacionais como o DBLP[17] e a ACL[18], também tenham desenvolvido projetos meritórios e que são usados no dia a dia dos informáticos e linguistas computacionais. Note-se a esse respeito também a Biblioteca Digital Brasileira de Computação[19], que tem, na nossa opinião, uma interação muito feliz em termos de usabilidade precisamente na navegação de autores.

Diferentemente do nosso objetivo, os autores acima citados preocupam-se primordialmente com o estabelecimento de identificações únicas e com a desambiguação entre vários autores que publicam com o mesmo nome, usando geralmente mais informação que o simples nome. O nosso projeto tem sido muito mais modesto e a sua principal preocupação é averiguar o comportamento dos utilizadores para a estudar a própria necessidade deste tipo de processo, no universo português.[20]

Convém, a esse propósito, referir que a tese de mestrado de Luís Miguel Cabral (Cabral, 2007) foi um trabalho valioso, em português, sobre os variados problemas da gestão de um catálogo de referências bibliográficas.

## *10 Observações finais*

Este projeto começou há um ano e meio e tem sido desenvolvido a meio tempo, e o processo de configurar o ımodus faciendi, ao mesmo tempo estabelecendo uma cooperação produtiva com várias outras instituições e grupos (KEEP, Serviços de documentação da Universidade do Minho, e o grupo paralelo do RCAAP na FCCN) com outras prioridades e objetivos, levou também o seu tempo, o que nos leva a afirmar que apenas podemos concluir que nos encontramos no início de um trabalho que, a ser continuado, permite muito mais estudos.

Por um lado, é óbvio que deveremos proceder à análise de outras vertentes da procura: assunto, título, e composição em geral da procura (quais os campos utilizados e como), que poderão ser igualmente elucidativos quer sobre a atividade dos utilizadores como como melhorar o ambiente de procura. A análise de co-autores, e de auto-citações se tivermos acesso aos próprios artigos, permitirá uma especificação muito mais fina dos autores do lado do RCAAP.

Com efeito, e como já discutido informalmente em Santos (2011), o trabalho até agora feito sobre a possibilidade de melhorar a identificação dos nomes dos autores não parece produzir uma melhoria considerável na usabilidade (e consequente satisfação) dos que pesquisam no portal do RCAAP, visto que a parte teoricamente melhorável corresponde a uma fatia percentualmente pequena dos casos.

Contudo, se isso pode indicar que o processamento da língua, mesmo em casos muito específicos, ainda não é suficientemente robusto para ser utilizado na prática, o problema pode ser o oposto – o de que a área ou questão dos nomes dos autores é mínima em relação à melhoria que poderíamos obter se processássemos o texto todo e não só essa parte semi-estruturada e bastante rígida que são os nomes.

É preciso também mencionar que o sistema de

[16] `http://csxstatic.ist.psu.edu/about`

[17] `http://www.informatik.uni-trier.de/~ley/db/`

[18] `http://www.aclweb.org/`, veja-se sobretudo a ACL Anthology.

[19] `http://www.lbd.dcc.ufmg.br/bdbcomp/`

[20] A este respeito, é interessante observar que os dados dos repositórios brasileiros com que o RCAAP funciona já desambiguam o autor colando-lhe ao nome a instituição a que pertence, o que pode indicar tanto uma quantidade muito maior de objetos como uma situação linguística em que os nomes são mais pequenos e mais ambíguos.

procura usado no RCAAP (pelo menos o usado na procura simples que é a mais utilizada) é baseado no modelo "saco de palavras", ou seja, não foi desenhado especialmente para o tipo de objetos e procuras que serve.

Um dos grandes desafios da Linguateca é conseguir convencer – com dados práticos e não apenas recorrendo a retórica – os desenvolvedores informáticos de que o "modelo único de recolha de informação" não é sempre a melhor maneira de obter resultados, e que conhecimento da estrutura do universo dos objetos e da forma de os referir poderá produzir resultados mais adequados.

No fundo, não é mais do que isso o que pretendemos conseguir com o trabalho aqui iniciado. Mas não podemos deixar de reconhecer que estamos muito longe de o concluir, e que pouco mais podemos apresentar do que pistas para futuras ações.

## *Agradecimentos*

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e presentemente pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

Agradecemos vivamente aos membros do projeto RCAAP a colaboração e a cedência dos materiais e do acesso aos seus diários. Sem essa colaboração o trabalho não teria sido possível.

Agradecemos também ao Alberto Simões a inestimável ajuda que nos deu na adaptação do Jspell, assim como todo o apoio sobre o programa.

## *Referências*

Almeida, José João e Ulisses Pinto. 1994. Manual de Utilizador do JSpell. Relatório técnico, Departamento de Informática, Universidade do Minho. `http://www.di.uminho.pt/~jj/pln/jspellman.ps.gz`.

Almeida, José João e Ulisses Pinto. 1995. Jspell – um módulo para análise léxica genérica de linguagem natural. Em *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*, pp. 1–15, 6-8 de Outubro de 1994, 1995.

Artiles, Javier, Julio Gonzalo, e Satoshi Sekine. 2009. WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. Em *18th International World Wide Web Conference*, 20-24 de Abril, 2009.

Artiles, Javier, Satoshi Sekine, e Julio Gonzalo. 2008. Web People Search - Results of the first evaluation and the plan for the second. Em *17th International World Wide Web Conference*, pp. 1071–1072, 21-25 April, 2008.

Cabral, Luís Miguel. 2007. SUPeRB - Sistema Uniformizado de Pesquisa de Referências Bibliográficas. Tese de Mestrado, Faculdade de Engenharia da Universidade do Porto, Março, 2007. `http://www.linguateca.pt/documentos/DissertacaoLuisCabral-SUPeRB.pdf`.

Cabral, Luís Miguel, Diana Santos, e Luís Fernando Costa. 2008. SUPeRB - Gerindo referências de autores de língua portuguesa. Em *VI Workshop Information and Human Language Technology (TIL'08)*, 28-29 de Outubro, 2008.

Carvalho, José, João Mendes Moreira, Eloy Rodrigues, e Ricardo Saraiva. 2010. O Repositório Científico de Acesso Aberto de Portugal : origem, evolução e desafios. Em Maria João Gomes e Flávia Rosa, editores, *Repositórios institucionais : democratizando o acesso ao conhecimento*, pp. 127–152. EDUFBA.

Costa, Luís, Cristina Mota, e Diana Santos. 2012. SIGA, a Management System to Support the Organization of Information Retrieval Evaluations. Em *Proceedings of PROPOR'2012*. Springer. No prelo.

Couto, Francisco M., C. Pesquita, T. Grego, e Paulo Veríssimo. 2009. Handling self-citations using Google Scholar. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, 13(1).

Dervos, Dimitris A., Nikolaos Samaras, Georgios EvangelisSundis, Jaakko P. Hyvarinen, e Ypatios Asmanidis. 2006. The Universal Author Identifier System(UAI Sys). Em *Proceedings 1st International Scientific Conference*.

Ferreira, Anderson A., Adriano Veloso, Marcos André, e H. F. Laender. 2010. Effective Self-Training Author Name Disambiguation in Scholarly Digital Libraries. Em *Proceedings of the 10th annual joint conference on Digital libraries*, pp. 39–48. ACM, 21-25 de Junho, 2010.

Gong, Jun e Douglas W. Oard. 2009. Determine the Entity Number in Hierarchical Clustering

for Web Personal Name Disambiguation. Em *18th International World Wide Web Conference*, 20-24 de Abril, 2009.

Han, Hui, Lee Giles, Hongyuan Zha, Cheng Li, e Kostas Tsioutsiouliklis. 2004. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. Em *JCDL'2004*, pp. 296–305, 7-11 Junho, 2004.

Han, Xianpei e Jun Zhao. 2009. CASIANED: Web Personal Name Disambiguation Based on Professional Categorization. Em *18th International World Wide Web Conference*, 20-24 de Abril, 2009.

He, Daqing e Ayse Göker. 2000. Detecting session boundaries from web user logs. Em *In Proceedingsof of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pp. 57–66.

Kern, Roman, Mario Zechner, e Michael Granitzer. 2011. Model Selection Strategies for Author Disambiguation. Em *22nd International Workshop on Database and Expert Systems Applications*, pp. 155–159, 29 de Agosto a 2 de Setembro, 2011.

Moreira, João Mendes, José Carvalho, Ricardo Saraiva, e Eloy Rodrigues. 2010. Repositório Científico de Acesso Aberto de Portugal : uma ferramenta ao serviço da ciência portuguesa. Em *Congresso nacional de bibliotecários, arquivistas e documentalistas, 10, Guimarães, Portugal, 2010 Políticas de informação na sociedade em rede : actas*. APBD.

Mota, Cristina. 2012. Resultados págicos: participação, resultados e recursos. *Linguamática*, 4(1). No prelo.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, 31 de Dezembro, 2008.

Qian, Yanan, Yunhua Hu, Jianling Cui, Qinghua Zheng, e Zaiqing Nie. 2011. Combining Machine Learning and Human Judgment in Author Disambiguation. Em *20th ACM Conference on Information and Knowladge Management*, 24-28 de Outubro, 2011.

Santos, Diana. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 1(1):25–59, Maio, 2009.

Santos, Diana. 2011. Compreensão de linguagem natural: voltando à carga, 18 de Julho, 2011. Palestra convidada na Universidade de Aveiro, `http://www.linguateca.pt/Diana/download/SantosAveiro2011.pdf`.

Santos, Diana e Nuno Cardoso, editores. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 12 de Novembro, 2007.

Santos, Diana e Ana Frankenberg-Garcia. 2007. The corpus, its users and their needs: a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12:335–374, Maio, 2007.

Santos, Diana e Fernando Ribeiro. 2010. Estudando os autores: Trabalho referente à colaboração com o RCAAP. Relatório técnico, Linguateca, FCCN, 6 de Agosto, 2010. `http://www.linguateca.pt/Diana/download/SantosRibeiroRelRCAAP6Ago2010.pdf`.

Santos, Diana e Fernando Ribeiro. 2011. Estudando os nomes dos autores no RCAAP: relatório do primeiro ano. Relatório técnico, FCCN, 4 de Junho, 2011. `http://www.linguateca.pt/documentos/SantosRibeiroEstudandoNomesAutoresRCAAP.pdf`.

Simões, Alberto, Cristina Mota, e Luís Costa. 2012. A wikipédia em português no Págico: adaptação e avaliação. *Linguamática*, 4(1). No prelo.

Simões, Alberto Manuel e José João Almeida. 2002. jspell.pm – um módulo de análise morfológica para uso em Processamento de Linguagem Natural. Em *Actas da Associção Portuguesa de Linguística (APL2001)*.

Spink, Amanda e Bernard J. Jansen. 2004. Searching for people on Web search engines. *Journal of Documentation*, 60(3):266–278.

Sun, Xiaoling, Jasleen Kaur, Lino Possamai, e Filippo Menczer. 2011. Detecting Ambiguous Author Names in Crowdsourced Scholarly Data. Em *SocialCom 2011*.

Tan, Yee Fan, Min-Yen Kan, e Dongwon Lee. 2006. Search engine driven author disambiguation. Em *JCDL'2006*, June, 2006.

Tsatsaronis, George, Iraklis Varlamis, Sunna Torge, Matthias Reimann, Kjetil Norvig, Michael Schroeder, e Matthias Zschunke. 2011. How to Become a Group Leader? or Modeling Author Types Based on Graph Mining. Em *International Conference on Theory and Practice of Digital Libraries*, Setembro, 2011.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, Cambridge, Mass.

# Corpus multimedia *VEIGA* inglés-galego de subtitulación cinematográfica

Patricia Sotelo Dios
Universidade de Vigo
psotelod@uvigo.es

**Resumo**

Neste artigo presento un proxecto de investigación que consiste na compilación e na explotación do corpus *Veiga*, un corpus multimedia de subtítulos en inglés e en galego. Trátase dun proxecto en fase de desenvolvemento que pretende servir como ferramenta para o estudo e a investigación de certos aspectos relacionados coa práctica da subtitulación intralingüística en inglés e da subtitulación interlingüística do inglés cara ao galego. O *Veiga*, inda que forma parte do corpus paralelo *CLUVI*, transcende o plano textual propio dos demais subcorpus do *CLUVI* e permite observar os subtítulos no seu estado natural, isto é, como parte dun produto audiovisual. Amais de cuestións relacionadas coa construción do corpus e co sistema de buscas, mencionarei algunha das posibles utilidades deste corpus para a práctica, a investigación e a formación en subtitulación.

