Editores
Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Número I - Maio 2009

linguaMATICA

ISSN: 1647-0818

Universidade de Vigo
Universidade do Minho
25 Anos Associação Portuguesa Para a Inteligência Artificial

Número 1 – Maio 2009

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

A revista Linguamática *pretende colmatar uma lacuna na comunidade de processamento de linguagem natural para as línguas ibéricas. Tem como principal objectivo a publicação de artigos que visem o processamento de alguma destas línguas, e escritos também numa destas línguas.*

Co fin de fomentar a investigación nesta área de traballo, Linguamática *quere ser unha revista completamente aberta. Os artigos publícanse en versión electrónica e son postos ao dispor de toda a comunidade de xeito totalmente gratuíto coa licenza* Creative Commons.

Este es el primer volumen de la revista. El proceso de aceptación, análisis, evaluación y selección de las propuestas fue lento y se vio sujeto a varios contratiempos, debidos fundamentalmente a la poca experiencia de los editores en la gestión de una revista científica. La revista recibió 11 contribuciones, de las que se seleccionaron 5. Durante el proceso de evaluación, detectamos algunos problemas que pueden haber motivado el rechazo de artículos válidos. Somos conscientes de estos problemas y estamos elaborando nuevas estrategias para que el segundo volumen sea mejor.

A més a més dels articles enviats per l'iniciativa dels mateixos autors i autores, els editors van encarregar dos articles convidats per tractar de manera monogràfica dos casos d'èxit en el processament de les llengües ibèriques: el sistema de traducció automàtica de codi obert Apertium *i la* Linguateca*, un centre de recursos per al processament computacional de la llengua portuguesa.*

Os editores agradecem a todas as pessoas que ajudaram nesta edição, aos autores e autoras que contribuíram com artigos (seleccionados ou não) e aos revisores e revisoras que leram e comentaram os artigos submetidos. O nosso muito obrigado, moitas grazas, moltes gràcies, muchas gracias.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís, Universidade de Vigo
Alberto Simões, Universidade do Minho
Álvaro Iriarte Sanroman, Universidade do Minho
Antón Santamarina, Universidade de Santiago de Compostela
António Teixeira, Universidade de Aveiro
Belinda Maia, Universidade do Porto
Carmen García Mateo, Universidade de Vigo
Diana Santos, SINTEF ICT
Gael Harry Dias, Universidade Beira Interior
Joaquim Llisterri, Universitat Autònoma de Barcelona
José João Almeida, Universidade do Minho
José Paulo Leal, Universidade do Porto
Joseba Abaitua, Universidad de Deusto
Iñaki Alegria, Euskal Herriko Unibertsitatea
Lluís Padró, Universitat Politècnica de Catalunya
Maria Antònia Martí Antonín, Universitat de Barcelona
Maria das Graças Volpe Nunes, Universidade de São Paulo
Mercè Lorente Casafont, Universitat Pompeu Fabra
Mikel Forcada, Universitat d'Alacant
Nieves R. Brisaboa, Universidade da Coruña
Salvador Climent Roca, Universitat Oberta de Catalunya
Xavier Gómez Guinovart, Universidade de Vigo

Dossier

Apertium: traducció automàtica de codi obert per a les llengües romàniques

Mikel L. Forcada
Universitat d'Alacant / Prompsit Language Engineering
mlf@ua.es

Resum

Es descriu breument la plataforma de traducció automàtica Apertium (www.apertium.org). Apertium és programari de codi obert, és a dir, programari lliure, que serveix per a construir sistemes de traducció automàtica, que funciona especialment bé en el cas de llengües emparentades com les romàniques, i que està disponible des de 2005. Després d'una breu introducció a la traducció automàtica i a les especials característiques de la traducció automàtica de codi obert, s'expliquen els principis de disseny de la plataforma Apertium, se'n fa una breu descripció tecnològica, es descriu la comunitat de desenvolupadors que s'hi ha format al voltant i es dóna notícia de la recerca realitzada sobre aquesta plataforma. Més avant s'explica el compromís d'Apertium amb les llengües de la Romània, des dels inicis amb els parells espanyol-català i espanyol-gallec fins a la situació actual, amb molts altres parells de llengües romàniques disponibles i en desenvolupament, il·lustrant-lo amb l'aplicació de la plataforma a la llengua occitana.

1. Introducció

Aquest article descriu breument la plataforma de traducció automàtica Apertium (www.apertium.org). Apertium és programari de codi obert, és a dir, programari lliure, que serveix per a construir sistemes de traducció automàtica, que funciona especialment bé en el cas de llengües emparentades com les romàniques, i que està disponible des de 2005. Després d'una breu introducció a la traducció automàtica i a les especials característiques de la traducció automàtica de codi obert (secció 2), s'expliquen els principis de disseny de la plataforma Apertium, se'n fa una breu descripció tecnològica, es descriu la comunitat de desenvolupadors que s'hi ha format al voltant i es dóna notícia de la recerca realitzada sobre aquesta plataforma (secció 3). La secció 4 explica el compromís d'Apertium amb les llengües de la Romània, des dels inicis amb els parells espanyol-català i espanyol-gallec fins a la situació actual, amb molts altres parells de llengües romàniques disponibles i en desenvolupament, il·lustrant-lo amb l'aplicació d'aquesta a la llengua occitana. La secció 5 tanca l'article amb uns comentaris finals.

2. Traducció automàtica de codi obert

2.1 Traducció automàtica

2.1.1 Qué és

La traducció automàtica tracta amb textos escrits, i, en particular, amb **textos informatitzats**, és a dir, amb documents de text emmagatzemats en un mitjà informàtic, documents com els que es poden generar o editar amb processadors de textos. És *automàtica* perquè la realitzen *sistemes informàtics*, és a dir, ordinadors amb el programari adequat instal·lat. Entenem per **traducció automàtica** la transformació, usant un sistema informàtic, d'un text informatitzat escrit en la *llengua origen,* en un altre text informatitzat escrit en la *llengua meta,* que anomenarem *traducció en brut*.

2.1.2 Limitacions de la traducció automàtica

La traducció automàtica (TA) té **limitacions**. En general, les traduccions en brut produïdes pels sistemes de traducció automàtica solen ser molt diferents a les produïdes pels professionals de la traducció i poden no ser adequades per a alguns propòsits comunicatius. Aquesta inadequació està causada per diversos factors, entre els quals podem comptar l'*ambigüitat* dels textos humans (que contenen moltíssims mots amb més d'un sentit o frases amb més d'una estructura sintàctica), les divergències sintàctiques entre la llengua origen i la llengua meta, etc. Aquestos problemes s'aborden amb mètodes que, en general, fan simplificacions bastant radicals del procés de traducció. Aquestes simplificacions, d'una banda, permeten la formulació de regles mecàniques senzilles per a poder construir sistemes de traducció automàtica ràpids i compactes en un temps raonable, però, d'una altra, fa que les solucions estiguen lluny de ser òptimes.

2.1.3 Què podem esperar de la traducció automàtica?

En vista d'aquestes limitacions podem esperar que un bon sistema de TA ens allibere de la part més mecànica (o "mecanitzable") de la tasca de traducció, però, per bo que siga, no podem esperar que comprenga el text, resolga sempre les ambigüitats correctament i produïsca textos en una variant genuïna de la llengua meta.

2.1.4 Aplicacions

Hi ha **dos grans grups d'aplicacions** de la traducció automàtica. El primer grup el formen les aplicacions per a l'**assimilació,** és a dir, l'ús de la traducció automàtica per a comprendre el sentit general de documents (per exemple, textos publicats en Internet) escrits en altra llengua. Un altre exemple de traducció automàtica per a l'assimilació és la traducció de converses en un *xat* o *chat*, de manera que cada persona que hi participa pot usar la seua llengua i llegir les contribucions dels altres participants traduïdes també a la seua llengua. En aquest tipus d'aplicacions la traducció automàtica ha de ser molt ràpida, idealment instantània, i s'usa directament, en brut; hi ha vegades que ni tan sols es llig completament, i normalment no es conserva ni guarda després d'haver-la llegit. Aquesta aplicació de la traducció automàtica no està relacionada amb la traducció professional.

En el segon grup, hi ha les aplicacions per a la **disseminació**. Es diuen així perquè comporten l'ús de la traducció automàtica com a pas intermedi en la producció d'un document en la llengua meta que serà publicat o disseminat; per tant, la traducció en brut es conserva perquè l'ha de revisar i corregir, o com se sol dir, *posteditar,* una persona especialitzada. Simplificant, podem dir que la traducció automàtica seguida de postedició constituirà una alternativa a la traducció professional només si el seu cost conjunt és menor que el de la traducció professional tradicional. De vegades, per a estalviar postedició (especialment quan es tradueix a més d'una llengua meta) es pot fer una miqueta de *preedició* del text original que es traduirà automàticament, evitant problemes coneguts del sistema de traducció automàtica concret que s'estiga usant. Una alternativa a la preedició en el cas que s'han de crear i després traduir molts documents de naturalesa similar és que els autors usen *llenguatges controlats*, és a dir, que escriguen evitant lèxic i construccions que haurien estat posteditades.

2.1.5 Dos grans grups de tecnologies de traducció

També hi ha **dos grans grups de tecnologies de traducció**. Des dels primers intents de fa uns 50 anys fins al decenni dels noranta, l'aproximació dominant a la traducció automàtica ha sigut l'anomenada *traducció automàtica basada en regles*: equips amb informàtics i experts en traducció compilen diccionaris en forma electrònica, programen analitzadors morfològics i sintàctics, definixen regles de transformació gramatical, etc. Des de principis dels noranta assistim a un creixement de l'anomenada *traducció automàtica basada en corpus* (de text): els programes de traducció automàtica "aprenen a traduir" (per exemple usant complexos models estadístics) a partir d'enormes corpus de textos bilingües on centenars de milers de frases en una llengua s'han alineat amb la seua traducció en l'altra llengua.

Aquest article presenta **Apertium**, un sistema de traducció automàtica basada en regles.

2.2 Què és el programari de codi obert?

Revisem breument el concepte de *programari de codi obert* (*open-source software*), o, si usem el seu nom històric encara en ús, *programari lliure (free software*). El programari lliure (podeu trobar una definició a http://www.gnu.org/philosophy/free-sw.html) és programari que (a) pot ser usat lliurement amb qualsevol propòsit, (b) pot ser examinat lliurement per veure com funciona i pot ser modificat lliurement per adaptar-lo a una necessitat nova o a una nova aplicació (per això, el codi font ha de ser disponible, d'ací el nom alternatiu *de codi obert*), (c) pot ser redistribuït lliurement a qualsevol, i (d) pot ser millorat lliurement i alliberat al públic de manera que la comunitat sencera d'usuaris se'n beneficie (el codi font ha de ser disponible per a això també). La *Open Source Initiative* («Iniciativa de codi obert») estableix una definició (http://www.opensource.org/docs/definition.php) que és més o menys equivalent per als propòsits d'aquest article. En aquest article, use la denominació *codi obert* perquè el meu grup ho ha fet tradicionalment, i no perquè, com altres, vulga evitar les connotacions polítiques o ètiques

associades a la denominació *lliure*, les quals compartisc.

2.3 Programari de traducció automàtica: obert o tancat?

2.3.1 Peculiaritats del programari de traducció automàtica

El programari de traducció automàtica (TA) és especial perquè depén fortament de les dades. La traducció automàtica basada en regles (TABR) depèn de dades lingüístiques com ara diccionaris morfològics, diccionaris bilingües, gramàtiques i arxius de regles de transferència estructural; la traducció automàtica basada en corpus (com ara la traducció automàtica estadística, per exemple) depén, directament o indirectament, de la disponibilitat de text paral·lel alineat frase a frase. En els dos casos, s'hi poden distingir tres components: un *motor* (descodificador, recombinador, etc.), *dades* (dades lingüístiques o corpus paral·lels), i, opcionalment, *eines* per mantenir aquestes dades i convertir-los en un format adequat perquè els use el motor.

2.3.2 La traducció automàtica comercial, normalment tancada

HhhLa majoria dels sistemes de traducció automàtica comercials són basats en regles (tot i que han començat a aparéixer sistemes de traducció automàtica amb un fort component basat en corpus[1]). La majoria dels sistemes de TABR usen motors amb tecnologies privatives o de propietat (*proprietary*) que no es revelen completament (de fet, la majoria de les empreses consideren aquestes tecnologies de propietat com el seu principal avantatge competitiu). Les dades lingüístiques no són plenament modificables tampoc; en la majoria de casos, la persona usuària només pot afegir paraules noves o els seus glossaris als diccionaris del sistema, i potser afegir-hi algunes regles senzilles, però no és possible construir un conjunt complet de dades lingüístiques per a un parell de llengües nou i utilitzar-lo amb el motor.

Que un sistema es puga usar en Internet no vol dir que siga obert. Per exemple, hi ha sistemes de TA en la xarxa que poden ser utilitzats lliurement (amb algunes restriccions); alguns són versions de prova de sistemes comercials, mentre que alguns altres sistemes lliurement disponibles, però tancats, no són ni tan sols comercials.[2]

2.3.3 Traducció automàtica de codi obert

D'una banda, perquè un sistema de traducció automàtica basat en regles siga de «codi obert», el codi font del motor i de les eines han de ser distribuïts així com el «codi font» de les dades lingüístiques pels parells de llengües desitjats. És més facil que les persones usuàries de la traducció automàtica de codi obert canvien les dades lingüístiques que que modifiquen el motor de traducció automàtica; a més, perquè les dades lingüístiques millorades puguen ser utilitzades amb el motor, les eines per mantenir-les també haurien de ser accessibles. D'altra banda, si el sistema de traducció automàtica és estadístic, el codi font tant dels programes que aprenen els models estadístics de traducció a partir del text paral·lel així com dels descodificadors que utilitzen aquests models de llengua per generar les traduccions més probables de frases noves haurien de distribuir-se *conjuntament amb els corresponents textos paral·lels alineats frase a frase.*

Recentment, han començat a aparéixer sistemes de traducció automàtica de codi obert. El sistema Apertium que es descriu en aquest article és un d'ells. Es dóna el cas que fins i tot una empresa que es dedicava al negoci de la TA comercial ha començat a distribuir els seus productes com a codi obert.[3]

2.3.4 Avantatges de la TA de codi obert

Els sistemes de TA de codi obert tenen avantatges específics sobre els sistemes comercials de codi tancat. En particular, m'agradaria destacar-ne dos:

1. **Increment de la perícia i dels recursos lingüístics.** Quan s'intenta construir un sistema de traducció automàtica de codi obert per un parell de llengües nou, cal un procés de reflexió sobre les llengües implicades que porta a l'explicitació i a la subsegüent fixació i codificació de coneixement monolingüe i

1 AutomaticTrans (http://www.automatictrans.es), Language Weaver (http://www.languageweaver.com), i, més recentment, Google Translate (http://translate.google.com).

2 Aquest és el cas, per exemple, de dos sistemes de traducció automàtica no comercials però lliurement disponibles entre espanyol i català: interNOSTRUM (http://www.internostrum.com), el qual té milers d'usuaris diaris, i un sistema menys conegut però molt potent anomenat SisHiTra (González et al. 2006).

3 LOGOS ha alliberat recentment el codi font del seu sistema de TA, ara OpenLogos (www.logos-os.dfki.de).

bilingüe. Així doncs, d'una banda, la perícia lingüística resultant, en un escenari de codi obert, queda disponible per a les comunitats lingüístiques interessades. D'altra banda, es generen recursos nous, disponibles de manera oberta per a la comunitat de parlants de les llengües implicades, i que poden ser usats per a nous parells de llengües, o fins i tot per a altres aplicacions de tecnologia lingüística a més de la traducció automàtica.

2. **Augment de la independència.** Un efecte secundari interessant és que la disseminació de coneixement obert i programari de codi obert fa que els usuaris de les comunitats lingüístiques corresponents siguen menys dependents d'un proveïdor comercial particular de programari de codi tancat, no només quant a tecnologies de traducció, sinó potser també quant a d'altres aplicacions de tecnologia lingüística que se'n podrien derivar.

La secció 3.2.3 explica amb més detall les raons per les quals la plataforma de traducció automàtica Apertium es desenvolupa i distribueix com a codi obert.

2.3.5 Reptes de la TA de codi obert

Per a poder gaudir d'aquests avantatges, les comunitats lingüístiques implicades han de fer front a una sèrie de reptes:

1. **Neutralització de les actituds «tecnofòbiques».** Moltes vegades, els experts que podrien ajudar a crear nous sistemes de traducció automàtica desconfien de les tecnologies, potser a causa de la seua visió idealitzada de la llengua i la comunicació humana, i de la seua poca estima pels usos no formals o no literaris.[4] També hi poden intervenir *barreres afectives* que interferisquen amb l'aprenentatge i la subsegüent adopció de les tecnologies de la llengua.
2. **Organització del desenvolupament comunitari.** És comú, i desitjable, que el desenvolupament de programari de codi obert es produïsca de manera comunitària, al voltant del que normalment s'anomena un *projecte*. Per organitzar un projecte, cal, d'una banda, un punt comú d'encontre, un servidor en el qual els desenvolupadors puguen millorar el programari o contribuir dades lingüístiques noves i que permeta els usuaris de la comunitat lingüística implicada descarregar o executar l'última versió del sistema. Però, d'altra banda, calen estructures de coordinació (administradors del projecte, coordinadors de cada parell de llengües, coordinadors del motor de traducció, etc.). Són possibles organitzacions més centralitzades i jerarquitzades o més "horitzontals", depenent del projecte.
3. **Elicitació del coneixement lingüístic**. Aquest és un dels reptes més importants, especialment per a llengües per a les quals la perícia lingüística és escassa o fragmentària. Perquè siga útil per a codificar dades lingüístiques, el coneixement intuïtiu de la llengua per part dels parlants s'ha de fer explícit, és a dir, ha de ser *elicitat*. En la mesura que siga possible, el nivell de coneixements lingüístics necessari per a ser capaç de construir un nou sistema de traducció automàtica nou hauria de ser el mínim possible.
4. **Estandardització i documentació dels formats de dades lingüístiques.** S'ha de definir amb claredat i precisió un format sistemàtic per a cada font de dades lingüístiques utilitzada pel sistema. Una de les millors maneres de definir formats de dades lingüístiques és basar-se en el llenguatge extensible de marcatge XML:[5] els formats resultants són bastant autodescriptius, és possible comprovar automàticament si són vàlids per a l'aplicació abans d'usar-los i es facilita notablement l'intercanvi de les dades amb altres tecnologies i aplicacions lingüístiques.
5. **Modularitat.** Perquè el motor i les dades lingüístiques de traducció automàtica de codi obert siguen útils per a parells de llengües diferents o per a altres aplicacions de tecnologia lingüística, convé que siguen modulars. Per exemple, tenir un analitzador

4 Heus ací una altra explicació possible per algunes d'aquestes actituds tecnofòbiques: molts d'aquests professionals de llengua tendeixen a centrar-se normalment en fenòmens improbables que són propis de la idiosincràsia d'una llengua particular (les «joies» de la llengua), que els sistemes de traducció automàtica tendeixen a tractar incorrectament, en comptes de centrar-se en com aquests sistemes tracten estructures i paraules comunes que constitueixen el 95% dels textos de cada dia (els «maons» de la llengua).

5 http://www.w3c.org/XML/. XML són les sigles d'***ex**tensible **m**arkup **l**anguage*.

morfològic independent i el corresponent diccionari morfològic independent per una certa llengua permet que s'usen en un altre motor de traducció automàtica que té la mateixa *llengua origen* (o *llengua de partida*) i una *llengua meta* (o *llengua d'arribada*) diferent.

3. Apertium

Apertium[6] és una plataforma de traducció automàtica de codi obert, inicialment concebuda per a parells de llengües emparentades (en particular, llengües romàniques), però que ha estat recentment expandida per a poder tractar parells de llengües més divergents (com ara anglès–català). La plataforma proporciona

- un *motor* de traducció independent de les llengües (vegeu la secció 3.3);
- *eines* per a gestionar les dades lingüístiques necessàries per a construir un sistema de traducció automàtica per a un parell de llengües donat o per a adquirir automàticament («aprendre») regles de transferència estructural (Caseli et al. 2006; Sánchez-Martínez et al. 2008) i de desambiguació a partir de textos (Sánchez-Martínez et al. 2008);
- *dades lingüístiques* per a un nombre creixent de parells de llengües (vegeu les seccions 3.4 i 4).

3.1 Rerefons

El disseny inicial està basat en el de sistemes que ja havia desenvolupat pel grup Transducens de la Universitat d'Alacant, com ara interNOSTRUM[7] (espanyol–català), i Traductor Universia[8] (espanyol–portugués). Aquestes tecnologies, inicialment dissenyades per a parells de llengües relacionades, han estat esteses per a tractar parells de llengües que no estiguen tan relacionades.

3.2 La filosofia sobre la qual es fonamenta Apertium

3.2.1 Simplicitat de disseny i modularitat

Per a generar traduccions que siguen raonablement intel·ligibles i fàcils de corregir entre llengües relacionades com l'espanyol (es) i el català (ca) o el portugués (pt), etc., només cal millorar la traducció mot per mot amb: processament lèxic robust (incloent-hi unitats lèxiques multi-mot), desambiguació lèxica categorial (*part-of-speech tagging*) i processament estructural local basat en regles simples i ben formulades per a transformacions estructurals freqüents (reordenació, concordança).

Per a parells de llengües més difícils, no tan relacionats, hauria de ser possible estendre aquest model senzill i generalitzar-ne els conceptes de manera que la complexitat es mantinguera tan baixa com fóra possible, tal com s'ha discutit en 2.3.5.

Apertium té un disseny modular basat en conceptes lingüístics senzills, que es detalla en la secció 3.3.

3.2.2 Separació eficient de motor i dades

D'una banda, hauria de ser possible generar un sistema complet de traducció automàtica a partir de dades lingüístiques (diccionaris monolingües i bilingües, regles gramaticals), especificades de manera *declarativa*. Aquesta informació hauria d'estar en un format interoperable; per exemple, basat en XML (vegeu la secció 2.3.5).

D'altra banda, hauria de ser possible tenir un motor de traducció únic (independent de la llengua) que llegiria dades específiques per a cada parell de llengües («separació d'algorismes i dades»). Les dades lingüístiques del parell de llengües haurien de ser preprocessades de manera que el sistema siga ràpid (més de 10.000 mots per segon) i compacte; per exemple, les transformacions lèxiques es farien amb transductors d'estats finits (TEFs).

Apertium pot ser usat per a construir sistemes de traducció automàtica per a una gran varietat de parells de llengües; per a això, Apertium usa formats senzills basats en XML per a codificar les dades lingüístiques necessàries (fetes a mà o per conversió de dades existents) que es compilen, amb les eines que es proveeixen, en els formats de gran velocitat usats per un motor únic, independent del parell de llengües concret.

Aquests són els quatre tipus bàsics de dades d'Apertium:

- regles (independents de la llengua) per a tractar els diferents formats de text
- especificació del desambiguador lèxic categorial
- diccionaris morfològics i bilingües i diccionaris de regles de transformació ortogràfica

6 http://www.apertium.org
7 http://www.internostrum.com
8 http://traductor.universia.net

- regles de transferència estructural

3.2.3 Desenvolupament i distribució com a codi obert

Aquestes són les raons que van inspirar el desenvolupament d'Apertium en codi obert:

- Donar a tothom accés lliure i il·limitat a les millors tecnologies possibles de traducció automàtica.
- Establir una plataforma modular, documentada i oberta per a la traducció automàtica de transferència superficial i per a altres tasques de processament automàtic de la llengua.
- Afavorir l'intercanvi i la reutilització de les dades lingüístiques existents, tant per a crear nous sistemes de traducció automàtica com per a usar-los en altres tecnologies lingüístiques.
- Facilitar la integració amb altres tecnologies de codi obert.
- Beneficiar-se del desenvolupament col·laboratiu del motor de traducció i de les eines de dades per a parells de llengües existents o nous per part de la indústria, de les universitats o d'organitzacions de suport de llengües menors.
- Promoure el canvi de model de negoci en TA, del model basat en llicències (obsolescent) a un model basat en serveis.
- Garantir radicalment la reproducibilitat de la recerca en TA (vegeu la secció 3.7).
- Perquè no té sentit usar diners públics per a desenvolupar programari no lliure i de codi tancat.

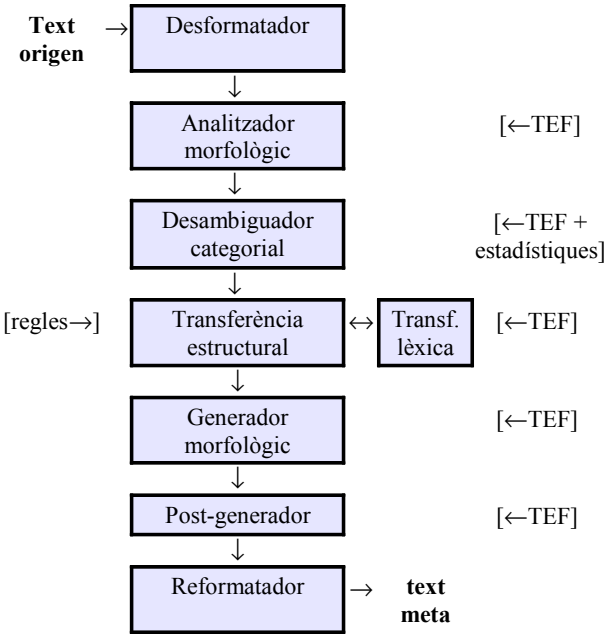
Apertium és, en el moment d'escriure aquest article, un dels pocs sistemes de TA de codi obert (basat en regles[9]) que poden ser utilitzats per a propòsits reals.[10]

3.3 Com funciona Apertium?

Apertium usa un motor de traducció de transferència superficial completament modular que processa el text d'entrada en etapes, com en una cadena de muntatge: desformatatge, anàlisi morfològica, desambiguació categorial, transferència estructural superficial, transferència lèxica, generació morfològica i reformatatge. La comunicació entre els mòduls que s'encarreguen de cada una d'aquestes etapes es fa en forma de text (usant les típiques canonades o *pipelines* d'Unix). Aquest esquema té avantatges clars: simplifica la diagnosi i la depuració d'errors, permet la modificació de dades entre dos mòduls, usant, per exemple, filtres, i facilita la inserció de mòduls alternatius (crucial per a la recerca i el desenvolupament, vegeu la secció 3.7).

Apertium és capaç de traduir textos en els formats de text més comuns (text pla, HTML, RTF, ODF, `.sxw` d'OpenOffice.org, etc.).

La següent figura resumeix el funcionament d'Apertium. Apertium usa transductors d'estats finits (en la figura, TEF) per a les operacions de processament lèxic (anàlisi i generació morfològica, transferència lèxica), models ocults de Markov (basats en estadístiques i tècniques d'estats finits) per a la desambiguació categorial i *chunking* (anàlisi sintàctica superficial) multietapa basat en patrons detectats mitjançant tècniques d'estats finits per a les regles de transferència superficial.



Segueix una breu descripció dels mòduls:

- El **desformatador** separa el text de la informació de format. Actualment hi ha desformatadors disponibles per a text pla, HTML, RTF, ODF, i `.sxw` d'OpenOffice.org. El funcionament està basat en tècniques d'estats finits. La majoria dels desformatadors es generen (usant un full d'estil XSLT) a partir d'un fitxer XML que especifica el seu funcionament per a cada format.

9 El sistema de TA de codi obert basada en corpus més usat és probablement Moses (http://www.statmt.org/moses/).

10 Com s'ha esmentat abans, hi ha també OpenLogos (http://ww.logos-os.dfki.de). Un altre sistema interessant és Matxin (http://matxin.sourceforge.net/), bastant relacionat amb Apertium.

- L'**analitzador morfològic** segmenta el text en llengua origen (LO) en *formes superficials* (FSs), assigna a cada FS una o més *formes lèxiques* (FLs), cada una amb lema, categoria lèxica o part de l'oració, i informació de flexió morfològica. És capaç de processar contraccions i unitats lèxiques multi-mot que poden ser invariables (`es`: *con cargo a*, *de suerte que*) o variables (`es`: *echaría de menos* → *echar de menos*). El mòdul lliga transductors d'estats finits *compilats* a partir d'un diccionari morfològic en XML.
- El **desambiguador lèxic categorial** tria una de les FLs corresponents a cada FS ambigua (al voltant del 30% en llengües romàniques) segons el context. Usa models de Markov ocults (preferències estadístiques) i restriccions escrites a mà. S'entrena usant corpus representatius per a la llengua origen (desambiguats manualment o no) o, més recentment, usant models estadístics de la llengua meta (Sánchez-Martínez et al. 2008, vegeu la secció 3.7). El seu comportament està controlat per un arxiu XML.
- El **mòdul de transferència estructural** reconeix *xuncs* o *chunks* (patrons de FLs de la LO) usant tècniques d'estats finits (d'esquerra a dreta i elegint el patró concordant més llarg), i executa les accions associades a cada patró en el fitxer de regles (de la forma *patró—acció*) per a generar el patró de FLs corresponent en la llengua meta. El fitxer de regles de transferència XML es preprocessa perquè siga interpretat més ràpidament. Per a parells de llengües "més difícils", hi ha disponible una transferència estructural en tres etapes:
  - Es detecten, processen i marquen patrons de FLs (*xuncs*)
  - Es detecten i processen patrons de *xuncs* : aquest processament *inter-xunc* permet transformacions sintàctiques d'abast més llarg
  - Els *xuncs* d'eixida son reprocessats si és necessari i les FLs que contenen s'envien a l'eixida.
- El mòdul de **transferència lèxica** llig cada FL de la LO i genera la FL corresponent en llengua meta (LM); usa transductors d'estats finits *compilats* a partir de diccionaris bilingües en XML, i és invocat quan és necessari pel mòdul de transferència estructural.
- El **generador morfològic** genera, flexionant adequadament cada FL en LM, la FS corresponent. Usa transductors d'estats finits *compilats* a partir de diccionaris morfològics en XML
- El **post-generador** realitza transformacions ortogràfiques com ara contraccions (`ca`: *de +els → dels* ; `en`: *can + not → cannot*), o inserció d'apòstrofs (`ca`: *de + amics → d'amics*), etc.; es basa en transductors d'estats finits *compilats* a partir de diccionaris de regles senzilles de post-generació.
- El **reformatador** reintegra la informació de format en el text traduït. Com el desformatador, es basa en tècniques d'estats finits i es genera a partir d'un fitxer d'especificació per a cada format. S'usa també per a modificar els URLs dels enllaços per a la modalitat *navegar i traduir*.

3.4 Dades lingüístiques (parells de llengües)

El projecte Apertium acull el desenvolupament col·laboratiu de dades per a un gran nombre de parells de llengües, amb un èmfasi especial sobre les llengües romàniques. Vegeu l'epígraf 4 per a més detalls.

3.5 Finançament

Des del 2004, Apertium ha estat finançat per nombroses institucions, sense les quals no hauria estat possible:

- Els ministeris d'Indústria, Turisme i Comerç, d'Educació i Ciència i de Ciència i Tecnologia d'Espanya
- La Secretaria de Telecomunicacions i Societat de la Informació (STSI) de la Generalitat de Catalunya
- El Ministeri d'Assumptes Exteriors de Romania
- La Universitat d'Alacant

Empreses: Prompsit Language Engineering, ABC Enciklopedioj, imaxin|software, Eleka Ingeniaritza Linguistikoa, Eolaistriu, etc.

3.6 La comunitat d'Apertium

Al voltant dels desenvolupadors originals (contractats amb el finançament descrit en la secció anterior), s'ha format una comunitat internacional de desenvolupadors (*instigada* fonamentalment per Francis Tyers). En

l'actualitat, hi ha 85 desenvolupadors inscrits en el projecte[11], molts de fora del grup original; el codi i les dades s'actualitzen molt freqüentment (centenars d'actualitzacions cada mes). Un *wiki* mantingut col·lectivament[12] documenta els components d'Apertium, mostra l'estat actual de desenvolupament i dóna consells per als desenvolupadors de dades lingüístiques o de programes. També s'han desenvolupat eines i codi externament: la interfície gràfica d'ús `apertium-tolk`, i l'eina de diagnòstic `apertium-view`; *plugins* per a OpenOffice.org, per al missatger Pidgin (abans Gaim), o per al sistema de gestió de continguts Wordpress; una versió dels diccionaris bilingües per a mòbils amb Java, i, recentment, per a PDA Palm (`TinyLex`); una aplicació para la traducció de subtítols de pel·lícules (`apertium-subtitles`), versions preliminars per al sistema operatiu Windows, etc. Molts dels desenvolupadors es troben en el canal de *xat* IRC `#apertium` (del servidor `irc.freenode.net`), per a discutir en línia assumptes d'Apertium de manera més o menys formal.

Des de fa dos anys els paquets estables estan disponibles com a part de la distribució Debian de GNU/Linux (i per tant, en la popular distribució Ubuntu Linux).

3.7 Apertium com a plataforma d'investigació

La plataforma de traducció automàtica (TA) de codi obert Apertium ha estat utilitzada com a plataforma d'investigació per a la implementació de nous mètodes que permeten el desenvolupament més ràpid i eficient d'alguns dels recursos necessaris per a la construcció de nous parells de llengües. De fet, recentment s'ha defensat una tesi doctoral en el marc del projecte (Sánchez-Martínez 2008).

Entre les recerques en què ha participat el grup Transducens de la Universitat d'Alacant, cal esmentar, a més de la tesi adés referida, els següents treballs:

- Caseli et al. (2006) proposen un mètode per a la inferència de recursos bilingües a partir de bitextos (textos en un idioma, juntament amb la seua traducció a un altre idioma). Els recursos obtinguts comprenen tant diccionaris bilingües com regles de transferència estructural superficial similars a les utilitzades en Apertium per a la TA entre llengües romàniques. El programari usat en aquest treball és també de codi obert[13] i s'ha usat per a iniciar el desenvolupament d'alguns diccionaris bilingües en Apertium.
- Sánchez-Martínez i Forcada (2009) fan ús de tècniques de TA estadística per a la inferència de regles de transferència estructural superficial a partir de bitextos; en aquest cas, no s'infereix cap diccionari bilingüe, sinó que se n'usa un d'existent. El mètode descrit per Sánchez-Martínez and Forcada (2009) ha estat implementat i alliberat com a codi obert dins d'Apertium de tal forma que s'integra fàcilment en el procés de desenvolupament de nous parells de llengües per a Apertium, ja que genera regles en el format XML utilitzat pel mòdul de transferència estructural.
- Sanchez-Martínez et al. (2008) han desenvolupat un nou mètode que permet l'entrenament dels desambiguadors lèxics categorials (*part-of-speech taggers*) basats en models ocults de Markov usats en Apertium de forma completament no supervisada mitjançant l'ús de textos tant en llengua origen com en llengua meta. Aquest mètode, que proporciona resultats clarament millors que els obtinguts pels mètodes d'entrenament no supervisats clàssics, ha estat alliberat com codi obert i s'integra plenament en el procés de desenvolupament de nous parells de llengües per a Apertium.

També hi ha recerques realitzades per investigadors externs:

- Homola i Kuboň (2008)descriuen un experiment realitzat amb Apertium sobre el parell portugués—espanyol, suggereixen una modificació de l'arquitectura del sistema que asseguren que millora la qualitat de traducció i discuteixen les implicacions de la millora de l'arquitectura per al disseny de recursos lingüístics per als sistemes de transferència sintàctica superficial com Apertium.
- Tyers i Donnelly (2009), com s'ha esmentat més amunt, descriuen un sistema obert de TA gal·lés-anglés basat en Apertium, pensat per a l'assimilació d'informació, n'avaluen els

11 En http://sourceforge.net/projects/apertium/

12 http://wiki.apertium.org

13 El programari forma part del projecte ReTraTos, i té l'adreça http://retratos.sourceforge.net/ .

resultats i discuteixen els avantatges del desenvolupament comunitari de sistemes basats en regles per a les llengües marginalitzades.

El fet que aquestes investigacions s'hagen fet sobre una plataforma oberta i disponible, facilita enormement la seua reproducibilitat a d'altres investigadors.

4. Apertium i les llengües romàniques

4.1 El grup de llengües millor representat

Entre els parells *estables*[14] disponibles a hores d'ara en la plataforma Apertium hi ha: **espanyol** ↔ **català**, **espanyol** ↔ **gallec**, **espanyol** ↔ **portugués**, **portugués** ↔ **català**, **portugués** ↔ **gallec**, anglés ↔ **català**, **francés** ↔ **català**, anglés ↔ **espanyol**, anglés ↔ **gallec francés** ↔ **espanyol**, **occità** ↔ **català**, **occità** ↔ **espanyol**, **romanés** → **espanyol**, **espanyol** → esperanto, **català** → esperanto, anglés → esperanto, basc → **espanyol** i gal·lés → anglés.[15] A més, hi ha un nombre creixent de parells de llengües en desenvolupament. Com es pot veure, la majoria dels parells estables inclouen una llengua romànica (en negretes). Això és perquè, de fet, la breu història d'Apertium (cinc anys) està molt lligada a les llengües romàniques, i la naturalesa col·laborativa del projecte ha atret desenvolupadors de procedències molt diverses, com veurem a la secció 4.2.

La taula "Parells de llengües d'Apertium..." dóna notícia de la data de l'última versió estable dels parells de llengües que inclouen una o dues llengües romàniques (a 15 de febrer de 2009). S'ha de tenir en compte que molts dels parells continuen en desenvolupament actiu encara que no se n'haja publicat cap versió estable recentment.