## 1. Introdución[1]

*Non resultaría contraditorio crear unha base de datos ou un corpus de diálogos cinematográficos e os seus correspondentes subtítulos, sen imaxes, coa pretensión de estudar a tradución audiovisual?*[2] (Gambier, 2006).

Non son poucos os proxectos desenvolvidos ata o de agora relacionados coa creación e coa análise de corpus multimedia e multimodais. Ao mesmo tempo, tanto a investigación en tradución audiovisual coma os estudos de tradución baseados en corpus demostran estar gozando de moi boa saúde nesta última década. Con todo, a maioría dos estudos de tradución interlingüística tratan aspectos puramente lingüísticos e culturais, como é a tradución dos distintos rexistros ou do humor verbal, deixando de lado a natureza multisemiótica do texto audiovisual. Concordo con Gambier (2008) en que, se cadra, cumpriría revisar, estender e reformular certos aspectos dos estudos de tradución no tocante á tradución audiovisual, como o concepto de texto e de unidade de tradución e o deseño e a análise de corpus, e non hai dúbida de que precisamos unha nova metodoloxía para tratar fenómenos como a multimodalidade e a multimedialidade. Se consideramos que os filmes, consonte a definición de Kress e van Leeuwen (2001), son articulacións multimodais de múltiples discursos integrados e recoñecemos que non se producen subtítulos para seren lidos de forma illada, a resposta á pregunta de Gambier é afirmativa: abofé que sería unha contradición crear un corpus de subtítulos formado unicamente por texto cando o que queremos é analizar os anteditos subtítulos tal e como os definen Baker & Saldanha (2009), é dicir, en canto transcricións de diálogos cinematográficos ou televisivos que se presentan de forma simultánea na pantalla, ou estudarmos trazos da subtitulación como son as limitacións espaciais e temporais.

Hogano, inspirados pola crecente abundancia de datos en formato dixital[3] e coa axuda de novos métodos, técnicas e ferramentas informáticos e demais adiantos no deseño de corpus multimedia, tanto investigadores como desenvolvedores de corpus teñen a posibilidade de abordar cuestións e tratar de atopar respostas que resultarían impensables cos métodos máis tradicionais. Porén, para superarmos os obstáculos que inda existen na creación e na explotación de corpus multimedia, cómpre contar cun maior número deste tipo de corpus, nos distintos xéneros, e con máis estudos de carácter tanto teórico como empírico.

Na seguinte sección presentarei os datos dos que consta o *Veiga* neste momento e describirei brevemente os procesos de aliñamento e anotación do corpus e o sistema de busca. O terceiro apartado recolle algunhas reflexións acerca dos posibles usos do devandito corpus en canto ferramenta para a investigación, a práctica e a formación en subtitulación. Menciónanse tamén algunhas das súas limitacións, que serán as que guíen os nosos

[1] Este traballo forma parte do proxecto «Desenvolvemento e explotación de recursos integrados da lingua galega» (INCITE08PXIB302185PR), financiado pola Consellería de Innovación e Industria da Xunta de Galicia.

[2] A tradución das citas é obra da autora do artigo.

[3] Cómpre mencionar a base de datos de subtítulos gratuíta e multilingüe OpenSubtitles.org, a partir da cal se constrúe o corpus OpenSubtitles, que forma parte, pola súa vez, do proxecto OPUS (http://opus.lingfil.uu.se/index.php).

pasos máis inmediatos; e rematamos con varias conclusións e apuntamentos relativos aos retos e ao traballo futuros.

## *2. O corpus* Veiga

O corpus *Veiga* consta na actualidade de 24 filmes de produción estadounidense, británica e australiana subtitulados en inglés (subtitulación intralingüística[4]) e en galego (subtitulación interlingüística[5]) e distribuídos para DVD, cine e internet. En total contén arredor de 300.000 palabras, inda que está previsto incluir máis filmes, entre eles produtos televisivos, e subtítulos noutras linguas distintas do galego.

Desenvolvido ao abeiro do *CLUVI*[6], que contén, pola súa vez, varios corpus paralelos de distintos ámbitos (xurídico, científico, xornalístico, tecnolóxico...) e combinacións lingüísticas, o *Veiga* naceu como un corpus textual. Non obstante, en canto contamos coas ferramentas axeitadas para procesar os datos e poñelos a disposición do público, decidimos dar o paso e convertelo nun corpus multimedia.

A diferenza das demais seccións do *CLUVI*, o *Veiga* non se pode considerar estritamente un corpus paralelo, como tampouco se podería catalogar de corpus comparable, segundo a definición tradicional de ambos os conceptos. Baker (1995), por exemplo, emprega o termo «corpus paralelos» para se referir a textos orixinais nunha lingua A e as súas traducións correspondentes nunha lingua B. Por outra banda, Sammut e Webb (2010) definen «corpus comparable» como un conxunto de documentos formado por dous ou máis subconxuntos, cada un nun idioma distinto, mais coa característica común de pertenceren todos a un mesmo campo temático. Un exemplo típico de corpus comparable podería ser un conxunto de artigos xornalísticos redactados en distintas linguas pero que dan conta dos mesmos feitos; non constitúen traducións uns dos outros, mais comparten gran parte de contido semántico.

Xa que logo, o *Veiga* semella ocupar unha especie de «terra de ninguén» entre corpus paralelo e comparable. Dunha banda, os subtítulos en galego non se poden considerar traducións dos subtítulos en inglés, inda que podería ser o caso que as/os subtituladoras/es partisen da versión inglesa á hora de realizar a tradución ao galego. E, doutra banda, ambas as versións subtituladas comparten algo máis que contido semántico: as dúas parten dun mesmo texto audiovisual orixinal. Poderiamos dicir, polo tanto, que o orixinal, os subtítulos en inglés e mais os subtítulos en galego manteñen unha relación de tipo triangular. O auténtico e estrito paralelismo sería aquel que se establece entre o texto audiovisual orixinal e cada un dos conxuntos de subtítulos en inglés e en galego. Os subtítulos ingleses corresponderíanse cunha clase particular de transcrición coñecida como subtitulación intralingüística, e os galegos serían froito dunha modalidade de tradución tamén moi peculiar denominada subtitulación interlingüística. Así e todo, non semella desatinado pensar que tamén existe certa relación de paralelismo entre as dúas versións de subtítulos, na medida en que ambas son «subprodutos» do mesmo texto orixinal. En resumo, poderíase establecer un dobre paralelismo unidireccional entre o orixinal e ambos os subtítulos e, asemade, unha correlación bidireccional entre as propias dúas versións de subtítulos, tal e como se mostra na figura 1.

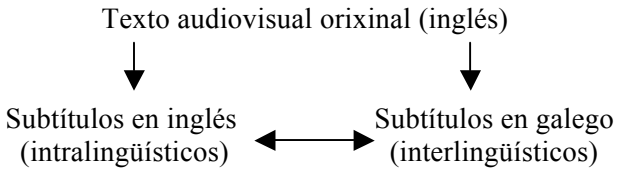



Figura 1: Relación triangular entre orixinal e subtítulos.

Obviamente, cun corpus de subtítulos composto unicamente de texto non sería posible observar paralelismo ningún entre o orixinal e os dous conxuntos de subtítulos. O feito de proporcionármoslles ás/aos usuarias/os acceso ao produto audiovisual orixinal permite que estas/es poidan establecer comparacións e paralelismos e, en definitiva, explorar as múltiples dimensións da subtitulación inter- e intralingüística.

### 2.1 Deseño do corpus

Como xa mencionamos, o corpus *Veiga* alóxase no *CLUVI* (Corpus Lingüístico da Universidade de Vigo), que está a ser desenvolvido polo SLI (Seminario de Lingüística Informática). O *CLUVI* funciona como un repositorio de subcorpus paralelos de distintos tamaños e áreas temáticas construídos mediante as mesmas técnicas de procesamento. Esta macroestrutura implica que todas/os as/os desenvolvedoras/es deben seguir os mesmos procesos de aliñamento e de anotación, o

[4] Os subtítulos intralingüísticos están escritos no mesmo idioma que o diálogo orixinal e os seus receptores adoitan ser persoas con problemas auditivos e estudantes de lingua.

[5] Na subtitulación interlingüística, os subtítulos non substitúen o texto orixinal senón que os dous están presentes, de forma sincronizada, no produto subtitulado.

[6] O corpus paralelo *CLUVI* pódese consultar no enderezo http://sli.uvigo.es/CLUVI/.

que supón unha vantaxe para elas/es e mais para as/os usuarias/os, que poden así acceder a varios corpus a partir dunha única páxina web, cunha interface e unha páxina de busca e de resultados idénticas.

Non obstante, o *Veiga* multimedia precisa dun procesamento máis complexo: amais de anotarmos fenómenos como as omisións, as adicións e os cambios de orde (pouco frecuentes na subtitulación por mor da necesidade de sincronía) das unidades de tradución[7], todos os subtítulos inclúen os tempos de entrada e de saída e un indicador de salto de liña, co que as/os usuarias/os poden observar particularidades inherentes á práctica da subtitulación tales como as limitacións espaciais e temporais, a segmentación ou a condensación, entre outras. Alén disto, a versión multimedia do *Veiga*[8] incorpora un capítulo extra: as/os usuarias/os teñen a posibilidade de visualizar os vídeos onde aparecen as unidades de tradución nas dúas linguas, co que accederían ao cotexto na súa forma multisemiótica orixinal. É dicir, que xunta os resultados da súa busca en formato texto aparece unha ligazón aos vídeos cos subtítulos correspondentes en cada unha das dúas linguas (inglés e galego).

Como é de imaxinar, os procesos da creación dos corpus multimedia son bastante complexos e levan moito tempo. Con todo, o imparable avance das tecnoloxías dixitais e da internet abre novos camiños cara ao tratamento de datos multimediais e multimodais. Xa que logo, agardamos poder contar cun maior número de corpus multimedia nun futuro próximo. A seguir, ofrecemos unha breve descrición do procesamento e do sistema de busca do *Veiga* multimedia.

## 2.2 Aliñamento, anotación e edición de vídeo

Bowker (2002) define 'aliñamento' como o proceso de comparar un texto orixinal e a súa tradución, emparellar os correspondentes segmentos e unilos en canto unidades de tradución nunha memoria de tradución. O aliñamento de corpus paralelos fundaméntase, entre outras cousas, na utilización que se lle quere dar ao corpus e no tipo de textos que conforman a base de datos, en función do cal ha de variar o seu nivel de segmentación (palabras, oracións, parágrafos...).

Como sinalan Guinovart e Sacau (2004), a unidade de segmentación definida para o aliñamento dos textos paralelos do *CLUVI* é a frase ortográfica do texto orixinal. Xa que logo, a correspondencia entre o texto orixinal e a súa tradución sempre será do tipo 1:n. Normalmente, a correspondencia entre oracións do texto orixinal e oracións da tradución é do tipo 1:1. Porén, pódese dar o caso de que a oración orixinal non dispoña de tradución (1:0), ou de que a oración orixinal se corresponda con parte da oración da tradución (1:1/2) ou con dúas oracións na versión traducida (1:2), ou mesmo que unha oración da tradución non se corresponda con ningunha oración no orixinal (0:1). Esta falta de correspondencia entre texto orixinal e traducido (1:0, 0:1) representan omisións ou adicións (sempre respecto do texto orixinal) e etiquétanse mediante unha versión adaptada dos elementos <hi> e <ph> que forman parte da especificación TMX. Con independencia do aplicativo que se empregue para aliñar os textos, todos os subcorpus do *CLUVI* deben seguir a citada especificación. No caso do *Veiga*, empregamos a ferramenta Trans Suite 2000 Align para segmentar e aliñar os textos nas dúas linguas de forma automática, inda que se revisaron todos os aliñamentos e as anotacións foron engadidas de xeito manual.

| Nome do filme (UT) | Subtítulo en inglés | Subtítulo en galego |
|---|---|---|
| PAR (32) | They said they found Travis. | Atoparon a Travis. |
| PAR (33) | [[hi type='**supr**']] Oh, no.[[/hi]] | [[---]] |
| PAR (34) | What are you going to do? | -E que vas facer? |

Táboa 1: Exemplo de omisión.

| Nome do filme (UT) | Subtítulo en inglés | Subtítulo en galego |
|---|---|---|
| BAB (1201) | In other news... | Noutras noticias... |
| BAB (1202) | [[---]] | [[hi type='**incl**']]Aos meus fillos, María Eladia e Eliseo...[[/hi]] |
| BAB (1203) | [[---]] | [[hi type='**incl**']]...as luces máis brillantes na noite máis escura.[[/hi]] |

Táboa 2: Exemplos de adición.