Parells de llengües d'Apertium que inclouen una llengua romànica

Parell de llengües	Última v. estable	Data de l'última versió estable
anglés↔**espanyol**	0.6	19 març 2008
anglés↔**català**	0.8.4	19 març 2008
anglés↔**gallec**	0.5.1	19 novembre 2008
basc→**espanyol**	0.3.0	11 novembre 2008
català→esperanto	0.9.0	20 febrer 2008
espanyol↔català	1.0	28 març 2006
espanyol↔gallec	1.0	7 octubre 2007
espanyol↔portugués	1.0.3	3 octubre 2007
espanyol→esperanto	0.9.0	20 febrer 2008
francés↔català	1.0	5 octubre 2007
francés↔espanyol	0.8.0	14 febrer 2008
occità↔català	1.0.5	12 juliol 2008
occità↔espanyol	1.0.5	12 juliol 2008
portugués↔català	0.8.0	18 juny 2008
portugués↔gallec	0.9.0	10 juny 2008
romanés→espanyol	0.7	8 octubre 2007

4.2 Breu història

Apertium naix, tal com s'esmenta a la secció 3.1, com una reescriptura en codi obert de les tecnologies de traducció existents en el grup Transducens de la Universitat d'Alacant. Aquestes tecnologies s'aplicaven aleshores a la traducció entre llengües romàniques: espanyol ↔ català i espanyol ↔ portugués. Aquesta reescriptura es va realitzar en el marc d'un projecte finançat pel Ministeri d'Indústria, Turisme i Comerç espanyol, en col·laboració amb universitats i empreses de tot Espanya. El resultat va ser un nou motor de traducció, completament redissenyat, i les dades per als parells espanyol ↔ català i espanyol ↔ gallec.

Més avant, amb suport de la Secretaria de Telecomunicacions i Societat de la Informació (STSI) de la Generalitat de Catalunya, es van llançar els parells francés ↔ català i català ↔ occità (inicialment, aranés), conjuntament amb l'anglés ↔ català. El cas de l'occità es descriu amb més detall en la secció següent.

Quasi paral·lelament, amb suport del Ministeri d'Assumptes Exteriors de Romania, i en un projecte dirigit per la Prof. Catalina Iliescu de la Universitat d'Alacant, es va començar a treballar en el parell romanés ↔ espanyol. Els problemes plantejats pel joc de caràcters del romanés van

14 L'ús de la denominació *estable* no fa referencia a la qualitat del traductor corresponent, sinó al fet que Apertium ha publicat paquets informàtics per a aquestes llengües, a punt per a poder-los instal·lar fàcilment.

15 Vegeu Tyers i Donnelly (2009)

motivar l'adaptació d'Apertium a Unicode (joc de caràcters universal, vàlid per a totes les llengües); això ha permés l'inici del desenvolupament de parells de llengües amb sistemes d'escriptura diferents (com el macedoni).

El parell espanyol ↔ portugués és també de la mateixa època. Aquest és, sens dubte, un dels parells de llengües romàniques més gran (darrere, potser, del parell espanyol ↔ francés). El grup Transducens va decidir muntar un paquet de dades (Armentano-Oller et al. 2006) a partir del coneixement que li havia permés desenvolupar el traductor Universia (http://traductor.universia.net/), ara comercial.

El 2006 es crea l'empresa Prompsit Language Engineering, amb programadors i lingüistes d'Apertium. Un dels primers parells que s'hi inicien, per encàrrec de l'empresa Eleka Ingeniaritza Linguistikoa, és l'espanyol ↔ francés, el qual continua en desenvolupament.

El 2007, la Universitat Pompeu Fabra i l'empresa ABC Enciklopedioj desenvolupen els sistemes espanyol → esperanto i català → esperanto. D'altra banda, Armentano-Oller i Forcada (2008) publiquen el primer prototip portugués ↔ català, construït a partir dels parells espanyol ↔ portugués i espanyol ↔ català.

El 2008, l'empresa imaxin|software publica el traductor portugués ↔ gallec, muntat a partir de les dades espanyol ↔ portugués i espanyol ↔ gallec.

També a finals de 2008, usant dades procedents del projecte Matxin,[16] la Universitat d'Alacant llança el primer prototip traductor basc → espanyol.

Actualment hi ha dos parells més de llengües en desenvolupament actiu en el projecte: espanyol—italià, finançat i desenvolupat per la Universitat d'Alacant, i bretó—francés cofinançat i desenvolupat per la Universitat d'Alacant i L'Ofis ar Brezhoneg (Oficina del Bretó).

4.3 Un exemple: Apertium i l'occità

El desenvolupament de TA per a l'occità per part de la Universitat d'Alacant i la Universitat Pompeu Fabra va començar en 2006 amb el parell aranés—català, finançat per la STSI de la Generalitat de Catalunya. Aquest parell connectava una llengua *mitjana* (el català, amb uns 6.000.000 parlants) i una variant estandarditzada *molt menuda* (l'aranés, amb uns 6.000 parlants) d'una llengua més gran, l'occità, amb potser 1.000.000 parlants. El desenvolupament (Armentano-Oller i Forcada 2006) es va iniciar partint de dades existents (espanyol—català), un exemple clar de reutilització de dades obertes.

Més avant, el 2007 les empreses alacantines Prompsit i Taller Digital guanyen un concurs públic i són contractades per la Generalitat de Catalunya per a construir els traductors oficials occità ↔ català i occità ↔ espanyol, tant per a l'aranés com per a l'occità general (*occitan larg*).

Un dels principals problemes d'aquest treball rau en l'estandardització de l'occità general, que avança molt lentament. Això convertia la iniciativa en autènticament pionera. Per a definir quin seria el model de llengua que produirà el sistema, es va crear una comissió d'experts lingüístics de *quasi* tot Occitània (2 experts per *regió*) amb participació d'una experta d'Apertium (Gema Ramírez). El model de llengua elegit (no sense llargues discussions) està basat en el dialecte llenguadocià.

En l'actualitat, amb un sistema bidireccional, completament operatiu, que es pot descarregar o usar en línia, i que té el 95% de cobertura i una taxa d'error del 10% per a la traducció aranés—català i del 25% d'error per a la traducció *occitan larg*—català (clarament millorable), es poden començar a produir els efectes següents:

- La quantitat de text en occità en la web, generat mitjançant traducció automàtica seguida de postedició, pot augmentar la visibilitat de la llengua.
- L'existència de traducció automàtica de qualitat pot promoure la difusió de les variants de l'occità elegides.
- La comunitat occitana general (la majoria a França) pot crear un traductor occità—francés a partir de les dades occità—català o occità—espanyol i francés—català o francès—espanyol ja existents en Apertium.
- Les dades públiques i obertes disponibles per a l'occità poden ser útils per a crear altres aplicacions de tecnologia lingüística per a aquesta llengua.

Els sistemes de traducció occità ↔ català i occità ↔ espanyol resultants, són, des del 5 de

16 http://matxin.sourceforge.net

novembre de 2008, els oficials de la Generalitat de Catalunya.[17]

5. Comentaris finals

El llançament, fa quatre anys, de la plataforma de traducció automàtica de codi obert Apertium (www.apertium.org) ha facilitat el desenvolupament col·laboratiu de sistemes de traducció automàtica oberts (i de tecnologia lingüística oberta, a punt per a ser transferida a d'altres aplicacions) per a moltes llengües, però molt especialment per a les llengües romàniques, per a les que va ser inicialment concebut. Això ha estat possible principalment gràcies al suport d'institucions públiques, però també d'empreses interessades a oferir serveis de traducció automàtica en el model de negoci emergent que possibilita el programari obert.

Crec que Apertium pot contribuir a una comunicació més fluida entre les comunitats de la Romània: d'una banda, ajudant en la producció de traduccions que es poden fer públiques amb poc esforç de correcció, i, d'altra, ajudant els internautes a llegir documents escrits en altres llengües romàniques per mitjà de traduccions aproximades instantànies.

En el cas particular de la llengua occitana, encara queda per avaluar quin serà l'impacte d'Apertium en l'estandardització pendent d'aquesta llengua.

Agraïments: Com ja he dit més amunt, Apertium ha estat finançat, des de 2004, pels governs espanyol, català i romanés, per la Universitat d'Alacant, i per nombroses empreses. Apertium (i aquest article) no serien possibles sense l'ajuda de molts investigadors i desenvolupadors, com Carme Armentano-Oller, Enrique Benimeli, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mireia Ginestí-Rosell, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Míriam A. Scalco, Francis M. Tyers, i molts altres.

Referències

Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A. (2005) "Open-source Portuguese-Spanish machine translation", in *Lecture Notes in Computer Science* **3960** (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006) 13-17 de maig de 2006, Itatiaia, Rio de Janeiro, Brasil., p. 50-59.

Armentano-Oller, C., Forcada, M.L. (2006) "Open-source machine translation between small languages: Catalan and Aranese Occitan", in Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages) (organitzat conjuntament amb l'LREC 2006 (22-28.05.2006)), p. 51-54.

Armentano-Oller, C., Forcada, M.L. (2008) "Reutilización de datos lingüísticos para la creación de un sistema de traducción automática para un nuevo par de lenguas", *Procesamiento del Lenguaje Natural* **41**, 243-250.

Caseli, H. M., M. G. V. Nunes, M. L. Forcada (2006). "Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation". *Machine Translation* **20**(4)227-245. Publicat el 2008.

González, J., Lagarda, A.L., Navarro, J.R., Eliodoro, L., Giménez A., Casacuberta, F., de Val, J.M., Fabregat, F. (2006) "SisHiTra: A Spanish-to-Catalan hybrid machine translation system". In *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages",* Gènova, Itàlia, 23 maig 2006; pp.69-73

Homola, P., Kuboň, V. (2008). "Improving Machine Translation Between Closely Related Romance Languages". In *Proceedings of the European Association of Machine Translation*, p. 72—77.

Sánchez-Martínez F. (2008). "Using unsupervised corpus-based methods to build rule-based machine translation systems". Tesi Doctoral, Departament de Llenguatges i Sistemes Infomàtics, Universitat d'Alacant.

Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. (2008). "Using target-language information to train part-of-speech taggers for machine translation". *Machine Translation*, **22**(1-2) 29-66.

Sánchez-Martínez, F., Forcada, M.L. (2009). "Inferring shallow-transfer machine translation rules from small parallel corpora". *Journal of Artificial Intelligence Research* (accepted).

Tyers, F. M. and Donnelly, K. (2009) "apertium-cy - a collaboratively-developed free RBMT system for Welsh to English". *The Prague Bulletin of Mathematical Linguistics* **91**: 57-66.

17 http://traductor.gencat.cat/

Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva

Diana Santos
Linguateca, SINTEF ICT
Diana.Santos@sintef.no

Resumo

Este artigo faz um balanço pessoal do percurso da Linguateca, uma organização virtual em demanda de uma maior facilidade e qualidade no processamento da língua portuguesa, nos últimos dez anos.

Inicio o artigo por uma curta perspectiva histórica para explicar o contexto em que a Linguateca surgiu e quais os objectivos iniciais para o progresso da área. Avalio de seguida resumidamente a situação actual no que respeita a esses objectivos iniciais, bastante vagos, identificando o que foi cumprido e perspectivando o que ficou por fazer.

Aproveito também a oportunidade para apresentar as variadas inflexões que o projecto tomou, num percurso que não foi linear.

Faço depois uma breve excursão pelos principais pontos atingidos, mas sem a preocupação de ser exaustiva, dado que o texto não pretende ser um relatório, mas sim uma reflexão crítica sobre o processo e os resultados, tentando relacioná-la, sempre que possível, com a discussão pública que teve lugar dez anos volvidos no Encontro *Linguateca: 10 anos*, em Aveiro a 11 de Setembro de 2008.

Embora o artigo seja centrado sobre a Linguateca, tento fazer numa última secção algumas pontes com outro trabalho em processamento do português, de forma a não transmitir a ideia errada de que teríamos sido os únicos a trabalhar na área ou a progredir neste período.

Termino o artigo com uma breve secção com algumas sugestões para projectos que possam continuar o espírito da Linguateca ou reforçar as contribuições da Linguateca para o objectivo mais geral da dignificação e da melhoria do processamento da língua portuguesa.

O processo de tornar o processamento do português mais percorrido e mais agradável assemelha-se ao desbravamento de vários caminhos num emaranhado de questões e problemas semelhante a uma selva ou país – daí o título deste texto referir o "mapa da portuguesificação". Ao invés de considerar o trabalho concluído, ponho a tónica no muito ainda que é preciso fazer nesta área, em que a acção da Linguateca é (ou foi) comparável, apenas, à criação de alguns caminhos. Também por isso indico neste texto aquelas sendas que acabaram em becos sem saída, mas que aumentaram a nossa experiência ou nos convenceram de que não devíamos seguir por ali.

1 Apresentação

A Linguateca foi um projecto político-científico financiado pelas autoridades na área da ciência e da tecnologia em Portugal para tratar do processamento computacional da língua portuguesa, área que tinha sido considerada prioritária.

Em vez de um projecto científico para fazer investigação, era um projecto de infraestrutura e de **serviço** à comunidade.

Após dez anos de diversas formas de financiamento e de bastante trabalho realizado, encontramo-nos numa situação de transição e de reflexão que tanto pode ser o início de uma nova fase da Linguateca como corresponder à sua conclusão.

Urge assim fazer um balanço de todo o processo e das várias fases e intenções que tivemos ao longo do tempo. Faço-o em meu nome pessoal porque fui a única que assisti e liderei este projecto desde o início, mas com o apoio de muitos e tomando em consideração todo o retorno recebido ao longo dos anos, quer dos muitos colaboradores quer da comunidade em geral, além de colher os frutos do encontro de reflexão pública em Aveiro em Setembro de 2008.

Outros textos ou apresentações sobre diferentes fases da Linguateca e sobre eventuais diferentes tónicas postas ao longo do tempo nas várias actividades podem ser consultados no catálogo de publicações da Linguateca. Saliento aqui como especialmente representativos de fases diferentes os seguintes textos (Santos, 2000; Santos, 2002b; Santos e Costa, 2005; Santos, 2007a), que serão brevemente resumidos na secção 3.2. Os vários relatórios anuais ou "finais" da Linguateca permitem dar outro tipo de visão complementar, mais concreta, cf. Santos (2003a), Santos (2005), Santos

(2006b) e Costa (2008).

2 A concepção: missão, estrutura, e ponto de partida

A Linguateca surgiu como uma forma de contrabalançar, ou resolver, muitos dos problemas ou limitações identificados durante o período da escrita do contributo para o livro branco (Santos, 1999b), há mais de dez anos, e que serão aqui repetidos esquematicamente.

Esse texto inicial, relativo à área como um todo, e de conteúdo essencialmente programático, foi uma das tarefas do projecto *Processamento Computacional do Português*[1]. Os pontos nele focados não eram para ser resolvidos na sua totalidade, ou mesmo abordados, em dez anos e por um projecto em rede. Contudo, estou convencida de que foi a nossa tentativa de não perder mais tempo e de começar logo a fazer o que era possível, ainda no âmbito do próprio projecto Processamento Computacional do Português, que levou à aprovação da Linguateca[2] nos anos que se seguiram.

É claro que os objectivos da Linguateca como projecto tiveram de ser mais concretos e realistas, embora desenhados e motivados pelos problemas que queríamos resolver e pelas metas que queríamos atingir, directa ou indirectamente. De qualquer maneira, faz todo o sentido utilizar os pontos mencionados em Santos (1999b) como uma bitola para comparar a actividade e os resultados obtidos, desde que nunca se esqueça que esse texto era dedicado à comunidade e não apenas aos membros de um projecto futuro que se viria a constituir.[3]

Vejamos então o que esse texto dizia. Antes disso, contudo, importa recordar e insistir no seguinte ponto: a área discutida e equacionada correspondia ao processamento da nossa língua e não à engenharia da linguagem em geral, veja-se Santos (1999a), o que veio a ser um dos principais cavalos de batalha da Linguateca.

Santos (1999b) mencionava as seguintes condições necessárias a um progresso significativo na área do processamento da língua portuguesa (note-se que, por conveniência da exposição, a ordem foi invertida em relação à original):

1. Transparência, participação e colaboração de todos
2. Desenvolvimento de aplicações relacionadas com o trabalho de todos os dias no sector da informação
3. Ligação da investigação fundamental com as tecnologias
4. Dinamização dos métodos empíricos
5. Serviços de desenvolvimento de recursos e ferramentas partilháveis (serviço de tradução, serviço de terminologia, rede de fala, rede de processamento da língua escrita)
6. Avaliação e controlo de qualidade em relação ao português
7. Disponibilização de recursos (nas suas múltiplas vertentes)
8. Definição do processamento do português como área prioritária

Passamos então a indagar se a Linguateca contribuiu algo para cada um destes pontos, tendo em consideração, repito, que a Linguateca foi desde o início definida como um projecto de **serviço à comunidade**, com a preocupação de não competir mas sim favorecer os actores existentes e futuros.

Mas, para o leitor incauto, convém primeiro indicar muito brevemente os pressupostos e estrutura inicial da Linguateca, ou seja, a sua espinha dorsal, antes de discutir a sua actuação e resultados.

A Linguateca, como um projecto de serviço e de apoio, foi idealizada, não através da contratação de investigadores, mas sim de "contratados" com tarefas específicas de manutenção, informação e apoio aos utilizadores, para fazer o que pomposamente se pode chamar "transferência de tecnologia" dos grupos (universitários, académicos) para o mundo exterior. Daí surgiu o conceito de **pólos** (da Linguateca), localizados em grupos ou ambientes a que faria sentido ajudar a disponibilizar o trabalho e reforçar a actividade.

Desde o início, a missão da Linguateca anunciou-se[4] como:

- facilitar o acesso aos recursos já existentes, através do desenvolvimento de serviços de acesso na rede, e mantendo um portal com informação útil,

[1] Financiado pela Agência de Inovação – organismo de financiamento português –, iniciado a 15 de Maio de 1998 no SINTEF, com a duração de dois anos.

[2] O nome *Linguateca* apenas surgiu em 2002. Do ponto de vista formal, o projecto aprovado em 2000 tinha o nome *Centro de Recursos – distribuído – para o processamento computacional da Língua Portuguesa, CRdLP*.

[3] Convém além disso esclarecer que, durante a escrita desse texto, não havia a mais remota previsão de que isso viria a acontecer, pelo menos da minha parte.

[4] De facto, esta formulação, patente na página inicial, foi pela primeira vez publicada, com algumas diferenças irrelevantes, a 9 de Agosto de 2000, como é possível verificar através do projecto Internet Archive (http://web.archive.org), ainda com o URL de www.portugues.mct.pt. A versão exacta, *ipsis verbis*, apareceu a 18 de Novembro de 2004.

- desenvolver, de forma harmoniosa, em colaboração com os interessados, os recursos considerados mais prementes,
- organizar avaliações conjuntas que envolvam a comunidade como um todo.

Assim, e ao contrário de um projecto de investigação, a nossa actividade – ou pelo menos o fundamento do nosso financiamento – repartiu-se (ou repartir-se-ia, conforme o plano) fundamentalmente entre:

- a formação de pessoal especializado em gestão, criação, disseminação e avaliação de recursos;
- o assegurar dos serviços básicos de repositório, distribuição e catálogo, de forma distribuída;
- o desenvolvimento de recursos públicos, em especial, recursos para avaliação ou calibragem;
- a manutenção do contacto e da comunicação entre os vários actores e clientes dos nossos serviços;
- a organização de avaliações conjuntas em torno de áreas chave.

Como será debatido na secção 3, de facto a Linguateca acabou por fazer muitas outras actividades não previstas inicialmente no seu desenho.

Passo então a considerar cada um dos pontos do documento original:

2.1 Transparência

A transparência foi, decididamente, uma das normas da Linguateca, embora uma questão fundamental, a da escolha dos pólos, tenha acontecido de uma forma quase aleatória, à medida que as pessoas se aproximavam de nós e se prontificavam a colaborar.

Uma das restrições (ou sugestões) que tinham sido impostas (ou recomendadas) no início era a da distribuição geográfica dos pólos, de forma a combater ou evitar a demasiada concentração de esforços num único local.

Também, do ponto de vista formal, houve ou havia restrições (inultrapassadas) no estabelecimento de pólos no estrangeiro ou em instituições privadas – o que nunca, contudo, impediu a cooperação e a formação de pólos informais, como foi o do VISL em Odense e o do COMPARA em Lisboa, ambos desde 2000.

Outra questão importante – que me parece agora explicar porque muitos grupos ou instituições não tentaram sequer obter um pólo da Linguateca – tinha a ver com a nossa filosofia de disponibilização pública dos recursos. Com efeito, fomos igualmente claros em afirmá-la, na página inicial da Linguateca, através das seguintes linhas mestras:

- Total abertura: Todas as actividades e trabalhos desenvolvidos pela Linguateca são públicos.
- Disponibilização livre: Os autores de recursos serão remunerados ou compensados de forma a não serem lesados, mas a Linguateca não se destina a desenvolver ou apoiar o desenvolvimento de recursos proprietários, mas sim a criar condições para a existência de recursos bons e gratuitos para a língua portuguesa.

Infelizmente, grande parte dos grupos na área não partilhavam ou partilham desta atitude.

Não obstante todas estas considerações, é inegável que o processo de constituição dos pólos dependeu em muitos casos da sorte, de os contactos terem sido feitos na altura certa, de as pessoas terem falado e de se terem entendido. Por isso, se a Linguateca for reaberta ou continuar, parece-nos mais correcto que todos os pólos sejam criados por concurso (aberto).

Não consideramos contudo que a primeira fase da Linguateca, por ter sido criada à medida das oportunidades que se ofereciam e dando total liberdade aos pólos – desde que com a filosofia de criarem recursos e avaliação para a comunidade – tenha sido errada ou demonstrado falta de transparência. Como é muitas vezes apontado, excesso de planeamento é geralmente sinónimo de falta de inovação (Chubin e Hackett, 1990), e ao podermos inovar, com base no material humano e tecnológico oferecido por cada pólo, fizemos muito mais do que seguir um plano rígido.

2.2 Trabalho de todos os dias

Esta é uma questão possivelmente genérica demais para ter uma concretização fácil, mas, se considerarmos que os trabalhadores nos sectores dos serviços (em que incluímos, aliás, os investigadores e desenvolvedores na nossa área) todos os dias escrevem, publicam, mandam mensagens de correio electrónico, procuram na rede e publicam na dita, além de mandarem mensagens pelo telemóvel e participarem em blogues e outras novas tecnologias, temos naturalmente de reconhecer que a actividade da Linguateca, embora com esse objectivo último, está longe de ter conseguido algum impacto, se excluirmos o círculo reduzidíssimo daqueles que pertencem ou comunicam com a Linguateca no âmbito do seu trabalho.

Assim, embora tenhamos, na medida das nossas possibilidades, apostado na promoção concreta do português através de

- sugestão de normas de redacção em português

- formas de referir publicações em língua portuguesa
- sugestões de terminologia e de desenho de sítios
- variadas intervenções em fóruns internacionais e nacionais sobre as diferenças e o respeito pela língua portuguesa
- localização e tradução para português sempre que necessário ou apropriado

não podemos considerar, de forma alguma, que esta missão – a de termos influenciado o trabalho de todos os dias das pessoas que usam o português – esteja próxima de ser cumprida.

Muito pelo contrário, cada vez mais somos instados por todos a render-nos à evidência de que o que é "internacional", isto é, escrito em inglês, é bom, e o que é nacional, isto é, escrito em português, é medíocre...

Assim, embora uma das palavras de ordem da Linguateca tenha sido a **portuguesificação**[5], demasiado ainda se encontra por fazer.

De facto, penso mesmo que estamos pior do que estávamos na altura do começo da Linguateca. Uma das convicções cada vez mais enraízadas nas camadas mais jovens – devida à forma como as agências de financiamento definem a qualidade – é que os melhores escrevem em inglês e os piores em português, o que leva naturalmente a que isso infelizmente aconteça.[6]

Alguns exemplos que demonstram claramente essa infeliz tendência são:

- o PROPOR – a conferência internacional sobre o processamento do português, com uma comissão de programa maioritariamente de lusofalantes, que desde 2003 é em inglês[7]
- a forma de avaliar os investigadores em Portugal e no Brasil: através de publicações "internacionais", mas esquecendo que o português – uma língua falada como língua materna, ou pelo menos oficial, nos cinco continentes – é uma língua internacional por excelência!
- a língua das teses e das defesas das mesmas em Portugal, que cada vez mais é o inglês em vez do português
- a língua nos sítios na rede dedicados ao processamento da língua, no Brasil e em Portugal, que cada vez mais é o inglês em detrimento do português

Veja-se, a este propósito, o valioso contributo de Gomes de Matos (1992) argumentando a favor do direito de ler e escrever na própria língua em ciência.

Por isso, parece-me evidente que a Linguateca tentou lutar contra a corrente mas que cada vez menos o português é a língua usada (ou apreciada) no local de trabalho de todos os dias.

2.3 Ligação da investigação fundamental com as tecnologias

Esta é uma atitude, mais do que uma medida: Achamos que nesta área não faz sentido uma separação, mas sim uma inter-relação entre desenvolvimento de sistemas e investigação com os mesmos.

Tentámos seguir sempre essa directiva, aliás pondo grande ênfase na questão da avaliação em tarefas práticas.

Contudo, pode ser que a linguística teórica e a informática teórica nos tenham ignorado sobranceiramente, como projecto aplicado e ateórico, e nesse aspecto a nossa intervenção tenha sido nula.

Em suma, é bastante possível que tenhamos nós mais teorizado sobre a nossa prática do que os teóricos tenham praticado graças à nossa actividade.

Não me parece, em resumo, que a Linguateca tenha de alguma forma intervindo neste aspecto, para além da sua própria prática. Que valha pelo menos o exemplo: insistimos sempre no estudo detalhado dos fenómenos da língua que poderiam estar subjacentes a um dado resultado, ou desempenho, em vez de nos ficarmos por simples medidas quantitativas deste.

2.4 Dinamização dos métodos empíricos

Neste ponto, pelo contrário, penso poder afirmar que a Linguateca contribuiu indiscutivelmente para esta dinamização, quer através da sua actividade quer através da criação de recursos que tornassem os métodos empíricos possíveis na prática.

Neste momento, na área do processamento do português, há muito mais avaliação (através de métodos empíricos) e muito maior consciência desta.

Contudo, muitas das medidas que preconizei estão longe (se calhar ainda mais longe) de serem uma realidade, senão veja-se:

> Obrigar a que todos os projectos financiados publicamente tenham uma parte de

[5] E não o aportuguesamento, ou seja, ir buscar coisas (ideias, técnicas, ferramentas) lá fora e adaptá-las ao português.

[6] Note-se que eu não estou a advogar publicação exclusiva em português, mas sim um balanço entre divulgação internacional e divulgação, didáctica e documentação na nossa língua.

[7] Na altura, a justificação avançada para esta mudança foi a de que a editora Springer concedia qualidade às publicações, e exigia o inglês como língua internacional.

avaliação (ou seja, esteja descrito na proposta como avaliar, e quando), de preferência controlável independentemente (ou seja, que a avaliação possa ser repetida por observadores externos).

Certamente que, se houve algo que não correu bem, foi a forma como o financiamento dos projectos nesta área foi atribuído em Portugal durante a existência da Linguateca – e que, acentue-se, foi sempre realizado de forma totalmente independente desta.[8]

De uma forma superficial, dir-se-ia que este foi concebido como precisamente uma compensação aos actores da área com filosofias e práticas mais distantes da Linguateca, ou seja, quanto mais "afastados" da Linguateca, mais financiamento receberiam.

Parece um critério politicamente defensável, mas os resultados práticos não o são necessariamente. Sobretudo se envolvem a repetição de esforços ou o financiamento duplo de algo já existente, como é convicção minha que aconteceu não poucas vezes.

2.5 Serviços de desenvolvimento de recursos e ferramentas partilháveis

Embora uma das áreas em que a Linguateca mais tenha investido tenha sido o desenvolvimento de serviços na rede (veja-se a secção 4.3 abaixo), tal não tomou o caminho descrito no documento preparatório. Convém talvez reflectir sobre as causas ou explicações dessa diferença aqui.

Com efeito, tínhamos preconizado a necessidade ou o interesse de desenvolver as seguintes redes de recursos:

- serviço de tradução
- serviço de terminologia
- rede de fala
- rede de processamento da língua escrita

A posteriori, parece-nos que a Linguateca se tornou **a** rede de processamento da língua escrita, e que, quanto aos outros serviços, ou foram implementados de forma completamente separada ou nunca chegaram a ser uma realidade.

Convém aqui indicar que, embora a intenção inicial da Linguateca fosse cobrir e apoiar tanto o processamento da língua escrita como da falada, tal nunca se realizou, e, após uma tentativa falhada de, logo em 2000, criar um pólo associado à fala – que nunca se materializou porque não houve candidatos a essa posição – acabámos por dirigir a nossa atenção apenas para a parte escrita.

2.5.1 Tradução automática

No início da dinamização da avaliação chegámos a criar uma lista associada à tradução automática, e vários pólos da Linguateca fizeram algum trabalho na área, mas de forma de tal maneira distinta que aparentemente não chegou nunca sequer a haver colaboração:

- O pólo do Porto dedicou-se ao estudo de ferramentas já existentes e ao trabalho necessário de pós-edição, numa perspectiva essencialmente linguística ou mesmo de estudos de tradução (Sarmento et al., 2007; Maia e Barreiro, 2007).
- O pólo de Braga dedicou-se a vários problemas tecnológicos associados ao paradigma da tradução automática por exemplos, desenvolvendo ferramentas para algumas dessas tarefas (Simões e Almeida, 2007) ou estudando a tecnologia de memórias de tradução (Almeida e Simões, 2007).
- Também se pode mencionar que implicitamente a criação do COMPARA (Frankenberg-Garcia e Santos, 2002) foi decisiva para estudos de tradução envolvendo o par de línguas português e inglês,
- assim como o pólo de Lisboa no LabEL (Barreiro e Ranchhod, 2005) produziu também algum trabalho na área.

Pese embora tanta actividade, não se chegou, pelo menos até agora, a atingir um estádio em que houvesse sistemas de tradução automática envolvendo o português desenvolvidos no âmbito da Linguateca (ou com o seu apoio) e que pudessem ser usados, embora haja algumas propostas nesse sentido, e um sistema incipiente de paráfrase (que poderá ser estendido a uma versão bilingue) foi posto ao serviço da comunidade (Barreiro, 2008).

2.5.2 Terminologia

Pior ainda, pelo menos aparentemente, foi o que aconteceu com a terminologia, visto que, embora a Linguateca tivesse desenvolvido um sistema de raiz para trabalho sério na área, o Corpógrafo (Sarmento, Maia e Santos, 2004; Maia, Sarmento e Santos, 2005; Maia, 2008b), aliás com mais de 1600 utilizadores espalhados por todo o mundo, não foi aparentemente possível congregar outras pessoas relacionadas com a área de terminologia, em Portugal ou no Brasil, de forma a trabalhar em rede.

Uma possível explicação para esse facto poderá ser a de já existirem a nível internacional várias

[8] Poderia imaginar-se que um projecto concebido para a disponibilização e avaliação de recursos poderia ser envolvido ou ser-lhe pedido um parecer quanto a novos projectos na área, com vista a garantir uma sua sustentação posterior. Cabe por isso documentar que tal nunca sucedeu.

redes de terminologia envolvendo o português[9], e como tal, em vez de criar mais uma, seria útil sim produzir sistemas que ajudassem a esse trabalho. Parece-me assim que será fundamental tentar entronizar o Corpógrafo como uma ferramenta a considerar nesses ambientes internacionais, em vez de repetir trabalho e aparecer como concorrente em vez de serviço.

Uma das questões que terá nesse caso de ser equacionada é a questão da terminologia bilingue, que, embora tenha estado na agenda do Corpógrafo desde o primeiro momento (veja-se por exemplo Maia (2003) ou Maia e Matos (2008)), ainda não tem suficiente tratamento nesse ambiente. Aliás, seria de todo o interesse aproximar (em vez de afastar) os terminólogos brasileiros, com uma longa tradição de excelência na área, note-se, e tentar na medida do possível fazer terminologia científica comum nas áreas em que isso faça sentido – a linguística e o processamento computacional da língua são, na minha opinião, uma delas.

Saliente-se, contudo, que houve algum trabalho de extracção de terminologia bilingue no âmbito da Linguateca através da tese de doutoramento de Alberto Simões (Simões, 2008).

O fosso entre abordagens linguísticas e informáticas, ao contrário do que seria a minha intenção, também ocorre(u) dentro da própria Linguateca, nunca tendo havido sinergia entre os pólos de Braga e do Porto nesse domínio.

Esse fosso, aliás já discutido por ocasião do debate em 1999[10], e que tentámos reduzir durante e através da Primeira Escola de Verão, reapareceu como não resolvido, no entender de Paulo Gomes (Gomes, 2008) ou de Belinda Maia (Maia, 2008a). Convém a esse respeito relembrar que Fernando Pereira, em 1999, tinha instado para que se criassem pessoas interdisciplinares ao contrário de equipas interdisciplinares. Ainda parece haver, no entanto, muitíssimo a fazer para que esse objectivo seja atingido.

2.6 Avaliação e controlo de qualidade em relação ao português

Em relação a este ponto, penso que a Linguateca deu um contributo decisivo, tendo-se de facto transformado no serviço preconizado em 1999:

> Seria, pois, vantajoso ter um serviço público de "portuguesificação" (por oposição a aportuguesamento) da tecnologia, incumbido de organizar as conferências de avaliação e de informar a comunidade, de garantir a distribuição dos recursos, de levar a cabo ou encomendar testes de qualidade e representar o país em órgãos internacionais

A única coisa que não aconteceu foi a "representação do país", mas dado que isso seria um trabalho sobretudo político, foi certamente preferível que esse trabalho não fosse misturado com o trabalho científico e tecnológico envolvido no resto das actividades da Linguateca, e que naturalmente nos deu muito trabalho e muito prazer.

De facto, mais do que isso: a questão "país" foi sempre substituída por "língua", tendo a Linguateca sempre defendido a língua portuguesa e não a língua dos portugueses, e tendo aliás conseguido muito boas parcerias com os investigadores brasileiros[11] exactamente por ter substituído a componente nacional por uma definida em termos da língua, que nos continua a parecer ser a única que faz sentido em termos do domínio de estudo e de prática: ou seja, no que respeita ao desenvolvimento de sistemas que lidem natural e inteligentemente com o português.

Assim, a organização de avaliações conjuntas e a sua motivação foi uma das actividades mais florescentes (e também mais absorventes) da Linguateca, como será descrito na secção 4.7.

2.7 Disponibilização de recursos (nas suas múltiplas vertentes)

Historicamente, a Linguateca foi aprovada com o nome bafiento e pouco imaginativo de *Centro de Recursos - distribuído - para a Língua Portuguesa (CRdLP)*, tendo como principal actividade a criação e distribuição de recursos.

Embora tenhamos mudado o nome e dedicado muito do nosso trabalho e empenho à avaliação, naturalmente que a criação e disponibilização de recursos – assim como a sua manutenção – foi o prato forte da actividade da Linguateca, como aliás será descrito no decurso do presente artigo.

É interessante a esse respeito ver o que foi considerado relevante em 1998 e contrastá-lo com o que temos agora (na Linguateca ou na comunidade mais vasta).

Em alguns casos, a lista referia produtos razoavelmente vagos, e noutros, demasiado específicos. Senão vejamos: Não temos provavelmente terminologias, mas temos sistemas que as permitem desenvolver; não temos dicionários com subcategorização, mas temos sistemas que permitem obtê-

[9] De facto, muito anteriores à Linguateca, como é o caso da RITERM, fundada em 1988, da TERMIP, de 1989, ou da Realiter, de 1993.

[10] cuja transcrição continua acessível do sítio da Linguateca

[11] Infelizmente, exceptuando alguns casos pontuais, a Linguateca não conseguiu (ainda) atingir ou colaborar com outros países de expressão portuguesa.

los a partir de corpos; não temos dicionários entre as variantes do português, mas temos sistemas de alinhamento que os podem eventualmente criar.

A própria terminologia também evoluiu (ou o nível de ambição): Em vez de tesauros, falamos agora de ontologias; em vez de corpos alinhados, de corpos paralelos; em vez de estudos de frequência, temos serviços que nos permitem fazê-los de forma não imaginada na altura.

Embora ainda haja certamente muitos recursos que podíamos e devíamos (como comunidade) criar, houve um claro progresso e pensamos poder afirmar que o português se encontra entre as línguas do mundo com mais recursos linguísticos públicos para o seu processamento.

Contudo, atentando nas propostas adiantadas para o conseguir, reparamos que fizemos a maior parte das coisas sozinhos, ou melhor, no âmbito da Linguateca, e não através dos meios propostos, que continuam, passados dez anos, a não passar do papel:

> a obrigatoriedade de inclusão de distribuidores e avaliadores de recursos nas próprias propostas de projectos a serem financiados, de forma a que cada centro ou grupo, além das actividades de desenvolvimento, investigação, ensino e divulgação também levasse a sério os serviços de teste, verificação e fornecimento de um serviço.

Isto continua a ser uma miragem, não há qualquer controlo de qualidade e disponibilidade dos resultados dos projectos financiados, pelo menos em Portugal.

Pelo contrário, a única coisa que se nos tornou clara em relação à disponibilização é que o nosso modelo público, **tudo grátis e sem entraves**[12], é a única maneira de chegar realmente a toda a comunidade e de evitar a mesquinhez dos tempos antigos.

Assim, como descrito na secção 4.4, comprámos o direito aos possuidores comerciais de disponibilizar recursos para todos, e isso foi um ovo de Colombo em que penso que fomos pioneiros.

Já quanto à parte da postura arquivística, também mencionada no mesmo item,

> Convém também referir que seria muito útil uma postura arquivística a respeito dos recursos, ou seja, para poder distribuir e descrever os recursos, há necessidade de criação (e de uso) de estruturas classificativas (taxonomias, tesauros classificativos); assim como se devia fomentar a codificação da informação em formatos partilháveis (tais como XML, TEI), ou pelo menos bem documentados.

temos de referir que não foi um sucesso, e isto por duas razões diferentes:

A primeira, passível de autocrítica, foi não termos tentado o suficiente. A catalogação foi sempre o parente pobre na Linguateca – ou seja, os nossos colaboradores, sem excepção, deram sempre menos prioridade a actualizar os diversos catálogos[13] do que a desenvolver sistemas ou programas ou serviços.

A segunda, no que tem a ver com a questão dos padrões, correspondeu a uma decisão pensada: considerámos sempre que o conteúdo era mais importante do que a forma, e que os padrões seriam definidos ou emergiriam do uso e não da estipulação exterior. Penso que tivemos razão, e que os padrões mencionados não são mais do que um embrulho que qualquer outro grupo pode aplicar, se precisar. Assim, os nossos padrões surgiram do trabalho que fizemos, não da adopção apriorística de regras na moda.

Em contrapartida, a documentação dos nossos produtos, serviços e recursos foi considerada de extrema importância, assim como a nossa presença na rede. Sentimos que a documentação em português era necessária quer para os falantes de português quer para a nossa identidade própria de desenvolvedores de sistemas para o processamento do português (ver secção 5.7).

2.8 Definição do processamento do português como área prioritária

Este ponto da proposta era muito vago e dirigido aos órgãos de financiamento ou organizações governativas. Até pelos percalços da actividade de governação, seria difícil de implementar ou garantir por governos sucessivos. Passe pois o conteúdo demagógico, e dediquemos apenas a atenção aos pontos concretos aventados, nomeadamente a questão da continuidade, da medida do peso da língua, a criação de um fórum, e de uma comissão internacional.

A parte ínfima que foi levada à prática foi a continuidade da própria Linguateca, no sentido em que conseguimos sobreviver dez anos e não os 2-3 anos mencionados e que continuam a constituir o prazo dos projectos de investigação.

Quanto à questão da avaliação da área, provavelmente no âmbito de um observatório estatal, nada foi para a frente que envolvesse o processamento da língua, nem mesmo a estipulação de me-

[12] No início do processo, não tínhamos esta percepção. De facto, até indico "Note-se que público não significa grátis" na respectiva secção de Santos (1999b).

[13] Como será referido em mais pormenor em 5.5.1.

didas a serem efectuadas. Contudo, existem outras instituições como a União Latina ou o Instituto Camões que poderiam tratar dessa questão. E de facto existe já há alguns anos o Observatório da Língua Portuguesa[14] que aparentemente faz alguns desses estudos.[15]

Quanto à criação de um fórum, no sentido de lista de discussão, já havia – e continua a haver – o forum-lp[16], mas que infelizmente apenas veicula anúncios (muitas vezes até em inglês!) e quase nunca discussão. Das muitas listas que a Linguateca foi criando ao longo dos anos sobre temáticas mais específicas, como avaliação conjunta, por exemplo, o mesmo resultado pode ser descrito: a comunidade portuguesa e brasileira na área do PLN não gosta nem costuma discutir questões científicas ou outras nas listas.

Se o fórum mencionado era uma conferência, temos o PROPOR, e agora no Brasil o (S)TIL e cada vez mais conferências em cada país. Mas como infelizmente o primeiro é em inglês, e o segundo não é restrito ao português, parece que ainda não existe a arena certa, ou pelo menos nenhuma especialmente dedicada e que permita a comunicação ideal dos assuntos tratados. Aparentemente, as associações de linguística de Portugal e do Brasil, APL e ABRALIN, embora ambas em países de língua portuguesa, não estabelecem fóruns comuns, e por isso também não parece possível usar nenhuma delas para dedicar ao processamento da língua portuguesa em geral, em português. Também não há (ainda?) nenhuma revista só em português sobre o seu processamento, embora a *Linguamática* seja um caso em que o mesmo é acarinhado, o que é de louvar.

Com o afã de publicação, temos de nos render à evidência: as pessoas querem publicar, não discutir nem mesmo convencer. Esse tal fórum seria ideal se fosse para as pessoas discutirem questões e da discussão sair a luz. O formato de publicação e comunicação que existe nos tempos presentes (e que não é exclusivo da nossa área ou dos nossos países) não favorece nada, contudo, esse resultado...

Finalmente, a menção de uma comissão internacional era um resquício da subserviência nacional à norma: "lá fora é melhor do que cá dentro", de que me congratulo sobremaneira não ter ido avante. No caso da língua, isso parece-me trivialmente falso. Na minha opinião, já existem demasiadas comissões internacionais de qualidade duvidosa a ameaçar a nossa soberania intelectual.

2.9 Balanço em relação ao enquadramento inicial

Santos (1999b), documento publicado na rede sem pretensões e discutido em 1999, era em muitos aspectos ingénuo e pouco fundamentado, mas apontava algumas questões concretas que era preciso atacar. Passados dez anos, é possível fazer planos muito mais concretos, e também ter muito maiores ambições quanto à área.

Agora já não falta (quase) tudo, como era o caso na altura, e a comunidade do processamento do português pode, se assim o desejar, fazer avaliação de qualidade e usar ou desenvolver recursos mais complexos. Nesse aspecto, e como aliás tentarei mostrar no resto do artigo, a actividade da Linguateca foi decisiva, embora não única.

Por outro lado, o que se passou nesta década demonstrou que, se era fácil ou possível melhorar a área no que se refere à investigação, era certamente muitíssimo mais complicado fazê-lo quanto ao impacto na sociedade em geral. Nesse ponto ainda está praticamente tudo por fazer. Voltarei a este assunto na secção 7, depois de esmiuçar as razões de satisfação – e preocupação – que o balanço da própria Linguateca me suscita.

Antes disso, porém, farei uma pequena história das várias inflexões que o projecto Linguateca sofreu, provocadas por um lado pela conjuntura político-científica distinta, e por outro por várias condicionantes pessoais da equipa da Linguateca: visto que a Linguateca são as pessoas que a compõem ou compuseram ao longo do tempo, com as suas forças e fraquezas específicas e com interesses individuais distintos.

3 A evolução

Podemos identificar alguns pontos de viragem, ou de nascimento de novas actividades, em vários momentos, não necessariamente redutíveis ao historial visível.[17]

Para referência, indica-se uma lista dos pólos[18]

[14] http://www.observatoriolp.com/

[15] O "aparentemente" deve-se ao facto de, a 30 de Março de 2009, o gráfico do "Conteúdo da Internet por línguas" se referir ao ano de 2001, e o das "Línguas da População em linha" se referir a Setembro de 2002, o que abona pouco quanto ao dinamismo e correcção de informação no dito sítio. As "Línguas de maior influência", por seu turno, referiam-se a Dezembro de 1997...

[16] Lista criada a 6 de Junho de 1997 pelo então denominado grupo "Glint - Grupo de Língua Natural DI/FCT/UNL/PT", do departamento de informática da FCT da Universidade Nova de Lisboa. Na perspectiva da Linguateca, contra a duplicação de esforços, era óbvio que devíamos apoiar e ajudar, usando, esta lista, em vez de tentar com ela competir, e temo-la usado desde sempre.

[17] No sítio da Linguateca, é possível consultar quer um historial quer uma lista de encontros organizados pela Linguateca.

[18] Conforme já indicado, muitos deles são ou foram pólos "informais" por razões administrativas. Para efeitos deste cômputo, desde que exista um doutorado associado à Linguateca, considero que um pólo existe, mesmo que a sua bolsa não seja paga pela Linguateca.

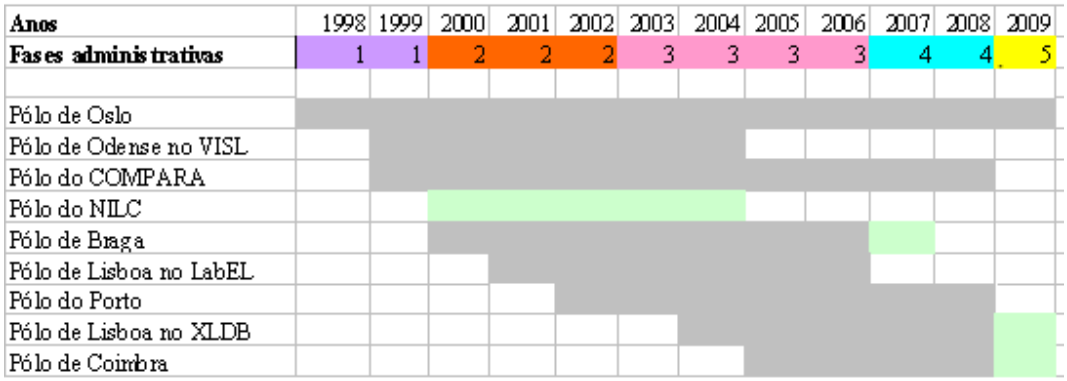


Figura 1: Actividade nos pólos, não necessariamente directamente financiada: a verde apresenta-se actividade exclusivamente no âmbito de doutoramentos

da Linguateca:

Pólo de Oslo Inicial, iniciado a 15 de Maio de 1998

Pólo do COMPARA Informalmente iniciado em 1999, formalmente transferido para a FCCN no início de 2007 e encerrado em Dezembro de 2008

Pólo de Odense Informalmente iniciado em 2000, desde 2004 apenas contando com Eckhard Bick como co-líder da Floresta

Pólo do NILC Iniciado em 2001 com o doutorado sanduíche da Rachel Aires e encerrado com a conclusão deste em 2005

Pólo de Braga Iniciado em 2000, sem pessoal afecto desde Outubro de 2007

Pólo de Lisboa no LabEL Iniciado em 2002, encerrado em Setembro de 2006

Pólo do Porto Iniciado em 2003, sem pessoal afecto desde Novembro de 2008

Pólo de Lisboa no XLDB Iniciado em Janeiro de 2004

Pólo de Coimbra Iniciado informalmente em Julho de 2005, e formalmente em Fevereiro de 2007

Além do cronograma institucional, na figura 1, e da lista dos recursos humanos com que contámos, na tabela 1, que iremos brevemente analisar na secção 3.4, podemos também mencionar actividades específicas de reunião de vários pólos num objectivo maior, e que foram fulcrais para a fertilização cruzada dos muitos ambientes distintos que compuseram a Linguateca ao longo dos tempos.

Durante os dois primeiros anos, além da preparação do documento discutido na secção 2, foram lançadas as sementes para a disponibilização dos corpos na rede (tanto o AC/DC (Santos e Bick, 2000) como o COMPARA (Frankenberg-Garcia e Santos, 2002) viram a luz do dia), e a primeira floresta para o português foi lançada, com três bolseiros em Odense (Afonso et al., 2001).

O primeiro grande acontecimento, que exigiu muito planeamento e muita discussão interna preliminar, foi o Encontro Preparatório sobre Avaliação conjunta (EPAv), com o objectivo de promover e iniciar o modelo da avaliação conjunta na comunidade do processamento computacional do português.

No ano seguinte ao EPAv, a parte de leão da actividade da Linguateca foi consagrada às Morfolimpíadas (Santos, Costa e Rocha, 2003), enquanto o pólo do Porto, o único pólo não envolvido nas ditas, dava os primeiros passos no desenvolvimento do Corpógrafo, ainda pré-baptizado "gestor de corpora" (Sarmento e Maia, 2003).

Em 2003, foi então sugerida uma expansão a nível das competências da Linguateca, que passava por ter mais formação (com a consequente atribuição de três bolsas de doutoramento), e foi integrada a área da recolha de informação, já presente desde o início do trabalho de doutoramento de Rachel Aires (Aires, 2005), através da criação de um pólo no XLDB em 2004.

Por essa altura também o CLEF (Rocha e Santos, 2007) passou a tomar um peso considerável na actividade da Linguateca, devido a estarmos nele tanto como organizadores como participantes (naturalmente, grupos ou indivíduos separados), e a sua periodicidade ser anual.

A questão das ontologias passou a ser mais uma actividade com que a Linguateca se preocupou, quer do foro geográfico quer com as ontologias lexicais criadas a partir das definições de um dicionário, o que levou à GeoNET (Chaves, Rodrigues e Silva, 2007) e ao PAPEL (Gonçalo Oliveira et al., 2008b).

A segunda actividade que congregou mais uma vez a Linguateca toda foi, contudo, o Primeiro HA-

REM, que se estendeu por quase dois anos desde o início dos preparativos até à publicação do livro a ele referente (Santos e Cardoso, 2007).

Outro acontecimento foi a (Primeira) Escola de Verão da Linguateca, que teve lugar no Porto em Junho de 2006, com todos os séniores (e alguns convidados) a disseminar o conhecimento e os recursos produzidos.[19]

Ao mesmo tempo, algumas actividades eram reduzidas ou paradas: foi o caso do serviço AnELL (Mota e Moura, 2003) no pólo do LabEL, que não chegou nunca a ter uma audiência significativa,[20] e da actividade de avaliação de tradução automática iniciada no pólo do Porto (veja-se Santos, Maia e Sarmento (2004)), que foi considerada demasiado difícil para ser continuada, com os recursos que tínhamos e as prioridades dos pólos. Também a actividade de busca inteligente, planeada como um cruzamento entre o conhecimento de terminologia e a recolha básica de informação, embora esboçada em Oliveira et al. (2005), nunca chegou a ser concretizada.

Outras ideias de projectos, ainda, não chegaram sequer a sair da fase de ideia, embora alguma publicidade lhes tivesse sido feita para obter novos colaboradores, mas em vão: um meta-dicionário (serviço na rede conjugando a consulta a muitas bases lexicais diferentes), a análise de diários às visitas ao sítio da Linguateca (e não só dos seus serviços), e interacção com fala.

Em 2006, uma nova proposta de continuação pôs a ênfase no reforço de alguns projectos com maturidade, nomeadamente o COMPARA e o HAREM (a sua segunda edição), cobrindo o resto do financiamento do programa POSC.[21]

3.1 Diferentes eixos

O modelo IRA (informação, recursos e avaliação), descrito desde o início como a trilogia fundamental da nossa actividade, foi passando a ser complementado, em novas versões da apresentação da Linguateca, com novos e variados eixos, à medida que nos compenetrávamos de tudo o que nos tínhamos comprometido a (ou tínhamos vontade de) fazer.

Senão vejamos: em Santos, Cabral e Costa (2006) ao fazer um balanço de sete anos da Linguateca, adicionámos as seguintes vertentes: manutenção de recursos, apoio, investigação (con-

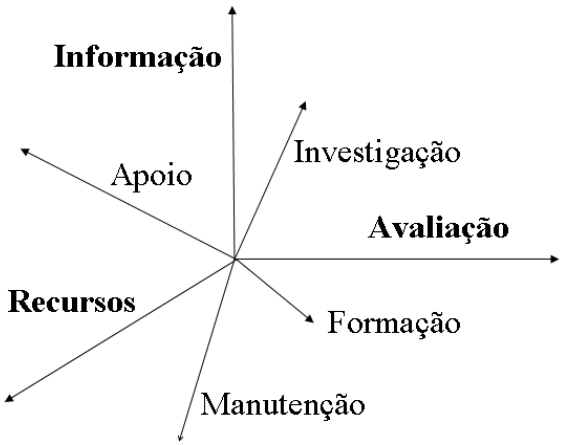


Figura 2: Eixos da actuação da Linguateca

substanciada nos doutoramentos e mestrados) e formação (relacionada com os vários simpósios doutorais e sobretudo com a (Primeira) Escola de Verão da Linguateca), veja-se a figura 2.

Ainda agora não tenho a certeza se o avançar por todos estes eixos foi uma boa ideia ou se resultou em alguma dispersão. Contudo, no âmbito da própria Linguateca, a Escola de Verão foi considerada por vários dos seus membros no encontro em Aveiro como um dos pontos altos da actividade. Possivelmente o facto de ter dado origem a – ou pelo menos influenciado positivamente – novas escolas ministradas em português: a I e II EBRalC[22], respectivamente em São Paulo e em São José do Rio Preto, e as futuras escolas que terão lugar ainda este ano de 2009, a primeira sobre "Aspectos do PLN em português", no Porto, e a III EBraLC, no Rio de Janeiro.

3.2 Formas de apresentação ao longo do tempo

Se compararmos a apresentação da Linguateca ao longo do tempo, vemos que a ênfase em catalogar e juntar os recursos acessíveis até à produção de ferramentas, sistemas ou avaliações conjuntas variou claramente.

Assim, numa leitura actual de Santos (2000), qua fazia o balanço dos dois primeiros anos de actividade, o que mais se destaca é a desproporção sobre o que, passados dez anos, fizemos em avaliação e o que pretendíamos ou imaginávamos poder fazer, em que até está mencionada a encomenda dessa actividades a actores fora da Linguateca. Assim como está bem patente a nossa esperança, depois frustrada, de incluir a fala.

Alguns pormenores interessantes mencionados, que saliento aqui, têm a ver com a preocupação de estabelecer uma metodologia (e formação) da

[19] À boa maneira da Linguateca, todo o material de ensino foi tornado público a seguir à escola, http://www.linguateca.pt/EscolaVerao2006/.

[20] Contudo, pode também interpretar-se como não ter sido totalmente implementado – de facto, outros serviços existem para o português, tais como o do VISL, http://visl.sdu.dk/, e o recente F-EXT-WS (Fernandes, Milidiu e Santos, 2009).

[21] Programa para a Sociedade do Conhecimento, activo em Portugal no período 2000-2008, http://www.posc.mctes.pt/.

[22] Escola Brasileira de Linguística Computacional

citação dos recursos criados pela Linguateca. Dada a explosão exponencial desses e doutros corpos no panorama do português, tivemos de nos render à evidência de que era quase impossível controlar ou dirigir a forma como nos citavam ou apresentavam exemplos de corpos.

Também já nessa altura pudemos apreciar que o repositório, ou seja, o serviço que iniciámos para que os investigadores que não tivessem possibilidade de o fazer tivessem uma prateleira para expor e disponibilizar os seus trabalhos na rede, não parecia muito interessante para a maioria da comunidade. Isto ainda veio a ser mais pertinente dado que a presença na rede de todas as instituições e actores passou a ser um dado adquirido, com o que aliás nos congratulamos vivamente.

Em Santos (2002b), publicado precisamente antes da escolha do nome *Linguateca*, é patente que já entrámos na espiral da avaliação conjunta, embora ainda tivéssemos a esperança de vir a ter pólos no Brasil, o que não foi nunca possível por questões políticas completamente fora do nosso alcance.

Santos e Costa (2005), por outro lado, ao apresentar a Linguateca numa revista de terminologia, põe a ênfase na publicitação dos vários recursos e projectos, constatando que, estando a infraestrutura montada, é altura de nos dedicarmos a tarefas mais complexas, de investigação aplicada. Essa previsão, e sobretudo a lista de tarefas apresentada, inspirada pelos assuntos que, na altura, se esperava que os novos doutorandos associados se dedicassem, não veio em geral a verificar-se. Mas o artigo é sintomático da fase por que passávamos (veja-se a próxima secção), que obrigava a que nos afirmássemos também como um projecto científico e não apenas de apoio e serviço à comunidade. Um foco interessante desse artigo é a descrição do levantamento feito na comunidade em 2002 sobre as áreas em que estariam interessados na avaliação, algo que foi realizado nessa altura mas nunca mais repetido ou actualizado.

Santos (2007a), por seu lado, é, até agora, o texto que melhor explica o conceito de avaliação conjunta, e a motivação para a Linguateca tomar a peito a sua divulgação e sobretudo implementação. Embora parcial porque só se refere a essa vertente, a da avaliação, foi escrito – em 2004, embora publicado em 2007 – para divulgar sem pressupor qualquer conhecimento desse paradigma de avaliação. E que muito brevemente exponho de novo aqui, para que os leitores possam compreender melhor as subsequentes referências às Morfolimpíadas, CLEF e HAREM: avaliação conjunta é a comparação do desempenho de vários sistemas com base numa tarefa comum, recursos comuns, e um aproximar de todos os interessados na área para o seu desenvolvimento e validação.

Finalmente, o presente artigo faz de novo um balanço, ao passar para uma nova fase: estou convencida de que o modelo da Linguateca tem de sofrer uma revisão substancial, e que a sua prática terá de ser mudada (ou transferida, ou encerrada) com base na reflexão que espero que este artigo possa suscitar.

3.3 Formas de apoio institucional à Linguateca (ou sua falta)

Parece-me que se deveria referir que a Linguateca não foi um projecto com um apoio estável ou com uma garantia de continuação sustentada se os seus resultados e o seu impacto fossem francamente bons – como aliás parece ser ser impossível num país da comunidade europeia ou da comunidade dos países de língua portuguesa.

Penso que, dado o financiamento e as restrições recebidas, os resultados foram bons, e a Linguateca merecia uma garantia de continuidade, mas isso não impediu a instabilidade e a total insegurança quanto à continuação do projecto em quase meia dezena de ocasiões, e aliás algumas interrupções reais de financiamento ocorridas, que não poucas vezes foram extremamente prejudiciais para os colaboradores mais jovens.

De facto, como todos os que lidaram de perto ou mesmo de longe com a nossa actividade sabem, a Linguateca materializou-se, do ponto de vista institucional, com uma sequência sempre precária e pouco reconhecida de "medidas" *in extremis* e a urgente necessidade de cumprir requisitos por vezes contraditórios de ano para ano, à medida que as fontes de financiamento foram surgindo ou mudando, assim como as regras a cumprir (de forma frequentemente inexplicável).

Se isso por um lado se deveu a diferentes governos, diferentes programas quadro e a diferentes reorganizações de tudo quanto é científico-tecnológico em Portugal e na Europa, extravasando claramente a insignificância da Linguateca e atingindo quase certamente toda a comunidade científica em todas as áreas,[23] por outro é preciso dar a ideia a quem não sabe que não fomos de forma alguma melhor tratados ou financiados do que qualquer outro projecto ou grupo em Portugal. De facto, foi elevada a percentagem de bolsas, contratos a recibos verdes, e trabalho voluntário para a Linguateca, assim como o expediente de considerar o contrato da Linguateca com o SINTEF como "investimento", de forma a garantir uma continui-

[23] Isto no que se refere ao financiamento da ciência. No que diz respeito à língua ou à cultura, ou melhor quanto à CPLP (e o seu IPLP) ou ao Instituto Camões, apesar de mais de dez anos de actividade da Linguateca, ainda não fomos reconhecidos sequer com um mero atalho nos sítios respectivos.

dade mínima (veja-se Santos (2008b) para os dados deste último).

Uma questão que foi discutida no Encontro dos 10 anos em Aveiro, mas que continua sem resolução, é exactamente que critérios de avaliação devem ser aplicados a uma iniciativa, ou organização virtual, como a Linguateca: que é ou foi concebida como um projecto de infraestrutura e não como um projecto científico.

Temos contudo e experiência negativa de em várias alturas a Linguateca ter sido avaliada (felizmente que positivamente) como se apenas de mais um projecto científico se tratasse (com critérios de número de publicações, por exemplo), o que demonstra mais uma vez um total desconhecimento ou falta de apoio dos organismos públicos que nos encomendaram a missão.

Em Costa e Cabral (2008), foram apresentados alguns indicadores sobre a Linguateca referentes a 2008, mas o estudo da verdadeira influência (ou falta dela) através de um estudo da literatura na área e áreas afins seria relevante para uma compreensão maior das consequências da nossa actividade.

3.4 O material humano associado à Linguateca

Na figura 1 apresento um quadro aproximado da ligação e trabalho efectivo dos variados membros afectos à Linguateca e pagos para tal.

Tornando a insistir na grande precariedade em que muitos elementos participaram na Linguateca, os "meses" são pois uma abstracção que se refere muitas vezes ao multiplicar e somar valores de contratos a prazo definidos à hora.

Se por um lado os mais de trinta elementos todos receberam mais ou menos formação – e pelo menos experiência – na manutenção e disponibilização de recursos e serviço continuado à comunidade, por outro as tarefas e as apetências de cada um variaram muito, conforme aliás o pólo em que estiveram envolvidas.

Se para alguns a Linguateca representou um acidente de percurso, estou convencida de que para muitos o espírito da Linguateca e o que aprenderam nela foi ou será importante para o seu futuro, e também penso que muito poucos lamentam a sua ligação.

É importante contudo salientar que escolhi fazer uma apresentação e balanço puramente pessoal – e não organizacional, como foi feito noutros casos, por exemplo em Santos et al. (2004) – e que este artigo deverá e poderá ser favoravelmente complementado pela apreciação que cada um dos séniores da Linguaetca, na sua versão pessoal, faz da sua pertença ou associação, pelo tempo que durou (no caso daqueles que já se retiraram), da vida do seu pólo e da integração ou não na Linguateca como um todo.

Diana Santos	120
Signe Oksefjell	14
Paulo Rocha	72
Tom Funcke	3
Susana Afonso	24
Miguel Oliveira	6
Rachel Marchi	18
Renato Haber	12
Alexsandro Soares	10
Rosário Silva	21
Pedro Moura	12
Anabela Barreiro	6
Luís Costa	57
Cristina Mota	22
Luís Sarmento	37
Alberto Simões	17
Luís Miguel Cabral	40
Débora Oliveira	12
Susana Inácio	50
Nuno Seco	10
Isabel Marcelino	12
Rui Vilela	26
Ana Sofia Pinto	12
Nuno Cardoso	38
António Silva	12
Ana Frankenberg Garcia	7
Sérgio Matos	12
Cláudia de Freitas	18
Hugo Oliveira	15
Pedro Martins Sousa	15
David Cruz	14
Paula Carvalho	13

Tabela 1: Colaboradores da Linguateca, por ordem de entrada (primeiro contrato), e seu contributo em meses de trabalho

Porque é preciso também relembrar que a Linguateca, mais do que a soma de todas as pessoas envolvidas, pode ser definida, estudada e explicada como a soma dos pólos, cada um deles envolvido em ambientes diferentes e com objectivos últimos diferentes.

4 Razões para satisfação e orgulho

De dez anos de trabalho em prol da comunidade, poder-se-ão naturalmente aduzir um grande número de razões para louvar e agradecer à Linguateca a sua actividade. Indico aqui as que, do meu ponto de vista, são as mais interessantes, embora não necessariamente as mais conhecidas.

Penso que em muitas destas coisas nós fomos até pioneiros a nível mundial, embora com a ressalva de que, sem a bênção da publicação interna-

cional, tal nunca será provavelmente reconhecido.

4.1 A importância da rede

Fomos dos primeiros a medir, de uma forma motivada pelo conhecimento da nossa língua, a dimensão da rede (em inglês, "Web") em português (Aires e Santos, 2002). Além disso, preocupámonos com a recolha de informação nesse contexto, em vez de usar colecções de textos jornalísticos. A primeira tese de doutoramento na Linguateca (Aires, 2005) foi pois pioneira de várias formas, e em particular pela sua intransigência determinada em recusar substitutos que não a própria rede para estudar e para desenvolver protótipos.

Também ajudámos ou incentivámos os motores de pesquisa na nossa língua e/ou cultura ao disponibilizar, e/ou ao ajudar à criação de colecções da rede disponíveis para investigação e desenvolvimento de sistemas para a língua portuguesa. A WBR-99 (Calado, 1999), a WPT-03 (Cardoso et al., 2007) e a WPT-05 são assim recursos relevantes para quem quer estudar a linguagem e a morfologia da rede em português.

Além disso temos usado cada vez mais – ao longo de uma era em que a rede cada vez mais explode em géneros e contribuições – material proveniente da vida virtual de cada um em todos os materiais de avaliação que temos tido a ocasião de criar. Assim, veja-se que, se nas Morfolimpíadas o texto da rede correspondia a menos de 10%, no Primeiro HAREM essa percentagem passou para 20% e no Segundo HAREM para 85%.[24]

Não foi também por acaso que outras teses de doutoramento se tenham concentrado em textos na rede: tanto Chaves (2008) como Cardoso (2008b), embora de forma muito diferente, lidam primordialmente com a informação geográfica na rede. Com se verá na secção seguinte, também o sistema de RAP desenvolvido na Linguateca, o Esfinge (Costa, 2005), usa a redundância da rede como um elemento principal.

Finalmente, o próprio uso da rede como recurso para outro tipo de dados, por exemplo para a compilação de corpos paralelos, também foi investigado pelo pólo de Braga desde muito cedo, como se pode apreciar em Almeida, Simões e Castro (2002).

4.2 Novos modelos de resposta automática a perguntas

Estou também convencida de que a Linguateca deu uma contribuição importante à área da resposta automática a perguntas, RAP – e não só à existência de vários sistemas e grupos interessados nessa aplicação para o português.

Com efeito, desde 2004 que somos co-responsáveis pela organização da pista de RAP do CLEF, QA@CLEF, incluindo o português, veja-se por exemplo Vallin et al. (2005) e Forner et al. (2009), e o que é um resultado indiscutível do CLEF é que já em 2007 o português foi a língua com mais sistemas participantes de RAP.

Contudo, a Linguateca também foi autora de uma proposta inovadorada de RAP colaborativa (Santos e Costa, 2007); da disponibilização de colecções sintacticamente anotadas para teste e treino de sistemas de RAP (Santos e Rocha, 2005); de um sistema desenvolvido de raiz para o português em código aberto, o Esfinge (Costa, 2005; Costa, 2006); e duma avaliação conjunta pioneira, o GikiP (Santos et al., 2009), seguido pelo GikiCLEF, em progresso neste momento.[25]

Além disso, embora indirectamente, esperamos contribuir para a existência de mais trabalhos de investigação na área ao incluirmos perguntas na colecção do Segundo HAREM, conforme explicado em Carvalho et al. (2008).

Ao contrário de muito do trabalho corrente em RAP, cuja preocupação é melhorar alguns pontos percentuais no desempenho de sistemas, sem entrar em conta com a realidade e/ou pertinência da tarefa ou com a validade linguística dos modelos empregues (veja-se por exemplo a tarefa de detecção do tipo de resposta descrita em Roberts e Hickl (2008)), a nossa actuação tentou sempre pautar-se por trazer a RAP para a realidade das necessidades do utilizador e não de uma comunidade científica específica.

4.3 Recursos realmente acessíveis

O que fizemos com o projecto AC/DC foi de facto pioneiro – colocar todos os corpos que pudemos disponibilizar acessíveis de uma maneira idêntica, para facilitar o seu uso e manipulação com um mínimo (ou nenhum) conhecimento informático (Santos e Bick, 2000; Santos e Sarmento, 2003).

Convém relembrar que na altura não havia nenhum sistema de procura ou acesso a corpos em português, e os poucos corpos existentes eram levantados em conjunto (ou seja, por "download").

Depois disso, muitas outras instituições – algumas sem sequer nos mencionar ou citar (Bacelar do Nascimento, Mendes e Pereira, 2004; Aluisio et al., 2004), outras explicitamente explicando que o nosso modelo não lhes convinha (Aluísio, Oliveira e Pinheiro, 2004) – puseram os seus corpos também acessíveis na rede.

Outros ainda criaram novos corpos e novas interfaces, o Corpus Informatizado do Português Me-

[24] No caso do Segundo HAREM, estou a contar apenas a colecção dourada, visto que a colecção do Segundo HAREM foi obtida a partir dessa e da colecção CHAVE. Para mais pormenores, ver Santos et al. (2008).

[25] Veja-se http://www.linguateca.pt/GikiCLEF/.

dieval (Xavier et al., 1998), o Corpus do Português (Davies e Preto-Bay, 2008), o Corpus Brasileiro (Berber Sardinha, Moreira Filho e Alambert, 2008). De facto, podemos agora afirmar que não existe efectivamente falta de material anotado sobre o português, embora eu ache que do ponto de vista da documentação, o material da Linguateca é ainda incomparavelmente superior – o que não significa que não possa ser melhorada.[26] Por outro lado, no que respeita à usabilidade e à experiência de interacção proporcionada ao utilizador, estamos decididamente bem atrás destes três projectos.

Não é possível, naturalmente, pronunciar-me sobre se todas estas iniciativas teriam existido na mesma sem a Linguateca, ou se, pelo contrário, apareceram como uma resposta, positiva ou negativa, à nossa actividade.

4.4 Modelos económicos

Uma questão em que a Linguateca sempre insistiu foi a de não dever haver diferença entre usos comerciais e usos académicos. Tal distinção foi, aliás, considerada um dos principais entraves à fertilização cruzada entre investigação e produtos com impacto no dia a dia.

Assim, o CETEMPúblico (Rocha e Santos, 2000) foi negociado com o jornal PÚBLICO exactamente nessa base, assim como o PAPEL (Gonçalo Oliveira et al., 2008b) e o CLASSLPPE, com a Porto Editora, o foram também. Estes casos são aliás a prova cabal de que não há uma distinção de mentalidades entre empresas e universidades. De facto, e ao contrário da tese "as companhias privadas só querem o proveito próprio, enquanto os universitários estão conscientes do seu papel social", as empresas foram em geral mais receptivas a disponibilizar do que muitos grupos ou investigadores individuais.

Talvez também seja de realçar que, mais uma vez ao contrário do que poderia ser esperado, foram sempre sistemas comerciais ou semicomerciais que venceram as avaliações conjuntas que organizámos: nomeadamente o PALAVRAS (Bick, 2000), o CorTex (Aranha, 2007) e o sistema da Priberam (Amaral et al., 2008). Não se pode, pois, partir de uma hipótese definitivamente não corroborada para continuar a defender a excelência académica por oposição à cegueira empresarial: no contexto da língua portuguesa, isto simplesmente não é verdade.

[26] Veja-se por exemplo a documentação sobre a revisão da anotação morfossintáctica da parte portuguesa do COMPARA (Inácio e Santos, 2008), que pretende indicar todas as opções tomadas em algo que é obviamente não trivial.

Tipo de texto	Abs.	Tam.	Rel.
Texto traduzido	444	723807	61,34
Texto original	258	818553	31,52

Tabela 2: Diferença entre texto original e traduzido no que se refere a *already* no COMPARA 13.1.4.

Expressão	Freq. absoluta	Freq. relativa
já	3121	2,17
já - already	811	0,56
already	916	0,59

Tabela 3: Ocorrências de *já* e de *already* no COMPARA, versão 13.1.4.: a frequência relativa é por mil palavras da língua respectiva

4.5 Corpos paralelos

Outra área em que a Linguateca muito fez foi na disponibilização e divulgação de corpos paralelos através do COMPARA (Frankenberg-Garcia e Santos, 2002) e, mais tarde, do CorTrad[27]. Que eu saiba, o COMPARA é o maior corpo paralelo revisto morfossintacticamente no mundo inteiro, e tem algumas funcionalidades únicas, tal como a procura por notas de tradução e a distribuição cruzada (Santos, 2002a). Além disso tem anotação semântica revista (Santos, Silva e Inácio, 2008), algo que também é raro, senão único, em corpos paralelos.

Ainda podemos salientar o facto de uma das primeiras análises quantitativas da interacção dos utilizadores com um corpo paralelo ter sido feita no COMPARA (Santos e Frankenberg-Garcia, 2007).

Contudo, um erro cometido no âmbito do COMPARA foi a dependência demasiada em relação a autorizações demasiado específicas de algumas editoras, o que implica (ou implicará, num futuro próximo, dependente de cada autorização) o retirar dos pares de textos respectivos do acesso público. É minha convicção agora que não deveríamos ter investido tanto trabalho (de revisão e anotação) em textos que teriam uma vida pública breve.

De qualquer maneira, noto que o DISPARA facilitou enormemente a obtenção de dados e de pesquisas num corpo paralelo: por exemplo, para obter a informação de que *already* é mais frequente em texto traduzido do que em texto original (ver tabela 2), ou de que *já* corresponde mais a *already* do que *already* a *já* (ver tabela 3), tabelas laboriosamente obtidas durante o meu doutoramento, e referidas entre outros em Santos (1995) ou Santos (2008c), basta um simples comando no DISPARA.

[27] O CorTrad é um subprojecto do projeto COMET - Corpus Multilíngüe para Ensino e Tradução, da Universidade de São Paulo, cuja disponibilização é feita através do sistema DISPARA, em parceria com a Linguateca e o NILC.

4.6 Análise gramatical

Outro dos pressupostos científicos da Linguateca, que pensamos ter sido completamente demonstrado, foi a inutilidade, e mesmo prejuízo, de focar em "POS tagging" (anotação da categoria gramatical em contexto) em vez de tentar uma análise sintáctica mais complexa. Como defendido em Santos (1999c), essa aplicação é boa para o inglês, mas pouco apropriada para línguas que, como o português, têm mais de setenta formas verbais diferentes, além de um sistema complexo de enclíticos e mesoclíticos. Claramente a ênfase no que é problemático (e fácil) na nossa língua é mais útil do que a importação acrítica de modelos criados para línguas diferentes.

É certo que o facto de termos um pólo em Odense levou a que a Linguateca favorecesse, no sentido de publicitasse, o PALAVRAS (Bick, 2000), mas não só é preciso indicar que isso se deveu ao desejo de Eckhard Bick colaborar com a Linguateca (uma colaboração que se afigurou vantajosa para ambas as partes), como não houve nem há nenhum outro sistema de análise gramatical comparável para o português, pelo menos de que eu tenha conhecimento. Por essa razão, existe de certa forma um monopólio do PALAVRAS para o processamento da língua portuguesa.[28]

Contudo, penso dever salientar que a Linguateca contribuiu para melhorar o PALAVRAS de várias formas distintas e não insignificantes: Por um lado, ao ter entrado em vários projectos conjuntos que incluíam o VISL, em particular a Floresta Sintá(c)tica (Afonso et al., 2001; Bick et al., 2007; Freitas, Rocha e Bick, 2008a), em que um dos objectivos principais era mesmo a melhoria do analisador sintáctico e das suas bases teóricas para a descrição do português real (ao congregar uma equipa de linguistas debruçada sobre os mais ínfimos pormenores), veja-se a secção 4.8. Por outro lado, a colaboração e uso do PALAVRAS em outros projectos, nomeadamente o AC/DC, o COMPARA, o Esfinge[29] e o CorTrad, levou a que fossem sendo enviados ao longo do tempo extensos relatórios de problemas ou de sugestões relativas à análise sintáctica computacional em português.