Á parte disto, debido á natureza particular do noso corpus, tamén se incluíron certos datos

[7] Véxase Baker e Saldanha (2009) para unha definición rigorosa do concepto de unidade de tradución (UT).

[8] O corpus multimedia de subtítulos *Veiga* está dispoñible no enderezo http://sli.uvigo.es/CLUVI/vmm.html. Cómpre lembrar que esta versión multimedia se atopa en proceso de construción e que, neste momento, só se poden consultar 10 dos 24 filmes dos que consta a versión textual.

propios da subtitulación como son a duración e a segmentación dos subtítulos.

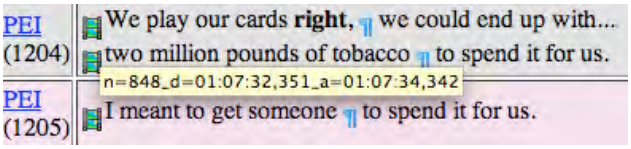



Figura 2: Anotación do código temporal e do salto de liña dos subtítulos.

Na figura 2 móstrase a parte do inglés de dous resultados de busca tal e como se visualizan na web do *Veiga*. A segmentación dos subtítulos vén indicada tanto polo símbolo do caldeirón (¶), que representa un salto de liña dentro dun subtítulo, como pola icona do celuloide, que marca un cambio de subtítulo. Tocante á duración, se pasamos o rato por riba da icona do celuloide poderemos ver os tempos de entrada e de saída do subtítulo en cuestión. A codificación TMX completa correspondente ao aliñamento e á anotación da versión inglesa e galega deste segmento que aparece na dita figura 2 é a seguinte:

```
<tu>
<tuv xml:lang="en">
<seg>[[s n="848" d="01:07:32,351"
a="01:07:34,342"]]We play our cards right,[[l/]]we
could end up with... [[s n="849" d="01:07:34,431"
a="01:07:36,023"]]two million pounds of tobacco[[l/]]to
spend it for us.</seg>
</tuv>
<tuv xml:lang="gl">
<seg>[[s n="808" d="01:07:33,271"
a="01:07:36,946"]]Podemos gastar 2 millóns en
tabaco[[l/]]e pósters de Pamela Anderson. </seg>
</tuv>
</tu>
<tu>
<tuv xml:lang="en">
<seg>[[s n="850" d="01:07:36,511"
a="01:07:39,025"]]I meant to get someone[[l/]]to spend
it for us.</seg>
</tuv>
<tuv xml:lang="gl">
<seg>[[s n="809" d="01:07:37,071"
a="01:07:39,539"]]Buscaremos alguén[[l/]]que o gaste
por nós. </seg>
</tuv>
</tu>
```

Como se pode ver, o código do TMX inclúe unhas indicacións do número (*n*) identificador do subtítulo, dos tempos inicial (*d*) e final (*a*) de cada un dos subtítulos e dos saltos de liña (*l*).

Amais do aliñamento paralelo das unidades de tradución e da anotación, que afecta sobre todo á dimensión textual do corpus, a compilación e o procesamento de material audiovisual require certos labores de edición de vídeo. Se temos en conta que a) estamos ante un corpus de subtítulos e b) unha das limitacións (e obxecto de estudo) principais da subtitulación é o tempo, un segundo paso lóxico foi o de someter o corpus a un novo proceso de segmentación e aliñamento, en que precisamente as unidades de segmentación estarían constituídas polos propios subtítulos. En primeiro lugar comprobamos que, efectivamente, os tempos dos diferentes subtítulos estivesen ben sincronizados co produto audiovisual correspondente. Nalgúns casos foi preciso editar os subtítulos —para o que empregamos o aplicativo gratuíto Subtitle Workshop— e modificar os tempos para axustalos á velocidade de reprodución da película.[9] A seguir, coa ferramenta gratuíta VirtualDubMod «pegamos» os subtítulos nas dúas linguas ao produto audiovisual; e, para rematar, editamos cada un destes filmes cos subtítulos nas dúas linguas e segmentámolos subtítulo a subtítulo. É dicir, todos os textos multimedia do *Veiga* (os filmes orixinais en inglés cos subtítulos en inglés por unha banda e en galego pola outra) foron cortados en clips, cada un dos cales se corresponde cun subtítulo. Como resultado, seguimos a contar con dous conxuntos de subtítulos, en inglés e en galego, compostos cada un deles de tantos clips como subtítulos ten cada produto. Ademais, como moitos dos subtítulos son de pouca duración (un ou dous segundos) e non se poderían visualizar de maneira axeitada, á hora de segmentar os filmes, a cada clip/subtítulo engadíronselle dez segundos extra (os cinco segundos anteriores ao tempo de entrada e os cinco posteriores ao tempo de saída), de xeito que a/o usuaria/o poida contar con certo contexto. Logo de obtermos dous conxuntos de videoclips subtitulados para cada un dos textos audiovisuais, estes almacénanse no sistema de ficheiros do servidor identificados polo título da película, a lingua do subtitulado do videoclip (inglés ou galego) e o número do videoclip correspondente[10], de maneira que cando se realiza unha busca no *Veiga* os resultados mostran non só a unidade de tradución aliñada nas dúas linguas, senón tamén unha ligazón ao vídeo onde aparece ese texto/subtítulo.

En resumo, o corpus multimedia *Veiga* sométese a dous procesos distintos de segmentación: un que afecta só á dimensión textual, isto é, aos subtítulos, e outro que afecta aos

[9] Sobre todo algúns dos subtítulos distribuídos a través da internet, que foron creados para un ficheiro de vídeo cun formato e propiedades concretos que non se corresponden necesariamente co que temos nós.

[10] No caso da figura 2 serían: peixe_en-848.flv, peixe_en-849.flv, peixe_gl-808.flv; e peixe_en-850.flv, peixe_gl-809.flv.

subtítulos e mais ao texto audiovisual orixinal ao que estes acompañan. No primeiro caso, a unidade de segmentación é a frase ortográfica (frase en inglés ↔ frase en galego), namentres que no segundo caso a segmentación se realiza ao nivel do subtítulo (texto audiovisual en inglés + subtítulo en inglés ↔ texto audiovisual en inglés + subtítulo en galego), co engadido dos cinco segundos anteriores e posteriores. No primeiro caso a segmentación efectúase de maneira automática, inda que se somete a unha revisión e corrección manual, e no segundo caso, á falta dunha ferramenta que segmente vídeo por subtítulos e engada os anteditos segundos de contexto automaticamente, a segmentación estase a facer de xeito manual. Tocante ao aliñamento, a relación entre as unidades textuais en inglés e en galego é do tipo 1:1, 1:0 ou 0:1, e estas unidades non teñen por que coincidir necesariamente coas unidades de segmentación que constitúen os subtítulos. De feito, o máis habitual é que unha única unidade textual se corresponda con máis dun clip/subtítulo, da mesma maneira que, como mostra a figura 3, un subtítulo en inglés non sempre se corresponde cun único subtítulo en galego.

| | | |
|---|---|---|
| EAR (340) | Carbon monoxide poisoning, ¶ Strychnine, suffocation, ¶ breaking the neck, and anal electrocution ¶ are some of the more common methods used. | Envelenamento por ¶ monóxido de carbono, estricnina ¶ abafamento, ¶ quebrar o pescozo, ¶ e a electrocución anal ¶ son algúns dos métodos usados. |
| EAR (341) | Removed from his or her cage **with** a heavy ¶ neck-pole, the animal is walked past ¶ the rows of bodies of slaughtered foxes, ¶ sables, raccoons and wolves, among others. | Sacado da súa gaiola ¶ cunha vara no pescozo, ¶ os animais camiñan pasando entre ¶ ringleiras de corpos de raposos, martas, ¶ mapaches, lobos e outros animais mortos. |
| EAR (342) | Death by anal electrocution is a crude ¶ process that requires a probe to be inserted ¶ in the rectum while the animal ¶ bites down on a metal conductor. | A morte por electrocución anal ¶ é un proceso cruel, ¶ no cal se insire un electrodo no recto, ¶ mentres o animal traba ¶ nun condutor de metal. |

Figura 3: Exemplo de resultados de busca. A correspondencia entre subtítulos non é do tipo 1:1.

## 2.3 Sistema de busca

Como xa mencionamos con anterioridade, tanto o corpus paralelo *CLUVI* como o *Veiga* se poden consultar en liña a través dun aplicativo PHP deseñado polo SLI. Un dos puntos fortes desta ferramenta de busca reside nas súas múltiples posibilidades de consulta. Permite facer buscas moi complexas de palabras illadas ou grupos de palabras, e mostra as equivalencias bilingües dos termos buscados en contexto. As buscas poden ir en calquera dirección (inglés>galego e galego>inglés) ou nas dúas ao mesmo tempo; é dicir, pódese buscar un termo en cada unha das dúas linguas de forma simultánea. Todos os corpus do *CLUVI* comparten a mesma interface de busca[11]: unha simple caixa onde as/os usuarias/os poden introducir a súa expresión de busca nunha ou en todas as linguas. Por cuestións de dereitos de autor, o sistema de busca detense logo de atopar as primeiras 1.500 concordancias. De conformidade co dereito de cita, os usuarios en ningún caso teñen acceso aos textos completos, senón que visualizan unicamente aquelas unidades de tradución e os subtítulos correspondentes que coinciden cos seus criterios de busca. Ademais, o texto que se ofrece é o resultado dun procesamento levado a cabo por nós; isto é, non se trata simplemente de texto orixinal, senón que estamos ante material manipulado en maior ou menor grao.

Imos ilustrar este sistema de busca cun exemplo: buscamos pares de unidades de tradución onde apareza tanto a palabra 'for' no inglés como a palabra 'por' no galego. O sistema atopa e mostra unha lista de 131 unidades que coinciden coa busca. Como vemos na figura 4, os resultados aparecen nunha táboa de cinco columnas que conteñen a seguinte información: código do filme e número de unidade de tradución, unidade en inglés, unidade en galego, icona dunha bobina e icona dunha frecha. Se prememos a icona da frecha, ábrese unha fiestra onde aparece o par en cuestión e mais o par anterior e posterior (cotexto). E a icona da bobina enlaza cunha páxina onde podemos visualizar os clips correspondentes tamén de forma paralela, é dicir, nas dúas linguas, e os clips anterior e posterior.

| | | | | |
|---|---|---|---|---|
| 1-CHU (195) | When you're 18, you can ¶ go to hell **for** all I care. | Canda teñas 18, **por** min ¶ como se vas para o inferno, | | |
| 2-CHU (258) | I've been a prisoner of my love ¶ **for** you **for** a very long time. | Levo moito tempo presa ¶ no amor que sinto **por** ti. | | |
| 3-CHU (436) | ¶ -I'd do anything **for** you. | ¶ - Faría calquera cousa **por** ti, | | |

Figura 4: Exemplo de resultados de busca.

# *3. Posibles usos e limitacións do corpus multimedia* Veiga

Xa o dicía Baldry (2004): precisamos acceder aos textos na súa forma *in vivo*, de xeito que a relación entre a pista de son e de vídeo se manteña intacta, porque a maneira principal en que un texto fílmico constrúe o seu significado é mediante a sincronización entre as fontes sonora e visual. Con todo, antes de falarmos das posibles vantaxes do *Veiga*, comecemos por recoñecer cales son as súas limitacións máis evidentes.

## 3.1 Limitacións

O primeiro hándicap atopámolo no reducido tamaño do corpus. No noso favor podemos alegar que, neste momento, só hai dúas persoas traballando no proxecto e que está previsto amplialo con filmes subtitulados emitidos pola televisión e poida que se inclúan outras linguas. Non cabe dúbida de que canto maior sexa o corpus, máis variado e extenso é o conxunto de fenómenos

[11] Cómpre indicar que o *Veiga* multimedia está en proceso de emigrar a un novo sitio web.

que podemos atopar, o que afectaría positivamente a fiabilidade dos posibles estudos baseados nos datos do *Veiga*. Non obstante, tal e como afirman McEnery & Wilson (2001), o tamaño non ten por que ser garantía de representatividade. Ademais, un corpus pequeno pero especializado tamén pode ser útil na investigación de fenómenos particulares.

Unha segunda limitación podería ser a heteroxénea procedencia e autoría dos subtítulos traducidos, o que semella demandar un subsecuente cambio de enfoque á hora de observar os datos e suscita a cuestión da calidade da tradución (e do corpus). Como xa se comentou, os ficheiros dos subtítulos en galego foron distribuídos en DVD, no cinema e na internet. En concreto, sete deles foron creados para DVD, é dicir, que é bastante probable que os seus autores fosen tradutoras/es profesionais e que pasasen algún tipo de control de calidade. Catorce deles foron creados para a gran pantalla e proxectados en ciclos de cine. Neste caso, a maior parte das/os subtituladoras/es son voluntarias/os[12], e estarían a medio camiño entre a primeira caste de tradutoras/es (profesionais) e a que vén a seguir (afeccionadas/os). E os outros tres grupos de subtítulos constituirían casos dun novo tipo de subtitulación que se vén practicando en España, e mais noutros países, coñecido no mundo anglosaxón co nome de *amateur subtitling* e descrito como unha práctica desenvolvida por non profesionais e que se rexe por unhas limitacións completamente diferentes ás da subtitulación profesional. O resultado final depende en gran medida do coñecemento desta/e afeccionada/o da lingua orixinal, o que probablemente derive en abundantes erros e interpretacións equivocadas. Así e todo, a calidade non foi un criterio que tivésemos en conta cando recompilamos o corpus.

E a terceira eiva constitúena os xa citados labores de procesamento e de edición do material audiovisual deste corpus, que, malia todos os adiantos tecnolóxicos dos que dispoñemos na actualidade, seguen a provocar unha gran demora no avance do proxecto.

En consecuencia, con independencia do uso que se queira facer dos resultados das buscas no corpus *Veiga*, cómpre ter sempre presentes as anteditas limitacións.

## 3.2 Posibles usos

Malia todo o anterior, pódense mencionar varias posibles utilidades do *Veiga* multimedia. Por unha banda, pódese empregar como un banco de exemplos, unha base de datos que as/os investigadoras/es poden usar para ilustrar os seus traballos e probar as súas hipóteses.

No eido pedagóxico, pódese usar en varios contextos, dende en cursos de lingua que versen sobre os rexistros e a xerga ata en cursos especializados en tradución audiovisual e subtitulación, pasando por cursos de tradución non necesariamente especializada onde se trate o transvasamento de referentes culturais, frases feitas, coloquialismos, etcétera (Zanettin *et al*., 2003). Na nosa opinión, é fundamental que as/os docentes de tradución audiovisual lle ofrezan ao seu alumnado material auténtico que se preste a análises contrastivas de textos orixinais e textos traducidos. Por outra banda, o uso de corpora como metodoloxía do ensino da tradución favorece o desenvolvemento de espíritu crítico, aprendizaxe autónoma, toma de decisións e outras competencias tanto propias dos estudos de tradución como transversais (Beeby *et al*., 2009). Ao mesmo tempo, tamén pode servir de ferramenta para a aprendizaxe en liña, xa que as/os alumnas/os contarían coa posibilidade de explorar as propiedades textuais en tanto que oen e ven os vídeos, que, amais, poden reproducir e pausar cando o desexen.

Para rematar, tamén os profesionais da subtitulación poderían facer uso deste corpus en canto unha colección de subtítulos xa feitos, no sentido de que poderían ver as solucións polas que optaron outras/os profesionais para resolver certas limitacións ou dificultades particulares da subtitulación.

Como xa comentamos, o tamaño limitado do corpus e a natureza híbrida dos subtítulos traducidos non permiten establecer xeneralizacións sobre a práctica da subtitulación inter- e intralingüística. De feito, cabería mesmo trazar máis distincións en función do xénero dos textos audiovisuais (películas, documentais, filmes para as/os cativas/os...) e do medio de distribución do produto. Emporiso, cómpre aclarar que a nosa pretensión é só fornecer unha ferramenta que poidan usar tanto investigadores como profesionais e docentes para ilustrar aspectos particulares da subtitulación. Por unha banda, o *Veiga* permite observar cuestións técnicas como a presentación dos subtítulos na pantalla (número de liñas, posición, cor, marcas de diálogo) e a súa duración (tempos de entrada e de saída, intervalo[13], cambios de plano e sincronización). Segundo Díaz Cintas e

[12] Son cinéfilas/os e afeccionadas/os ás versións orixinais subtituladas e forman parte da propia asociación organizadora dos ciclos de cine, mais non recibiron formación especializada en tradución para o subtitulado e realizan este labor de balde.

[13] De Linde e Kay (1999) distinguen entre *leading* e *lagging*. Cando o subtítulo precede o texto oral falamos de anticipación e cando o segue falariamos de demora.

Remael (2007), a práctica da subtitulación é bastante heteroxénea e varía de maneira considerable dun programa audiovisual, dunha empresa e mesmo dun país a outro. Malia os varios intentos que se fixeron de elaborar un conxunto de convencións ou pautas harmonizadas, semella que en cada idioma/cultura se subtitula consonte a práctica tradicional que lle é propia. Por outra banda, os subtítulos, tanto os intralingüísticos coma os interlingüísticos, constitúen versións condensadas do texto audiovisual. A subtitulación adoita implicar unha selección do material lingüístico, o que obriga ás/aos subtituladoras/es a tomaren continuas decisións acerca do que resulta importante e do que semella superfluo ou mesmo redundante. A redundancia, xa que logo, é un concepto esencial na subtitulación, xa que parte da información omitida nos subtítulos pode ser proporcionada por outros elementos do texto audiovisual, como a imaxe e/ou o son. Con todo, a redución é un fenómeno frecuente e decote inevitable que implica omisión de información. O *Veiga* multimedia non só sitúa os subtítulos ao carón do texto audiovisual orixinal, senón que coloca cara a cara as dúas versións subtituladas, co que podemos observar fenómenos, tales como a cohesión e a condensación, enraizados na semiótica da subtitulación.

## *4. Conclusións e traballo futuro*

Nestas páxinas presentamos o corpus multimedia de subtítulos inglés-galego *Veiga*, un proxecto en fase de elaboración que tenta facerse eco dunha idea formulada por varias/os autoras/es nos últimos anos: cómpre superar o enfoque tradicional da elaboración de corpus, onde prima o texto escrito, e convencerse da necesidade de deseñar corpora multimedia que amosen os aspectos polisemióticos do discurso fílmico e da subtitulación. Tamén tratamos cuestións relacionadas cos datos e coa construción do corpus; en concreto, coa condición «intermedia» dos subtítulos, que non son exactamente traducións uns dos outros. Comentamos algunha das eivas máis obvias do corpus como son o seu tamaño e as limitacións tecnolóxicas, que agardamos poder resolver no futuro. E, para rematar, mencionamos varias áreas da práctica, da investigación e da docencia da subtitulación onde o *Veiga* podería ser de certa utilidade.

O primeiro chanzo no camiño consistirá en incluír no sitio web datos acerca do medio de distribución (DVD, cinema, internet...) e/ou autoría dos distintos subtítulos, de maneira que as/os usuarias/os poidan facer un uso o máis informado posible do corpus. E como pasos futuros, amais de completar a transmutación dos subtítulos en formato texto ao formato audiovisual, e de aumentar o número de filmes subtitulados para o cinema e daqueles creados por afeccionados e distribuidos a través da internet, temos a intención de incluír tamén produtos televisados —emitidos principalmente pola CRTVG—. Deste xeito estaremos ampliando o tamaño do corpus, engadindo un novo modo de distribución —que se sumaría ao DVD, ao cinema e á internet— e diversificando o seu contido, xa que é moi probable que algúns deses novos produtos constitúan novos xéneros dentro dos textos audiovisuais. Tocante á extensión a outras linguas, considerámolo como unha posibilidade a máis longo prazo que habería que estudar chegado o momento; polo de agora, a nosa prioridade é seguir a traballar na combinación inglés-galego e mellorar o corpus de maneira que cada día sexan menos as súas limitacións e máis as súas aplicacións e posibilidades.

## *Referencias*


Baker, M. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7(2):223-243.

Baker, M. e G. Saldanha, editoras. 2009. *Routledge Encyclopedia of Translation Studies*, segunda edición, páxinas 244-245 e 304-306. Londres e Nova York: Routledge.

Baldry A. e C. Taylor. 2004. Multimodal concordancing and subtitles with MCA. En A. Partington, J. Morley e L. Haarman, editores, *Corpora and Discourse*, páxinas 57-70. Bern: Peter Lang.

Beeby, A., P. Rodríguez Inés e P. Sánchez-Gijón, editoras. 2009. *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam-Philadelphia: John Benjamins Publishing.

Bowker, L. 2002. *Computer-Aided translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

De Linde, Z. e N. Kay. 1999. *The Semiotics of Subtitling*. Manchester: St. Jerome Publishing.

Díaz Cintas, J. e G Anderman, editores. 2009. *Audiovisual translation: Language Transfer on Screen*. Londres: Palgrave Macmillan.

Díaz Cintas, J. e A. Remael. 2007. *Audiovisual Translation: Subtitling*. Manchester: St. Jerome Publishing.

Gambier, Y. 2008. Recent developments and challenges in audiovisual translation research. En D. Chiaro, C. Heiss e C. Bucaria, editoras, *Between Text and Image: Updating Research in Screen Translation*, páxinas 11-33. Amsterdam-Philadelphia: John Benjamins Publishing.

Gambier, Y. 2006. Multimodality and Audiovisual Translation. En *MuTra 2006: Audiovisual*

*Translation Scenarios: Conference Proceedings*. Copenhaguen. Dispoñible en liña no enderezo http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Gambier_Yves.pdf.

Gómez Guinovart, X. e E. Sacau Fontenla. 2004. Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). En T. Lino *et al*., editoras, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, páxinas 1179-1182. Lisboa. Dispoñible en liña no enderezo http://webs.uvigo.es/sli/arquivos/lrec2004.pdf.

Kress, G. e T. van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Nova York: Oxford University Press.

McEnery, T. e A. Wilson. 2001. *Corpus Linguistics*, segunda edición. Edimburgo: Edinburgh University Press.

Sammut, C. e G. I. Webb, editores. 2010. *Encyclopedia of Machine Learning*. Nova York: Springer.

Zanettin, F., S. Bernardini e D. Stewart, editores. 2003. *Corpora in Translator Education*. Manchester: St Jerome Publishing.

# Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português

Vera Vasilévski
Universidade Federal de Santa Catarina
sereiad@hotmail.com

Márcio José Araújo
Universidade Tecnológica Federal do Paraná
marciomjapr@gmail.com

**Resumo**

Este artigo apresenta um analisador morfológico automático de verbos do português, com destaque para seu desempenho no processamento das regras que regem esse sistema verbal e no tratamento das ambigüidades geradas. Nesta etapa, trabalha-se com as ambigüidades decorrentes da alomorfia dos sufixos modo-temporais e da possibilidade de esses morfemas serem zero (Ø) em alguns modos e tempos, nas três conjugações do português. Para esclarecer o trabalho feito com o analisador, traz um resumo das regras morfológicas do sistema de verbos do português. Obteve-se êxito no tratamento de muitas das ambigüidades que o programa registrou, as quais eram esperadas, uma vez que coincidem com as ambigüidades do sistema de verbos da língua portuguesa. A resolução da maioria delas fez-se com base em regras computacionais (estruturas de seleção) que consideram o contexto do enunciado. Conclui que a resolução de outras ambiguidades relacionadas a modo e tempo verbal somente será possível ao se levar em conta também os morfemas número-pessoais, que são objeto de outro trabalho.