Saliente-se também que os corpos anotados no âmbito da Floresta e do AC/DC estão acessíveis publicamente (nos casos em que os detentores do material no-lo permitiram), assim como o serviço SketchEngine[30] (Kilgarriff et al., 2005), que produz uma descrição automática das propriedades gramaticais e contextuais das palavras para efeitos lexicográficos, é grátis para o português – e só para o português – porque baseado nos corpos anotados da Linguateca.[31]

Esses corpos anotados deram aliás origem pelo menos a um analisador estatístico público para o português (Wing e Baldrige, 2006).

Outro lado da nossa aposta na anotação gramatical foram as várias tentativas de discutir e/ou de centrar a atenção em muitos aspectos da análise da língua portuguesa ainda pouco explorados, ilustrados por Santos e Gasperin (2002), Afonso (2003), Santos (2004), Afonso (2004) ou Inácio, Santos e Silva (2008).

Refiram-se também as várias acções pedagógicas e de explicação dos vários conceitos envolvidos, que foram realizadas em várias ocasiões (Santos, 2006a; Santos, 2008a) além da constante ajuda aos utilizadores dos vários projectos envolvendo anotação gramatical.[32]

Finalmente, a nossa "Bíblia florestal" (Freitas e Afonso, 2008) não pode deixar de ser referida como um dos trabalhos mais extensos e completos, baseados em texto, criados nos últimos tempos sobre a análise sintáctica do português, e cobrindo, além disso, as duas variantes da língua.

4.7 Avaliação conjunta

Quanto à avaliação conjunta, foi a área em que decididamente houve mais progresso no processamento computacional da língua portuguesa nestes dez anos:

Passámos de uma total ausência e desconhecimento desse paradigma até à implantação forte do modelo em (quase) toda a comunidade, e com o consequente reconhecimento da necessidade e utilidade de novas iniciativas.

Para isso a Linguateca foi absolutamente fundamental, desde a formação e divulgação até à concepção de iniciativas de reconhecido valor internacional e com pressupostos originais e únicos.

Visto que temos um livro expressamente dedi-

[28] Esse "monopólio" não é, contudo, obra da Linguateca: o PALAVRAS tem sido empregue por quase todos os grupos de PLN no Brasil ou Portugal, sem qualquer relação com a nossa actividade.

[29] Neste último caso, o PALAVRAS é usado apenas para a parte da referência anafórica, ver Cabral, Costa e Santos (2007).

[30] http://www.sketchengine.co.uk/

[31] Pelo menos foi essa a combinação feita com Adam Kilgarriff e Eckhard Bick quando nos foi pedida autorização para usar o CETEMPúblico e o CETENFolha. Não me pronuncio aqui sobre novas licenças e/ou formas de aceder a esse serviço que não incluam nem sejam baseadas em material da Linguateca, mas insisto em que a Linguateca não tem quaisquer objecções a que o material por nós criado seja usado por empresas ou para fins comerciais.

[32] Esta é uma actividade que é de certa forma invisível, a não ser para aqueles que a recebem directamente, mas que pode corresponder a uma diferença significativa em termos da utilidade para o exterior dos corpos e recursos disponibilizados. Pensamos que esta característica é especial da Linguateca, e que tal não acontece com a maior parte dos outros recursos ou serviços na rede, embora não tenhamos, naturalmente, dados objectivos para o afirmar.

cado a esse paradigma (e incluindo os participantes nas Morfolimpíadas) (Santos, 2007b), assim como dois outros livros referentes às duas edições do HAREM, Santos e Cardoso (2007) e Mota e Santos (2008), não me nos vou alongar aqui.

Gostava contudo de salientar três traços importantes desta actividade que nem sempre são óbvios para quem está de fora:

- a criação e disponibilização pública de ferramentas e serviços de avaliação (Seco et al., 2006; Gonçalo Oliveira et al., 2008a; Cardoso, 2008a);
- a documentação e reflexão sobre os recursos, também públicos, de avaliação (Santos e Barreiro, 2004; Barreiro e Afonso, 2007; Cardoso e Santos, 2007);
- a congregação de comunidades até aí inexistentes mas que se dedicam a uma mesma tarefa (Santos, 2007a).

Além disso, convém também apontar que o ReRelEM (Freitas et al., 2008; Freitas et al., 2009), a tarefa de detecção de relações entre entidades mencionadas proposta no Segundo HAREM, ao conseguir um cruzamento entre a detecção automática de referência anafórica, tal como por exemplo analisada pelo MUC (Chinchor e Robinson, 1998) ou pelo ARE (Orăsan et al., 2008) e a detecção de relações em texto típica da extracção de informação constitui um desafio original, embora com parecenças com o ACE (NIST e ACE, 2007), que coloca o português entre as línguas que desbravam o processamento da linguagem natural.

4.8 A floresta mais complexa do mundo?

Embora a Floresta Sintá(c)tica não tenha tido o sucesso ou impacto – em termos de utilizadores – que esperaria, penso que foi um projecto inovador e de grande qualidade que possivelmente criou uma das primeiras florestas com informação sintáctica complexa para qualquer língua.

Porque este me parece um caso paradigmático de falta de impacto na comunidade apesar de um esforço considerável para o contrário, refiro que a equipa tentou "tudo" para congregar o máximo de actores à volta dela, senão vejamos: i) apelámos ruidosamente no início do desenvolvimento da Floresta para que fosse um projecto de colaboração entre toda a comunidade, a quem pedíamos para sugerir e prover novos textos e novos analisadores automáticos; ii) temos feito ao longo dos tempos sempre muita divulgação em departamentos de linguística e de computação no Brasil e em Portugal; iii) temos insistido em que se pode obter dados mais simples (tal como sintagmas nominais não complexos) para (avaliar) tarefas que apenas precisem de análise superficial; iv) a Floresta existe numa quase dezena de formatos diferentes "ao gosto do freguês" (Vilela et al., 2005), e com variada informação, semântica, anafórica, de discurso, etc. (criada pelo VISL), (v) finalmente, está integrada em diversos ambientes de processamento internacionais, tal como o NLP toolkit[33], assim como foi usada em avaliações conjuntas internacionais, como o CoNLL.

Muitas das opções tomadas e das ferramentas desenvolvidas no âmbito da Floresta também me parece terem sido originais: Por exemplo, o Picapau (Haber, 2001) está bem à frente dos sistemas desenvolvidos para lidar com florestas, como aliás se vê pela resenha e descrição feita em Lai e Bird (2004), que infelizmente também não menciona o Águia (Santos, 2003b).[34]

Convém reflectir sobre a Floresta Sintá(c)tica e sobre a pertinência da sua criação: O que é certo é que existe um recurso, por enquanto muito pouco explorado, mas que permite uma enorme riqueza de estudos e pesquisas ainda por estabelecer. A que ponto é que tal riqueza seria necessária em 2000 (ou agora)? Deveríamos antes ter começado pelas coisas mais simples? Isto é algo que tem sido bastante discutido pela comunidade que nos cerca.

A minha opinião é que teria sido redutor não tentar ambos os caminhos, apostando assim no servir o máximo de público e de colaboradores interessados, embora não desprezando outras formas de produzir recursos menos ambiciosos. Veja-se uma discussão inicial sobre o assunto em Inácio e Santos (2006), contrastando a revisão do COMPARA com a criação da Floresta. Para outras achegas para o debate em torno da Floresta consulte-se as apresentações de balanço no Encontro "Um Passeio na Floresta Sintáctica", e os novos rumos e interfaces do projecto (Freitas, 2008; Freitas, Rocha e Bick, 2008b).

4.9 Publicar e catalogar em português

Uma das questões mais óbvias que se nos deparou no nosso trabalho interno de todos os dias foi a falta de qualidade dos sistemas de gestão de referências "internacionais" para lidarem com os falantes, e autores, de língua portuguesa, o que levou a que acabássemos por ter de gizar de raiz um sistema para garantir esse (algum) controlo de qualidade, o SUPeRB (Cabral, 2007; Cabral, Santos e Costa, 2008).

Em paralelo, a nossa experiência convenceu-nos

[33] http://www.nltk.org/

[34] É que, como aliás voltarei ao assunto mais à frente, na minha opinião também existe, na comunidade de língua inglesa, o preconceito de que "o que não está ainda feito para o inglês, não existe", mesmo que publicado em inglês.

também de que a actualização manual de um sítio, sem ajuda automática, é muito pouco eficiente e possivelmente condenada ao insucesso (veja-se, por exemplo, a discussão em Pekar e Evans (2007) sobre os catálogos na rede), e que o ideal são sistemas supervisionados em que o processamento automático é depois validado por especialistas: aliás uma opção que nos parece fazer sentido em quase todas as áreas de PLN.

Assim, ao mesmo tempo que tentávamos aplicar a tecnologia e o conhecimento do processamento da nossa língua na nossa actividade quotidiana, nomeadamente na catalogação (das publicações) da área, desenvolvemos um serviço e um sistema que poderia extravasar claramente a área da engenharia da linguagem e ser utilizado por todos os membros da comunidade científica lusofalante, ou seja, um SUPeRBibliotecário desenvolvido de raiz para o português mas com consciência e conhecimento do mundo da publicação em inglês e noutras línguas (por agora, apenas europeias).

Este sistema, além de ser subjacente ao catálogo de publicações da Linguateca (na área), e às variadas páginas de publicações de cada subprojecto (criadas automaticamente), foi usado no desenvolvimento e preparação dos vários livros e artigos desenvolvidos na Linguateca, e encontra-se, quer como serviço, quer como programa em código aberto, acessível publicamente.

4.10 A contribuição das Morfolimpíadas

Parece-me importante retirar do esquecimento as Primeiras Morfolimpíadas para o português, porque, embora não tenha havido seguimento nem aparentemente resultados baseados em estudos sobre os recursos tornados acessíveis, várias coisas ficaram claras:

Por um lado, a existência de fortes divergências teóricas e de diferente importância dada a diferentes fenómenos entre grupos que desenvolveram ou desenvolviam sistemas de análise morfológica.

Por outro lado, uma medição concreta – e extremamente significativa – das diferenças em relação à atomização praticada por cada grupo (Santos, Costa e Rocha, 2003).

Mais uma vez penso que estas medidas foram as primeiras para qualquer língua, embora naturalmente outras medidas e outros problemas tivessem sido privilegiados para o alemão (Hausser, 1996), a língua em que a primeira avaliação conjunta relacionada com morfologia computacional foi levada a cabo. Basta, contudo, reconhecer que esta última língua tem o problema dos compostos para se compreender que outras questões e outras medidas fazem sentido nas duas línguas.

Finalmente, parece-me que também ficou claro que, por ser uma tarefa demasiado teórica, ou seja, dependente de uma separação arbitrária entre níveis ou estratos de língua, muitas das opções ficaram por avaliar, visto que não se encontravam inseridas numa tarefa concreta com resultados consensuais, independentes do modelo teórico.

5 Razões para preocupação

Não gostava contudo de terminar este balanço sem indicar que também houve muita coisa que correu mal, ou que poderia ter corrido melhor. Apresento aqui estes variados pontos para ajudar a fazer não só uma apreciação justa da nossa actividade, como para permitir a outros ou a nós, a começar de novo, não cometer os mesmos erros ou pelo menos ter logo em conta os riscos apontados.

Os quatro primeiros itens têm a ver com a aceitação ou relação da Linguateca com o seu contexto, e podem pois considerar-se do foro sociológico. O quinto ponto refere críticas que nos foram feitas e com que concordo total ou parcialmente, ou que pelo menos considero importante reconhecer a sua existência. Os últimos pontos discutem questões reconhecidamente difíceis mas com cujo tratamento não me considero, de qualquer maneira, totalmente satisfeita.

5.1 Pouco impacto

Atingimos muito poucas pessoas das que poderíamos ter atingido. A grande maioria das pessoas relacionadas com a língua portuguesa ou com a cultura portuguesa nunca ouviu falar da Linguateca. Isso reflecte-se tanto em alunos de doutoramento em Portugal e Brasil como em pesquisadores brasileiros ou portugueses em áreas centrais ou próximas. Ainda agora nos aparecem pessoas que "encontraram o nosso sítio por acaso".

Se isso de certa forma constituiu uma escolha nossa, por termos definido como base de utilizadores (e beneficiários) as pessoas que trabalhavam em ou com o processamento do português (ou seja, a área do PLN, da engenharia da linguagem ou da linguística computacional), e não com a área da língua portuguesa em geral, parece-nos de qualquer maneira que o nosso impacto (e consequente utilidade) deveria ter sido maior.

Da mesma forma, em áreas em que a nossa actividade poderia ter abrangido muito mais gente, como é o caso da publicação científica em geral, e em particular a criação de listas bibliográficas em português ou incluindo correctamente autores de língua materna portuguesa, aparentemente ninguém sabe que fizemos algo que lhes pode ser útil, e que está público. Daí existirem muitos e variados projectos e iniciativas, até de criar bibliografias relacionadas com a área (por exemplo de linguística), que poderiam beneficiar de interacção,

colaboração e troca de dados e das próprias ferramentas desenvolvidas, mas que não utilizam aquilo que oferecemos ou poderíamos oferecer.[35]

Isto demonstra que a colaboração com outras instituições e o reuso de materiais ou trabalho feito por um dado projecto é algo muito mais complexo e exige muito mais atenção do que ingenuamente supusemos.

5.2 Pouco reconhecimento

Uma questão que está relacionada com o pouco impacto e que talvez contribua para ele mesmo é a falta de reconhecimento público aos serviços ou recursos desenvolvidos ou providenciados pela Linguateca.

Penso que não é exagero dizer que mesmo as pessoas que têm bom conhecimento da Linguateca não fazem em geral qualquer esforço para a citar como deve ser, pese embora a nossa continuada insistência em providenciar modelos e até explicitamente indicar como os recursos ou o nosso trabalho devem ser citados. De facto, temos na lista de perguntas já respondidas a informação de como citar cada recurso, assim como muitas vezes na própria página do dito recurso. No entanto, a maior parte das pessoas, se citam, dizem simplesmente "o corpus do Público" (ou "da Folha") ou até os "corpos da Linguateca".

Mesmo as pessoas dentro da Linguateca demonstram o espírito "fora é melhor", porque dá publicação internacional, como se pode ver pela apresentação do Mário J. Silva no encontro que fez um balanço da Linguateca passados dez anos (Silva, 2008b). Segundo ele, o trabalho feito pela Linguateca no CLEF foi muito mais útil e importante que o por exemplo do HAREM, mesmo que a participação de grupos de processamento da língua portuguesa tenha sido mais reduzida[36] e a influência e qualidade do trabalho feito em relação ao português seja incomparavelmente menor[37], dado que a exposição internacional é muito superior no primeiro.

Mas, se esse espírito continua na comunidade do processamento do português, por definição impede que o português atinja a maioridade científica, o que era exactamente uma das intenções da Linguateca: demonstrar que, para o processamento da língua portuguesa, os próprios membros da comunidade que conheciam a língua como sua língua materna eram naturalmente os melhores para essa tarefa.

De facto, a questão do português na comunidade internacional é de alguma forma interessante problematizar: não só considero (Santos, 2007c) bastante pernicioso para o próprio PLN em geral, como disciplina que não haja investigação feita de novo para outras línguas – em particular a nossa – como é muito mais fácil publicar dados empíricos errados ou mal interpretados quando a comissão de programa não percebe a língua. Além disso, convém não esquecer que a maioria dos nossos colegas anglofalantes têm arreigada uma concepção completamente errada, na minha opinião, da área, e que se traduz no seguinte: "todas as inovações começam no inglês", donde a história da área faz-se com base sempre, ou quase sempre, na história da cultura anglo-americana.

No entanto, se os portugueses e brasileiros continuarem sem citar nem mencionar os seus pares na comunidade do processamento do português, e se projectos como a Linguateca não receberem a menção que deveriam ao ter contribuído para o trabalho descrito, está-se a perpetuar essa percepção na comunidade internacional, e na da língua portuguesa.

5.3 Falta de confiança?

Embora a Linguateca tenha dito desde o primeiro dia que queria servir a comunidade, a nossa oferta de disponibilizar os corpos de outras instituições foi recebida com desconfiança (quase) total, e essas instituições foram desenvolver e criar as suas próprias soluções (com o seu próprio financiamento ou com financiamento público), o que teria sido muito mais bem empregue em parceria connosco em vez de contra nós.

Com efeito, nós oferecemo-nos para disponibilizar todos os corpos de português existentes (através do projecto AC/DC). Contudo, muitos projectos para fazer exactamente isso foram iniciados e levados a cabo depois. Dado que nós oferecíamos a tecnologia e o nosso saber-fazer, e muitas dessas instituições até eram académicas e não especialmente interessadas em tecnologia ou disponibilização, é difícil compreender a rejeição, ou ignorância voluntária, dessa oferta.

Outra dessas manifestações é a procura de uma dada ferramenta e/ou serviço, que depois, ao descobrirem que não existe para a língua portuguesa, ou pelo menos não na Linguateca, acaba numa proposta de projecto que, regra geral, não inclui como colaboração ou parceria, ou sequer consultoria, a Linguateca.

Não seria melhor para todos se também se acon-

[35] Veja-se a título de exemplo a Bibliografia Corrente de Linguística do Português, http://dupond.ci.uc.pt/celga/, com apenas dezassete entradas de linguística computacional em Abril de 2009.

[36] Na pista geral do CLEF e no GeoCLEF, em cinco anos e portanto cinco edições participaram apenas quatro grupos diferentes, brasileiros ou portugueses, entre os mais de quarenta. No HAREM participaram vinte em duas edições.

[37] Como pode ser facilmente apreciado, sendo preciso discutir e chegar a consenso com uma miríade de co-organizadores encarregados das outras línguas.

selhassem, ou perguntassem a nossa opinião sobre uma possível colaboração ou participação no desenho dos requisitos, em vez de apenas nos utilizarem como bibliotecários especializados? Mais uma vez, penso que essa forma de proceder não é a melhor para a comunidade como um todo, porque dá prioridade aos interesses específicos de um dado grupo.

Outra possibilidade aventada para explicar este comportamento é a questão do protagonismo. É melhor fazer as coisas sozinho, para receber todos os louros, e o reconhecimento de ser primeiro ou original, do que em colaboração com outros, aliás porque o financiamento é por competição.

De facto, uma das coisas que se tornou mais clara para mim é que muitas pessoas preferem independência a colaboração, e que não são movidas por um desejo de avançar a área como um todo, mas sim de se tornarem os líderes incontestados num determinado nicho ou sub-área.

Será preciso reflectir se esta atitude é saudável ou se é preciso reforçar a interdependência ou, pelo contrário, proceder a uma distribuição de feudos por diferentes actores para estimular o progresso.

De qualquer forma, a única afirmação que é indiscutível é que, mesmo sempre nos apresentando como um serviço, muitos houve que não quiseram partilhar a fama ou os trabalhos connosco.

Outra questão que é preciso mencionar e que é de grande importância tem a ver com o facto de a Linguateca ter sido um projecto iniciado por Portugal e de nunca se ter conseguido (ainda?) pôr de pé os mecanismos formais para criar pólos no Brasil, assim como uma estrutura paralela ou geminada. Isto faz ou fez com que de facto seja muito mais difícil estabelecer projectos comuns com grupos brasileiros e/ou sobretudo obter financiamento para tal.

Ora exactamente para aproveitar o facto de que em português nos entendemos seria essencial promover um apoio, por exemplo, à participação em avaliações conjuntas especialmente promovidas para estimular o progresso do processamento do português, assim como à realização e promoção de fóruns, conferências, encontros, escolas, em português para discutir a língua e o seu processamento.

5.4 Livros difíceis de obter?

Um dos resultados mais fácil de medir objectivamente é a actividade de organização de livros no âmbito da Linguateca: quatro livros distintos sobre a actividade da Linguateca vieram à luz (Santos, 2007b; Santos e Cardoso, 2007; Costa, Santos e Cardoso, 2008; Mota e Santos, 2008).[38]

Mas, além de tal actividade se ter demonstrado muito complexa, tenho fortes dúvidas de que os resultados sejam positivos no cômputo geral: Com efeito, o objectivo de organizarmos nós próprios os livros é podermos ter o controlo total da qualidade, e aliás dos assuntos tratados. No entanto, se esses livros não receberem um canal de publicação apropriado e não forem portanto passíveis da divulgação por nós desejada, não cumprirão o seu objectivo.

Em relação ao primeiro livro, não só se revelou um processo complicadíssimo obter uma saída editorial (atrasando mais de três anos a distribuição do seu conteúdo), como a opção por uma editora comercial impediu a fácil divulgação dos textos. No segundo e terceiro casos, a opção de publicar directamente na rede, embora resultando numa divulgação muito mais rápida, diminuiu claramente o valor científico-comercial do produto, e possivelmente mesmo a sua longevidade.

Neste momento, dado que nenhuma alternativa parece ser realmente satisfatória, ainda nos encontramos num processo de reflexão no que se refere à publicação da quarta obra.

5.5 Críticas variadas

Não posso naturalmente deixar de reconhecer que muitas das críticas que nos foram feitas, aliás por ocasião do balanço dos dez anos, são justas e merecem que as reconheçamos como pontos em que falhámos.

5.5.1 Egocentrismo institucional

Uma das missões da Linguateca era a de catalogar a área, construindo um portal de entrada para tudo o que existisse na rede e pudesse ser útil ao processamento computacional do português.

Contudo, é fácil de ver que o nosso sítio (do qual se apresenta um ecrã na figura 3) está muito mais centrado na nossa actividade do que na da catalogação (Nunes, 2008). Com efeito, ao lado dos catálogos de recursos, ferramentas, actores e publicações, que reflectem ou deviam reflectir a área como um todo, temos muitíssimas outras opções para seduzir o visitante incauto ou interessado, que não vá já com um objectivo determinado.

Em primeiro lugar, damos "Acesso a recursos" da Linguateca primeiro que ao catálogo em geral, "Catálogo de recursos", e iniciamos a lista de opções no menu da esquerda pela pouca modesta apresentação (da Linguateca); depois juntamos, além dos catálogos e de informação interessante, a rubrica "Avaliação conjunta" em que também tivemos um papel fundamental.

Em segundo lugar, os itens "sistemas de procura" e "perguntas já respondidas", que são uti-

[38] Outros livros também organizados parcialmente no âmbito da Linguateca foram Almeida (2003) e Peters et al. (2008).

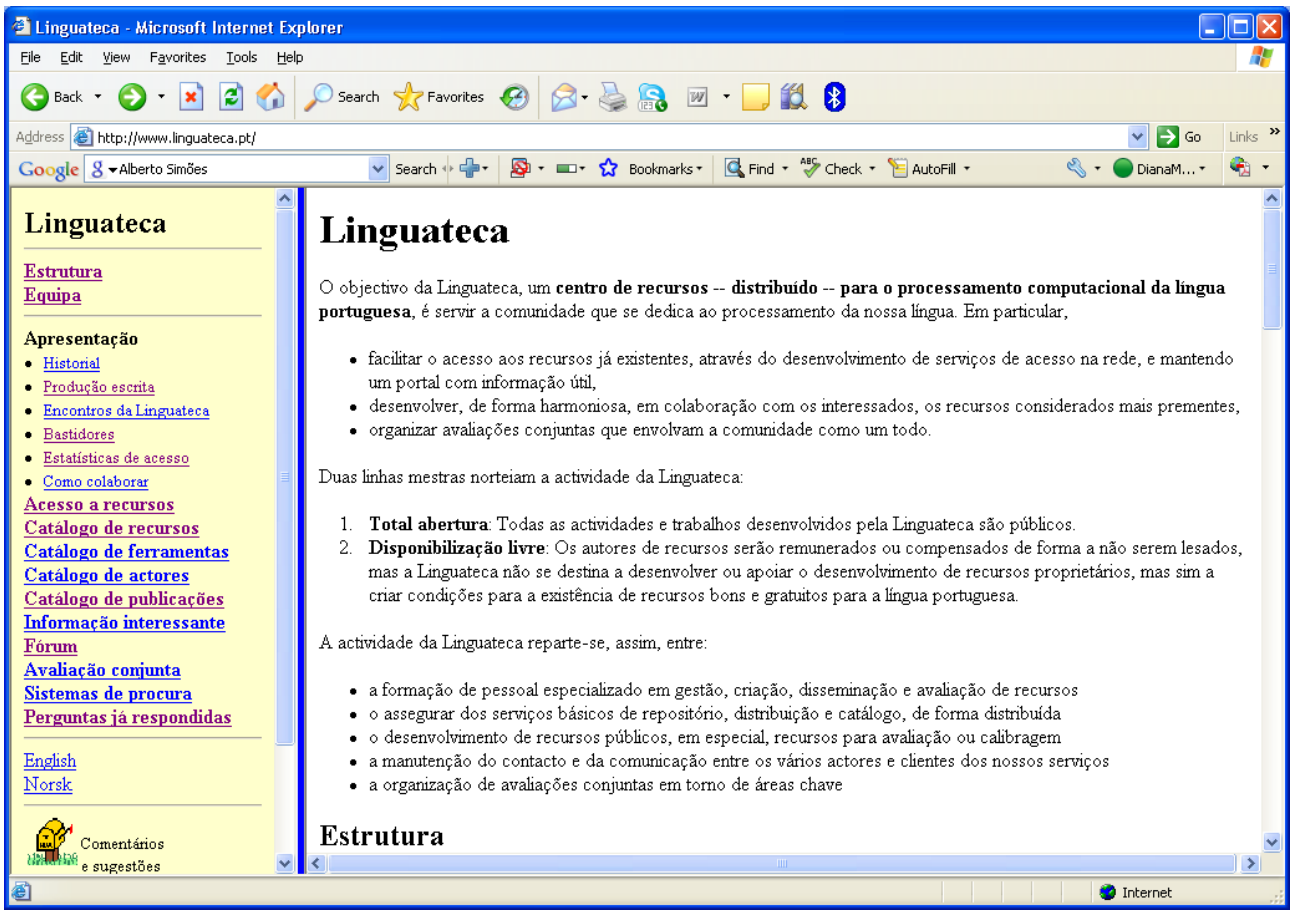


Figura 3: Ecrã da página de entrada da Linguateca

litários associados ao sítio da Linguateca (cujo desenho não é óbvio) pendem claramente para o lado da Linguateca e não da área em geral. Ou seja, as perguntas são exclusivamente sobre a Linguateca e os seus recursos, e os sistemas de procura têm como universo (ou base) todas as páginas apontadas pelo sítio da Linguateca mais as próprias páginas criadas por nós, o que significa, por definição, que incluem muito mais informação sobre a Linguateca do que sobre qualquer outro projecto na área.

Por um lado, isto pode compreender-se dado que é assim que funcionam todos os sistemas de busca locais (quem quer procurar de forma global e não local, usaria os motores gerais), mas, por outro lado, o objectivo de criar um sistema de busca na área, melhor do que os outros para esta área específica, porque informado por mais conhecimento, claramente falhou redondamente. Não por desígnio propositado, mas por o trabalho nessa ferramenta ter sido sempre preterido em relação a outros que pareciam mais urgentes ou que tinham utilizadores mais exigentes.

Provavelmente, este é um caso ovo-galinha clássico: nunca tivemos um sistema suficientemente bom para motivar utilizadores, donde estes nunca puxaram por nós, e por isso o sistema nunca foi desenvolvido como deveria.

Neste caso, a decisão e planeamento de quais as prioridades levou a que esse caminho ficasse atrofiado, muito embora a Linguateca até tenha aberto um pólo no grupo especializado nessa área em Portugal, o XLDB.

Voltando ao ponto de partida, é verdade que o sítio da Linguateca não se conseguiu impor como um catálogo actualizado, dinâmico e interessante para a área. Pelo contrário, a grande maioria dos nossos visitantes foram utilizadores dos recursos que criámos ou participantes nas actividades que organizámos.

Talvez também associado a esta questão, raríssimos foram os membros da comunidade que nos contactaram para incluirmos os seus recursos ou projectos no nosso sítio.

5.5.2 Falta de directivas

Embora tenhamos ganho muita experiência ao fazer e organizar avaliações conjuntas, medições de área e panorâmicas, não propagámos suficientemente (ou nada) como é que isso se deve fazer, como referido por Ferreira e Teixeira (2008).

Tal neste caso foi inocentemente motivado por imaginarmos que a Linguateca seria sempre o núcleo dessa organização, que grupos individuais

não se sentissem com motivação para levar aos ombros esse tipo de tarefa. Mas fica a chamada de atenção de que seria interessante tentar ensinar como fazer – refira-se que em Ferreira et al. (2009) já os mesmos autores demonstram a vantagem de o fazer no domínio da medicina.

5.5.3 Falta de ligação à comunidade empresarial

Outra crítica que nos foi feita, de formas variadas, foi que a Linguateca não olhou especialmente nem dedicou nenhuma vertente aos actores comerciais: assim, não só não nos preocupámos em ganhar dinheiro nem ajudar outros que connosco colaborassem a ganhá-lo, ou que quisessem colaborar connosco se nós os ajudássemos a ganhar dinheiro.

Embora eu não tenha a certeza de que concorde que isto deva ser visto como crítica – e de facto o testemunho de Braga e Dias (2008) pareça indicar que fomos, seja como for, úteis para algumas empresas, reconheço que é profundamente verdade.

Nós não dedicámos atenção diferente a nenhum tipo de actor e assumimos que a nossa actividade seria benéfica para todos por igual. Esta questão merece ser equacionada à luz destas críticas ou observações:

Seria aceitável ou (mais) produtivo se alguma actividade da Linguateca fosse dirigida (e mesmo paga) por actores comerciais, como aventado por Daniela Braga no encontro em Aveiro?

Seria natural transformar a Linguateca numa incubadora de empresas cujo objectivo seria rentabilizar e disseminar recursos públicos, como proposto por Anabela Barreiro no mesmo encontro?

Ficam as perguntas, e o repto de que esses modelos teriam de ser propostos e equacionados também por esses mesmos actores.

Aliás, e dada a (na minha opinião, triste) conversão progressiva das próprias universidades em máquinas de ganhar dinheiro, esta questão pode ser expandida a todos os modelos de colaboração com instituições no futuro.

O que não me parece fazer sentido, é propor que a Linguateca seja ela transformada numa actividade lucrativa.

5.6 Ferramentas em código aberto

Voltando a carregar na tecla "Casa de ferreiro, espeto de pau", o facto de o primeiro pólo da Linguateca em Portugal, o de Braga, ser especialista em código aberto e na disponibilização desse tipo de ferramentas não foi suficiente para conseguir que a Linguateca tivesse uma actividade consequente, profissional e de impacto profundo, quer na dita comunidade, quer em geral.

Com efeito, embora todo o código que tenhamos criado tenha vindo, melhor ou pior, a ser disponibilizado publicamente (o que não significa que tenha sido usado ou disseminado como deve ser), toda a cultura de desenvolvimento de código aberto não foi aproveitada, nem nós aproveitámos as possibilidades que teríamos de teste aos programas pela comunidade.

Por um lado, isso deveu-se ou deve-se à grande quantidade de linguagens de programação e ambientes usados, donde qualquer opção ou escolha nossa iria apenas satisfazer (ou melhor, apenas satisfez) um fragmento ou fracção da comunidade.[39]

Por outro lado, tivemos muitas vezes a impressão de que a maioria dos membros da comunidade preferiam obter programas a funcionar (e nesse caso como serviços na rede) do que estar a programar ou mexer em código de outrem. Os verdadeiros programadores, por outro lado, não abdicavam de programar tudo outra vez (de raiz) e estavam mais interessados em recursos ou ideias.

De qualquer maneira, temos de dar a mão à palmatória e confirmar que não conseguimos, nestes dez anos de actividade, produzir sistemas computacionais que fossem usados e manipulados por uma faixa grande de membros da nossa comunidade. Conseguimos isso em relação aos recursos, mas não a programas informáticos.

Embora também o NLP registry[40] seja um caso desses que parece não ter conseguido descolar[41], e que a maior parte dos programas de código aberto, mesmo no SourceForge, não têm sucesso (Feitelson, Heller e Schach, 2006), nós estamos claramente conscientes de que nos faltou uma estratégia nesse aspecto, assim como uma actividade de produção e manutenção dos sistemas já disponibilizados.[42] De facto, tal questão já tinha sido abordada criticamente em Santos (2000), mas não foi por isso resolvida.

Alguns exemplos de má prática:

O atomizador da Linguateca foi distribuído como um módulo do PLNbase pelo Alberto Simões, a cavalo noutro atomizador por ele desenvolvido (mas sem qualquer informação sobre as diferenças entre os dois). A primeira edição do atomizador e separador de frases foi publicada em 2004; desde essa altura e embora na Linguateca problemas pontuais e pequenas melhorias tenham

[39] A título anedótico, refira-se que, só dentro do âmbito da Linguateca, têm sido desenvolvidos e tornados públicos programas nas seguintes e diversas linguagens de programação: Perl, Java, PHP, C, R, Lisp, awk, Groovy e JavaScript.

[40] http://registry.dfki.de/

[41] Embora já na sua quarta versão, contém pouquíssimas entradas, e em muitas delas a informação sobre disponibilidade é simplesmente: "to negotiate".

[42] Tanto o catálogo de ferramentas, como o Jardim de Ferramentas, nunca tiveram de facto cobertura, publicidade e atenção suficientes para se tornarem eles próprios ferramentas úteis.

continuado a ser efectuadas, tal nunca (até agora) foi reflectido na versão pública.[43]

O Corpógrafo foi disponibilizado em código aberto antes de ser instalado em Barcelona,[44] mas o código ainda estava cheio de problemas e de questões não resolvidas, e só em fins de 2008 uma nova versão mais estável foi colocada ao dispor da comunidade. Este exemplo demonstra o que é bem sabido por todos os produtores comerciais: às vezes é preciso publicar ou pôr nas bancas um produto por razões que não são a de estar perfeito ou acabado. No nosso caso, foi para garantir que o produto seria tratado como código aberto pela instituição na qual foi instalado.

O código do Esfinge também foi disponibilizado desde 2006, veja-se Costa (2007), mas sem a garantia que as novas versões deste sistema, pioneiro para a língua portuguesa, estivessem logo acessíveis para a comunidade. Como só as pessoas que desenvolvem programas podem saber, não é trivial a documentação e manutenção de sistemas que evoluem ao longo de anos de trabalho, e existe sempre uma diferença entre uma versão estável e documentada e o programa do momento.

Finalmente, a questão da disponibilização de sistemas complexos ainda provoca mais dificuldade devido à questão das dependências: não faz sentido começar a fazer tudo do nada, mas, se se inclui outros sistemas, como seria natural e boa prática, obriga-se o utilizador incauto a instalar e ter de levar em conta muitos outros programas desenvolvidos por terceiros e que podem eles próprios ser difíceis de instalar ou compreender.

5.7 Documentação – a sempre vilipendiada

Há duas leis na informática: a de que a documentação é essencial, e a de que a documentação nunca está actualizada. Todos os projectos lutam com estas duas leis, e embora no caso da Linguateca tenhamos feito um esforço não irrisório de boa doumentação, não conseguimos também escapar à segunda lei, de que ainda falta documentar ou melhorar muita coisa.

Ao contrário do que certas pessoas pregam, de que um programa ou sistema bom ou bem desenhado não precisa de explicação ou documentação, tal parece-me completamente errado no caso da área do processamento de uma língua. Não vou pois argumentar em geral, mas apenas no domínio em que trabalhamos.

Dando alguns exemplos concretos:

- qual a utilidade de saber quantos substantivos ou adjectivos há num texto, sem saber quais os critérios de classificação de uma e outra categoria?
- qual a utilidade de saber quais as palavras mais frequentes, ou a frequência de um conjunto de palavras, sem se saber qual a base (os textos) usada para essas contagens?
- que vantagem tem um sistema que anota um texto, sem que se saiba os critérios de anotação usados?

Ou: como é que se pode avaliar um dado sistema se não se consegue interpretar a sua saída? Como é que se pode usar um sistema para fazer uma coisa quando foi desenhado para outra?

Em todos os casos de trabalho sério, é preciso saber como é que cada tarefa ínfima é feita – ou ter a possibilidade de o saber. Sem isso, estamos no reino da "banha da cobra", e não estamos a criar recursos ou ferramentas que possam contribuir para o progresso e que possam ser melhorados por outros. Estamos apenas a tentar vender, no sentido de convencer a usar, um produto de forma irresponsável.

Este aspecto da documentação e da explicação de como é que os recursos foram criados, e quais os pressupostos envolvidos na sua criação, é uma das tónicas mais importantes postas pela Linguateca no seu trabalho.

Outra questão – menos crítica – é a remoção de assuntos ou páginas claramente desactualizadas ou irrelevantes, que tendem a ficar perdidas ou penduradas num sítio da rede em vez de activamente limpas ou reescritas pelos gestores do sítio. Embora isto faça parte do manual dos gestores de sítios, é preciso reconhecer ou relembrar que as principais capacidades da Linguateca não são a de gestão profissional de sítios. Apenas muito recentemente, há menos de um ano, passámos a gerir uma parte (ínfima) das nossas actividades em wiki, como se pode ver em relação à página do GikiCLEF. Tal deveu-se, mais uma vez, a não haver pessoal com apetência especial para manutenção de sítios e ao facto de termos já uma quantidade de programas e rotinas desenhadas para gerir o sítio da Linguateca, e que reconvertê-las levaria a muito trabalho – que seria afinal só cosmético.