## *1. Introdução*

Os computadores e, a partir deles, a Lingüística Computacional tornaram possíveis o armazenamento e a análise de quantidade nunca antes conhecida de dados da comunicação verbal, em tempo realmente curto. Isso possibilita descrições, comparações e generalizações com base em uma massa de dados bastante densa. No entanto, apesar do desenvolvimento computacional voltado para trabalho com língua escrita e falada que se constata atualmente, ainda é reduzido o número de aplicativos dessa natureza disponíveis e efetivamente utilizados por usuários, por diversos motivos. Em especial, a metalinguagem vê-se carente de recursos eletrônicos específicos para auxiliar pesquisas em várias áreas da Lingüística.

Dentre os motivos para o pouco uso desses aplicativos, além da falta de divulgação, está a dificuldade do usuário de utilizar esses programas, às vezes, porque não consegue compreender como eles funcionam. Agrava essa situação o fato de muitos desses aplicativos serem em inglês ou feitos para essa língua, o que reduz sua eficiência para o português (Vasilévski, 2010). Ainda, não raro, seu nível de interatividade é precário (Vasilévski, 2010), o que desestimula seus potenciais usuários. Faz-se necessário remediar essa situação.

Este estudo apresenta um recurso computacional especialmente desenvolvido para automatizar o sistema de verbos do português, a partir de suas regras morfológicas, e debate algumas implicações advindas dessa tarefa. Ressalta a automatização dos morfemas modo-temporais, os casos ambíguos dela decorrentes e discute sua desambiguação. Essa ferramenta foi desenvolvida como parte do projeto *Análise morfológica Automática do Português* (Scliar-Cabral, 2009), cujo objetivo maior é depreender uma gramática automática do português brasileiro, mediante análise do córpus pau003.cha – que é constituído por 10688 enunciados e está disponível para ser baixado (Childes, 2011) –, levando-se em conta a fala dos adultos quando conversam entre si, a fala dirigida à criança e a fala da criança ao se comunicar com os adultos. Desenvolve-se esse projeto em parceria com o projeto Childes (MacWhinney, 2011). Etapas desse projeto têm sido apresentadas (Vasilévski 2010, 2011a, 2011b, 2011c; Scliar-Cabral e Vasilévski, 2011; Scliar-Cabral, 2011), e a que se expõe aqui se refere à análise morfológica automática de verbos conjugados,

portanto, em situações de uso, encontradas em um enunciado.

Buscas foram feitas em bases de artigos científicos e diretamente na Rede Mundial de Computadores, na tentativa de encontrar documentos sobre outros analisadores morfológicos de verbos do português similares a este que ora se apresenta, para aprimorá-lo, bem como para compará-los, mas não se obteve sucesso. Isso não significa que não existam trabalhos dessa natureza, no entanto, eles não estão disponíveis ou suficientemente divulgados.

A automatização que se demonstra nesta ocasião restringe-se aos verbos regulares, mas a finalidade da ferramenta computacional é analisar morfologicamente também os verbos irregulares do português. É possível testar-se qualquer forma verbal no programa, pois todos os tempos verbais e pessoas gramaticais da língua portuguesa foram inseridos em seu algoritmo. Para dar suporte às etapas do projeto, criou-se um programa que abriga várias ferramentas, as quais funcionam em conjunto e em interface com outros aplicativos. Esse programa chama-se Laça-palavras (Vasilévski e Araújo, 2010), e o analisador morfológico enfocado aqui é uma de suas ferramentas.

A partir disso, este artigo aborda o referencial teórico básico utilizado para se compreender o sistema de verbos do português e suas regras, o programa Laça-palavras, os princípios de funcionamento do analisador morfológico, suas principais convenções e seu desempenho.

## *2. O sistema de verbos do português*

Para se desenvolver o analisador morfológico automático para verbos do português, foi necessário conhecer a fundo as regras gramaticais que regem o sistema de verbos dessa língua, o que se obteve na literatura pertinente ao tema (Câmara Jr., 1986, 1976; Scliar-Cabral, 2003, 2007, 2008). Parte dessa literatura já foi discutida (Vasilévski, Scliar--Cabral e Araújo, 2012), quando se tratou especificamente do comportamento da vogal temática, mas cabe revisitá-la e complementá--la com teoria específica que fundamente o assunto deste artigo.

O analisador morfológico em foco baseia-se em regras gramaticais, e não em aprendizado de máquina, ou seja, não gera regras automaticamente, a partir de um dicionário de treino. Ocorre sim que regras gramaticais foram convertidas em algoritmos e testadas no córpus. O projeto é criado e coordenado por cientistas da língua, e conta com o suporte indispensável da computação.

O sistema de conjugação de verbos do português é considerado, de certa forma, simples e previsível (Câmara Jr., 1986), o que respalda a criação de uma ferramenta computacional baseada em suas regras. O sistema de verbos do português compreende três conjugações, assinaladas pela vogal temática. Há três vogais temáticas, conforme a notação escrita usual: "a" para a primeira conjugação (1.ªC), "e" para a segunda conjugação (2.ªC) e "i" para a terceira conjugação (3.ªC). Compõem esse sistema verbos ditos regulares – que seguem o paradigma fixo da conjugação a que pertencem e são maioria em português – e os irregulares – que se desviam do paradigma regular.

O tema do infinitivo é a forma básica do verbo regular. Assim, dado um verbo regular em sua forma infinitiva, é possível conjugá-lo com facilidade, nas seis pessoas gramaticais, sobretudo nos tempos do modo indicativo. Em contrapartida, tomar uma forma verbal conjugada e dela extrair os morfemas que a compõem, a fim de desvendar tempo, modo, pessoa e número em que está flexionada, não é tão fácil.

Em português, há três modos verbais finitos, com seus tempos simples (indicativo (seis tempos), subjuntivo ou conjuntivo (três tempos) e imperativo (afirmativo e negativo), além do infinitivo pessoal e das formas nominais infinitivo, gerúndio e particípio. Na seção 4, expõem-se tais tempos, da maneira como foram inseridos no algoritmo morfológico. Cabe destacar que o pretérito mais-que-perfeito é pouco usado no Brasil. Na fala coloquial, ele restringe-se a frases feitas, e é raramente usado na língua escrita. Também, é preservado, sobretudo, na literatura e em músicas, e aparece esporadicamente no falar jornalístico, por exemplo, em editoriais. Quanto às formas nominais, o infinitivo é a forma mais genérica do verbo, que de maneira

mais ampla e vaga resume sua significação, sem noções de tempo, modo e aspecto. Por sso, ele é usado para designar o nome do verbo, e pode funcionar como substantivo.

Entre o gerúndio e o particípio há oposição de aspecto, pois o primeiro é imperfeito (processo inconcluso) e o segundo é perfeito (processo concluso). Morficamente, o particípio desvia-se da natureza do verbo, pois pode tornar-se um adjetivo verbal. É a única forma verbal que assume gênero e número, além da categoria de voz passiva. Assim, morfologicamente, ele pertence aos adjetivos, embora tenha valor verbal no âmbito sintático e semântico (Câmara Jr., 1986). Como verbo, o particípio entra na formação dos tempos compostos com o auxiliar "ter" e "haver" (quando permanece invariável em gênero e número) e com o auxiliar "ser", na construção da voz passiva analítica, além de núcleo do predicado de orações reduzidas. Já o gerúndio é morfologicamente verbal, assim, não admite flexão de gênero e número (Câmara Jr., 1986), e entra na formação das formas progressivas e também como núcleo do predicado de orações reduzidas.

É válido mencionar que, por irregulares, entendem-se os verbos cujos temas das formas primitivas são distintos entre si – essa é a principal característica de sua irregularidade. São formas primitivas os temas da primeira pessoa do singular e as segundas pessoas do presente do indicativo; o tema da segunda pessoa do singular do pretérito perfeito do indicativo; e o tema do infinitivo impessoal, os quais dão origem aos outros tempos verbais. Por exemplo, o tema da primeira pessoa do singular do presente do indicativo dá origem ao presente do subjuntivo. O verbo "ser", na primeira pessoa do singular do presente do indicativo, tem o tema *so-*, enquanto, na segunda pessoa do singular do pretérito perfeito do indicativo, seu tema é *fo-*, portanto, ele é irregular. Ainda, além das formas primitivas distintas, o verbo "ser" apresenta irregularidades nas derivações, como no presente do subjuntivo, cujo tema é *sej-*, em todas as pessoas. No entanto, os verbos irregulares não o são em todos os tempos e pessoas gramaticais. Nos tempos futuros (do presente e do pretérito) do modo indicativo, há pouquíssima irregularidade. Nesses tempos, verbos irregulares como "ser", "estar", "vir" são perfeitamente analisados pelo programa, como formas regulares. A exceção fica por conta dos verbos "fazer" e "trazer", cujas formas nesses tempos, para serem regulares, deveriam ser *"fazerá", *"trazerá" e *"fazeria", *"trazeria", as quais são incorretas.

Do grupo dos verbos irregulares fazem parte os verbos auxiliares, que figuram nas conjugações compostas. Em uma cadeia podem entrar vários verbos auxiliares, sendo o último verbo o verbo principal, aquele que carrega a significação externa, sempre em uma forma nominal (infinitivo, gerúndio ou particípio). É ele quem dita a regência, por isso, o primeiro auxiliar da cadeia se flexionará em pessoa, número, tempo e modo, conforme tal verbo principal determinar. Por exemplo: "**ia** entrar", "**deve estar** havendo muitas suspeitas", "**podem**-se esperar vitórias" e "**tinha** feito".

O verbo é, em português, o vocábulo flexional por excelência, dada a complexidade e a multiplicidade de suas flexões. As noções gramaticais de tempo e modo e de pessoa e número que a forma verbal indica correspondem a duas desinências (sufixos flexionais) chamadas de sufixo modo-temporal (SMT) e sufixo número-pessoal (SNP), que se aglutinam e se ligam ao tema. O tema constitui-se do radical seguido da vogal temática da conjugação correspondente. No padrão geral, o radical é invariável e dá a significação lexical do verbo. Assim, a fórmula geral da estrutura do vocábulo verbal português – na qual RAD indica radical do verbo; VT, vogal temática; e SF, sufixos flexionais – é (Câmara Jr., 1986):

TEMA (RAD+VT) + SF (SMT + SNP)

Levando-se em conta a alomorfia de cada um dos sufixos flexionais e a possibilidade de ser zero (Ø) para um deles ou ambos, essa fórmula dá a regra geral da constituição morfológica do verbo em português, além de indicar a ordem obrigatória dos morfemas. A aglutinação em um único morfema das noções de tempo e modo determina, evidentemente, 13 morfemas modo-temporais, nos quais só esporadicamente ocorre alomorfia, isto é, a variação de um morfema condicionada pelo contexto onde ele

ocorre. No analisador, são levados em conta apenas os alomorfes do sistema escrito.

A complexidade para a interpretação do morfema flexional propriamente verbal, em português, ou seja, o modo-temporal, decorre, em primeiro lugar, justamente da cumulação dessas duas noções, além da noção suplementar de aspecto, que às vezes se inclui na noção de tempo (Câmara Jr., 1986). De maneira muito resumida, pois o assunto é complexo, o tempo verbal refere-se ao momento de ocorrência do processo, visto do momento da comunicação; já o modo refere-se a um julgamento implícito do falante a respeito da natureza, subjetiva ou não, da comunicação que faz. No entanto, é comum em português, assim como em outras línguas, o emprego modal dos tempos verbais, que já foi chamado de metafórico. Não obstante, a apreciação do modo em português tem de se firmar, inicialmente, nas formas modais propriamente ditas, deixando à margem o emprego metafórico dos tempos (Câmara Jr., 1986).