Assim, embora a documentação e a apresentação sejam de certa forma acessórias ao verdadeiro trabalho da Linguateca, são requisitos necessários para que este seja compreendido e usado. Sistemas ou serviços sem documentação, são completamente inúteis – ou até perigosos, se induzirem

[43] A reforçar o já dito anteriormente sobre as linguagens de programação, uma total reescrita do mesmo atomizador noutra linguagem foi recentemente disponibilizada por Nuno Cardoso no âmbito do seu sistema REMBRANDT (Cardoso, 2008c).

[44] No âmbito da colaboração entre o CLUP/Linguateca e o grupo de Teresa Cabré no Institut Universitari de Lingüística Aplicada (IULA) na Universitat Pompeu Fabra.

as pessoas em erro.

Mas sistemas e serviços que devido à sua má apresentação assustam ou repelem os utilizadores a quem foram destinados também constituem um entrave sério ao impacto da Linguateca e à nossa possibilidade de sermos úteis à comunidade.

5.8 A usabilidade e preocupação com os utilizadores

De facto, uma outra área que é preciso mencionar, é a usabilidade, ou seja, a preocupação da Linguateca com os utilizadores dos vários programas que desenvolvemos, avaliamos ou estudamos. Pese embora a nossa consciencialização sobre o assunto, e uma tentativa de actuação variada, o cômputo geral parece mais negativo do que positivo.

Esta preocupação pode apreciar-se em vários ramos diferentes da nossa intervenção na área do processamento da língua:

Por um lado, refira-se o estudo sério de necessidades de informação como preliminar para o desenvolvimento posterior do sistema de recolha de informaçao na rede de Rachel Aires (Aires e Aluísio, 2003), que aliás fez girar toda a problemática da sua tese à volta da formalização e detecção das necessidades do utilizador, e efectuou testes com utilizadores para avaliar o sistema implementado.

Por outro, tivemos sempre uma atitude muito crítica em relação à forma como algumas tarefas foram definidas no CLEF, pondo-nos no lugar de utilizadores de língua portuguesa, ou de simples pessoas interessadas em recolha de informação cruzada (Santos e Rocha, 2005; Santos e Cardoso, 2005). Em muitas ocasiões, fomos de certa forma os primeiros a gritar que "o rei vai nu": muitas das hipóteses tomadas como óbvias num ambiente anglofalante caem pela base ao considerar outras línguas, no nosso caso o português.

Como já mencionado, fomos dos primeiros a nível internacional a levar a cabo, e a publicar, dados sobre utilizadores de um serviço de corpos, o COMPARA (Santos e Frankenberg-Garcia, 2007), em que explicitamente aplicamos métodos de investigação não-obstrusiva da actividade dos utilizadores aos diários de interacção com o serviço.

Fomos também dos primeiros a executar estudos dos diários de procura na rede com base no instantâneo da rede portuguesa WPT03 para efeitos de processamento da língua ou recolha de informação (Seco e Cardoso, 2006).

Finalmente, a um nível completamente diferente, implementámos um serviço cooperativo de resposta aos utilizadores de forma a dar sempre resposta às mais variadas questões, como mencionado na secção anterior.

Contudo, a aparência dos nossos serviços e informação na rede foi sempre o nosso calcanhar de Aquiles e, nas palavras críticas de um dos leitores do presente artigo:

> É uma imagem que me transporta para meados dos anos 90. (...) qualquer utilizador banal vai pensar que o site não é actualizado há anos e que não vai encontrar lá nada de útil. Transmite a ideia de site criado por amadores, sem conhecimentos de informática.

Numa altura em que todas as empresas, pelo menos as associadas a meios de comunicação social ou editorial, aplicam rotineiramente análise de diários e de comportamento de utilizadores para melhorar a sua presença na rede, a Linguateca, embora possivelmente à frente na comunidade científica do processamento da língua, está muito atrás da realidade da vida de todos os dias.

5.9 Publicação em nome da Linguateca

Embora a Linguateca possa apregoar um grande número de publicações e apresentações produzidos ao longo destes dez ou onze anos – trezentas a quatrocentas, não podemos infelizmente garantir ou confirmar que todos os textos publicados com a chancela da Linguateca tenham sido verificados em termos de qualidade ou mesmo de oportunidade.

A existência de cerca de trinta colaboradores ao longo do tempo e o facto de as publicações não estarem prontas na maior parte das vezes a tempo suficiente antes da data final de entrega levou a uma publicação muito descentralizada e que não usufruiu, na maior parte dos casos, das vantagens que poderia colher ao ser redigida no seio de um equipa de peritos.

Isso, aliás, é claramente patente na ausência, na maior parte dos artigos, de agradecimentos a revisão cruzada de outros elementos da Linguateca. Não dizendo que isto é um problema específico da nossa equipa, falhou claramente, na maior parte dos casos, também entre nós a possibilidade de retorno e de discussão científica séria antes da publicação.

Idealmente, deveríamos ter definido normas mais concretas tanto quanto à divulgação da Linguateca em geral como ao posicionamento do trabalho relatado no plano geral da nossa actividade, assim como deveríamos ter estipulado um certo conjunto de normas de qualidade, empíricas, a que os artigos da Linguateca como Linguateca deviam obedecer, e que em alguns casos teriam levado a uma reescrita ou à não publicação do artigo como trabalho realizado no âmbito da Linguateca. Se viermos a continuar como instituição virtual, parece-me que isto tem de ser decididamente contemplado no futuro, até porque teria sido uma forma relati-

vamente fácil de obter maior impacto.

Que é possível empenhar a equipa – e mesmo elementos de fora da Linguaetca mas que possam rever-se como pertencendo ao círculo da mesma – foi patente em relação ao presente texto, o qual foi extraordinariamente melhorado devido ao excelente retorno e problematização de várias afirmações e opiniões patentes em versões anteriores, por mais de uma dezena de leitores interessados.

6 A saúde do processamento computacional do português

Embora este artigo seja sobre a Linguateca, não posso deixar de chamar aqui a atenção sobre outras vitórias nesta área durante o período coberto por esta reflexão, completamente independentes da nossa acção. Não gostava de forma nenhuma de parecer estar a afirmar que, sem nós, nada teria acontecido, ou que, excepto nós, ninguém fez nada.

Assim, gostava de salientar – sem quaisquer pretensões de exaustividade, visto que tal assunto poderia e deveria constituir um artigo novo – alguns acontecimentos ou sistemas que me parece fazerem a diferença, ou seja, serem vitórias incontornáveis do português no campo internacional:

- o primeiro detector automático de metáforas foi desenvolvido para o português – e depois aplicado ao inglês – por Tony Berber Sardinha (Berber Sardinha, 2006; Berber Sardinha, 2007);
- o primeiro sistema automático para produção de livros auditivos foi criado por uma parceria entre o INESC e a FCUL (Serralheiro et al., 2003);
- o primeiro serviço automático com classificação semântica foi feito no VISL para o português (Bick, 2006; Bick, 2007)[45];
- o primeiro motor de procura sobre a rede completa de um país foi efectuado pela equipa do tumba! (Gomes e Silva, 2005);
- a primeira legendagem automática de telejornais para deficientes auditivos foi realizada pelo projecto Tecnovoz (Meinedo, Viveiros e Neto, 2008);
- a primeira geração de fala para fórmulas matemáticas ou equações foi descrita em Rolo e Serralheiro (2008).

[45] É preciso notar que embora Eckhard Bick tenha uma relação estreita com a Linguateca, a grande maioria dos rtrabalhos efectuados pelo projecto VISL são completamente independentes desta. O que também se aplica ao grupo do XLDB ou outros que sejam mencionados nesta secção.

Mesmo quando não estamos a falar de primeiros para qualquer língua, não queremos deixar de chamar a atenção, que, para o português, houve naturalmente muitíssimos "primeiros" sem qualquer relação com a Linguateca.

Por exemplo, os três seguintes sistemas ou recursos nasceram no NILC:

- o primeiro sistema de sumarização automática para o português (Pardo e Rino, 2002);
- a primeira ontologia lexical para o português inspirada pelo método da WordNet (Oliveira, Dias da Silva e Moraes, 2002);
- o primeiro detector da estrutura retórica de um texto para o português (Pardo, Nunes e Rino, 2004).

E outros primeiros foram:

- o primeiro sistema de RAP em português baseado em análise sintáctica, pelo VISL (Bick, 2003);
- o primeiro sistema completo de síntese de base articulatória suportada em estudos de produção para o português, pelo IETA em Aveiro (Oliveira, 2009);
- o primeiro sistema de desenvolvimento de ontologias a partir de texto pela PUC-RS (Gasperin, 2001);
- o primeiro modelo cognitivo quantitativo para o estudo da evolução diacrónica de variedades do português (Silva, 2008a).

Tal é sinal evidente de que o processamento do português tem boas pernas para andar. Penso que – de preferência com a colaboração de todos – poderemos ir longe na investigação e desenvolvimento de sistemas computacionais que lidem perfeitamente com a nossa língua.

7 Comentários finais

Neste artigo, comecei por comparar as intenções iniciais e o ponto de situação efectuado no começo da actividade da Linguateca, como um exercício salutar de avaliação, dez anos passados. Apresentei brevemente a história da Linguateca, depois salientei sucintamente as actividades ou áreas de intervenção em que penso que a Linguateca foi útil para a comunidade do processamento do português e nem só, passando a indicar os problemas ou áreas em relação aos quais a Linguateca não conseguiu, na minha opinião, dar um contributo suficientemente positivo.

Tentei mostrar que ao longo da nossa história muito de bom aconteceu, apresentando alguns casos de maturidade e de inovação na área. Também

considero, contudo, que muito mais podia ter sido feito se tivesse havido confiança na Linguateca e um espírito de colaboração entre os vários grupos ou instituições dedicados à área, especialmente em Portugal. Espírito esse que foi apanágio de muito dos nossos colegas brasileiros, que cooperaram, produziram recursos para o repositório, e aproveitaram (como nós queríamos) o nosso trabalho, e a quem estou particularmente grata por isso.

Se pudesse começar de novo, e mais uma vez esta é uma visão muito pessoal, continuaria a organizar avaliações conjuntas e a criar recursos de avaliação em conjunto com membros da comunidade, mas não tentaria catalogar a área ou observá-la, tentando fixá-la num sítio megalómano. Pelo contrário, tentaria que todos discutissem e comunicassem através de listas de discussão e da troca de ideias e, claro, da participação em avaliações conjuntas.

Assim como temos um serviço de resposta a todas as perguntas que nos fazem (mas que são limitadas e muitas vezes fora do contexto da própria Linguateca), tentaria fazer com que essas perguntas fossem feitas e respondidas num verdadeiro fórum de todos os interessados na área (como acontece por exemplo na lista `corpora`), permitindo a interacção, o conhecimento dos intervenientes, e uma resposta cooperativa que ajuda a quem perguntou mas também aos outros que estão a ouvir porque fazem parte da comunidade.

Tentaria também oferecer a Linguateca como um serviço de avaliação no sentido de podermos ajudar a criar materiais de teste ou mesmo métricas para avaliar trabalhos ou sistemas de empresas ou académicos, devido à nossa experiência no assunto.

Finalmente, se fosse a continuação da Linguateca que estava em jogo, e nos fossem concedidos mais dez anos, seria essencial focar-nos em projectos com impacto nacional ou internacional (em língua portuguesa, claro), tal como o Museu da Pessoa, a procura inteligente nas obras da(s) Biblioteca(s) Nacional(is), a procura na rede, o arquivo da rede portuguesa e brasileira, e sistemas de tradução automática com respeito pelo português, não descurando, também, toda a parte cultural e multimodal associada à procura em imagens, vídeo e sons, e em meios mistos.

É minha convicção de que uma Linguateca futura teria de ter uma componente prática muito maior envolvendo empresas e instituições, e o seu fito deveria ser aplicar a tecnologia existente à realidade de todos os dias.

Não faz sentido a continuação da Linguateca como é agora, apenas com parceiros académicos e com impacto na comunidade científica: a Linguateca para merecer sobreviver e poder continuar a ser útil, terá de se "praticalizar", ou seja, tomar em mãos aspectos e projectos claramente práticos.

Agradecimentos

Este artigo foi escrito no âmbito da Linguateca, contrato número 339/1.3/C/NAC, financiado pelo governo português e pela União Europeia.

A existência da Linguateca deve-se, em primeiro lugar, ao interesse do então ministro da Ciência e da Tecnologia, José Mariano Gago, pela questão da língua, que levou à inclusão deste assunto no Livro Verde e depois no Livro Branco, e, em segundo lugar, ao apoio constante, institucional e pessoal, do presidente da FCCN[46], Pedro Veiga.

Agradeço a todos os membros da Linguateca, a todas as pessoas que colaboraram com a Linguateca, a todos os que contribuíram, com as suas perguntas, pedidos ou sugestões, para a melhoria do nosso projecto, e finalmente a todos os que comentaram, criticaram e enriqueceram o presente texto.

Referências

Afonso, Susana. 2003. Clara e sucintamente: um estudo em corpus sobre a coordenação de advérbios em -mente. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 27–36, Lisboa, 2-4 de Outubro, 2003. APL.

Afonso, Susana. 2004. Estudo dos argumentos verbais e ambiguidade dos sintagmas preposicionais através do Águia. Relatório técnico, Linguateca, 21 de Abril, 2004. `http://www.linguateca.pt/documentos/ArgumentosambiguidadeAfonso2004.pdf`.

Afonso, Susana, Eckhard Bick, Renato Haber, e Diana Santos. 2001. Floresta sintá(c)tica: um treebank para o português. Em Anabela Gonçalves e Clara Nunes Correia, editores, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*, pp. 533–545, Lisboa, Portugal, 2-4 de Outubro, 2001. APL.

Aires, Rachel e Diana Santos. 2002. Measuring the Web in Portuguese. Em Brian Matthews, Bob Hopgood, e Michael Wilson, editores, *Euroweb 2002 conference*. pp. 198–199, 17-18 Dezembro, 2002.

Aires, Rachel Virgínia Xavier. 2005. *Uso de marcadores estilísticos para a busca na Web em por-*

[46] A FCCN é a instituição portuguesa que, em termos jurídicos, é "executora" do projecto Linguateca desde 2000.

tuguês. Tese de doutoramento, ICMC - USP - São Carlos, Agosto, 2005.

Aires, Rachel Virgínia Xavier e Sandra Maria Aluísio. 2003. Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em português. *Revista Ciência da Informação*, 32(1):5–16.

Almeida, José João, editor. 2003. *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)*. Universidade do Minho, Braga.

Almeida, José João e Alberto Simões. 2007. XML::TMX - Processamento de Memórias de Tradução de Grandes Dimensões. Em José Carlos Ramalho, João Correia Lopes, e Luís Carriço, editores, *XML: Aplicações e Tecnologias Associadas (XATA2007)*, pp. 83–93. Universidade do Minho, 15-16 de Fevereiro, 2007.

Almeida, José João, Alberto Manuel Simões, e José Alves Castro. 2002. Grabbing parallel corpora from the web. *Sociedade Española para el Procesamiento del Lenguaje Natural*, 29:13–20.

Aluisio, Sandra, Gisele Montilha Pinheiro, Aline M. P. Manfrin, Leandro H. M. de Oliveira, Luiz C. Genoves Jr., e Stella E. O. Tagnin. 2004. The Lácio-Web: Corpora and tools to advance Brazilian Portuguese language investigations and computational linguistic tools. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 1779–1782, 26-28 de Maio, 2004.

Aluísio, Sandra Maria, Leandro H.M. de Oliveira, e Gisele Montilha Pinheiro. 2004. Os tipos de anotações, a codificação, e as interfaces do Projeto Lácio-Web: Quão longe estamos dos padrões internacionais para córpus? Em *II Anais do TIL - Workshop de Tecnologia da Informação e Linguagem Humana*, pp. 1–10, 5 a 6 de Agosto, 2004.

Amaral, Carlos, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto, e Tiago Veiga. 2008. Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Aranha, Christian Nunes. 2007. O Cortex e a sua participação no HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 113–122.

Bacelar do Nascimento, Maria Fernanda, Amália Mendes, e Luísa Pereira. 2004. Providing online access to portuguese language resources: corpora & lexicons. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 1825–1828, 26-28 de Maio, 2004.

Barreiro, Anabela. 2008. ParaMT: a Paraphraser for Machine Translation. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190. Springer Verlag, pp. 202–211, 8-10 de Setembro, 2008.

Barreiro, Anabela e Susana Afonso. 2007. Construção da lista dourada para as primeiras Morfolimpíadas do português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 107–118.

Barreiro, Anabela e Elisabete Ranchhod. 2005. Machine Translation Challenges for Portuguese. *Linguisticæ Investigationes*, 28(1):3–18.

Berber Sardinha, Tony. 2006. An online program for tagging metaphors in corpora. Em S. Zyngier, V. Viana, e A. M. Spallanzani, editores, *Linguagens e Tecnologias: Estudos Empíricos*, pp. 165–182, Rio de Janeiro, Brasil. Editora da UFRJ.

Berber Sardinha, Tony. 2007. *Metáfora*. Parabola, São Paulo, Brasil.

Berber Sardinha, Tony, J. L. Moreira Filho, e E. Alambert. 2008. O corpus brasileiro. Comunicação ao VII Encontro de Lingüística de Corpus, 2008, UNESP, São José do Rio Preto, SP, Brasil.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento, Aarhus University, Aarhus, Denmark, Novembro, 2000.

Bick, Eckhard. 2003. A Constraint Grammar Based Question-Answering System for Portuguese. Em Fernando Moura Pires e Salvador Abreu, editores, *Progress in Artificial Intelligence: 11th Portuguese Conference on Artificial Intelligence, EPIA 2003. Beja, Portugal, December 2003, Proceedings*, pp. 414–418, Berlin/Heidelberg. Springer.

Bick, Eckhard. 2006. Noun sense tagging: Semantic prototype annotation of a portuguese treebank. Em Jan Hajic e Joakim Nivre, editores, *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*, 1-2 de Dezembro, 2006.

Bick, Eckhard. 2007. Automatic semantic role annotation for portuguese. Em *TIL, V Workshop em Tecnologia da Informação e da Linguagem Humana*, pp. 1715–1719, 30 de Junho a 6 de Julho, 2007.

Bick, Eckhard, Diana Santos, Susana Afonso, e Rachel Marchi. 2007. Floresta Sintá (c)tica: Ficção ou realidade? Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 291–300.

Braga, Daniela e Miguel Sales Dias. 2008. Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de fala na Microsoft. Em Luís Costa, Diana Santos, e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, pp. 29–33.

Cabral, Luís Miguel. 2007. SUPeRB - Sistema Uniformizado de Pesquisa de Referências Bibliográficas. Tese de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Março, 2007.

Cabral, Luís Miguel, Luís Fernando Costa, e Diana Santos. 2007. Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach. Em Alessandro Nardi e Carol Peters, editores, *Working Notes for the CLEF 2007 Workshop (CLEF 2007)*, pp. s/pp, 19-21 de Setembro, 2007.

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.

Calado, Pável. 1999. The WBR-99 Collection: Description of the WBR-99 Web collection data-structures and file formats. Relatório técnico, LATIN - Laboratório para o Tratamento de Informação, Departamento de Computação, Universidade Federal de Minas Gerais. `http://www.linguateca.pt/Repositorio/WBR-99/wbr99.pdf`.

Cardoso, Nuno. 2008a. Apêndice H: SAHARA - Serviço de Avaliação HAREM Automático. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Cardoso, Nuno. 2008b. Novos rumos para a recuperação de informação geográfica em português. Em Luís Costa, Diana Santos, e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*. Linguateca, pp. 71–85.

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

Cardoso, Nuno, Bruno Martins, Daniel Gomes, e Mário J. Silva. 2007. WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 279–288.

Cardoso, Nuno e Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 211–238.

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, e Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Chaves, Marcirio, Catarina Rodrigues, e Mário J. Silva. 2007. Data Model for Geographic Ontologies Generation. Em José Carlos Ramalho, João Correia Lopes, e Luís Carriço, editores, *XML: Aplicações e Tecnologias Associadas (XATA2007)*, pp. 47–58. Universidade do Minho, 15-16 de Fevereiro, 2007.

Chaves, Marcirio Silveira. 2008. *Uma Metodologia para Construção de Geo-Ontologias*. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa, Dezembro, 2008.

Chinchor, Nancy e P. Robinson. 1998. MUC-7 Named Entity Task Definition (version 3.5). Em *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, EUA.

Chubin, Daryl E. e Edward J. Hackett. 1990. *Peerless Science: Peer Review and U.S. Science*

Policy. State University of New York Press, Nova Iorque, EUA.

Costa, Luís. 2005. Esfinge - Resposta a perguntas usando a Rede. Em José María Gutiérrez, Flavia Maria Santoro, e Pedro Isaías, editores, *Proceedings da conferência IADIS Ibero-Americana WWW/Internet 2005*, pp. 616–619. IADIS Press, 18-19 de Outubro, 2005.

Costa, Luís. 2006. Esfinge - A Question Answering System in the Web using the Web. Em *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pp. 127–130, 3-7 de Abril, 2006.

Costa, Luís. 2007. Question answering beyond CLEF document collections. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multimodal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, pp. 405–414.

Costa, Luís. 2008. Resumo da actividade da Linguateca de 16 de Dezembro de 2006 a 31 de Dezembro de 2008. Relatório técnico, Linguateca, Dezembro, 2008. Com a colaboração (por ordem alfabética) de Ana Frankenberg-Garcia, Anabela Barreiro, Cláudia Freitas, Cristina Mota, David Cruz, Diana Santos, Hugo Oliveira, Luís Cabral, Nuno Cardoso, Paula Carvalho Paulo Rocha, Sérgio Matos, `http://www.linguateca.pt/documentos/RelatorioLinguateca20072008.pdf`.

Costa, Luís e Luís Miguel Cabral. 2008. Medindo a Linguateca, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprCostaCabralL10.pdf`.

Costa, Luís, Diana Santos, e Nuno Cardoso, editores. 2008. *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca.

Davies, Mark e Ana Maria Raposo Preto-Bay. 2008. The Corpus do Português and the Routledge frequency dictionary of Portuguese: New tools for learners and teachers. Em *TaLC 8 Lisbon: Proceedings of 8th Teaching and Language Corpora Conference (3-6 July 2008)*. Associação de Estudos e de Investigação Científica do ISLA - Lisboa, pp. 96–99.

Feitelson, Dror G., Gillian Z. Heller, e Stephen R. Schach. 2006. An empirically-based criterion for determining the success of an open-source project. Em *Australian Software Engineering Conference*, pp. 363–368, Abril, 2006.

Fernandes, Eraldo R., Ruy L. Milidiu, e Cicero N. Santos. 2009. Portuguese language processing service. Em *18th International World Wide Web Conference*, 20-24 de Abril. 2009.

Ferreira, Liliana, Cesar Telmo Oliveira, António Teixeira, e João Paulo Silva Cunha. 2009. Extracção de informação de relatórios médicos. *Linguamática*, 1, Maio, 2009.

Ferreira, Liliana e António Teixeira. 2008. Linguateca e Processamento de Linguagem Natural na Área da Saúde: Alguns Comentários e Sugestões. Em Luís Costa, Diana Santos, e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, pp. 43–48, 11 de Setembro, 2008.

Forner, Pamela, Anselmo Peñas, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, e Erik Tjong Kim Sang. 2009. Overview of the CLEF 2008 Multilingual Question Answering Track. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer.

Frankenberg-Garcia, Ana e Diana Santos. 2002. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, IX(1):61–79.

Freitas, Cláudia. 2008. A Floresta Sintáctica no Ensino de Português, 3 de Julho, 2008. `http://www.linguateca.pt/documentos/FreitasWorkshopTaLC2008.pdf`.

Freitas, Cláudia e Susana Afonso. 2008. Bíblia Florestal: Um manual lingüístico da Floresta Sintá (c)tica. `http://linguateca.dei.uc.pt/Floresta/BibliaFlorestal/`.

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2008a. Um mundo novo na Floresta Sintá (c)tica - o treebank para Português. *Calidoscópio - Revista de Pós Graduação em Lingüística Aplicada da Unisinos, Rio Grande do Sul*, 6(3), Set / Dezembro, 2008.

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2008b. Um mundo novo na Floresta Sintá

(c)tica - o treebank para Português. *Calidoscópio - Revista de Pós Graduação em Lingüística Aplicada da Unisinos, Rio Grande do Sul*, 6(3), Set / Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2008. Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2009. Relation detection between named entities: report of a shared task. Em *Proceedings of Semantic Evaluations Workshop*, 4 de Junho, 2009.

Gasperin, Caroline Varaschin. 2001. Extração automática de relações semânticas a partir de relações sintáticas. Tese de Mestrado, Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

Gomes, Daniel e Mário J. Silva. 2005. Characterizing a National Community Web. *ACM Transactions on Internet Technology*, 5(3):508–531, Agosto, 2005.

Gomes, Paulo. 2008. Linguateca: Polo de Coimbra - Plantando Florestas e Criando Papel, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprGomesL10.pdf`.

Gomes de Matos, Francisco. 1992. O cientista de língua portuguesa e seus direitos linguísticos. *Revista Internacional de Língua Portuguesa*, 7:79–81.

Gonçalo Oliveira, Hugo, Cristina Mota, Cláudia Freitas, Diana Santos, e Paula Carvalho. 2008a. Avaliação à medida no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes, e Nuno Seco. 2008b. PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, pp. 31–40. Springer Verlag, 8-10 de Setembro, 2008.

Haber, Renato Ribeiro. 2001. Pica-pau: Um protótipo de ferramenta para visualização e edição de árvores sintáticas. Texto produzido no âmbito da Floresta Sintá (c)tica, `http://www.linguateca.pt/treebank/Picapau.html`.

Hausser, Roland, editor. 1996. *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994*. Max Niemeyer Verlag.

Inácio, Susana e Diana Santos. 2006. Syntactical Annotation of COMPARA: Workflow and First Results. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*, volume LNAI 3960, pp. 256–259, Berlin/Heidelberg, 13-17 de Maio, 2006. Springer.

Inácio, Susana e Diana Santos. 2008. Documentação da anotação morfossintáctica da parte portuguesa do COMPARA, Dezembro, 2008. Primeira versão: 9 de Dezembro de 2005, `http://www.linguateca.pt/COMPARA/DocAnotacaoPortCOMPARA.pdf`.

Inácio, Susana, Diana Santos, e Rosário Silva. 2008. COMPARAndo cores em português e inglês. Em Sónia Frota e Ana Lúcia Santos, editores, *Artigos seleccionados do XXIII Encontro da Associação Portuguesa de Linguística (APL)*, pp. 271–286, 1-3 de Outubro de 2007, 2008.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, e David Tugwell. 2005. The Sketch Engine. Em *Proc. Euralex*. pp. 105–116, Julho, 2005.

Lai, Catherine e Steven Bird. 2004. Querying and updating treebanks: A critical survey and requirements analysis. Em *In Proceedings of the Australasian Language Technology Workshop*, pp. 139–146.

Maia, Belinda. 2003. Constructing comparable and parallel corpora for terminology extraction - work in progress. Em Dawn Archer, Paul Rayson, Andrew Wilson, e Tony McEnery, editores, *Proceedings of the Corpus Linguistics 2003 conference (CL2003)*, 28-31 de Março. 2003.

Maia, Belinda. 2008a. Alice no País das Maravilhas ou as aventuras e desventuras de uma linguista no mundo do PLN, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprMaiaL10.pdf`.

Maia, Belinda. 2008b. Corpógrafo V4 - Tools for Educating Translators. Em Elia Yuste Rodrigo,

editor, *Topics in Language Resources for Translation and Localisation*. John Benjamins Pub. Co, Amsterdam/Philadelphia, pp. 57–70, Novembro, 2008.

Maia, Belinda e Anabela Barreiro. 2007. Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 205–216, 20 de Março, 2007.

Maia, Belinda e Sérgio Matos. 2008. Corpógrafo V4 - Tools for Researchers and Teachers using Comparable Corpora. Em Pierre Zweigenbaum, Éric Gaussier, e Pascale Fung, editores, *LREC 2008 Workshop on Comparable Corpora (LREC 2008)*. European Language Resources Association (ELRA), pp. 79–82, 31 de Maio, 2008.

Maia, Belinda, Luís Sarmento, e Diana Santos. 2005. Introduzindo o Corpógrafo - um conjunto de ferramentas para criar corpora especializados e comparáveis e bases de dados teminológicas. *Terminómetro*, 7:61–62. Número especial - A terminologia em Portugal e nos países de língua portuguesa em África.

Meinedo, Hugo, Márcio Viveiros, e João Paulo da Silva Neto. 2008. Evaluation of a live broadcast news subtitling system for Portuguese. Em *Interspeech 2008*. ISCA, Setembro, 2008.

Mota, Cristina e Pedro Moura. 2003. ANELL: A Web System for Portuguese Corpora Annotation. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 184–188, Berlin/Heidelberg. Springer Verlag.

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

NIST e ACE. 2007. Automatic Content Extraction 2008 Evaluation Plan (ACE08) – Assessment of Detection and Recognition of Entities and Relations within and across Documents. Relatório técnico, NIST. `http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf`.

Nunes, Maria das Graças Volpe. 2008. Relato sobre a parceria Linguateca-NILC, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprNunesL10.pdf`.

Oliveira, Catarina Alexandra Monteiro de. 2009. *Do Grafema ao Gesto: Contributos Linguísticos para um Sistema de Síntese de Base Articulatória*. Tese de doutoramento, Universidade de Aveiro.

Oliveira, Débora, Luís Sarmento, Belinda Maia, e Diana Santos. 2005. Corpus analysis for indexing: when corpus-based terminology makes a difference. Em Pernilla Danielsson e Martijn Wagenmakers, editores, *Proceedings from the Corpus Linguistics 2005 Conference Series*, volume 1, 14-17 de Julho. 2005.

Oliveira, Mirna, Bento C. Dias da Silva, e Helio Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 189–196, Berlin/Heidelberg, 23-26 de Junho, 2002. Springer.

Orăsan, Constantin, Dan Cristea, Ruslan Mitkov, e Antonio Branco. 2008. Anaphora resolution exercise: An overview. Em *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marraqueche, Marrocos, 28 - 30 de Maio, 2008.

Pardo, Thiago A. S., Maria das Graças Volpe Nunes, e Lúcia H. M. Rino. 2004. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. Em Ana L.C. Bazzan e Sofiane Labidi, editores, *Advances in Artificial Intelligence. XVII Brazilian Symposium on Artificial Intelligence (SBIA'04)*, Lecture Notes in Computer Science, pp. 224–234, Berlin/Heidelberg, 29 de Setembro - 1 de Outubro, 2004. Springer Verlag.

Pardo, Thiago A. S. e Lucia H. M. Rino. 2002. DMSumm: Review and Assessment. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 263–274, Berlin/Heidelberg, 23-26 de Junho, 2002. Springer.

Pekar, Viktor e Richard Evans. 2007. Discovery of language resources on the web: Information extraction from heterogeneous documents. *Literary and Linguistic Computing*, 22(3):329–343.

Peters, Carol, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras, e Diana Santos, editores. 2008. *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF*

2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, volume 5152 of *Lecture Notes in Computer Science*. Springer, Berlin.

Roberts, Kirk e Andrew Hickl. 2008. Scaling answer type detection to large hierarchies. Em *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. ELRA, 28-30 Maio, 2008.

Rocha, Paulo e Diana Santos. 2007. CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 143–158.

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.

Rolo, Carlos Juzarte e António Joaquim Serralheiro. 2008. An approach to natural language equation reading in digital talking books. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume 5190. Springer Verlag, pp. 268–271.

Santos, Diana. 1995. On grammatical translationese. Em Kimmo Koskenniemi, editor, *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*. pp. 59–66, 29-30 de Maio, 1995.

Santos, Diana. 1999a. Porquê processamento computacional do português e não processamento de linguagem natural?, 24 de Março, 1999. `http://www.linguateca.pt/branco/Porque.html`.

Santos, Diana. 1999b. Processamento computacional da língua portuguesa: Documento de trabalho. Versão base de 9 de Fevereiro de 1999; revista a 13 de Abril de 1999, `http://www.linguateca.pt/branco/index.html`.

Santos, Diana. 1999c. Towards language-specific applications. *Machine Translation*, 14(2):83–112, Junho, 1999.

Santos, Diana. 2000. O projecto Processamento Computacional do Português: Balanço e perspectivas. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. ICMC/USP, São Paulo, pp. 105–113, 19-22 de Novembro, 2000.

Santos, Diana. 2002a. DISPARA, a system for distributing parallel corpora on the Web. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 209–218, Berlin/Heidelberg. Springer.

Santos, Diana. 2002b. Um centro de recursos para o processamento computacional do português. *DataGramaZero - Revista de Ciência da Informação*, 3(1), Fevereiro, 2002.

Santos, Diana. 2003a. Relatório Linguateca 2000-2003. Relatório técnico, Linguateca, Setembro, 2003. `http://www.linguateca.pt/documentos/RelatorioLinguateca2000-2003Revisto.pdf`.

Santos, Diana. 2003b. Timber! Issues in treebank building and use. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 151–158, Berlin/Heidelberg. Springer.

Santos, Diana. 2004. Aonde vamos em relação a aonde. *the ESPecialist*, 25(1):85–103.

Santos, Diana. 2005. Relatório da Linguateca de 15 de Maio de 2004 a 14 de Maio de 2005. Relatório técnico, Linguateca, 2 de Junho, 2005. `http://www.linguateca.pt/documentos/RelatorioLinguatecaMaio2005.pdf`.

Santos, Diana. 2006a. Desenho, construção e utilização de corpora, 10 de Julho, 2006. `http://www.linguateca.pt/escolaverao2006/Corpora/CorporaEscolaVerao.pdf`.

Santos, Diana. 2006b. Resumo da actividade da Linguateca de 15 de Maio de 2003 a 15 de Dezembro de 2006. Relatório técnico, Linguateca, Dezembro, 2006. Com a colaboração (por ordem alfabética) de Alberto Simões, Ana Frankenberg-Garcia, Belinda Maia, Luís Costa, Luís Miguel Cabral, Luís Sarmento, Marcirio Chaves, Mário J. Silva, Nuno Cardoso, Paulo Gomes e Rui Vilela, `http://www.linguateca.pt/documentos/RelatorioLinguateca2003-2006.pdf`.

Santos, Diana. 2007a. Avaliação conjunta. Em Diana Santos, editor, *Avaliação conjunta: um*

novo paradigma no processamento computacional da língua portuguesa. IST Press, Lisboa, Portugal, pp. 1–12, 20 de Março, 2007.

Santos, Diana, editor. 2007b. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal.

Santos, Diana. 2007c. Computational linguistics beyond the processing of english. `http://www.linguateca.pt/Diana/download/FirstWords2007.pdf`.

Santos, Diana. 2008a. Curso avançado de estudos contrastivos usando o COMPARA como ferramenta, 3-5 de Novembro, 2008. Módulo na EBraLC, Segunda Escola Brasileira de Linguística Computacional, `http://www.linguateca.pt/documentos/cursoCOMPARASantosEBRALC2008.pdf`.

Santos, Diana. 2008b. Linguateca 10 anos: festejo ou luto?, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprSantosL10.pdf`.

Santos, Diana. 2008c. Perfect mismatches: Result in English and Portuguese. Em Margaret Rogers e Gunilla Anderman, editores, *Incorporating Corpora: The Linguist and the Translator*. Multilingual matters, Clevedon, pp. 217–242.

Santos, Diana e Anabela Barreiro. 2004. On the problems of creating a consensual golden standard of inflected forms in. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 483–486, 26-28 de Maio, 2004.

Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, e Gregory Stainhauer, editores, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 205–210, 31 de Maio - 2 de Junho, 2000.

Santos, Diana, Luís Miguel Cabral, e Luís Costa. 2006. Linguateca: seven years working for the computational processing of Portuguese, 23 de Novembro, 2006. `http://www.linguateca.pt/Diana/download/AprLinguatecaNov2006.pdf`.

Santos, Diana e Nuno Cardoso. 2005. Portuguese at CLEF 2005: Reflections and Challenges. Em Carol Peters, editor, *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005)*, pp. s/pp, Viena, Áustria, 21-23 de Setembro, 2005. Centromedia.

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.

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, e Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer.

Santos, Diana e Luís Costa. 2005. A Linguateca e o projecto 'Processamento Computacional do português'. *Terminómetro*, 7:63–69. Número especial - A terminologia em Portugal e nos países de língua portuguesa em África.

Santos, Diana e Luís Costa. 2007. QolA: fostering collaboration within QA. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multimodal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *Lecture Notes in Computer Science*, pp. 569–578, Berlin / Heidelberg. Springer.

Santos, Diana, Luís Costa, e Paulo Rocha. 2003. Cooperatively evaluating Portuguese morphology. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 259–266, Berlin/Heidelberg. Springer Verlag.

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(3):335–374, Maio, 2007.

Santos, Diana, Cláudia Freitas, Hugo Gonçalo Oliveira, e Paula Carvalho. 2008. Second HAREM: new challenges and old wisdom. Em

António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pp. 212–215. Springer Verlag.

Santos, Diana e Caroline Gasperin. 2002. Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. ELRA, Paris, pp. 597–604, 29-31 de Maio, 2002.

Santos, Diana, Belinda Maia, e Luís Sarmento. 2004. Gathering empirical data to evaluate MT from English to Portuguese. Em Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrún Magnúsdóttir, Anna Samiotou, e Khalid Choukri, editores, *Proceedings of LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*. pp. 14–17, 25 de Maio, 2004.

Santos, Diana e Paulo Rocha. 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, e Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, pp. 821–832.

Santos, Diana e Luís Sarmento. 2003. O projecto AC/DC: acesso a corpora/disponibilização de corpora. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 705–717, Lisboa, 2-4 de Outubro de 2002, 2003. APL.

Santos, Diana, Rosário Silva, e Susana Inácio. 2008. What's in a colour? Studying and contrasting colours with COMPARA. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. s/pp. European Language Resources Association (ELRA), 28-30 de Maio, 2008.