Outrossim, há 06 sufixos número-pessoais, para indicar os falantes (1.ª pessoa do discurso), os ouvintes (2.ª pessoa do discurso) e as entidades sobre quem se fala (3.ª pessoa do discurso) (Câmara Jr., 1986). No português do Brasil (PB), a segunda pessoa do discurso pode se valer da terceira pessoa gramatical. As pessoas gramaticais são designadas por 1, 2 e 3 do singular (S) e do plural (P), assim, tem-se: 1S (eu), 2S (tu), 3S (ele, ela, você), 1P (nós), 2P (vós) e 3P (eles, elas, vocês). Como visto, no PB, usam-se conjugadas como 3S e 3P "você" e "vocês", respectivamente, o que aumenta o nível de ambigüidade do sistema de verbos, pois apesar das flexões da terceira pessoa, essas formas referem-se à segunda pessoa do discurso. Ainda, 1P (nós) pode ser substituída por "a gente", quando então assume as flexões de 3S ou, muito mais raramente, 1P. Os últimos casos são tão potencialmente ambíguos, que as pessoas gramaticais sempre estão expressas, o que facilita a desambiguação pelo contexto escrito.

Estudo recente (Scliar-Cabral, 2008) propõe o refinamento da fórmula anterior para:

TEMA (RAD+VT) + SF (SMT**A** + SNP + **SPF**)

com a inclusão do acento ou suprafixo (SPF) e da categoria de aspecto (A). Essa inclusão do acento de intensidade com a função de assinalar diferenças aspectuais tem sido negligenciada na literatura sobre aquisição da linguagem, o que causa problemas para a codificação automatizada ainda não tratados, como a queda do morfema *-r* do infinitivo na pronúncia. Contudo, a ferramenta que aqui se expõe lida com a língua escrita, não foca, por enquanto, esse ponto. Estima-se abordar essa questão com apoio da fonologia, em outra fase do projeto, tarefa para a qual o programa Laça-palavras já está preparado.

De posse desse conhecimento, levaram-se em conta todas essas considerações e transformaram-se essas regras – e outras não detalhadas aqui – em algoritmos. Para tanto, fizeram-se ajustes e complementações que o ambiente computacional exige, obviamente, e isso implicou a criação de novas regras. Depois, estudou-se o comportamento do aplicativo, a fim de se observarem ambigüidades geradas e resolvê-las, bem como para criar um léxico verbal automático para o córpus de trabalho. A criação do léxico – que já foi demonstrada (Vasilévski, Scliar-Cabral e Araújo, 2012) – resolveu as ambigüidades geradas pela alomorfia da vogal temática, nas três conjugações, e pela harmonia vocálica que ocorre no radical de verbos da 3.ªC, uma vez que a harmonia vocálica que ocorre no radical de verbos da 1.ªC e 2.ªC conjugações não é registrada no sistema escrito. Ainda, esse léxico, associado a instruções computacionais, resgata radicais regulares que sofrem transformações ditadas pelos valores grafêmicos, pelos quais: "g", quando vem antes de "e" e "i", é escrito "gu", para preservar o valor de /g/, como "**lig**ar" → "**ligu**ei"; "c", antes de "e" e "i", é escrito "qu", para preservar o valor de /k/, como "**fic**ar" → "**fiqu**ei"; e "c", antes de "o", "a" e "u", é escrito "ç", para preservar o valor de /s/, como "**esquec**er" → "**esqueç**o", "**esqueç**a". Trabalho a ser publicado detalha esse processo e outros semelhantes.

Antes de passar-se à ferramenta para análise morfológica dos verbos do português, cabe resumir o funcionamento e os recursos do programa Laça-palavras, que é o ambiente no qual está inserida essa ferramenta.

## *3. O programa Laça-palavras*

O funcionamento geral do programa Laça--palavras foi relatado anteriormente (Scliar--Cabral e Vasilévski, 2011), bem como algumas de suas ferramentas (Vasilévski, 2010, 2011a, 2011b, 2011c) e resultados oriundos de sua implementação parcial (Vasilévski, 2011d; Costa e Scliar-Cabral, 2011).

O Laça-palavras (LP) surgiu da necessidade de haver flexibilidade dos dados de trabalho maior do que a oferecida pelo programa Clan, disponibilizado pela Plataforma Childes e usado para se ler o córpus. Foi preciso se disporem os dados de diferentes formas e se extraírem deles informações que não eram possibilitadas pelo Clan. O Laça-palavras volta-se para arquivos em português, trabalha em conjunto com o Clan e também disponibiliza recursos próprios.

As interfaces do Laça-palavras com o programa Clan e as diretrizes dessa interação já foram descritas (Scliar-Cabral e Vasilévski, 2011). Então, cabe apenas lembrar suas principais ferramentas: 1) pesquisa no córpus, com marcação das linhas de seus enunciados com o tipo de discurso – de adulto para criança (ad-chi), de criança para adulto (chi-ad) e de adulto para adulto (ad-ad) –, resgate de palavras específicas – ou grupos de palavras – para trabalho com classes gramaticais, geração de relatório estatístico; 2) criação no córpus de uma linha denominada %pho, mediante interface com o programa Nhenhém (Vasilévski, 2008), para fazer a transcrição fonológica automática, com marcação das sílabas tônicas de determinado enunciado do arquivo, com ajuste da transcrição fonológica para fonética; 3) criação de uma linha para tradução morfológica automática dos verbos, chamada %mor, cuja ferramenta que a controla é foco deste estudo. Apesar de estar dentro do Laça-palavras e de isso facilitar sobremaneira seu uso, o analisador morfológico, quando estiver concluído em todas suas etapas, poderá ser instalado diretamente no computador, sem obrigatoriedade de haver também instalado o Laça-palavras. Não obstante, a integração entre ferramentas traz vantagens ao usuário do analisador, já que elas se comunicam entre si e compartilham resultados.

O processamento automático das unidades morfológicas dos enunciados do córpus coloca à disposição dos pesquisadores que trabalham com a morfologia do português uma eficiente ferramenta para análises quantitativas e qualitativas. No plano teórico, contribui em nível explicativo para melhor compreensão da construção das gramáticas do PB, particularmente, do sistema de verbos, e amplia o entendimento sobre o papel do *input* na construção de tais gramáticas (Scliar--Cabral, 2008), além de demonstrar a intuição do adulto, ao utilizar um registro adequado ao nível da criança.

## *4 Padrões e convenções do analisador*

O correto funcionamento do programa depende da metodologia empregada, sobretudo, na delimitação das tarefas que ele deve executar e na criação de códigos para as categorias que ele deve controlar. O analisador está preparado para carregar e ler arquivos criados no programa Clan, então, adotaram-se convenções estipuladas por esse programa para as classes gramaticais e para anotar córpus de língua oral (MacWhinney, 2000), bem como se criaram outras convenções específicas para o analisador morfológico.

### 4.1 Preparação do córpus

Para a pesquisa, mostrou-se relevante anotar os verbos diretamente no córpus, na linha do enunciado, no sistema Clan, com @v (verbos regulares – *default*), @vi (verbos irregulares) e @va (verbos auxiliares), para possibilitar, no Laça-palavras, a pesquisa (resgate e filtragem de dados), a análise morfológica automática e, conseqüentemente, a criação da linha %mor no arquivo original a ser lido pelo Clan. No entanto, para fins de clareza e limpeza do texto, esses símbolos, bem como outros símbolos do Clan, podem ser omitidos na pesquisa feita pelo LP, a critério do usuário. Todos os verbos auxiliares são irregulares, mas a decisão de assinalá-los separadamente se deve ao fato de preparar a computação, posteriormente, das locuções verbais e dos tempos compostos (Scliar-Cabral e Vasilévski, 2011).

## 4.2 Nomenclatura

Além da notação no córpus com @v, @vi e @va, usou-se um código para cada um dos tempos verbais do português, em seus respectivos modos. Assim, inseriram-se no programa os seguintes códigos: PI – Presente do Indicativo, PII – Pretérito Imperfeito do Indicativo, PPI – Pretérito Perfeito do Indicativo, PMI – Pretérito Mais-que-perfeito do Indicativo, FPI – Futuro do Presente do Indicativo, FPPI – Futuro do Pretérito do Indicativo, PS – Presente do Subjuntivo, PIS – Pretérito Imperfeito do Subjuntivo, FS – Futuro do Subjuntivo, IMA – Imperativo Afirmativo, IMN – Imperativo Negativo, INF – Infinitivo, GER – Gerúndio, PAR – Particípio.
Da mesma forma, as pessoas gramaticais, como visto, são assim designadas: 1S, 2S, 3S, 1P, 2P e 3P.

# *5 Análise morfológica automática*

Como mencionado, para a criação da linha %mor, foi desenvolvida uma ferramenta específica, o analisador, cujo algoritmo contém as regras das três conjugações verbais, em seus respectivos modos e tempos (Vasilévski, 2011b). Cabe esclarecer que o sistema não conjuga verbos, mas sim analisa entradas, que devem ser formas verbais escritas corretamente flexionadas.

## 5.1 Regras gerais

O primeiro conjunto de regras gramaticais desenvolvido foi relativo às vogais temáticas, seguido das regras dos morfemas modo-temporais e então das regras dos morfemas número-pessoais, para os verbos regulares. Tais regras foram formalizadas, para posterior inserção no programa. As regras da vogal temática foram objeto de trabalho anterior (Vasilévski, Scliar-Cabral e Araújo, 2012) e as regras específicas das pessoas gramaticais serão objeto de trabalho futuro. Aqui, cabe detalhar o segundo conjunto, ou seja, os sufixos ou desinências de modo e tempo, que se aglutinam em português. Exemplificam-se algumas dessas regras.

| SMT | se realiza | como | em contexto | Exemplos |
|---|---|---|---|---|
| \|-va-\| | → | -ve- / -va- | a__i / ... | cantá**ve**is<br>louva**va**,<br>ligá**va**mos |

Figura 1: Esquema de regras alomórficas do sufixo flexional modo-temporal da 1.ª C para o pretérito imperfeito do modo indicativo.

O esquema da Figura 1 mostra as regras alomórficas da desinência modo-temporal do pretérito imperfeito do modo indicativo, para a 1.ªC, no qual se nota que somente há alomorfia (de *-va-* para *-ve-*) na segunda pessoa do plural (vós), na qual o SMT está entre as vogais "a" e "i", contexto que condiciona a alomorfia. Nas demais pessoas (...), permanece o SMT inicial. Aliás, a forma pessoal "vós" tem uso restrito no Brasil, mas é usada em outros países em que se fala português. Preservam-se no programa formas pouco usadas no Brasil – por estarem consagradas na literatura e ainda em uso no discurso atual religioso, bem como nas músicas desse teor, por exemplo –, pois um sistema automático deve abranger todas as possibilidades oferecidas pela língua, sejam elas pouco ou muito usadas, e isso vale para os tempos verbais.

| SMT | se realiza | como | em contexto | Exemplos |
|---|---|---|---|---|
| \|-ia-\| | → | -ie- / -ia- | __i / ... | venc**ie**is<br>aplaud**ia**m |

Figura 2: Esquema de regras alomórficas do sufixo flexional modo-temporal da 2.ªC e 3.ªC para o pretérito imperfeito do modo indicativo.

O esquema da Figura 2 mostra que, no pretérito imperfeito do modo indicativo, na 2.ªC e 3.ªC, somente há alomorfia em 2P.
O esquema da Figura 3, a seguir, mostra que, na 1.ªC, 2.ªC e 3.ªC, no futuro do presente do modo indicativo, o morfema respectivo *-re-* sofre alomorfia para *-rá-*, em fim de vocábulo (#) e antes de "s" em fim de vocábulo, ou seja, em 2S e 3S, e sua vogal aberta "a" recebe til diante da vogal "o" em fim de vocábulo, ou seja, em 3P. Da mesma forma, a partir do esquema da Figura 4, observa-se que, no futuro do pretérito do modo indicativo, para as três

conjugações do português, somente há alomorfia da desinência *-ria-* em 2P, ou seja, diante da vogal "i".

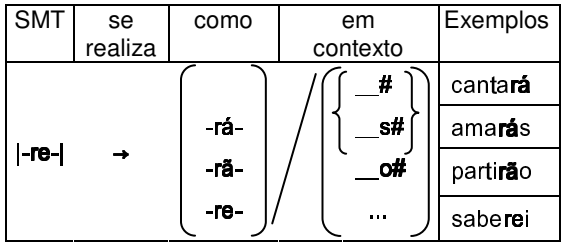



Figura 3: Esquema de regras alomórficas do sufixo flexional modo-temporal da 1.ªC, 2.ª C e 3.ªC para o futuro do presente do modo indicativo

O analisador contém as regras dos casos em que acentos gráficos ocorrem na vogal temática ("lig**á**vamos") e nas desinências, como mostram a figuras anteriores. Então, se o usuário omitir o assento gráfico do vocábulo verbal a ser analisado, o sistema poderá acusar erro, por não encontrar uma regra em que tal vocábulo se encaixe, ou encaixá-lo em uma regra incorreta.

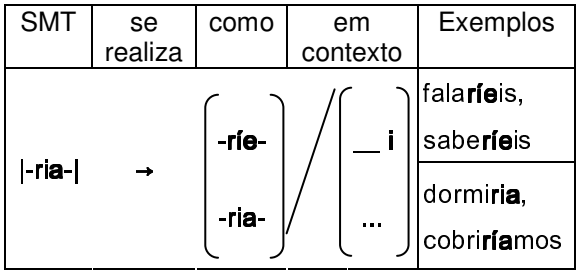



Figura 4: Esquema de regras alomórficas do sufixo flexional modo-temporal da 1.ªC, 2.ª C e 3.ªC para o futuro do pretérito do modo indicativo.

A partir das regras formalizadas e com apoio da literatura, fez-se um quadro geral do comportamento dos morfemas modo-temporais, com suas respectivas alomorfias, em parênteses, para os tempos verbais do português:

Quadro 1: Regras alomórficas dos sufixos modo-temporais do português.

| **MT** | **SMT** | | | Onde há alomorfia |
|---|---|---|---|---|
| | **1.ª C** | **2.ª C** | **3.ª C** | |
| PI | Ø | Ø | Ø | - |
| PII | va (ve) | ia (ie) | ia (ie) | 2P |
| PPI | Ø (ra) | Ø (ra) | Ø (ra) | 3P |
| PMI | ra (re) | ra (re) | ra (re) | 2P |
| FPI | re (rá, rã) | re (rá, rã) | re (rá, rã) | 2S,3S,3P |
| FPPI | ria (ría,ríe) | ria (ría, ríe) | ria (ría, ríe) | 1P e 2P |
| PS | e | a | a | - |
| PIS | sse | sse | sse | - |
| FS | r (re) | r (re) | r (re) | 2S |
| IMA | e (Ø) | a (Ø) | a (Ø) | 2S e 2P |
| IMN | e | a | a | - |
| INF | r (re) | r (re) | r (re) | 2S |
| GER | ndo | ndo | ndo | - |
| PAR | do | do (to) | do | - |

Cabe destacar que o pretérito imperfeito do subjuntivo e o gerúndio são tempos não ambíguos, pois seus morfemas são exclusivos e não há alomorfes. Os futuros do presente e do pretérito do indicativo têm alomorfes, os quais são exclusivos desses tempos, o que também torna esses tempos não ambíguos. Os demais tempos estão sujeitos a ambigüidades em alguma conjugação. Esse assunto voltará a foco.

## 5.2 Regras dos verbos irregulares

O algoritmo que contém as regras verbais está em fase de aprimoramento, para que dê conta dos verbos irregulares do PB. Assim, é válido esboçar algumas diretrizes que guiarão tal trabalho.

Verbos irregulares, na verdade, são formas irregulares, que devem ser entendidas como desvios do padrão geral morfológico, que não deixam de ser regulares, no sentido de que são suscetíveis a uma padronização. Trata-se de pequenos grupos de verbos, com certos padrões comuns, que podem ser explicitados (Câmara Jr., 1986). Tais irregularidades podem ser referir aos sufixos, mas, quando ocorrem no radical, são muito mais relevantes para a análise morfológica automática, pois se cria uma série de padrões morfológicos. Ainda, nesses verbos ocorre constantemente a supressão da vogal temática, o que acontece também na segunda e terceira conjugações com os verbos regulares, e provoca entrave na análise morfológica automática, pois se perde a conjugação do verbo, o que conseqüentemente dificulta o resgate da forma infinitiva desse verbo. A partir disso, entende-se que poderá ser bem-sucedida a decomposição morfológica automática desses verbos.

## *6. Desempenho*

O analisador morfológico verifica os verbos contidos em um córpus previamente preparado, carregado no sistema Laça-palavras, que o abriga. As formas verbais anotadas são automaticamente lidas e analisadas, e o resultado é mostrado. Internamente, ocorre que, ao identificar uma forma verbal, o analisador morfológico a compara com suas regras internas, para decompô-la em morfemas.

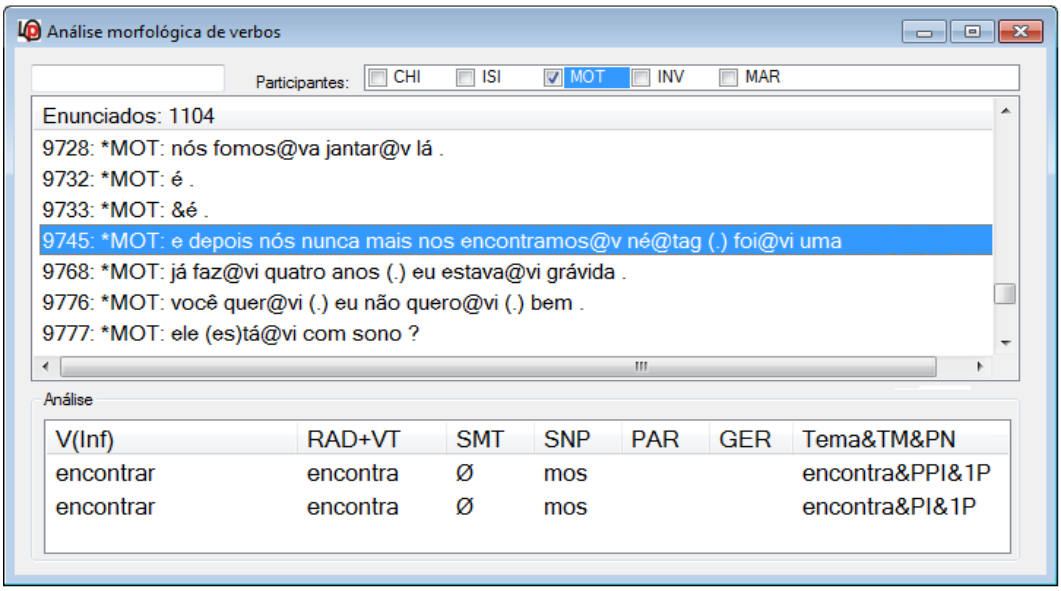



Figura 5: Tela principal do analisador morfológico automático.

Assim, a saída do programa é a realização da fórmula geral da estrutura do vocábulo verbal português e a depreensão de sua forma infinitiva. Cabe esclarecer que a criação do léxico dos verbos do córpus resolveu problemas de sobregeração de formas infinitivas, causada por conflito entre as regras do sistema de verbos do português. Por exemplo, a entrada "cantava" gerava as opções de formas infinitivas "cantar", *"cantavar", *"cantaver" e *"cantavir", das quais somente a primeira é correta e existe em português. Com a criação do léxico, o programa faz a análise e obtém todas as formas possíveis, mas, antes de mostrá-las, compara-as com o conteúdo do léxico verbal. Então, escolhe a forma que está contida no léxico e mostra somente ela (Vasilévski, Scliar-Cabral e Araújo, 2012).

A Figura 5 traz o resultado da análise morfológica da forma verbal "encontramos", que ocorre no enunciado da linha 9745 do córpus – o arquivo pau003.cha – e é proferido pelo participante MOT (a mãe da criança), quando se dirige a outro adulto. À medida que o cursor desce pelos enunciados, cada forma verbal identificada é analisada automaticamente pelo programa. Ao encontrar mais de um verbo na mesma linha do enunciado – por exemplo, na linha 9728, em que há "fomos" e "jantar" –, o programa analisa-o abaixo do verbo anterior. O campo Participantes permite ao usuário escolher os participantes cujos enunciados ele quer verificar. No córpus, há cinco participantes, sendo CHI a criança, os demais são adultos. Ao carregar o arquivo, o programa preenche esse campo com os participantes automaticamente, para o usuário selecionar os que deseja checar. Após essa seleção, aparecerá, no campo Enunciados, a quantidade de enunciados encontrados referentes ao(s) participante(s) selecionado(s), e abaixo, os enunciados propriamente ditos, no formato do Clan.

Para verificar os verbos de um enunciado, o usuário clica sobre ele, e o campo Análise fornecerá a análise morfológica do verbo que aparece no enunciado selecionado. A primeira informação morfológica fornecida é a forma infinitiva do verbo em questão (V(Inf)), seguida pela depreensão do tema (RAD+VT), desinências modo-temporal (SMT) e número-pessoal (SNP), formas nominais particípio (PAR) e gerúndio (GER). Finalmente, essa análise morfológica é traduzida em uma

seqüência de morfemas separados pelo caractere & – que nesse caso indica inserção de sufixo –, a qual repete o tema e fornece tempo/modo e pessoa/número verbais: Tema&TM&PN.

Assim, a Figura 5 mostra que a forma verbal "encontramos" é da 1.ªC, pois sua forma infinitiva (Inf) é "encontrar" e sua vogal temática é "a", que, portanto, não sofreu alomorfia nem desapareceu; não há morfema modo-temporal, o que é assinalado por Ø; o sufixo número-pessoal embutido nela é "mos"; e não se trata de particípio nem de gerúndio. Tudo isso junto diz que essa forma verbal está conjugada em 1P do presente ou do pretérito perfeito, ambos do modo indicativo.

Observam-se, nos enunciados que aparecem na Figura 5, outras anotações usadas no córpus, em palavras que não são verbos, como, por exemplo: *, que indica que a palavra seguinte designa um participante; &, no início de palavras que devem ser descartadas da computação, por serem imitações ou hesitações; @tag, que refere partículas interrogativas que pedem confirmação no final de um enunciado, como né@tag; e parênteses, que denotam fonemas que foram omitidos na fala, como em p(r)ato.

## 6.1 Ambigüidades do sistema de verbos do Português

Os casos em que as regras são ambíguas se revelam na resposta do programa. Isso era de se esperar, pois a reprodução pelo programa das ambigüidades do sistema de verbos do português mostra que seu algoritmo corresponde a este sistema. Cabe documentar aqui as ambigüidades do sistema de verbos do português relacionadas aos sufixos modo--temporais.

As ambigüidades cuja resolução é complicada são justamente causadas pela ausência de morfemas específicos que distingam formas verbais. Aliás, quando essas formas ocorrem em um texto, nem sempre é claro para o leitor o tempo verbal em que elas estão. A Figura 5 reproduz a ambigüidade do sistema de verbos no que se refere à ausência de sufixo modo--temporal tanto para 1P-PI como para 1P-PPI. Nesse caso, somente o contexto poderá desambiguar a forma verbal.

Por exemplo, quando a ambigüidade ocorre com o modo imperativo, o contexto pode facilitar a desambiguação ou encarregar-se dela. As formas imperativas afirmativas normalmente ocorrem no início do enunciado ou logo após um vocativo, ao qual sucede uma vírgula, e ocorrem com a segunda pessoa do discurso, isto é, segunda ou terceira pessoas gramaticais. Elas também ocorrem após "por favor" – que não ocorre no córpus de trabalho – e após outras poucas expressões semelhantes. Os morfemas modo-temporais do imperativo negativo coincidem com os do presente do subjuntivo, no entanto, o imperativo negativo, além de estar no mesmo contexto do imperativo afirmativo, sempre vem acompanhado do advérbio de negação "não", de modo que se facilita a resolução da ambigüidade pelo contexto.