Santos, Diana, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela, e Susana Afonso. 2004. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. Em Guillermo De Ita Luna, Olac Fuentes Chávez, e Mauricio Osorio Galindo, editores, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA 2004)*, pp. 147–154, Novembro, 2004.

Sarmento, Luís, Anabela Barreiro, Belinda Maia, e Diana Santos. 2007. Avaliação de Tradução Automática: alguns conceitos e reflexões. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 181–190.

Sarmento, Luís e Belinda Maia. 2003. Gestor de corpora - Um ambiente Web integrado para Linguística baseada em Corpora. Em José João Almeida, editor, *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)*, pp. 25–30, Braga, 3 de Junho, 2003. Universidade do Minho.

Sarmento, Luís, Belinda Maia, e Diana Santos. 2004. The Corpógrafo - a Web-based environment for corpora research. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*. pp. 449–452, 26-28 de Maio, 2004.

Seco, Nuno e Nuno Cardoso. 2006. Detecting user sessions in the tumba! web log. Relatório técnico, Linguateca, Março, 2006. `http://eden.dei.uc.pt/~nseco/tumba.pdf`.

Seco, Nuno, Diana Santos, Rui Vilela, e Nuno Cardoso. 2006. A Complex Evaluation Architecture for HAREM. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*, volume LNAI 3960, pp. 260–263, Berlin/Heidelberg. Springer Verlag.

Serralheiro, A., I. Trancoso, D. Caseiro, T. Chambel, L. Carriço, e N. Guimarães. 2003. Towards a repository of digital talking books. Em *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology (Interspeech'2003)*. Genebra, Suíça, Setembro, 2003.

Silva, Augusto Soares. 2008a. Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos da Linguagem*, 16(1):49–81.

Silva, Mário J. 2008b. Pólo XLDB da Linguateca: 4 anos, 11 de Setembro, 2008. Apresentação no Encontro Linguateca: 10 anos, `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprMJSilvaL10.pdf`.

Simões, Alberto. 2008. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Tese de doutoramento, Faculdade de Engenharia da Universidade do Minho, Braga, Março, 2008.

Simões, Alberto e José João Almeida. 2007. Parallel Corpora based Translation Resources Extraction. *Procesamiento del Lenguaje Natural*, 39:265–272, Setembro, 2007.

Vallin, Alessandro, Bernardo Magnini, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov, e Richard Sutcliffe. 2005. Overview of the CLEF 2004 Multilingual Question answering track. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, e Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, pp. 371–391.

Vilela, Rui, Alberto Manuel Simões, Eckhard Bick, e José João Almeida. 2005. Representação em XML da Floresta Sintáctica. Em José Carlos Ramalho, Alberto Simões, e João Correia Lopes, editores, *3ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas (XATA 2005)*, pp. 351–361. Departamento de Informática, Universidade do Minho.

Wing, Benjamin e Jason Baldrige. 2006. Adaptation of Data and Models for Probabilistic Parsing of Portuguese. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)*, volume LNAI 3960, pp. 140–149, Berlin/Heidelberg. Springer.

Xavier, Maria Francisca, Maria de Lurdes Crispim, Graça Vicente, A. Castro, Alexandra Fiéis, Maria Cristina Silva, e M. Lobo. 1998. Utilizações informáticas de corpora textuais medievais. Em Palmira Marrafa e Maria Antónia Mota, editores, *Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística*. Colibri, Lisboa, pp. 347–358.

Artigos de Investigação

Anotación morfosintáctica do Corpus Técnico do Galego

Xavier Gómez Guinovart
Universidade de Vigo
xgg@uvigo.es

Susana López Fernández
Universidade de Vigo
susanalopez@uvigo.es

Resumo

Neste traballo preséntanse a metodoloxía e os criterios empregados na anotación lingüística (etiquetaxe categorial e lematización) do Corpus Técnico do Galego, un corpus elaborado na Universidade de Vigo con textos monolingües especializados do galego contemporáneo nos eidos do dereito, da informática, da economía, das ciencias ambientais, da socioloxía e da medicina.

1. *Introdución*

O Corpus Técnico Anotado do Galego (CTAG) é a versión categorizada e lematizada do Corpus Técnico do Galego (CTG), unha colección de córpora do galego contemporáneo composta de textos monolingües especializados nos eidos do dereito, da informática, da economía, das ciencias ambientais, da socioloxía e da medicina, dispoñible en Internet desde 2006 para libre consulta (Gómez Clemente e Gómez Guinovart, 2006-2009). Cunha extensión actual de 12,5 millóns de palabras, o CTG reúne textos do ámbito xurídico-administrativo (2.516.846 palabras), textos de informática e telecomunicacións (2.027.816 palabras), textos de ecoloxía e ciencias ambientais (2.349.362 palabras), textos de economía (2.055.837 palabras), textos de socioloxía (2.442.765 palabras) e textos de medicina (1.154.071 palabras, aínda en fase de recompilación). A anotación do Corpus CTAG non é totalmente automática, senón que ten unha primeira fase na que se lle aplica un programa etiquetador e lematizador, e unha segunda fase na que se revisan manualmente os resultados deste procesamento automático. Os traballos de anotación lingüística do CTAG, en fase avanzada de elaboración, lévanse a cabo no marco de dous proxectos de investigación en curso[1], aínda que os seus resultados iniciais xa se poden consultar en Internet (Gómez Guinovart, 2006-2009). En concreto, xa se atopa dispoñible en Internet unha sección do CTAG de máis de 2 millóns de palabras, correspondente ao ámbito especializado da ecoloxía e das ciencias ambientais.

A etiquetaxe inicial do CTAG levouse a cabo empregando unha adaptación modificada do analizador morfolóxico do galego que forma parte do par español-galego do tradutor Apertium (Armentano Oller et al., 2006; Alegría Loinaz et al., 2006), con cambios no seu etiquetario, no tratamento das contraccións e no manexo das formas non normativas do galego. De maneira xeral, o conxunto de etiquetas deseñado para a anotación do CTAG constitúe unha adaptación ás características propias do galego dos principios elaborados polo grupo EAGLES (Leech e Wilson, 1996) para a creación dun estándar europeo de anotación morfosintáctica de léxicos e córpora. De maneira máis específica, o conxunto normalizado de etiquetas utilizado para o CTAG elaborouse tendo en conta as propostas realizadas por Civit para a lingua castelá (Civit, 2003) e adoptadas con algunhas modificacións no etiquetador morfolóxico do Freeling (Atserias et al., 2006). Nos seguintes apartados deste traballo, presentarase polo miúdo o etiquetario empregado na anotación do corpus, e as cuestións de deseño relacionadas coa codificación das formas anotadas e co tratamento das contraccións, formas enclíticas e formas non normativas presentes nos textos.

2. *Etiquetaxe do Corpus CTAG*

2.1. Codificación

A anotación do Corpus CTAG ten en conta todas as formas léxicas (galegas e non galegas, normativas e non normativas) que aparecen nos textos, e mais as cifras, abreviaturas, símbolos e signos de puntuación. Cada forma etiquetada consta de tres partes: a forma que aparece no texto, o

[1]Este traballo foi financiado polo Ministerio de Educación y Ciencia e o Fondo Europeo de Desenvolvemento Rexional (FEDER), dentro do proxecto *Deseño e implementacion dun servidor de recursos integrados para o desenvolvemento de tecnoloxías da lingua galega (RILG)* do Plan Nacional de I+D+I, 2006-2009 (ref. HUM2006-11125-C02-01/FILO); e pola Consellaría de Innovación e Industria da Xunta de Galicia, dentro do proxecto *Desenvolvemento e aplicación de recursos integrados da lingua galega* do Plan galego de investigación, desenvolvemento e innovación tecnolóxica (Incite), 2008-2011 (ref. INCITE08PXIB302185PR). Ambos son proxectos coordinados da Universidade de Vigo (Grupo TALG) coa Universidade de Santiago de Compostela (Instituto da Lingua Galega).

lema (ou representación abstracta da clase flexiva) e a etiqueta categorial, consonte o seguinte esquema: *^forma/lema_etiqueta*. Deste xeito, o adxectivo *transxénicos* vai ser anotado no corpus como *^transxénicos/transxénico_A0MP*.

2.2. Etiquetario

Para cada categoría inclúense dúas táboas. Na primeira táboa, recóllense as características lingüísticas ou atributos pertinentes para cada categoría (segunda columna), cos seus posibles valores (terceira columna), a abreviatura ou codificación dos valores na etiqueta (cuarta columna), e o lugar ou posición (primeira columna) que cada un dos valores vai ocupar na etiqueta resultante. Na segunda táboa, recóllese o inventario completo de etiquetas para cada categoría, cun exemplo de palabra e lema para cada caso. Esta descrición esquemática do etiquetario do CTAG, empregando táboas, está baseada no sistema utilizado en Civit (2003).

2.2.1. Nomes

NOMES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nome	N
2	Tipo	Común	C
		Propio	P
3	Xénero	Masculino	M
		Feminino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Grao	Apreciativo	A

Táboa 1: Etiquetas para nomes

Forma	**Lema**	**Etiqueta**
neno	neno	NCMS0
nenos	neno	NCMP0
nena	neno	NCFS0
nenas	neno	NCFP0
xornalista	xornalista	NCCS0
xornalistas	xornalista	NCCP0
microondas	microondas	NCMN0
Breogán	Breogán	NP000
neniño	neno	NCMSA
neniños	neno	NCMPA
neniña	neno	NCFSA
neniñas	neno	NCFPA

Táboa 2: Exemplos de nomes

O lema dos nomes vai ser sempre a forma masculina singular (*neno*) ou a forma singular común se o nome é de xénero común (*xornalista*). Nos nomes invariables, isto é, naqueles que presentan a mesma forma tanto no singular coma no plural (*microondas*), o lema e a forma van ser sempre coincidentes.

O atributo *grao* con valor **A** especifícase nos nomes con sufixación apreciativa (aumentativos, diminutivos, pexorativos, etc.) (*neniño, nenón*). No resto de nomes, o valor do atributo *grao* é de non especificado ou **0**.

Finalmente, os nomes propios levan no CTAG a etiqueta NP000, cos valores de xénero, número e grao sen especificar.

2.2.2. Adxectivos

ADXECTIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adxectivo	A
2	Grao	Apreciativo	A
3	Xénero	Masculino	M
		Feminino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N

Táboa 3: Etiquetas para adxectivos

Forma	**Lema**	**Etiqueta**
febles	feble	A0CP
feble	feble	A0CS
ecolóxicas	ecolóxico	A0FP
ecolóxica	ecolóxico	A0FS
ecolóxicos	ecolóxico	A0MP
ecolóxico	ecolóxico	A0MS
choromicas	choromicas	A0CN
grandiñas	grande	AAFP
grandiña	grande	AAFS
grandiños	grande	AAMP
grandiño	grande	AAMS

Táboa 4: Exemplos de adxectivos

O lema dos adxectivos vai ser sempre a forma masculina singular (*ecolóxico*) ou a forma singular común se o adxectivo é de xénero común (*fértil*). Nos adxectivos invariables (*choromicas*), o lema e a forma van ser sempre coincidentes.

O atributo *grao* especificarase para os adxectivos con grao comparativo (*meirande*) ou superlativo (*altísimo*), ou con sufixación apreciativa (diminutivos, aumentativos, pexorativos, etc.) (*pequeniño, fermosón*). Estes dous tipos de adxectivos vanse distinguir porque o valor do segundo atributo da etiqueta vai ser **A**, mentres que no resto de adxectivos vai ser sempre **0**.

2.2.3. Verbos

O lema dos verbos é sempre o infinitivo. O atributo de xénero só afecta aos participios. Nas formas de infinitivo e xerundio non conxugados non se especifican os atributos de tempo, persoa, número e xénero, polo que o seu valor vai ser

sempre de **0**. Só os participios e os xerundios poden levar o atributo de apreciativo, nos resto dos casos o valor na etiqueta é **0**.

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Modo	Indicativo	I
		Subxuntivo	S
		Imperativo	M
		Infinitivo	N
		Xerundio	X
		Participio	P
3	Tempo	Presente	P
		Copretérito	I
		Futuro	F
		Pretérito	S
		Pospretérito	C
		Antepretérito	A
4	Persoa	Primeira	1
		Segunda	2
		Terceira	3
5	Número	Singular	S
		Plural	P
6	Xénero	Masculino	M
		Feminino	F
7	Grao	Apreciativo	A

Táboa 5: Etiquetas para verbos

Tempo	**Forma**	**Lema**	**Etiqueta**
Pres. Ind.	canto	cantar	VIP1S00
	cantas	cantar	VIP2S00
	canta	cantar	VIP3S00
	cantamos	cantar	VIP1P00
	cantades	cantar	VIP2P00
	cantan	cantar	VIP3P00
Copretérito	cantaba	cantar	VII1S00
	cantabas	cantar	VII2S00
	cantaba	cantar	VII3S00
	cantabamos	cantar	VII1P00
	cantabades	cantar	VII2P00
	cantaban	cantar	VII3P00
Pret. Ind.	cantei	cantar	VIS1S00
	cantaches	cantar	VIS2S00
	cantou	cantar	VIS3S00
	cantamos	cantar	VIS1P00
	cantastes	cantar	VIS2P00
	cantaron	cantar	VIS3P00
Fut. Ind.	cantarei	cantar	VIF1S00
	cantarás	cantar	VIF2S00
	cantará	cantar	VIF3S00
	cantaremos	cantar	VIF1P00
	cantaredes	cantar	VIF2P00
	cantarán	cantar	VIF3P00
Pospretérito	cantaría	cantar	VIC1S00
	cantarías	cantar	VIC2S00
	cantaría	cantar	VIC3S00
	cantariamos	cantar	VIC1P00
	cantariades	cantar	VIC2P00
	cantarían	cantar	VIC3P00
Antepretérito	cantara	cantar	VIA1S00
	cantaras	cantar	VIA2S00
	cantara	cantar	VIA3S00
	cantaramos	cantar	VIA1P00
	cantarades	cantar	VIA2P00
	cantaran	cantar	VIA3P00
Pres. Subx.	cante	cantar	VSP1S00
	cantes	cantar	VSP2S00
	cante	cantar	VSP3S00
	cantemos	cantar	VSP1P00
	cantedes	cantar	VSP2P00
	canten	cantar	VSP3P00
Pret. Subx.	cantase	cantar	VSI1S00
	cantases	cantar	VSI2S00
	cantase	cantar	VSI3S00
	cantásemos	cantar	VSI1P00
	cantásedes	cantar	VSI2P00
	cantasen	cantar	VSI3P00
Fut. Subx.	cantar	cantar	VSF1S00
	cantares	cantar	VSF2S00
	cantar	cantar	VSF3S00
	cantarmos	cantar	VSF1P00
	cantardes	cantar	VSF2P00
	cantaren	cantar	VSF3P00
Imperativo	canta	cantar	VM02S00
	cante	cantar	VM03S00
	cantemos	cantar	VM01P00
	cantade	cantar	VM02P00
	canten	cantar	VM03P00
Infinitivo	cantar	cantar	VN00000
Xerundio	cantando	cantar	VX00000
	cantandiño	cantar	VX0000A
Participio	cantada	cantar	VP00SF0
	cantado	cantar	VP00SM0
	cantadas	cantar	VP00PF0
	cantados	cantar	VP00PM0
	cantadiña	cantar	VP00SFA
	cantadiño	cantar	VP00SMA
	cantadiñas	cantar	VP00PFA
	cantadiños	cantar	VP00PMA
Inf. conxugado	cantar	cantar	VN00000
	cantares	cantar	VN02S00
	cantar	cantar	VN00000
	cantarmos	cantar	VN01P00
	cantardes	cantar	VN02P00
	cantaren	cantar	VN03P00
Xer. conxugado	cantándomos	cantar	VX01P00
	cantándodes	cantar	VX02P00

Táboa 6: Exemplos de verbos

2.2.4. Adverbios

A indicación de **R** no CTAG serve para etiquetar tanto os adverbios coma as locucións adverbiais. Por outra banda, os adverbios rematados en *-mente*, derivados de adxectivos, manteñen como lema a súa forma derivada.

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Grao	Apreciativo	A

Táboa 7: Etiquetas para adverbios

Forma	Lema	Etiqueta
xa	xa	R0
hoxe	hoxe	R0
sempre	sempre	R0
tecnoloxicamente	tecnoloxicamente	R0
ambientalmente	ambientalmente	R0
de_acordo	de_acordo	R0
ao_chou	ao_chou	R0
non	non	R0
loguiño	logo	RA
a_modiño	a_modiño	RA

Táboa 8: Exemplos de adverbios

2.2.5. Numerais

NUMERAIS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Numeral	M
2	Tipo	Cardinal	C
		Ordinal	O
		Partitivo	P
3	Grao	Apreciativo	A
4	Xénero	Masculino	M
		Feminino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N

Táboa 9: Etiquetas para numerais

A diferenza da proposta de Civit (2003), na que os numerais se inclúen entre os determinantes, no etiquetario do CTAG aparecen como unha categoría de seu, consonte coa tradición gramatical galega e coas recomendacións de EAGLES (Leech e Wilson, 1996). Como no caso dos adxectivos e do resto das categorías posuidoras de flexión, para os numerais con xénero e número morfoloxicamente marcado, o lema indicado no CTAG vai ser a forma masculina singular.

Forma	Lema	Etiqueta
un	un	MC0MN
unha	un	MC0FN
tres	tres	MC0CN
primeiras	primeiro	MO0FP
primeira	primeiro	MO0FS
primeiros	primeiro	MO0MP
primeiro	primeiro	MO0MS
primeiriñas	primeiro	MOAFP
primeiriña	primeiro	MOAFS
primeiriños	primeiro	MOAMP
primeiriño	primeiro	MOAMS
medio	medio	MP0MS
media	medio	MP0FS

Táboa 10: Exemplos de numerais

2.2.6. Determinantes

No etiquetario do CTAG, só se inclúen na categoría dos determinantes as formas do artigo definido. A categoría de artigo indeterminado (*un*) trátase dentro da dos pronomes indefinidos. Tampouco se inclúen entre os determinantes os demostrativos, posesivos, indefinidos, relativos, exclamativos ou interrogativos, sendo todos eles tratados como categorías independentes.

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Artigo	G
2	Xénero	Masculino	M
		Feminino	F
		Común	C
3	Número	Singular	S
		Plural	P

Táboa 11: Etiquetas para determinantes

Forma	Lema	Etiqueta
o	o	GMS
os	o	GMP
a	o	GFS
as	o	GFP
@s	o	GCP

Táboa 12: Exemplos de determinantes

2.2.7. Pronomes

Malia que Civit (2003) inclúe nesta categoría os pronomes demostrativos, posesivos, indefinidos, relativos, interrogativos, exclamativos e numerais, na etiquetaxe do CTAG todas estas categorías considéranse categorías independentes, resérvandose a categoría pronominal do etiquetario para os denominados tradicionalmente pronomes persoais. Na anotación do CTAG, o atributo de cortesía, marcado con valor **P**, especifícase soamente para as formas *vostede* e *vostedes*.

PRONOMES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronome	P
2	Persoa	Primeira	1
		Segunda	2
		Terceira	3
3	Xénero	Masculino	M
		Feminino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	Nominativo	N
		Nom/Recto	B
		Acusativo	A
		Dativo	D
		Oblicuo	O
		Acus/Dat/Reflex	C
		Reflexivo	R
6	Cortesía	Cortés	P

Táboa 13: Etiquetas para pronomes

Forma	**Lema**	**Etiqueta**
eu	eu	P1CSN0
min	min	P1CSO0
nós	nós	P1CPB0
nosoutros	nosoutros	P1MPB0
nos	nos	P1CPC0
te	te	P2CSA0
che	che	P2CSD0
ti	ti	P2CSB0
vostede	vostede	P3CSBP
vostedes	vostede	P3CPBP
vós	vós	P2CPB0
vos	vos	P2CPC0
el	el	P3MSB0
ela	el	P3FSB0
elas	el	P3FPB0
eles	el	P3MPB0
a	o	P3FSA0
as	o	P3FPA0
o	o	P3MSA0
os	o	P3MPA0
lle	lle	P3CSD0
lles	lle	P3CPD0
se	se	P3CNR0
si	si	P3CNO0

Táboa 14: Exemplos de pronomes

2.2.8. Posesivos

No CTAG, o atributo de *posuidor* utilízase cos pronomes posesivos para marcar o número do posuidor: singular para *meu* e *teu*, plural para *noso* e *voso*. Os pronomes en que o posuidor é unha terceira persoa (*seu*) reciben como valor **0** para este atributo, dada a dificultade de distinguir o seu número gramatical singular ou plural, isto é, se fai referencia a *el/ela* ou a *eles/elas*.

POSESIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Posesivo	X
2	Persoa	Primeira	1
		Segunda	2
		Terceira	3
3	Xénero	Masculino	M
		Feminino	F
4	Número	Singular	S
		Plural	P
5	Posuidor	Singular	S
		Plural	P

Táboa 15: Etiquetas para posesivos

Forma	**Lema**	**Etiqueta**
miña	meu	X1FSS
miñas	meu	X1FPS
meus	meu	X1MPS
meu	meu	X1MSS
nosa	noso	X1FSP
nosas	noso	X1FPP
noso	noso	X1MSP
nosos	noso	X1MPP
súa	seu	X3FS0
súas	seu	X3FP0
seu	seu	X3MS0
seus	seu	X3MP0
túa	teu	X2FSS
túas	teu	X2FPS
teu	teu	X2FSS
teus	teu	X2MPS
vosa	voso	X2FSP
vosas	voso	X2FPP
voso	voso	X2MSP
vosos	voso	X2MPP

Táboa 16: Exemplos de posesivos

2.2.9. Demostrativos

Forma	**Lema**	**Etiqueta**
aquelas	aquel	DFP
aquela	aquel	DFS
aqueles	aquel	DMP
aquel	aquel	DMS
aquilo	aquel	DNS
esas	ese	DFP
esa	ese	DFS
eses	ese	DMP
ese	ese	DMS
iso	ese	DNS
estas	este	DFP
esta	este	DFS
estes	este	DMP
este	este	DMS
isto	este	DNS

Táboa 17: Exemplos de demostrativos

DEMOSTRATIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Demostrativo	D
2	Xénero	Masculino	M
		Feminino	F
		Neutro	N
3	Número	Singular	S
		Plural	P

Táboa 18: Etiquetas para demostrativos

2.2.10. Interrogativos

INTERROGATIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interrogativo	T
2	Xénero	Masculino	M
		Feminino	F
		Común	C
3	Número	Singular	S
		Plural	P
		Invariable	N
4	Grao	Apreciativo	A

Táboa 19: Etiquetas para interrogativos

Forma	**Lema**	**Etiqueta**
cal	cal	TCS0
cales	cal	TCP0
que	que	TCN0
canto	canto	TMS0
cantos	canto	TMP0
canta	canto	TFS0
cantas	canto	TFP0
cantiño	canto	TMSA

Táboa 20: Exemplos de interrogativos

2.2.11. Relativos

RELATIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Relativo	Q
2	Xénero	Masculino	M
		Feminino	F
		Común	C
3	Número	Singular	S
		Plural	P
		Invariable	N
4	Grao	Apreciativo	A

Táboa 21: Etiquetas para relativos

Forma	**Lema**	**Etiqueta**
cal	cal	QCS0
cales	cal	QCP0
canta	canto	QFS0
cantas	canto	QFP0
canto	canto	QMS0
cantos	canto	QMS0
cantiño	canto	QMSA
que	que	QCN0

Táboa 22: Exemplos de relativos

2.2.12. Indefinidos

Con esta categoría etiquétanse tamén no CTAG os artigos indeterminados (*un*), alén dos catalogados tradicionalmente como pronomes indefinidos.

INDEFINIDOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Indefinido	I
2	Xénero	Masculino	M
		Feminino	F
		Neutro	N
3	Número	Singular	S
		Plural	P
4	Grao	Apreciativo	A

Táboa 23: Etiquetas para indefinidos

Forma	**Lema**	**Etiqueta**
algo	algo	IMS0
alguén	alguén	IMS0
algunha	algún	IFS0
algunhas	algún	IFP0
algún	algún	IMS0
algúns	algún	IMP0
calquera	calquera	INS0
mesma	mesmo	IFS0
mesmas	mesmo	IFP0
mesmo	mesmo	IMS0
mesmos	mesmo	IMP0
mesmiño	mesmo	IMSA
mesmiña	mesmo	IFSA
mesmiñas	mesmo	IFPA
nada	nada	IMS0
nadiña	nada	IMSA
ninguén	ninguén	INS0
ningunha	ningún	IFS0
ningunhas	ningún	IFP0
ningún	ningún	IMS0
ningúns	ningún	IMP0
pouca	pouco	IFS0
poucas	pouco	IFP0
pouco	pouco	IMS0
poucos	pouco	IMP0
pouquiño	pouco	IMSA
unha	un	IFS0
unhas	un	IFP0
un	un	IMS0
uns	un	IMP0
varias	varios	IFP0
varios	varios	IMP0

Táboa 24: Exemplos de indefinidos

2.2.13. Preposicións

PREPOSICIÓNS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Preposición	S

Táboa 25: Etiquetas para preposicións

Forma	Lema	Etiqueta
a	a	S
de	de	S
ante	ante	S
baixo	baixo	S
con	con	S
cara_a	cara_a	S

Táboa 26: Exemplos de preposicións

2.2.14. Conxuncións

CONXUNCIÓNS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Conxunción	C
2	Tipo	Coordinativa	C
		Subordinativa	S

Táboa 27: Etiquetas para conxuncións

Forma	Lema	Etiqueta
e	e	CC
e_mais	e_mais	CC
nin	nin	CC
ou	ou	CC
pero	pero	CC
senón	senón	CC
aínda_que	aínda_que	CS
porque	porque	CS
pois	pois	CS
que	que	CS
se	se	CS
xa_que_logo	xa_que_logo	CS

Táboa 28: Exemplos de conxuncións

2.2.15. Interxeccións

INTERXECCIÓNS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interxección	O

Táboa 29: Etiquetas para interxeccións

Forma	Lema	Etiqueta
ou	ou	O
xe	xe	O
bo	bo	O
vaites	vaites	O

Táboa 30: Exemplos de interxeccións

2.2.16. Puntuación

A respecto da puntuación, o etiquetario do CTAG segue o utilizado no FreeLing (Atserias et al., 2006), baseado en Civit (2003).

PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

Táboa 31: Etiquetas para puntuación

Forma	Lema	Etiqueta
¡	¡	Faa
!	!	Fat
,	,	Fc
[	[	Fca
]	]	Fct
:	:	Fd
"	"	Fe
-	-	Fg
¿	¿	Fia
?	?	Fit
{	{	Fla
}	}	Flt
.	.	Fp
(	(	Fpa
)	)	Fpt
«	«	Fra
»	»	Frc
...	...	Fs
%	%	Ft
;	;	Fx
_	_	Fz
+	+	Fz
=	=	Fz

Táboa 32: Exemplos de puntuación

2.2.17. Cifras

As cifras etiquétanse no CTAG co código **Z**. Con esta categoría, abránguense anos, enderezos, números de teléfono, etc.

CIFRAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Cifra	Z

Táboa 33: Etiquetas para cifras

Forma	Lema	Etiqueta
10'2	10'2	Z
1.998	1.998	Z

Táboa 34: Exemplos de cifras

2.2.18. Abreviaturas

No CTAG emprégase a etiqueta **Y** para as abreviaturas de resolución incerta e tamén para os enderezos electrónicos e indicacións de unidades de temperatura (^oC) e outras. Porén, etiquétanse como nomes propios formas como M^a ou siglas que corresponden a entidades propias e individualizadas, como *CEE* ou *EEUU*; como nomes comúns formas abreviadas como n^o ou *ex.* ou siglas do tipo *SA*, *SP* ou *PEMES* (sic); e como numerais ordinais as abreviaturas como 1^o ou 3^a.

ABREVIATURAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Abreviatura	Y

Táboa 35: Etiquetas para abreviaturas

Forma	Lema	Etiqueta
ºC	graos Celsius	Y
sli.uvigo.es	sli.uvigo.es	Y
Mª	Mª	NP000
1º	1º	NO0MS
S.A.	S.A.	NCFS0
PEMES	PEMES	NCFP0

Táboa 36: Exemplos de abreviaturas

2.2.19. Símbolos

Inclúense na categoría dos símbolos todas as formas abreviadas que representan símbolos químicos da táboa periódica e formas compostas por eles. O lema vai coincidir coa forma plena estándar que corresponde a cada símbolo.

SÍMBOLOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Símbolo	L

Táboa 37: Etiquetas para símbolos

Forma	Lema	Etiqueta
Fe	ferro	L
Ni	níquel	L
O	osíxeno	L
ClH	ácido clorhídrico	L

Táboa 38: Exemplos de símbolos

2.2.20. Estranxeirismos

Todos os estranxeirismos pertencentes a calquera lingua distinta do galego etiquétanse no CTAG como **E**, sen especificar o idioma de orixe.

ESTRANXEIRISMOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Estranxeirismo	E

Táboa 39: Etiquetas para estranxeirismos

Forma	Lema	Etiqueta
monsieur	monsieur	E
and	and	E

Táboa 40: Exemplos de estranxeirismos

2.2.21. Palabras non clasificadas

As formas que resultan descoñecidas ou de difícil clasificación codifícanse no CTAG coa etiqueta **U**.

NON CLASIFICADAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Non clasificada	U

Táboa 41: Etiquetas para palabras non clasificadas

Forma	Lema	Etiqueta
R50	R50	U
LOUFungi	LOUFungi	U

Táboa 42: Exemplos de palabras non clasificadas

2.3. Contraccións e enclises

O galego ofrece moitas posibilidades de contraccións, por iso cómpre precisión á hora de describilas, tendo en conta as especificacións de cada un dos seus compoñentes.

O sistema de anotación do CTAG equipara formalmente a codificación dos diversos casos onde se produce a unión de dúas ou máis formas como ocorre, por exemplo, na segunda forma do artigo, na enclise dos pronomes átonos, nas contraccións propias das preposicións, ou na contracción con artigo da conxunción comparativa *ca*.

Dun modo xeral, o método de codificación das contraccións e enclises no CTAG é o seguinte: se *F* é unha forma contracta ou enclítica formada pola unión das palabras *P1+P2+...+Pn*, sendo *L1,L2...Ln* os lemas das palabras compoñentes e *C1,C2...Cn* as súas etiquetas categoriais, a forma codificada xenérica da forma contracta sería *F/L1_ C1 ∼/L2_ C2 ... ∼/Ln_ Cn*, como se ilustra a seguir na etiquetaxe das formas contractas *facelas*, *nesoutra* e *entrámbolos*:

- facelas/facer_VN0000 ∼/o_P3FPA0
- nesoutra/en_S ∼/ese_DFS ∼/outro_IFS0
- entrámbolos/entre_S ∼/ambos_IMP0 ∼/o_GMP

Xa que logo, as formas contractas e enclíticas están analizadas no CTAG como secuencias de palabras aptas para a posterior análise sintáctica. O til (∼) indica que a forma fónica da palabra está subsumida na contracción anterior.

A mesma codificación aplícase coherentemente ás enclises da segunda forma do artigo determinado, como se amosa nos seguintes exemplos:

- face-lo/facer_VN0000 ∼/o_GMS
- perdíche-los/perder_VIS2S00 ∼/o_GMP
- collémo-la/coller_VIP1P00 ∼/o_GFS
- protexe-las/protexer_VN0000 ∼/o_GFP
- tódolos/todo_IMP0 ∼/o_GMP
- mailas/mais_CC ∼/o_GFP
- nó-los/nós_P1CPB0 ∼/o_GMP

A mesma codificación aplícase aos pronomes enclíticos, mesmo cando estes van seguidos dunha segunda forma do artigo determinado:

- lévannos/levar_VIP3P0 ∼/nos_P10PC0
- permítenlles/permitir_VIP3P0 ∼/lle_P30PD0

- débellelo/deber_VIP3S0 ∼/lle_P30PD0 ∼/o_P3MSA0
- dóuvo-la/dar_VIP1S00 ∼/vos_P20PC0 ∼/o_GFS
- quitóulle-las/quitar_VIS3S00 ∼/lle_P30PD0 ∼/o_GFP

O mesmo método de anotación utilízase tamén para as unións dos pronomes en dativo co acusativo de terceira persoa:

- cho/che_P2CSD0 ∼/o_P3MSA0
- nola/nos_P10PC0 ∼/o_P3FSA0
- lla/lle_P30SD0 ∼/o_P3FSA0
- llela/lle_P30PD0 ∼/o_P3FSA0

Tamén nos diversos casos de contracción de preposicións con artigos determinados e indeterminados, demostrativos, pronomes persoais, indefinidos, etc.:

- das/de_S ∼/o_GFP
- polos/por_S ∼/o_GMP
- coa/con_S ∼/o_GFS
- no/en_S ∼/o_GMS
- cara á/cara a_S ∼/o_GFS

E tamén nos casos de contracción da conxunción comparativa *ca* coas diferentes formas do artigo determinado:

- cá/ca_CS ∼/o_GFS
- cás/ca_CS ∼/o_GFP
- có/ca_CS ∼/o_GMS
- cós/ca_CS ∼/o_GMP

Isto é, o CTAG utiliza un método uniforme analítico para a etiquetaxe de toda a ampla variedade de formas enclíticas e contractas do galego.

2.4. Problemas normativos

Non é infrecuente atopar nos textos do corpus CTAG exemplos de palabras que non se adaptan á normativa ortográfica oficial para o galego, vixente desde o ano 2003 (Real Academia Galega e Instituto da Lingua Galega, 2003). Nalgúns casos, as formas non normativas identificadas son froito do descoñecemento da norma ou do *lapsus calami*; mais noutros casos trátase de formas documentadas en textos escritos en datas anteriores á reforma normativa de 2003, correctas na normativa vixente no momento en que foron escritas; ou mesmo de formas pertencentes a normativas distintas da oficial. En todos estes casos, a anotación do CTAG inclúe, ao carón da forma non normativa documentada, a forma normativa da palabra precedida do símbolo '#'. Doutra banda, cando esta corrección implica un cambio categorial na forma non normativa documentada, a anotación inclúe tamén a etiqueta morfolóxica da forma incorrecta precedida do símbolo '|'. Véxase a aplicación destas convencións na etiquetaxe do corpus CTAG nos seguintes exemplos:

- presencia/presencia#presenza_NCFS0
- productos/producto#produto_NCMP0
- efeitos/efeito#efecto_NCMP0
- meio/meio#medio_NCMS0
- desbroce/desbroce|NCMS0#roza_NCFS0
- aporte/aporte|NCMS0#achega_NCFS0
- fango/fango|NCMS0#lama_NCFS0
- promedio/promedio|NCMS0#media_NCFS0
- llo/llo|lle_P30SD0#lle_P30PD0 ∼/o_P3MSA0 (*llo* por *llelo*)

Outra causa frecuente de conflito coa normativa provén dos recursos gráficos utilizados en relación co uso non sexista da linguaxe. A efectos da etiquetación do corpus, as arrobas e as formas alternativas con barra inclinada tipo que se documentan en exemplos como *@s europe@s*, *o/a consumidor/a* ou *os/as destinatarios/as* son tratatadas como grafías que indican un xénero común do lema (inexistente na súa morfoloxía), xénero que se recolle para cada caso na etiqueta correspondente, como se pode observar nos seguintes exemplos:

- @s/o_GCP
- europe@s/europeo_NCCP0
- o\a/o_GCS consumidor\a/consumidor_NCCS0
- destinatarios\as/destinatario_AOCP

3. *Fragmentos ilustrativos*

Seguen algúns fragmentos ilustrativos dos principios metodolóxicos expostos neste artigo tirados do Corpus CTAG.

```
<frase>^A/O_GFS ^expansión/expansión_NCFS0 ^dos/de_S ^∼/o_GMP ^cultivos/cultivo_NCMP0 ^transxénicos/transxénico_AOMP ^ameaza/ameazar_VIP3S00 ^a/o_GFS ^diversidade/diversidade_NCFS0 ^xenética/xenético_AOFS ^pola/por_S ^∼/o_GFS ^simplificación/simplificación_NCFS0 ^dos/de_S ^∼/o_GMP ^sistemas/sistema_NCMP0 ^de/de_S ^cultivos/cultivo_NCMP0 ^e/e_CC ^a/o_GFS ^promoción/promoción_NCFS0 ^da/de_S ^∼/o_GFS ^erosión/erosión_NCFS0 ^xenética/xenético_AOFS ^./._Fp </frase>
```

Méndez, Lucía, "Queres comer alimentos transxénicos?". *Terra: Boletín da Federación Ecoloxista Galega*, 4, 1999.

```
<frase>^Por exemplo/Por exemplo_R0 ^non/non_R0 ^podemos/poder_VIP1P00 ^dicir/dicir_VN00000 ^que/que_CS ^Galicia/Galicia_NP000 ^sexa/ser_VSP3S00 ^moi/moi_R0 ^diversa/diverso_AOFS ^en/en_S ^aves/ave_NCFP0 ^,/,_Fc ^lévanse/levar_VIC3P00 ^∼/se_P3CNR0
```

```
^rexistradas/rexistrar_VP00PF0 ^unhas/un_IFP0
^250/250_Z ^habituais/habitual_A0CP ^ó/a_S
^∼/o_GMS ^longo/longo_A0MS ^dun/de_S ^∼/un_IMS0
^ano/ano_NCMS0 ^como moito/como moito_R0 ^,/,_Fc
^mentres que/mentres que_CS ^en/en_S ^toda/todo_IFS0
^Europa/Europa_NP000 ^hai/hai_VIP3S00 ^un-
has/un_IFP0 ^500/500_Z ^especies/especie_NCFP0
^e/e_CC ^en/en_S ^países/país_NCMP0 ^como/como_CS
^Perú/Perú_NP000 ^a/o_GFS ^cifra/cifra_NCFS0
^ascende/ascender_VIP3S00 ^a/a_S ^máis/máis_R0
^de/de_S ^1600/1600_Z ^para/para_S ^un/un_IMS0
^total/total_NCMS0 ^de/de_S ^9000/9000_Z
^aves/ave_NCFP0 ^de/de_S ^diferentes/diferente_A0CP
^especies/especie_NCFP0 ^existentes/existente_A0CP
^no/en_S ^∼/o_GMS ^planeta/planeta_NCMS0 ^./._Fp
</frase>
```

Vázquez Pumariño, Xabier, *Que é a biodiversidade.* Documento electrónico dispoñible na web da Asociación para a Defensa Ecolóxica de Galiza (ADEGA).

```
<frase>^Galicia/Galicia_NP000 ^é/ser_VIP3S00
^a/o_GFS ^primeira/primeiro_M00FS ^Comu-
nidade/Comunidade_NCFS0 ^Autónoma/Autónomo_A0FS
^pesqueira/pesqueira_A0FS ^do/de_S ^∼/o_GMS
^Estado/estado_NCMS0 ^español/español_A0MS
^,/,_Fc ^o/o_GMS ^sector/sector_NCMS0
^pesqueiro/pesqueiro_A0MS ^represen-
ta/representar_VIP3S00 ^o/o_GMS ^8/8_Z ^%/%_Ft
^do/de_S ^∼/o_GMS ^PIB/PIB_NCMS0 ^e/e_CC
^o/o_GMS ^5/5_Z ^%/%_Ft ^da/de_S ^∼/o_GFS
^poboación/poboación_NCFS0 ^activa/activo_A0FS
^,/,_Fc ^estas/este_DFP ^cifras/cifra_NCFP0 ^a
pesar de/a pesar de_CS ^estar/estar_VN00000
^en consonancia/en consonancia_R0 ^coa/con_S
^∼/o_GFS ^importancia/importancia_NCFS0
^do/de_S ^∼/o_GMS ^litoral/litoral_A0CS
^a/a_S ^nivel/nivel_NCMS0 ^mundial/mundial_A0CS
^,/,_Fc ^o/o_GMS ^40/40_Z ^%/%_Ft ^da/de_S
^∼/o_GFS0 ^poboación/poboación_NCFS0 ^do/de_S
^∼/o_GMS ^mundo/mundo_NCMS0 ^vive/vivir_VIP3S00
^nas/en_S ^∼/o_GFP ^zonas/zona_NCFP0
^costeiras/costeiro_A0FP ^,/,_Fc ^pre-
senta/presentar_VIP3S00 ^unhas/un_IFP0
^cifras/cifra_NCFP0 ^moi/moi_R0 ^por/por_S ^en-
riba/enriba_R0 ^de/de_S ^calquera/calquera_INS0
^dos/de_S ^∼/o_GMP ^outros/outro_IMP0 ^paí-
ses/país_NCMP0 ^comunitarios/comunitario_A0MP
^./._Fp </frase>
```

López Fernández, Alfredo, *Estatus dos pequenos cetáceos da plataforma de Galicia.* Tese de doutoramento, Universidade de Santiago de Compostela, 2003.