A ambigüidade causada pelo uso das flexões de 3S e 3P para as formas "você" e "vocês" é de resolução mais complicada em alguns casos, contudo, nesses casos, a pessoa gramatical normalmente é expressa no enunciado, de maneira que novamente o contexto facilita a desambiguação. Por exemplo, a forma verbal "mostra", do enunciado da linha 940 do córpus:

0940: *MOT: depois você mostra@v p(a)r(a) o papai .

gerava as duas saídas seguintes, das quais nenhuma era correta, pois o pronome subjetivo expresso na sentença não deixa dúvida de que não se trata de imperativo, mas se trata de 2S:

Quadro 2: Resposta inicial do programa à entrada "mostra".

| (RAD+VT) | | SMT | SNP | Tema&TM&PN |
|---|---|---|---|---|
| mostr | a | Ø | Ø | mostra&PI&3S |
| mostr | a | Ø | Ø | mostra&IMA&2S |

Observe-se que à forma verbal "mostra" não está agregado morfema modo-temporal nem número-pessoal – ambos são zero. Como diferenciar tempo/modo e pessoa/número, então, se a forma verbal não os expressa? Para resolver esse caso, criou-se uma rotina computacional que verifica o enunciado, à

procura das formas "você", "vocês" e "a_gente".
O pronome "você" ocorre 325 vezes no córpus, e a criança usa-o duas vezes. Por exemplo:

0044 *INV: ah@i (.) você acendeu@v a luz ?
5516 *CHI: vo(u)@va liga(r)@v p(r)a você .

O pronome "vocês" ocorre 14 vezes, e a criança não o usa. Por exemplo:

2855 *ISI: vocês conseguem@va sentar@v os dois juntos ou +...

A forma composta "a gente" somente é pronominal se as duas palavras que a compõem estiverem nessa seqüência e precederem um verbo ou precederem a partícula "se" e/ou um ou mais advérbios antes desse verbo. Para evitar ambigüidade, no córpus, as duas palavras que a compõem aparecem ligadas por "_". No córpus, ela aparece 41 vezes, todas nessa situação. Por exemplo:

1402 *INV: a_gente se diverte@v ,, né ?

O funcionamento dessa rotina computacional consta, em forma de fluxograma, na Figura 6. Depois dessa complementação, a resposta do programa à situação do Quadro 2 é: mostra&PI&2S.
Obter tal distinção nem sempre é fácil, mesmo porque há casos, como visto, em que a pessoa gramatical não é expressa ou está distante do verbo, o que não garante que ela seja seu sujeito. Apesar disso, a grande maioria dos casos fica resolvida com a verificação do contexto do enunciado. Na rotina computacional demonstrada no fluxograma anterior, foi implementada uma instrução para que seja ignorada a partícula "se" anteposta a um verbo, de forma que o verbo "diverte" do enunciado da linha 1402, do exemplo anterior, é corretamente analisado pelo programa: diverte&PI&1P. A análise completa do programa mostra que, nesse caso, SMT e SNP são Ø e que há alomorfia da VT da 3.ªC, que passa de "i" ("diver**t**i**r**") para "e" ("diver**t**e").

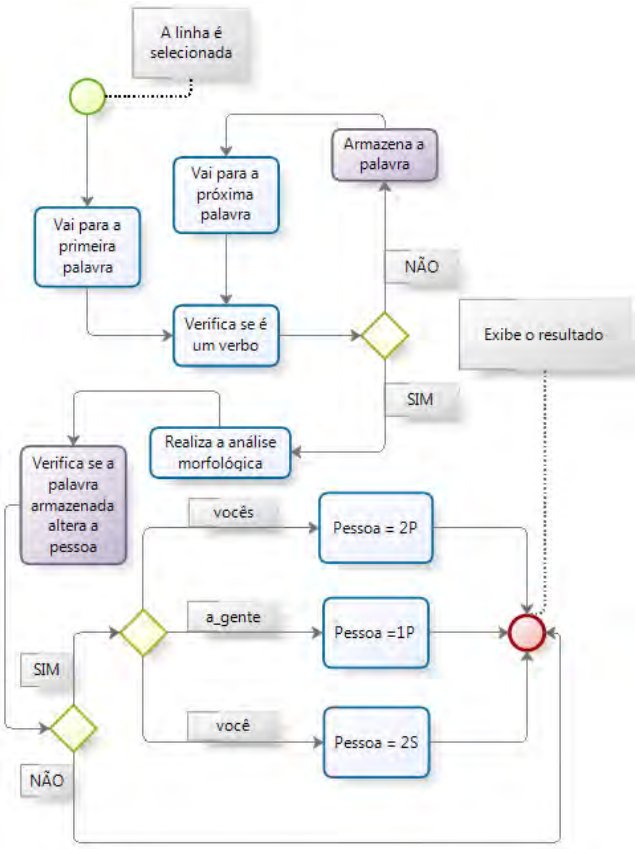



Figura 6: Fluxograma da função que verifica "você", "vocês" e "a_gente" no enunciado.

À medida que a programação avança, revelam-se mais questões a serem tratadas. Assim, a resolução completa das ambigüidades relacionadas às desinências modo-temporais passa pela consideração das desinências número-pessoais e das pessoas gramaticais em si, pois mesmo os tempos verbais não ambíguos dependem das flexões número-pessoais para sua correta e completa análise, bem como do reconhecimento das palavras de outras classes gramaticais que circundam o verbo. Somente a partir disso será possível reduzir ao mínimo ou, talvez, eliminar – a pesquisa dirá – as ambigüidades do sistema de verbos do português do Brasil ocasionadas pelo comportamento dos morfemas modo-temporais.

## *7 Conclusão e perspectivas*

A fase do analisador morfológico automático para verbos do português aqui documentada descreve a automatização das desinências modo-temporais, assim como aborda as ambigüidades provocadas pelo comportamento desses morfemas e apresenta soluções para a

maioria desses casos. No entanto, algumas ambigüidades persistem e, vale dizer, outras podem aparecer. A criação de regras computacionais para verificar o contexto do enunciado foi a principal solução adotada, e o próximo passo será estudar o comportamento dos morfemas número-pessoais, para complementar a desambiguação. Para essa tarefa, já se vislumbram novas regras, que levam em conta, sobretudo e novamente, o contexto do enunciado.

Como se percebe, o trabalho está em evolução constante, de forma que algumas respostas e conclusões somente poderão ser fornecidas ao fim de todas as etapas. Nesse trajeto, pode haver redefinições e redirecionamentos, frutos de aprendizado e testagem empírica.

## *Agradecimento*

Este trabalho é desenvolvido pelo Laboratório de Produtividade Lingüística Emergente da UFSC (LAPLE), com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), entidade do governo brasileiro voltada para a formação de recursos humanos, à qual agradecemos.

## *Referências*

Câmara Jr., Joaquim Mattoso. 1986. *Estrutura da língua portuguesa.* 26.ed. Petrópolis, RJ: Vozes.

Câmara Jr., Joaquim Mattoso. 1976. *História e estrutura da língua portuguesa*. 2.ed. Rio de Janeiro: Padrão.

Childes – Child Language Data Exchange System. 1991-2011. *Clan*: Computerized Language Analysis. http://childes.psy.cmu. edu/clan/

Childes. 2011. Index of Data. *pau003.cha*. http://childes.psy.cmu.edu/data/Romance/ Portuguese/ Florianopolis.zip

Costa, Richard Fernando S. & Scliar-Cabral, Leonor. 2011. Regularização do sistema verbal pela criança. *Anais do Simpósio Internacional Linguagens e Culturas: Homenagem aos 40 anos dos programas de Pós-graduação em Lingüística, Literatura e Inglês da UFSC (SILC)*, Florianópolis, Brasil.

MacWhinney, Brian. 2003-2011. Child Language Data Exchange System. http://childes.psy.cmu.edu/

MacWhinney, B. 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd. ed. Mahwah, NJ: Lawrence Erlbaum Associates. http://childes.psy.cmu.edu/manuals/chat.pdf

Scliar-Cabral, Leonor. 2011. Análise Automática da Morfologia do PB (Plataforma CHILDES): aquisição da morfologia verbal. Em *VII Congresso Internacional da Abralin*, Curitiba, Brasil.

Scliar-Cabral, Leonor. 2009. *Análise morfológica Automática do Português*. CAPES-UFSC, Florianópolis.

Scliar-Cabral, Leonor. 2008. Codificação da morfologia do PB e análise da fala dirigida à criança. *Fórum Lingüístico*, vol. 5(2), pp.69-82, Florianópolis.

Scliar-Cabral, Leonor. 2007. Emergência gradual das categorias verbais no Português brasileiro. *Alfa*, vol. 51(1), pp.223-234, São Paulo.

Scliar-Cabral, Leonor. 2003. *Princípios do sistema alfabético do português do Brasil*. São Paulo: Contexto.

Scliar-Cabral, Leonor & Vasilévski, Vera. 2011. Descrição do português com auxílio de programa computacional de interface. *Anais da II Jornada de Descrição do Português (JDP)*, Cuiabá, Brasil.

Vasilévski, Vera & Araújo, Márcio J. 2010-2011. *Laça-palavras:* sistema eletrônico para descrição do português brasileiro. LAPLE-UFSC, Florianópolis. https://sites.google.com/ site/sisnhenhem/

Vasilévski, Vera. 2011a. O hífen na separação silábica automática. *Revista do Simpósio de Estudos Lingüísticos e Literários - SELL*, vol. 1(3), pp.657-676, Uberaba.

Vasilévski, Vera. 2011b. Programa para processamento automático das unidades verbais do PB. Análise automática da morfologia do PB (Plataforma CHILDES): aquisição da morfologia verbal. Em *VII*

*Congresso Internacional da Abralin*, Curitiba, Brasil.

Vasilévski, Vera. 2011c. An automatic system for verb morphological analysis of BP. Em *VIII Encontro Inter-Nacional de Aquisição da Linguagem (ENAL)*, Juiz de Fora, Brasil.

Vasilévski, Vera. 2011d. Diferenças entre Input e Intake: evidências na aquisição de pronomes interrogativos. *Anais do Simpósio Internacional Linguagens e Culturas: Homenagem aos 40 anos dos programas de Pós-graduação em Lingüística, Literatura e Inglês da UFSC (SILC)*, Florianópolis, Brasil.

Vasilévski, Vera. 2010. Divisão silábica automática de texto escrito baseada em princípios fonológicos. *Anais do III Encontro de Pós-graduação em Letras da UFS (ENPOLE)*, São Cristóvão, Sergipe, Brasil.

Vasilévski, Vera. 2008. *Construção de um programa computacional para suporte à pesquisa em fonologia do português do Brasil*. Tese de doutorado, Florianópolis: UFSC.

Vasilévski, Vera; Scliar-Cabral, Leonor & Araújo, Márcio J. 2012. Automatic Analysis of Portuguese Verb Morphology: Solving Ambiguities Caused by Thematic Vowel Allomorphs. In *The 10th International Conference on the Computational Processing of Portuguese (PROPOR)*, Coimbra, Portugal, April 17-20.

# 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 2012
- Resultados da selecção até: 15 de Maio de 2012
- Versão final até: 31 de Maio de 2012
- Publicação da revista: Junho de 2012

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 2012
- Resultados da selección: 15 de maio de 2012
- Versión final: 31 de maio de 2012
- Publicación da revista: xuño de 2012

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 2012
- Resultados de la selección: 15 de mayo de 2012
- Versión final: 31 de mayo de 2012
- Publicación de la revista: junio de 2012

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 2012
- Resultats de la selecció: 15 de maig de 2012
- Versió final: 31 de maig de 2012
- Publicació de la revista: juny de 2012

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: 2012eko apirilak 15.
- Hautapenaren emaitzak: 2012eko maiatzak 15.
- Azken itzulpena: 2012eko maiatzak 31.
- Aldizkariaren argitarapena: 2012eko ekainean.

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

Volume 3, Número 2 - Dezembro 2011

linguaMATICA

ISSN: 1647-0818

## Dossier



## Artigos de Investigação



## Novas Perspectivas



## Apresentação de Projectos