4. Conclusións

Neste artigo presentamos as bases para a anotación lingüística (etiquetaxe categorial e lematización) do Corpus CTAG (Corpus Técnico Anotado do Galego) da Universidade de Vigo. Aínda que se trata dun proxecto en curso, algúns dos seus resultados xa se poden consultar libremente en Internet (Gómez Guinovart, 2006-2009) mediante unha interface web de consulta accesible en <http://sli.uvigo.es/CTAG/> que dá acceso a unha sección do corpus de máis de 2 millóns de palabras, constituída por textos pertencentes aos eidos da ecoloxía e das ciencias ambientais. Ao remate do proxecto, está prevista a dispoñibilización en Internet do resultado da anotación morfosintáctica da totalidade do Corpus Técnico do Galego.

Referencias

Alegría Loinaz, Iñaki, Iñaki Arantzabal, Mikel L. Forcada, Xavier Gómez Guinovart, Lluis Padró, José Ramom Pichel Campos, e Josu Waliño. 2006. Opentrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, 37:357–358.

Armentano Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí Bellot, Mikel L. Forcada, Mireia Ginestí Rosell, Sergio Ortiz Rojas, Juan Antonio Pérez Ortiz, Gema Ramírez Sánchez, Felipe Sánchez Martínez, e Miriam A. Scalco. 2006. Open-source portuguese-spanish machine translation. En *Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006)*, páxinas 50–59, Itatiaia, Rio de Janeiro.

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

Civit, Montserrat. 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español.* SEPLN (Colección Monografías, 3), Alacante.

Gómez Clemente, Xosé María e Xavier Gómez Guinovart, editores. 2006-2009. *Corpus Técnico do Galego.* Universidade de Vigo, Vigo. <http://sli.uvigo.es/CTG/>.

Gómez Guinovart, Xavier, editor. 2006-2009. *Corpus Técnico Anotado do Galego.* Universidade de Vigo, Vigo. <http://sli.uvigo.es/CTAG/>.

Leech, Geoffrey e Andrew Wilson. 1996. Recommendations for the morphosyntactic annotation of corpora. Eagles guidelines. <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>.

Real Academia Galega e Instituto da Lingua Galega. 2003. *Normas ortográficas e morfolóxicas do idioma galego.* RAG/ILG, Santiago de Compostela.

Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português

Eloize Rossi Marques Seno, Maria das Graças Volpe Nunes
NILC – ICMC – Universidade de São Paulo
São Carlos – SP, Brasil
{eloize,gracan}@icmc.usp.br

Resumo

A fusão de sentenças é uma tarefa que consiste em produzir, a partir de um conjunto de sentenças relacionadas, uma única sentença que resume as informações comuns apresentadas no conjunto. Essa tarefa é de grande interesse em diversas aplicações do Processamento de Língua Natural (PLN), tais como a Sumarização Automática, a Tradução Automática, os sistemas de Perguntas e Respostas, entre outros. No entanto, um dos principais desafios da fusão consiste em identificar as informações comuns entre as sentenças relacionadas. Este trabalho apresenta um sistema baseado em conhecimento lexical, sintático, semântico e em algumas regras de parafraseamento que permite o reconhecimento de seqüências de palavras distintas, mas com o mesmo significado em sentenças comparáveis do Português. Os experimentos realizados com o sistema mostraram um desempenho de 87% de Precisão, 83% de Cobertura e 85% de Medida-f. Os resultados estão de acordo com outros trabalhos reportados na literatura para outras línguas.

1. Introdução

A fusão de sentenças é uma tarefa de geração de texto a partir de texto (*text-to-text generation*, em inglês) que, dadas duas ou mais sentenças semanticamente relacionadas como entrada, produz uma nova sentença de saída, preservando as informações comuns entre elas (Barzilay, 2003; Barzilay and Mckeown, 2005). A fusão de sentenças é uma área de pesquisa emergente em Processamento de Língua Natural (PLN) e é motivada por aplicações práticas tais como a Tradução Automática (Pang et al., 2003), a Sumarização Automática (vide Barzilay and Mckeown, 2005), os sistemas de Perguntas e Respostas (vide Marsi and Krahmer, 2005; Krahmer et al. 2008), entre outras. Na sumarização multidocumento, por exemplo, o processo de fusão de informações comuns é de grande relevância para eliminar a redundância de informações nos sumários, especialmente no que diz respeito aos métodos de sumarização extrativos que identificam as sentenças (ou parágrafos) mais importantes de um conjunto de documentos e as extraem para compor o sumário. A repetição de informações influencia diretamente a qualidade dos sumários, prejudicando, principalmente, a coesão e a coerência. A fusão de várias sentenças que expressam uma mesma informação em uma única sentença pode minimizar esses problemas, eliminando a repetição de informações e, conseqüentemente, melhorando a textualidade dos sumários.

A Figura 1 apresenta um exemplo de sentença produzida a partir da fusão automática de três sentenças comparáveis sobre um mesmo assunto, porém de fontes distintas, extraídas do corpus de trabalho (Seção 3.1). No exemplo da figura, a sentença resultante da fusão corresponde à intersecção das sentenças [1], [2] e [3] e expressa somente os conceitos comuns a todas elas (em negrito).

[1] O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 da terça-feira e chegou a São Paulo às 18h45. **[2] A aeronave da TAM Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16** com destino a Congonhas. **[3] Um Airbus A320** com capacidade para 170 passageiros **partiu de Porto Alegre (RS) às 17h16** com destino a Congonhas.
Fusão das sentenças [1], [2] e [3]: O Airbus A320, vôo JJ 3054, partiu de Porto Alegre (RS) às 17h16.

Figura 1: Exemplo de Fusão de Sentenças[1]

[1] Essas sentenças foram identificadas automaticamente pelo sistema de clustering SiSPI (vide Seção 3.1), a partir de um conjunto de cinco documentos sobre o acidente envolvendo o Airbus A320, vôo JJ 3054, da TAM.

Nos trabalhos existentes na literatura (por exemplo, Pang et al. 2003; Barzilay and Mckeown, 2005 e Marsi and Krahmer, 2005) a fusão de sentenças é comumente dividida em três etapas, a saber: i) identificação de informações comuns, ii) fusão de informações e iii) linearização. A primeira etapa consiste em reconhecer informações semanticamente similares (por exemplo, paráfrases e sinônimos) que se repetem nas sentenças. A segunda etapa consiste em escolher os itens lexicais que irão compor a nova sentença e determinar o modo como eles serão combinados na sentença. A última etapa, por sua vez, consiste em realizar em língua natural a sentença obtida a partir da etapa anterior e envolve, portanto, aspectos gramaticais da sentença. A identificação dos elementos que expressam informações comuns e a combinação desses elementos para a geração da nova sentença consistem no maior desafio na construção de algoritmos de fusão.

Neste artigo apresenta-se um alinhador de conceitos similares baseado em informações lexicais, sintáticas e semânticas que permite o reconhecimento de informações comuns em sentenças comparáveis do português. Com base no alinhamento de duas ou mais árvores de dependência sintática que representam cada sentença de um conjunto de sentenças comparáveis, constrói-se uma floresta a partir da união de sentenças previamente alinhadas (ou seja, unindo as informações comuns a cada sentença). A união de todas as sentenças em uma única estrutura de dependência sintática possibilita que um subseqüente módulo de fusão e linearização gere todas as sentenças possíveis a partir da floresta. Um modelo probabilístico de língua é utilizado, posteriormente, para auxiliar na seleção da melhor sentença gerada, como proposto no trabalho de Barzilay and Mckeown (2005).

Em um trabalho anterior (Seno and Nunes, 2008a) foi apresentada uma versão preliminar do alinhador para a identificação de informações comuns entre pares de sentenças comparáveis (os resultados obtidos são sumarizados na Seção 4). No presente trabalho, são apresentadas diversas modificações realizadas ao sistema, por exemplo, a possibilidade de alinhamento de um conjunto de sentenças (e não somente de pares de sentenças), mudanças na estratégia de alinhamento, a inclusão de novos conhecimentos lingüísticos e modificações no pré-processamento, que resultaram em um melhor desempenho do sistema (vide Seção 4).

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta alguns trabalhos correlatos de alinhamento de informações comuns; a Seção 3 apresenta o alinhador proposto; a Seção 4 mostra alguns experimentos realizados e, por fim, a Seção 5 apresenta as conclusões e algumas possibilidades de trabalhos futuros.

2. Trabalhos Relacionados

As abordagens de alinhamento de informações similares existentes na literatura se distinguem em dois aspectos principais: i) quanto ao tipo de sentença de entrada e ii) quanto ao tipo de conhecimento usado. Quanto aos tipos de sentenças têm-se as sentenças comparáveis, que se referem a um mesmo fato ou evento, porém são de fontes de informação diferentes, e as sentenças paralelas, que são traduções distintas de uma mesma fonte para uma mesma língua alvo. Em relação aos tipos de conhecimento utilizados para a identificação de conceitos similares, destacam-se as informações sintáticas, por exemplo, as relações de dependência entre os constituintes sintáticos, as relações semânticas, os léxicos de sinônimos e de paráfrases.

Pang et al. (2003) alinham árvores sintáticas de sentenças paralelas usando somente informações de *part-of-speech* (*POS*). As palavras com o mesmo POS são tratadas como paráfrases. Embora essa abordagem tenha se mostrado satisfatória para trabalhar com sentenças paralelas, somente informações de POS não são suficientes para o reconhecimento de conceitos comuns em sentenças comparáveis, uma vez que as estruturas sintáticas dessas sentenças nem sempre são similares, como é o caso das sentenças paralelas. Já Shen et al. (2006) consideram, além das informações de POS, os traços de dependência dos constituintes sintáticos. O alinhamento ocorre somente entre palavras lexicalmente similares que compartilham o mesmo POS e o mesmo traço de dependência.

Ao contrário desses trabalhos, outras abordagens ignoram completamente as informações de POS.

Em Marsi and Krahmer (2005), por exemplo, o alinhamento envolvendo sentenças paralelas é baseado apenas na similaridade de suas correspondentes estruturas de dependência sintática e em relações semânticas (por exemplo, *restates* e *intersects*). O alinhamento entre duas palavras só se realiza se houver uma relação semântica entre elas. Os autores relatam uma precisão de 86% e uma cobertura de 84% (isto é, 85% de Medida-f) do sistema[2]. Contudo, a principal limitação desse método está na dificuldade de se construir *parsers* semânticos. Outra limitação, que também se pode observar nos trabalhos de Pang et al. (2003) e Shen et al. (2006), é que essas abordagens não tratam o reconhecimento de paráfrases multipalavras. Como exemplos desse tipo de paráfrases tem-se *mercado moscovita* com *mercado Cherskisov de Moscou* e *capital da Rússia* com *capital russa*.

Já em Barzilay and Mckeown (2005), o alinhamento de informações similares entre sentenças comparáveis ocorre em nível de palavras e de *phrases*. Assim como em Marsi and Krahmer (2005), os autores também consideram a similaridade entre as estruturas de dependência sintática das sentenças. A similaridade entre palavras é obtida a partir de um conjunto de sinônimos, enquanto a similaridade entre multipalavras é determinada com o uso de um léxico de paráfrases, induzido automaticamente a partir de corpora. Entretanto, a construção de um léxico representativo de paráfrases requer um grande volume de dados de treinamento (ou seja, de sentenças parafrásticas), um recurso praticamente inexistente para a maioria das línguas.

As paráfrases multipalavras são as mais freqüentes, principalmente em sentenças comparáveis (vide Seção 3.2) e são muito difíceis de se tratar automaticamente.

O método descrito neste trabalho faz uso de regras de parafraseamento, identificadas a partir da análise de corpora (vide Seção 3), e de conhecimentos lexical, sintático (ou seja, POS e traços de dependência) e semântico (isto é, relações de sinonímia) que possibilitam a identificação de palavras e multipalavras que conduzem informações semanticamente similares em sentenças do português.

3. Reconhecimento de Informações Comuns

Esta seção apresenta o alinhador de informações comuns, destacando as melhorias realizadas em relação à primeira versão do sistema. Antes, porém, as subseções 3.1 e 3.2 descrevem a construção do corpus de trabalho e a formulação das regras de parafraseamento a partir da análise do corpus, respectivamente.

3.1 Construção do Corpus

Para a construção do corpus de sentenças comparáveis, foram coletadas manualmente 50 coleções de documentos a partir de diversas agências de notícias brasileiras disponíveis na web. O corpus compreende textos de diferentes domínios, tais como ciência, cotidiano, esporte, mundo e política. Cada coleção é composta por aproximadamente 4 documentos relacionados a um mesmo assunto, totalizando 71 documentos e 1.153 sentenças em todo o corpus. Todos os documentos de uma mesma coleção foram publicados em uma mesma data, o que assegura uma maior similaridade do conteúdo apresentado nesses documentos.

Após a coleta dos textos, cada coleção de documentos foi submetida a um processo de agrupamento de sentenças, para a identificação das sentenças comparáveis de cada coleção. Para esse processo foi desenvolvido o sistema SiSPI (Seno and Nunes, 2008b), baseado em um método de agrupamento incremental e não supervisionado conhecido por *Single-pass* (Van Rijsbergen**,** 1979). A abordagem incremental tem a vantagem de não ser baseada em treinamento e, portanto, não requer grandes conjuntos de dados.

O Single-pass, como o próprio nome sugere, requer um único passo seqüencial sob todo o conjunto de sentenças a ser agrupado. Dado um conjunto de documentos como entrada, o primeiro grupo é criado selecionando-se a primeira sentença do primeiro documento do conjunto. A cada iteração, o algoritmo verifica se a nova sentença de entrada deve pertencer a algum grupo já existente ou se um novo grupo

[2] Outros trabalhos reportados aqui não relatam resultados sobre o processo de alinhamento de informações em específico, já que esse é um processo intermediário.

deve ser criado para aquela sentença. Essa decisão é baseada em uma condição previamente estabelecida para a função de similaridade adotada, ou seja, um limiar de similaridade. Duas funções distintas foram implementadas no sistema, para calcular a distância semântica entre uma sentença e um grupo. A primeira função é baseada na medida *Word-Overlap* (*Wol*) (Radev et al., 2008), que calcula o número de palavras em comum entre uma sentença *S* e um grupo *C*, normalizado pelo total de palavras de *S* e *C* (Fórmula 1). O valor de similaridade da *Wol* varia de 0 a 0,5. Quanto mais próximo de 0,5, maior é a similaridade entre a sentença e o grupo.

(1)

$$Wol(S,C) = \frac{\#PalavrasComuns(S,C)}{(|S| + |C|)}$$

A segunda função de similaridade é baseada na distância do co-seno (Salton and Allan, 1994) aplicada entre o vetor de freqüência de termos de uma sentença e o vetor que representa os termos mais importantes de um grupo, denominado centróide. O valor de similaridade dessa função varia de 0 a 1. Quanto mais próximo de 1, maior é a similaridade entre a sentença e o grupo.

O centróide de um grupo de sentenças é determinado a partir de duas medidas estatísticas. A primeira medida é uma adaptação do *TF-IDF* (*Term Frequency Inverse Document Frequency*) (Salton and Allan, 1994). O valor do *TF-IDF* de uma palavra *w* pertencente a um grupo *c*, denotado por *TF-IDF(w,c)*, é dado pela Fórmula 2, onde *TF(w,c)* representa a freqüência da palavra *w* no grupo *c*. Quanto maior o valor de *TF*, mais representativa do grupo a palavra *w* é. A freqüência de documento inversa de *w*, denotada por *IDF(w)*, é dada pela Fórmula 3, onde |*C*| representa o total de sentenças de toda a coleção de documentos e *DF(w)* representa o total de sentenças da coleção que contem *w*.

(2)

$$TF\text{-}IDF(w,c) = TF(w,c) * IDF(w)$$

(3)

$$IDF(w) = 1 + \log (|C| / DF(w))$$

A segunda medida usada para calcular o centróide de um grupo é a *TF-ISF* (*Term Frequency Inverse Sentence Frequency*) (Larocca Neto et al., 2000). Essa medida é similar ao *TF-IDF*, exceto que ela calcula a freqüência de sentença inversa para um grupo em específico, ao invés de calcular para todos os documentos da coleção. A freqüência de sentença inversa de *w*, denotada por *ISF(w)*, é dada pela Fórmula 4, onde |*C*| representa o total de sentenças do grupo e *SF(w)* é o total de sentenças do grupo que contêm *w*.

(4)

$$ISF(w) = 1 + \log (|C| / SF(w))$$

Para que uma palavra seja representativa de um determinado grupo, ela deve ter um alto valor de *TF* e um alto valor de *ISF* (ou *IDF*) e, portanto, um alto valor de *TF-ISF* (ou *TF-IDF*).

Para avaliação do método foram selecionadas aleatoriamente 20 coleções de documentos do corpus. Visando construir um corpus de referência de sentenças similares, cada sentença de uma coleção foi manualmente classificada, isto é, associada a um nome de grupo (daqui a diante, a classificação manual será referenciada por *classes* e o agrupamento automático será referenciado por *grupos*). Para a classificação manual adotou-se o conceito de similaridade proposto por Hatzivassiloglou et al. (1999) para a mesma tarefa de identificação de sentenças semanticamente similares. De acordo com Hatzivassiloglou et al., duas sentenças são semanticamente similares se elas se referem a um mesmo objeto ou evento e i) o objeto realiza a mesma ação em ambas as sentenças, ou ii) é sujeito da mesma descrição. Considere, por exemplo, as sentenças (a), (b) e (c), extraídas do corpus. Apesar de todas as sentenças se referirem a explosão de uma bomba caseira, as sentenças (a) e (b) focam na explosão ocorrida no Ministério Público, enquanto que (c) se refere à explosão ocorrida na Secretaria de Estado da Fazenda. Nesse caso, somente (a) e (b) são consideradas similares.

(a) Uma bomba caseira foi atirada contra a sede do Ministério Público (MP).
(b) Uma bomba caseira foi jogada contra o prédio do Ministério Público, na capital do estado.
(c) Uma bomba caseira atingiu o prédio da Secretaria de Estado da Fazenda, localizado na avenida Rangel Pestana, ao lado do Poupatempo Sé.

O desempenho do método de agrupamento foi avaliado usando as medidas de Precisão, Cobertura e Medida-f, redefinidas no domínio de clustering (vide Funch et al., 2003 e Steinbach et al., 2000).

Seja N o número total de sentenças a serem agrupadas, K o conjunto de classes, C o conjunto de grupos e n_{ij} o número de sentenças da classe $k_i \in K$ que estão presentes no grupo $c_j \in C$. A Precisão e a Cobertura de k_i e c_j, denotada por $P(k_i,c_j)$ e $C(k_i,c_j)$, respectivamente, são dadas pelas fórmulas 5 e 6. A Precisão representa o número de sentenças do grupo c_j que pertence a classe k_i e indica o quão o grupo c_j é homogêneo em relação a classe k_i. Similarmente, a Cobertura é dada pelo total de sentenças da classe k_i que estão presentes no grupo c_j, representando, assim, a completude do grupo c_j em relação à classe k_i. Por fim, a Medida-f mede a qualidade do grupo c_j em descrever a classe k_i , calculando a média harmônica entre a Precisão e a Cobertura.

(5)

$$P(k_i, c_j) = n_{ij} / |c_j|$$

(6)

$$C(k_i, c_j) = n_{ij} / |k_i|$$

(7)

$$F(k_i, c_j) = \frac{(2 * C(k_i, c_j) * P(k_i, c_j))}{C(k_i, c_j) + P(k_i, c_j))}$$

A Medida-f para cada classe de todo o conjunto de dados se baseia no grupo que melhor descreve cada classe k_i, ou seja, no grupo que maximiza o valor de $F(k_i,c_j)$ para todo j. Assim, o valor de Medida-f global de uma solução de agrupamento S, denotado por $F(S)$, é dado pela Fórmula 8. O valor de $F(S)$ varia de 0 (pior) a 1 (melhor).

(8)

$$F(S) = \sum_{k_i \in K} \frac{|k_i|}{N} \max_{c_j \in C} \{F(k_i, c_j)\}$$

Além das medidas de desempenho apresentadas anteriormente, foram usadas ainda duas métricas para avaliar a qualidade dos grupos de sentenças similares obtidos automaticamente. A primeira métrica, chamada Entropia (Steinbach et al., 2000), mede a organização de cada grupo, ou seja, como as várias classes de sentenças estão distribuídas em cada grupo. A solução de agrupamento ideal será aquela na qual todos os grupos contêm sentenças de uma única classe. Nesse caso, o valor de Entropia será 0. O cálculo da Entropia é baseado na distribuição de classes de cada grupo e é exatamente o que é feito pela medida de Precisão. Em outras palavras, a Precisão representa a probabilidade de uma sentença escolhida aleatoriamente de um grupo c_j pertencer a classe k_i. Desse modo, a Entropia de um grupo c_j, denotada por $E(c_j)$, é dada pela Fórmula 9. A Entropia global de uma solução de agrupamento S, denotada por $E(S)$, é dada pela soma das entropias de cada grupo c_j ponderada pelo tamanho do grupo, conforme a Fórmula 10. Quanto menor o valor de $E(S)$, melhor é a solução de agrupamento.

(9)

$$E(c_j) = -\sum_{k_i} P(k_i, c_j) \log P(k_i, c_j)$$

(10)

$$E(S) = \sum_{c_j} \frac{|c_j|}{N} E(c_j)$$

A segunda métrica usada para medir a qualidade dos grupos é a Pureza (Rosell et al., 2004), que mede o quão puro cada grupo de sentença é. Em outras palavras, a Pureza representa o percentual da classe mais freqüente de cada grupo. Assim, a Pureza de um grupo *cj*, denotada por *P(cj)*, é definida pela classe k_i que maximiza a Precisão do grupo *cj* (Fórmula 11). A Pureza global de uma solução de agrupamento S, denotada por $P(S)$, é dada pela soma dos valores de Pureza de cada grupo *cj* ponderada pelo tamanho do grupo (Fórmula 12). O valor de $P(S)$ varia de 0 (pior) a 1 (melhor).

(11)

$$P(c_j) = \max_{k_i} \{P(k_i, c_j)\}$$

(12)

$$P(S) = \sum_{c_j \in C} \frac{|c_j|}{N} P(c_j)$$

A fim de identificar o limiar de similaridade que melhor define o corpus de trabalho, cada função de similaridade foi avaliada com diferentes configurações de limiares, variando de 0,1 a 1, com exceção da função *Word Overlap* que varia de 0,1 a 0,5. A Tabela 1 apresenta os resultados

Tabela 1. Resultados médios obtidos com cada medida de avaliação para diferentes limiares de similaridade

	Similaridade	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
TF-IDF	**Entropia**	0,843	0,287	0,096	0,037	0,016	0,005	0,004	0,003	0,002	0,001
	Medida-f	0,603	0,814	**0,886**	**0,860**	0,841	0,828	0,812	0,799	0,775	0,736
	Pureza	0,549	0,808	0,907	**0,934**	**0,945**	0,945	0,942	0,940	0,941	0,938
TF-ISF	**Entropia**	1,759	0,900	0,319	0,101	0,043	0,013	0,004	0,003	0,002	0,002
	Medida-f	0,348	0,603	0,805	**0,864**	0,856	0,843	0,828	0,813	0,798	0,786
	Pureza	0,315	0,564	0,804	0,913	**1,000**	0,950	0,954	0,953	0,952	0,951
Word	**Entropia**	0,572	0,079	0,010	0,000	0,001	–	–	–	–	–
Overlap	**Medida-f**	0,695	**0,860**	0,838	0,809	0,786	–	–	–	–	–
	Pureza	0,654	0,908	**0,946**	0,943	0,941	–	–	–	–	–

obtidos por cada função de similaridade para cada medida de avaliação. No que diz respeito aos modelos *TF-IDF* e *TF-ISF*, os resultados foram gerados usando um centróide de tamanho 15, ou seja, considerando-se as 15 palavras mais importantes de cada grupo no cálculo da similaridade entre uma sentença e um grupo qualquer. O tamanho ideal do centróide foi obtido automaticamente a partir de experimentos com o corpus (Seno and Nunes, 2008b).

De acordo com a Tabela 1, os valores de Entropia melhoram consideravelmente na medida em que se aumenta o limiar de similaridade para todos os casos. O mesmo ocorre para os valores de Medida-f e Pureza, mas até certo ponto. A Medida-f alcança o seu valor máximo com um limiar de 0,2, 0,3 e 0,4 para *Word Overlap*, *TF-IDF* e *TF-ISF*, respectivamente. Em relação à Pureza, os valores melhoram até um limiar de 0,3 para *Word Overlap*, e um limiar de 0,5 para os modelos *TF-IDF* e *TF-ISF*.

Especificamente em relação aos valores de Entropia e de Pureza, esses se justificam pelo fato de que o número de grupos cresce na proporção em que se aumenta o limiar de similaridade, de modo que eles se tornam mais homogêneos, ou seja, a variedade de classes em cada grupo tende a diminuir. Além disso, como há muitas sentenças não similares no corpus, a tendência é de que esses valores melhorem ainda mais, uma vez que muitos grupos contêm somente uma sentença.

Em relação aos valores de Medida-f, apesar da tendência dos grupos de se tornarem mais homogêneos (aumentando a Precisão), à medida que o limiar de similaridade aumenta, torna-se mais difícil identificar sentenças semanticamente equivalentes, mas lexicalmente muito distintas. Dessa forma, os valores de Cobertura tendem a diminuir, prejudicando o desempenho global.

Em termos de bom desempenho do método de agrupamento e qualidade dos grupos de sentenças, o modelo *TF-IDF* com similaridade 0,4 (daqui a diante *TF-IDF-0,4*) se mostrou mais apropriado para o propósito deste trabalho. Além de obter uma Medida-f de 86% (a melhor Medida-f foi de 88,6% (*TF-IDF-0,3*)), ele obteve bons valores de Entropia (isto é, 0,037) e de Pureza (isto é, 93,4%), principalmente se comparado aos valores obtidos pelo *TF-IDF-0,3*, *TF-ISF-0,4* e *Word-Overlap-0,2*. Além do mais, o desvio padrão obtido pelo *TF-IDF-0,4* (0,07 para Medida-f, 0,06 para Pureza e 0,05 para Entropia) foi menor do que o obtido para o *TF-IDF-0,3* (0,08 para Medida-f, 0,07 para Pureza e 0,10 para Entropia), *TF-ISF-0,4* (0,09 para Medida-f, 0,08 para Pureza e 0,09 para Entropia) e *Word-Overlap-0,2* (0,08 para Medida-f, 0,06 para Pureza e 0,07 para Entropia). Portanto, para a construção do corpus de sentenças comparáveis utilizou-se o modelo *TF-IDF-0,4*.

Visando facilitar a formulação das regras de parafraseamento (Seção 3,2), para cada grupo identificado foram obtidas todas as possíveis combinações de pares de sentenças comparáveis, resultando aproximadamente em 670 pares em todo corpus.

3.2 Formulação de Regras de Parafraseamento

Para a formulação das regras de parafraseamento foram selecionados aleatoriamente 30 pares de sentenças comparáveis do corpus. Cada par foi

analisado e um total de 81 paráfrases foram identificadas manualmente em todo conjunto. A definição de paráfrases adotada nessa análise segue aquela proposta por Hoey (1991) em que duas seqüências distintas de palavras são ditas paráfrases se uma delas puder ser substituída pela outra, em um dado contexto, sem alterar significativamente o sentido do texto.

A Tabela 2 mostra alguns exemplos de ocorrência de paráfrases no corpus. Aproximadamente 26% dos casos identificados são paráfrases lexicais (isto é, ocorrem entre palavras), por exemplo, (a), (g) e (h). Os outros 74% das paráfrases são multipalavras (por exemplo, (b), (c), (d), (f) e (j)) ou ocorrem entre uma palavra e um segmento multipalavras (por exemplo, (e), (i)).

a. colisão ⇔ choque
b. tucano Geraldo Alckmin ⇔ candidato tucano Geraldo Alckmin
c. capital russa ⇔ capital da Rússia
d. direção da Câmara ⇔ Mesa Diretora da Câmara
e. acordo ⇔ acordo financeiro
f. mercado moscovita ⇔ mercado Cherskisov de Moscou
g. membro ⇔ integrante
h. arrasou ⇔ venceu
i. grupo ⇔ grupo criminoso
j. liderança do Grupo B ⇔ liderança do Grupo B da Liga
l. não chegaram a obter ⇔ não alcançaram

Tabela 2: Exemplo de paráfrases

27 regras de parafraseamento foram formuladas a partir da análise de corpus. Alguns exemplos de regras são apresentados na Tabela 3 (onde ADJ: adjetivo; ART: artigo; ADV: advérbio; V: verbo; N: substantivo; PRP: preposição; PROP: nome próprio; ?: indica zero ou uma ocorrência; |: indica alternativa (operador ou) e os números indicam as unidades lexicalmente similares). A regra R1 cobre os exemplos (c) e (f) da Tabela 2; R2 cobre os exemplos (e) e (i); R3 cobre o exemplo (b); R4 cobre o exemplo (l) e R5 cobre os exemplos (d) e (j). Para os exemplos (a), (g) e (h) não há regras, uma vez que são paráfrases lexicais. É importante observar que todas as regras preveem ao menos uma ocorrência de palavras similares em ambos os segmentos, conforme indicam os números subscritos em cada regra.

No caso de R5, por exemplo, dois segmentos S_1 e S_2 são considerados paráfrases se S_1 iniciar com um substantivo (N) e uma preposição (PRP), acompanhada ou não de um artigo (ART?), e finalizar com um nome próprio ou um substantivo ($PROP_1|N_1$) e S_2 iniciar com um nome próprio ou um substantivo (PROP|N) e uma preposição (PRP), que pode ser acompanhada ou não de um artigo (ART?), seguido de um outro nome próprio (similar ao de S_1, se existir) ou de um outro substantivo ($PROP_1|N_1$) que, por sua vez, pode ou não ser acompanhado por uma preposição (PRP?), um artigo (ART?) e mais um nome próprio ou substantivo ((PROP|N)?). A paráfrase (d) da Tabela 2, por exemplo, inicia-se com um substantivo (*direção*), seguido de um artigo e uma preposição (*de* + *a* = *da*), e termina com um nome próprio (*Câmara*). A sua correspondente, por sua vez, é iniciada por um nome próprio (*Mesa Diretora*), acompanhado de um artigo mais uma preposição (*de* + *a* = *da*), e finalizado por outro nome próprio (*Câmara*).

R1. N_1 ADJ ; N_1 PROP? PRP ART? PROP
R2. N_1 ; N_1 ADJ
R3. N $PROP_1$; N ADJ $PROP_1$
R4. ADV? V PRP V_1 ; ADV? V_1
R5. N PRP ART? ($PROP_1|N_1$) ; (PROP|N) PRP ART? ($PROP_1|N_1$) PRP? ART? (PROP|N)?

Tabela 3: Exemplo de regras de parafraseamento

3.3 Alinhamento

O alinhador de conceitos comuns é baseado em informações de *part-of-speech* (*POS*) e em relações de dependência sintática fornecidas pelo *parser* Palavras (Bick, 2000). Dessa maneira, as sentenças comparáveis são primeiramente processadas pelo *parser*, de modo a obter todo o conhecimento sintático necessário de entrada para o alinhador (vide Figura 2). Durante o processo de alinhamento, o sistema também faz uso da base de sinônimos Tep[3] (Maziero et al., 2008), desenvolvida no contexto do projeto Wordnet-Br (Dias-da-Silva et al., 2006), de um

[3] Disponível em: http://www.nilc.icmc.usp.br/tep2/download.htm (último acesso em 13/01/2009)

conjunto de regras de parafraseamento (Seção 3.2) e de uma *stoplist,* que permite a identificação das palavras irrelevantes ao alinhamento (vide Subseção 3.3.2). Como saída tem-se um conjunto de alinhamentos que representam as informações em comum entre as sentenças de entrada.

A versão preliminar do alinhador, descrita em Seno and Nunes (2008a), trabalha somente com pares de sentenças comparáveis. No atual sistema, é possível alinhar duas ou mais sentenças de entrada, conforme ilustra a Figura 2.

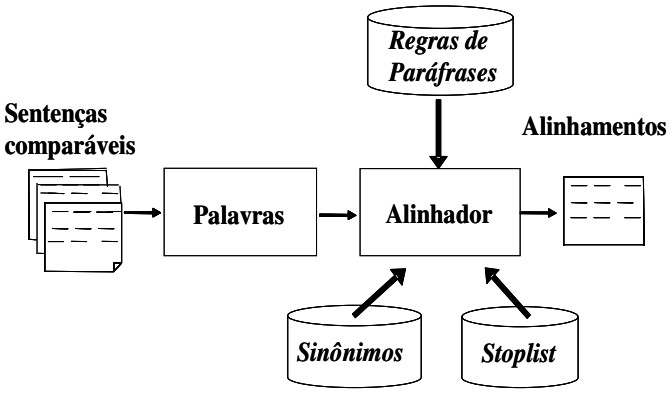


Figura 2: Ilustração do processo de alinhamento

A subseção a seguir descreve a etapa de pré-processamento das sentenças feita pelo Palavras para, então, apresentar o processo de alinhamento propriamente dito na Subseção 3.3.2.

3.3.1 Pré-processamento

O parser Palavras permite análises em diferentes formatos de saída, por exemplo, *Visl* e *TigerXML*, sendo que as informações de dependência sintáticas são obtidas apenas com o formato *Visl* (Bick, 2000). A Figura 3 apresenta um exemplo de análise de dependência sintática realizada pelo *parser* para a sentença "*O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 da terça-feira e chegou a São Paulo às 18h45,*" (sentença [1] da Figura 1). Os traços de dependência se realizam entre *tokens* e incluem relações entre sujeito e verbo, objeto e verbo, etc, No exemplo da figura, *Airbus A320* (*token* #2) é o sujeito (@SUBJ) do verbo (V) *partiu* (*token* #7) e #2->7 indica que o *token* #2 é dependente do *token* #7 (isto é, dependência entre sujeito e verbo)[4]. O *parser* também inclui o processo de lematização (os lemas de cada palavra estão apresentados entre colchetes).

```
O [o] <artd> DET M S @>N #1->2
Airbus=A320 [Airbus=A320] <V> PROP M S
@SUBJ> #2->7
$, #3->0
Vôo [vôo] <activity><np-close> N M S
@N<PRED #4->2
JJ=3054 [JJ=3054] <top> PROP M/F S
@APP #5->4
$, #6->0
partiu [partir] <predco><cjt-
head><fmc> <mv> V PS 3S IND VFIN @FS-
STA #7->0
de [de] PRP @<ADVL  #8->7
Porto=Alegre [Porto=Alegre] <civ> PROP
M S @P< #9->8
$, #10->0
a [a] <sam-> PRP @<ADVL #11->7
as [o] <-sam><artd> DET F P @>N #12-
>13
17h16 [17h16] <temp> N F P @P< #13->11
de [de] <sam-><np-close> PRP @N< #14-
>13
a [o] <artd><-sam> DET F S @>N #15->16
terça-feira [terça-feira] <temp> N F S
@P< #16->14
e [e] <co-fin><co-fmc><co-fin> KC @CO
#17->7
chegou [chegar] <nosubj><cjt><fmc><mv>
V PS 3S IND VFIN @FS-STA #18->7
a [a] PRP @<SA #19->18
São=Paulo [São=Paulo] <civ> PROP M S
@P< #20->19
a [a] <sam-> PRP @<ADVL  #21->18
as [o] <-sam><artd> DET F P @>N #22-
>23
18h45 [18h45] <temp> N F P @P< #23->21
$. #24->0
```

Figura 3: Análise de dependência sintática fornecida pelo Palavras (formato *Visl*)

Apesar de fornecer os traços de dependência entre os constituintes sintáticos, o formato *Visl* não fornece informações sobre os segmentos das sentenças como os sintagmas nominais e os sintagmas verbais, entre outros, Dessa forma, para recuperar as relações de dependência entre sintagmas, na versão preliminar do sistema (Seno and Nunes, 2008a) foram utilizadas algumas expressões regulares definidas com base nos traços de dependência entre os *tokens* (vide Figura 3).

[4] Vide http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html, para maiores informações sobre as etiquetas do Palavras (último acesso em 13/01/2009).

```
...
<terminals>
  <t id="s1_1" word="O" lemma="o" pos="art" morph="M S" sem="--" extra="--"/>
  <t id="s1_2" word="Airbus_A320" lemma="Airbus_A320" pos="prop" morph="M S"
sem="V" extra="--"/>
  <t id="s1_3" word="," lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
  <t id="s1_4" word="vôo" lemma="vôo" pos="n" morph="M S" sem="activity"
extra="np-close"/>
  <t id="s1_5" word="JJ_3054" lemma="JJ_3054" pos="prop" morph="M/F S" sem="--
" extra="top"/>
  <t id="s1_6" word="," lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
  <t id="s1_7" word="partiu" lemma="partir" pos="v-fin" morph="PS 3S IND VFIN"
sem="--" extra="predco predco  fmc mv"/>
   ...
</terminals>
<nonterminals>
  ...
  <nt id="s1_502" cat="np">
     <edge label="DN" idref="s1_1"/>
     <edge label="H" idref="s1_2"/>
     <edge label="DNc" idref="s1_503"/>
  </nt>
  <nt id="s1_503" cat="np">
     <edge label="H" idref="s1_4"/>
     <edge label="DNapp" idref="s1_5"/>
  </nt>
  ...
```

Figura 4: Exemplo de saída do Palavras no formato *TigerXML*

```
...
<tokens>
  <t id="1" word="O" lemma="o" pos="art" morph="M S" sem="--" extra="--"
traco="@>N " dep="2"/>
  <t id="2" word="Airbus_A320" lemma="Airbus_A320" pos="prop" morph="M S"
sem="V" extra="--" traco="@SUBJ> " dep="7"/>
  <t id="3" word="," lemma="--" pos="pu" morph="--" sem="--" extra="--"
traco="--" dep="--"/>
  <t id="4" word="vôo" lemma="vôo" pos="n" morph="M S" sem="activity"
extra="np-close" traco="@N<PRED  " dep="2"/>
  <t id="5" word="JJ_3054" lemma="JJ_3054" pos="prop" morph="M/F S" sem="--"
extra="top" traco="@APP " dep="4"/>
  <t id="6" word="," lemma="--" pos="pu" morph="--" sem="--" extra="--"
traco="--" dep="--"/>
  <t id="7" word="partiu" lemma="partir" pos="v-fin" morph="PS 3S IND VFIN"
sem="--" extra="predco predco  fmc mv" traco="@FS-STA " dep="0"/>
  ...
<phrases>
  <p id="502" phrase="1_2_3_4_5" pos-ph="S"/>
  ...
<dependencies>
  <d id="0" type="S-Verb" son="502" father="7"/>
   ...
```

Figura 5: Formato de entrada atual do alinhador com as relações de dependência entre *phrases*

Na versão atual do sistema, optou-se por modificar o formato dos arquivos de entrada, de modo a representar as relações de dependência entre os sintagmas. O novo formato, ilustrado na Figura 5, foi construído a partir de informações extraídas de duas saídas distintas do *parser* para a mesma sentença de entrada (sentença [1] da Figura 1), São eles: o *Visl* (Figura 3) e o *TigerXML* (Figura 4).

Enquanto o *Visl* fornece os traços de dependência entre os *tokens*, o *TigerXML* fornece as informações sobre os sintagmas das sentenças. No exemplo da Figura 4, o nó não-terminal *s1_502* (`nt id="s1_502"`) é um sintagma nominal (`cat="np"`) composto pelos *tokens* 1 e 2 (`idref="s1_1" e idref="s1_2"`), ou seja, "*o*" e "*Airbus_A320*", e por outro sintagma nominal (`id="s1_503"`) o qual é composto, por sua vez, pelos *tokens* 4 e 5 (`idref="s1_4" e idref="s1_5"`), ou seja, "*vôo*" e "*JJ_3054*". A partir do traço de dependência de cada *token* e da informação sobre qual sintagma ele pertence, é possível obter as relações de dependência entre sintagmas, como mostra o exemplo da Figura 5. Nesse exemplo, o sintagma nominal 502 (`id="502"`), que é composto pelos *tokens* de 1 a 5 (`phrase="1_2_3_4_5"`), ou seja, "*o Airbus_A320 , vôo JJ_3054*", é o sujeito (`pos-ph="S"`) da sentença e estabelece uma relação com o *token* 7 (`son="502" father="7"`), ou seja, "*partiu*", configurando a dependência entre sujeito e verbo (`type="S-Verb"`).

3.3.2 Estratégia de Alinhamento

Dado um conjunto de sentenças comparáveis como entrada (mínimo de duas sentenças), previamente processadas (conforme Figura 5), o algoritmo inicialmente identifica todos os alinhamentos possíveis entre as duas primeiras sentenças do conjunto. Então, as sentenças alinhadas são unidas em uma única estrutura de dependência sintática, denominada floresta. As demais sentenças são alinhadas uma a uma com a floresta e, incrementalmente, também são unidas a ela (isto é, ao término de cada alinhamento entre uma sentença e a floresta). Como resultado final, tem-se uma única estrutura de dependência sintática representando todas as sentenças do conjunto e as intersecções entre elas. A Figura 6 ilustra a floresta construída a partir da união de duas árvores de dependências sintáticas, correspondentes às sentenças "*O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 da terça-feira e chegou a São Paulo às 18h45,"* e "*A aeronave da TAM Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas,*" (sentenças [1] e [2] da Figura 1). As setas indicam as dependências entre cada nó terminal e seu nó pai. Por exemplo, o nó terminal *Porto Alegre* (Árvores 1 e 2) é dependente do nó não terminal *partir* e representa uma relação de dependência entre verbo (ver) e objeto (obj). As caixas de textos e as setas não tracejadas representam os nós alinhados, enquanto que as setas tracejadas indicam os nós sem alinhamento.

O alinhamento realizado é do tipo um-para-um, ou seja, cada segmento de uma sentença tem no máximo um segmento correspondente na outra sentença. É válido dizer que o processo de alinhamento descrito neste trabalho difere consideravelmente daquele realizado em outras tarefas do PLN (por exemplo, na Tradução Automática), pois algumas informações não estão presentes em ambas as sentenças, mas em apenas uma delas e, nesses casos, elas não são alinhadas. Além do mais, somente as palavras de classes abertas como os substantivos, os verbos, os advérbios e os adjetivos são alinhados. As palavras de classes fechadas (por exemplo, artigos, preposições e conjunções) participam somente dos alinhamentos envolvendo paráfrases multipalavras (por exemplo, *capital russa* e *capital da Rússia*) e por esse motivo elas foram omitidas da Figura 6.

Algoritmo incremental de alinhamento

Passo 1 (inicial): *Alinhamento de duas sentenças*

Dadas duas sentenças do conjunto de entrada (aqui denominadas de sentença fonte e sentença alvo), o algoritmo tenta encontrar o melhor alinhamento entre segmentos que compartilham a mesma informação semântica. Ao invés de analisar exaustivamente todo o espaço de busca dos alinhamentos possíveis, para cada palavra da sentença fonte, o algoritmo procura por possíveis candidatas ao alinhamento na sentença alvo. Para isso, são usadas como âncoras palavras sinônimas, cognatas ou que possuem o mesmo lema da palavra alvo. Além do mais, as palavras candidatas têm que ter o mesmo *POS* da palavra fonte, de modo a garantir um alinhamento mais confiável. As relações de sinonímia são obtidas

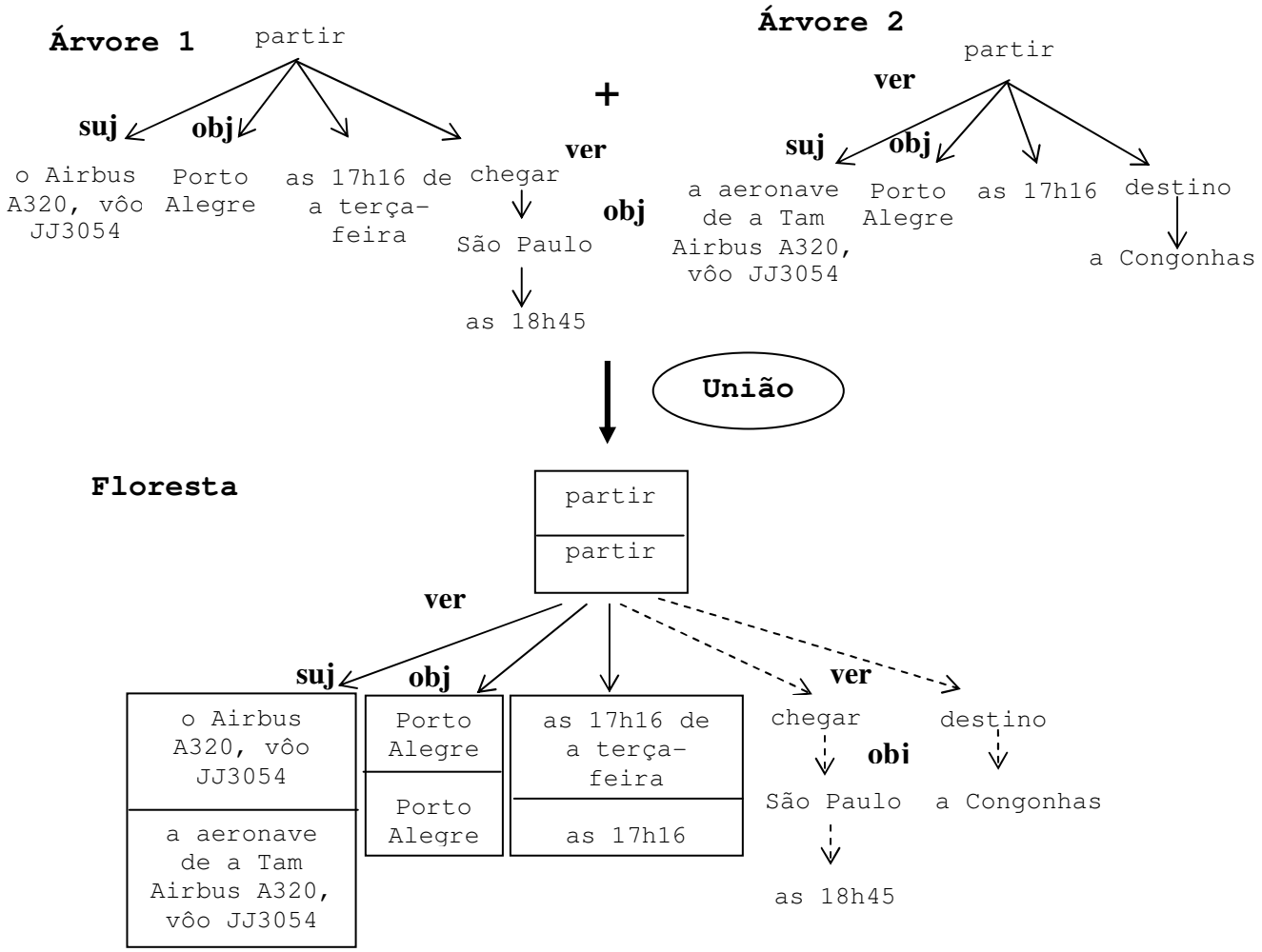


Figura 6: Exemplo de floresta obtida a partir do alinhamento de um par de árvores de dependências sintática

por meio de consultas à base Tep, enquanto que as palavras cognatas são identificadas com o uso de uma medida de similaridade conhecida como LCSR (*Longest Common Subsequence Ratio*, em inglês). O LCSR de duas palavras é calculado dividindo-se o comprimento da maior subseqüência de caracteres em comum entre elas pelo comprimento da maior palavra. Essa medida permite a identificação de palavras com algumas alterações de grafia (por exemplo, *Hezbolla* e *Hisbola*) e também o reconhecimento de diferentes formas de um mesmo nome próprio (por exemplo, *Rui Pimenta* e *Rui Costa Pimenta*). A LCSR só não é usada para os verbos, a fim de evitar casos como *correr* e *morrer* que, apesar do alto valor de LCSR (0,84), têm significados completamente distintos.

Após encontrar todas as candidatas, o algoritmo recupera os sintagmas correspondentes da palavra fonte e de cada palavra candidata, caso a palavra pertença a algum sintagma (por exemplo, *Airbus A320* pertence ao sintagma nominal *o Airbus A320, vôo JJ3054* (sentença [1] da Figura 1)). O sistema então calcula a probabilidade de alinhamento de cada palavra candidata e aquela que apresentar a maior probabilidade é alinhada com a palavra fonte.

Na versão preliminar do sistema (Seno and Nunes (2008a)), a probabilidade de alinhamento é igual a 1, em caso de segmentos idênticos, 0,5 em casos de paráfrases e 0,3 em casos de sinônimos ou cognatos. Esses valores foram determinados empiricamente e priorizam os alinhamentos de palavras e multipalavras literalmente idênticas. Os traços de dependência sintática são considerados somente no alinhamento de verbos. Ou seja, para os casos em que os sujeitos correspondentes aos verbos são similares (isto é, se eles foram previamente alinhados) a probabilidade de alinhamento dos

verbos é acrescida de 0,1, ou penalizada em 0,1, caso contrário. Portanto, nas primeiras iterações, o algoritmo prioriza o alinhamento de nomes próprios e substantivos, visando encontrar as correspondências entre os sujeitos. Por fim, o algoritmo tenta alinhar as palavras e multipalavras restantes ainda não alinhadas, para as quais nenhuma regra de parafraseamento pôde ser aplicada. Esses alinhamentos são realizados somente para os verbos e os sujeitos e se baseiam apenas nos seus traços de dependência sintática, Nos casos em que os sujeitos das sentenças fonte e alvo foram previamente alinhados e os verbos correspondentes ainda não foram alinhados, alinham-se os verbos, assumindo-se que há uma paráfrase entre eles. De maneira similar, se dois verbos foram previamente alinhados e os sujeitos correspondentes nas sentenças não foram, então eles também são alinhados.

No atual sistema, o cálculo da probabilidade de alinhamento foi modificado de modo a considerar não apenas a similaridade entre palavras e multipalavras (isto é, se eles são idênticos, sinônimos, cognatos ou paráfrases), mas também o papel sintático que cada um desempenha na sentença (por exemplo, sujeito, objeto direto, objeto indireto, etc,) e a similaridade entre seus dependentes (para todos os casos, e não somente para os verbos). Nos casos em que a palavra candidata e a palavra fonte têm a mesma função sintática, o sistema adiciona um bônus de 0,3 na probabilidade de alinhamento entre elas. A similaridade entre os dependentes sintáticos é verificada tanto para os verbos, quanto para os sujeitos e objetos das sentenças. Porém, como os verbos são alinhados por último, ao alinhar sujeitos e objetos, o algoritmo verifica se os verbos correspondentes são sinônimos ou paráfrases e, em caso positivo, aumenta a probabilidade de alinhamento em 0,3.

Outra modificação realizada ao sistema diz respeito aos valores de similaridade entre palavras e multipalavras. Para os casos de identidade e de paráfrases, a similaridade é 1, e para os cognatos e sinônimos, a similaridade é 0,5. Esses valores foram ajustados manualmente com base no corpus usado para a identificação das regras de parafraseamento (vide Seção 3.2).

Para que o alinhamento entre duas palavras (ou dois segmentos multipalavras) se concretize, a probabilidade máxima deve ser maior ou igual a 0,5. Esse limite foi estabelecido de modo a permitir também o alinhamento de segmentos que têm funções sintáticas e dependentes em comum, mas para os quais nenhuma regra de parafraseamento pôde ser aplicada.

Passo 2 (incremental): *Alinhamento entre uma sentença e a floresta*

O alinhamento entre uma sentença qualquer e a floresta é realizado de maneira similar ao alinhamento de duas sentenças. Assim, para cada palavra de uma sentença fonte, o algoritmo procura por possíveis candidatas ao alinhamento na floresta. A floresta é armazenada em um vetor associativo cujas chaves correspondem ao identificador de cada sentença do conjunto já alinhada à ela. Para cada chave de uma sentença, é mantido outro vetor associativo contendo cada palavra da sentença e, para cada palavra, por sua vez, são guardadas informações sobre o sintagma ao qual pertence e sobre o alinhamento, ou seja, as palavras (ou sintagmas) de outras sentenças que estão alinhadas a ela, em caso da palavra já ter sido alinhada anteriormente. Desse modo, a palavra fonte é comparada a cada palavra de uma sentença da floresta. Ao encontrar possíveis candidatas ao alinhamento, o algoritmo recupera os sintagmas correspondentes a cada uma delas (se houver) e, então, calcula a probabilidade de a probabilidade de alinhamento, conforme descrito anteriormente (Passo 1). Caso haja alguma candidata com probabilidade >= 0,5, ela é alinhada à palavra fonte (e a todas as outras que já foram previamente alinhadas a ela, se existir alguma) e a busca por novas candidatas é finalizada. Caso contrário, a busca procede na próxima sentença da floresta.

Para fins de ilustração, considere o alinhamento entre a floresta apresentada na Figura 6 e a sentença "*Um Airbus A320 com capacidade para 170 passageiros partiu de Porto Alegre (RS) às 17h16 com destino a Congonhas,*" (sentença [3] da Figura 1). Ao buscar na floresta possíveis candidatos ao alinhamento para o nome "*Airbus A320",* por exemplo, o algoritmo inicialmente analisa todas as palavras de uma determinada sentença da floresta. As sentenças são ordenadas de acordo com seu identificador, isto é, sua chave no vetor associativo, e são selecionadas em ordem.

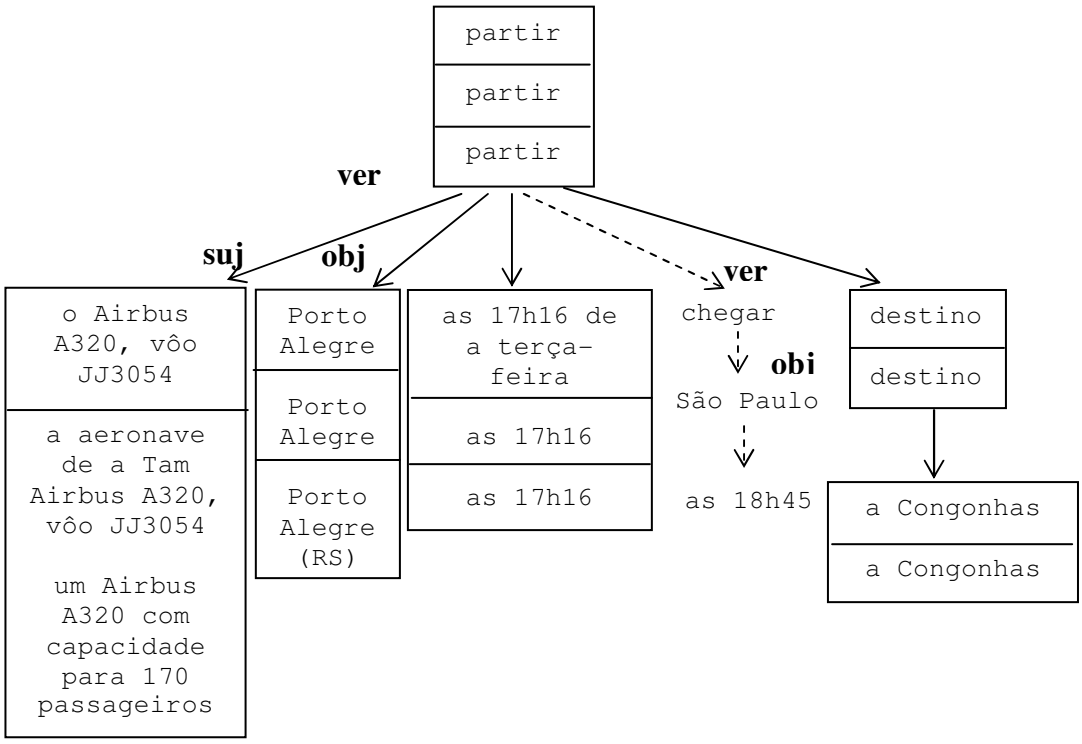


Figura 7: Exemplo de floresta obtida a partir do alinhamento de 3 árvores de dependências sintáticas

Supondo que a sentença da floresta em foco seja "*A aeronave da TAM Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas,*", somente um nome candidato será encontrado, "*Airbus A320*". O algoritmo então recupera os sintagmas correspondentes da sentença fonte e da sentença da floresta, ou seja, "*um Airbus A320 com capacidade para 170 passageiros*" e "*a aeronave da TAM Airbus A320, vôo JJ 3054*", respectivamente. Após recuperar os sintagmas, o sistema calcula a probabilidade de alinhá-los. Para esse exemplo, em particular, não há regras de parafraseamento. Desse modo, a probabilidade de alinhamento é igual a 0,6, uma vez que ambos os segmentos desempenham o papel de sujeito e os verbos correspondentes (*partir*) são similares. Portanto, eles são alinhados e a busca por novos candidatos em outras sentenças da floresta é finalizada. Como o sintagma da floresta ("*a aeronave da TAM Airbus A320, vôo JJ 3054*") já havia sido alinhado a outro sintagma ("*o Airbus A320, vôo JJ 3054*") (vide Figura 6), o novo correspondente "*um Airbus A320 com capacidade para 170 passageiros*" é adicionado ao mesmo alinhamento. A Figura 7 ilustra a floresta resultante do alinhamento entre a sentença [3] (Figura 1) e a floresta da Figura 6.

4. Experimentos

Com o propósito de verificar se as mudanças no pré-processamento das sentenças de entrada e na estratégia de alinhamento de fato contribuem para um melhor desempenho do sistema, foram avaliados somente os alinhamentos produzidos entre pares de sentenças comparáveis (e não a partir de um conjunto de sentenças). Uma vez que o alinhamento entre uma sentença qualquer e a floresta é similar ao alinhamento de um par de sentenças (vide Seção 3.3.2), acredita-se que o desempenho do sistema tanto no alinhamento de duas sentenças como no alinhamento de um conjunto de sentenças será equivalente.

A qualidade dos alinhamentos automáticos foi verificada com base em um corpus de referência composto por 20 pares de sentenças extraídos aleatoriamente do corpus comparável (Seção 3.1). É válido dizer que esse subcorpus é diferente daquele usado para a formulação das regras de parafraseamento.

Os 20 pares de sentenças foram manualmente alinhados por dois anotadores. Posteriormente, a concordância entre eles foi calculada com base no total de alinhamentos em comum dividido pelo total de alinhamentos produzidos pelos dois anotadores. Uma taxa de concordância de 87% foi obtida, indicando que os alinhamentos de referência são razoavelmente confiáveis.

Para a avaliação do sistema, foram usadas as medidas de Precisão, Cobertura e Medida-f. Seja *R* o conjunto de alinhamentos de referência, *A* o conjunto de alinhamentos produzidos automaticamente e $|A \cap R|$ o conjunto de alinhamentos automáticos corretamente produzidos. A Precisão representa a fração dos alinhamentos automáticos identificados corretamente, em relação a todos os alinhamentos automáticos produzidos (Fórmula 13). A Cobertura representa a fração dos alinhamentos automáticos identificados corretamente, em relação a todos os alinhamentos previstos no conjunto de referência (Fórmula 14). A Medida-f, por sua vez, representa a média harmônica entre a Precisão e a Cobertura (Fórmula 15).

$$\text{Precisão} = \frac{|A \cap R|}{|A|} \quad (13)$$

$$\text{Cobertura} = \frac{|A \cap R|}{|R|} \quad (14)$$

$$\text{Medida-f} = \frac{2 \times \textit{Precisão} \times \textit{Cobertura}}{\textit{Precisão} + \textit{Cobertura}} \quad (15)$$

O sistema proposto foi comparado com outros dois sistemas *baselines*. O *baseline* 1, que é baseado somente na similaridade lexical e semântica, alinha apenas segmentos idênticos, cognatos e sinônimos. O *baseline* 2 é uma extensão do *baseline* 1 que inclui, além dos sinônimos e cognatos, os traços de dependência sintática. O primeiro *baseline* tem como propósito avaliar a contribuição das regras de parafraseamento e das relações de dependência sintática para o processo de alinhamento, enquanto que o *baseline* 2 visa apenas verificar a contribuição das regras de parafraseamento.

A Tabela 4 apresenta os valores médios obtidos pelo alinhador proposto (versão 2,0) e por cada *baseline* para Precisão, Cobertura e Medida-f. Para fins de comparação, a tabela também resume os resultados obtidos com a versão preliminar do sistema (versão 1,0), apresentados em Seno and Nunes (2008a). Os *baselines* usados na versão 1,0 são equivalentes aos *baselines* descritos neste trabalho.

Sistema	Precisão	Cobertura	Medida-f
Versão 2,0			
Baseline 1	0,81	0,76	0,78
Baseline 2	0,81	0,75	0,78
Alinhador Proposto	0,87	0,83	0,85
Versão 1,0			
Baseline 1	0,77	0,72	0,74
Baseline 2	0,77	0,72	0,74
Alinhador Proposto	0,86	0,81	0,83

Tabela 4: Resultados do alinhamento automático obtidos para Precisão, Cobertura e Medida-f

Conforme os resultados apresentados na Tabela 4, o atual sistema obteve uma melhora de 2,4% no desempenho global em relação à sua primeira versão (isto é, 85% de Medida-f contra 83% de Medida-f) e um ganho de 9% comparado aos seus *baselines*. Os *baselines*, por sua vez, já obtiveram um desempenho bem elevado (isto é, 78% de Medida-f), o que era esperado devido às características do corpus (aproximadamente 72% dos alinhamentos identificados ocorrem entre segmentos literalmente idênticos).

É importante observar que os *baselines* atuais também apresentaram um desempenho de cerca de 5% melhor em relação aos *baselines* usados na avaliação do sistema anterior (ou seja, 78% de Medida-f contra 74% de Medida-f). Isso se deve principalmente às modificações no pré-processamento das sentenças que permitem recuperar de forma mais abrangente e confiável as dependências sintáticas entre os sintagmas.

Outro ponto importante a ser notado é que o *baseline* 2 não apresentou ganho de desempenho comparado ao *baseline* 1 (em ambas as versões), quando foram incluídos os traços de dependência entre os constituintes sintáticos. O ganho de desempenho apenas foi verificado ao se incluir

as regras de parafraseamento nos sistemas propostos (conforme mostrado na Tabela 4).

Com propósito de verificar a contribuição do sistema proposto para o alinhamento de paráfrases apenas (tanto lexicais, isto é, sinônimos e cognatos, quanto sintáticas), a Precisão, a Cobertura e a Medida-f foram calculadas considerando-se somente esses casos. Os resultados obtidos são mostrados na Tabela 5. Para fins de comparação, os resultados alcançados com a versão 1,0 do sistema também são mostrados na tabela.

De acordo com a Tabela 5, a segunda versão do alinhador apresentou um ganho de aproximadamente 21% em comparação a sua primeira versão (ou seja, 64% de Medida-f contra 53% de Medida-f). É válido notar que o ganho de Precisão e de Cobertura foi de 9,5% e 33,3%, respectivamente. Além do mais, o sistema obteve uma melhora substancial de desempenho em relação aos *baselines* (isto é, um aumento de 94% e de 178% comparado ao *baseline* 2 e ao *baseline* 1, respectivamente), quando considerados apenas os casos de paráfrases.

Sistema	Precisão	Cobertura	Medida-f
Versão 2,0			
Baseline 1	0,55	0,14	0,23
Baseline 2	0,53	0,24	0,33
Alinhador Proposto	0,69	0,60	0,64
Versão 1,0			
Baseline 1	0,63	0,12	0,20
Baseline 2	0,50	0,17	0,25
Alinhador Proposto	0,63	0,45	0,53

Tabela 5: Resultados do alinhamento automático obtidos para Precisão, Cobertura e Medida-f, considerando-se somente os casos de paráfrases

O uso das relações de dependência sintática no *baseline* 2 (versão 2,0) contribuiu para um aumento de cerca de 43% no desempenho global, em relação ao *baseline* 1 (sem relações de dependências), quando considerados apenas os alinhamentos de paráfrases. No entanto, como dito anterior, nenhuma melhora foi observada entre os *baselines* 1 e 2, quando considerados todos alinhamentos em ambas as versões dos sistemas (vide Tabela 4).

Esses resultados comprovam que a similaridade lexical, as relações de sinonímia e as relações sintáticas auxiliam no alinhamento de informações comuns, porém não são suficientes para tratar os casos mais complexos de paráfrases como é o caso das paráfrases sintáticas, parcialmente tratadas pelas regras de parafraseamento.

A Figura 8 mostra alguns exemplos de alinhamentos produzidos pelo algoritmo. A maioria deles foi identificado com o auxílio das regras de parafraseamento, como os exemplos (a), (b), (c), (d), (e), (f), (h) e (i). Alguns casos de paráfrases que não foram cobertos pelas regras são ilustrados na Figura 9.

(a) 44% das intenções de voto ⇔ 44% dos votos
(b) março ⇔ março de o ano que vem
(c) a agência Itar-Tass ⇔ a agência oficial russa Itar-Tass
(d) Luiz Inácio Lula da Silva ⇔ o presidente Luiz Inácio da Silva ⇔ Lula
(e) a cidade de Tampere ⇔ Tampere (FIN)
(f) o chefe de polícia do campus ⇔ o chefe de polícia da universidade
(g) afirmou ⇔ disse
(h) aconteceu ⇔ foi registrada
(i) bujão de gás ⇔ botijão de gás

Figura 8: Exemplos de alinhamentos automáticos

(a) os 69 deputados acusados pela CPI dos Sanguessugas de envolvimento ⇔ os deputados envolvidos
(b) os quatro menores ⇔ os quatro com menos de 18 anos
(c) o prédio de carga e descarga da companhia aérea ⇔ o prédio da TAM Express
(d) 23 pessoas ⇔ o grupo

Figura 9: Exemplos de paráfrases não identificadas pelas regras de parafraseamento

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma nova versão do alinhador descrito em Seno and Nunes (2008a), para a identificação de segmentos que conduzem a mesma informação semântica entre sentenças comparáveis do português.

Diversas melhorias realizadas ao sistema, como alterações no pré-processamento das sentenças de entrada, modificações na estratégia de alinhamento e a inclusão de novas relações sintáticas, resultaram em um aumento de desempenho de aproximadamente 21%, comparado com a primeira versão do sistema, quando avaliados somente os alinhamentos entre paráfrases (tanto lexical, quanto sintática). Quando considerados todos os alinhamentos (incluindo os casos de segmentos literalmente idênticos), o ganho no desempenho foi de 2,4%,

O resultado alcançado neste trabalho, ou seja, um desempenho de 85% de Medida-f considerando todos os alinhamentos, representa um ganho de 9% em relação aos *baselines* de comparação e está de acordo com outros resultados reportados na literatura (vide Seção 2).

Com relação ao alinhamento de paráfrases somente (isto é, excluindo-se os casos de segmentos idênticos), o método apresentou um ganho de até 178% no desempenho global, comparado aos *baselines*. Os trabalhos encontrados na literatura não reportam resultados para os casos de paráfrases apenas.

Os experimentos apresentados na seção anterior são preliminares e se referem apenas aos alinhamentos produzidos a partir de pares de sentenças. Entretanto, como a estratégia de alinhamento é independente do número de sentenças de entrada, acredita-se que o sistema obterá um desempenho similar no alinhamento de um conjunto de sentenças. Novos experimentos deverão ser realizados para comprovar essa hipótese. Além disso, estão previstos experimentos com corpora maiores e a indução automática de paráfrases a partir de corpus.

É importante notar que o alinhador foi projetado para trabalhar com sentenças semanticamente muito similares (ou seja, comparáveis ou paralelas monolíngües). Portanto, é natural que haja uma queda de desempenho do sistema ao tentar alinhar sentenças com pouca similaridade semântica.

Como continuação deste trabalho, os próximos passos incluem a implementação de um módulo de fusão e linearização, para a geração de novas sentenças a partir da fusão de informações comuns previamente alinhadas. Esse módulo já está em desenvolvimento atualmente e poderá ser usado em um futuro próximo para validar o processo de alinhamento de informações comuns, inclusive no que se refere ao alinhamento envolvendo mais de duas sentenças.

Agradecimento

Agradecemos ao CNPq (Conselho Nacional de Pesquisa e Desenvolvimento) pelo suporte financeiro.

Referências

Barzilay, R. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*, Phd, Thesis, Columbia University, New York, 221 p.

Barzilay, R, and McKeown, K. 2005. Sentence Fusion for Multi-document News Summarization, *Computational Linguistics*, Vol, 31, nº 3, pp, 297-327.

Bick, Eckhard. 2000. *The Parsing System "Palavras" - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press.

Dias-da-Silva, B.C., Di Felippo, A., and Hasegawa, R. 2006. Methods and Tools for Encoding the WordNet,Br Sentences, Concept Glosses and Conceptual-Semantic Relations. In: *Proceedings of the 7th Workshop on Computational Processing of the Portuguese Language - Written and Spoken -*

PROPOR (Lecture Notes in Artificial Intelligence, 3960), pp, 120-130.

Fung, B.C.M., Wang, K., Ester, M. 2003. Hierarchical Document Clustering using Frequent Itemsets. In: Barbará, D, Kamath, C, eds, *3rd SIAM International Conference on Data Mining*, pp, 59-70.

Hatzivassiloglou, V., Klavans, J. L., Eskin, E. 1999. Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In: *Proceedings of the Empirical Methods in Natural Language Processing and Very Large Corpora – EMNLP*, pp, 203-212.

Hoey, M. 1991. *Patterns of Lexis in Text*, Oxford: Oxford University Press,

Krahmer, E., Marsi, E. and van Pelt, P. 2008. Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion, In: *Proceedings of the Human Language Technology Conference – HLT/ACL*, pp, 193-196.

Larocca Neto, J., Santos, A.D., Kaestner, C.A.A., Freitas, A.A. 2000. Document Clustering and Text Summarization. In: *4th International Conference Practical Applications of Knowledge Discovery and Data Mining – PAAD*, pp, 41-55.

Marsi, E. and Krahmer, E. 2005. Explorations in Sentence Fusion. In: *Proceedings of the 10th European Workshop on Natural Language Generation – ENLG*, pp, 109-117.

Maziero, E.G., Pardo, T.A.S., Di Felippo, A., Dias-da-Silva, B.C. 2008. A Base de Dados Lexical e a Interface Web do TeP 2,0 - Thesaurus Eletrônico para o Português do Brasil. *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp, 390-392.

Pang, B., Knight, K. and Marcu, D. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In: *Proceedings of the Human Language Technology Conference – HLT/NAACL*, pp, 102-109.

Radev, D., Otterbacher, J., Zhang, Zhu. 2008. Cross-document Relationship Classification for Text Summarization. Disponível em: tangra,si,umich,edu/~radev/papers/progress/p1,ps (último acesso: 13/04/2009).

Rosell, M., Kann, V., Litton, J. 2004. Comparing Comparisons: Document Clustering Evaluation Using Two Manual Classifications. In: Sangal R, Bendre SM, eds, *International Conference on Natural Language Processing*, Allied Publishers Private Limited, pp, 207-216.

Salton, G. and Allan, J. 1994. Text Retrieval Using the Vector Processing Model. In: *Proceedings of the 3rd Symposium on Document Analysis and Information Retrieval*, University of Nevada, Las Vegas.

Seno, E.R.M. and Nunes, M.G.V. 2008a. Automatic Alignment of Common Information in Comparable Sentences of Portuguese. In: *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, pp, 331-335.

Seno, E.R.M. and Nunes, M.G.V. 2008b. Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In: *Proceedings of the International Conference on Computational Processing of Portuguese Language - PROPOR* (Lecture Notes in Artificial Intelligence, 5190), pp, 133-144.

Shen, S., Radev, D. R., Patel, A. and Erkan, G. 2006. Adding Syntax to Dynamic Programming for Aligning Comparable Texts for the Generation of Paraphrases. In: *Proceedings of the COLING/ACL*, pp, 747-754.

Steinbach, M., Karypis, G., Kumar, V. 2000. A Comparison of Document Clustering Techniques. In: *International Conference on Knowledge Discovery & Data Mining – KDD*.

Van Rijsbergen, C.J. 1979. *Information Retrieval*, 2nd edition, Butterworths, Massachusetts.

Extracção de Informação de Relatórios Médicos

Liliana Ferreira[1] César Telmo Oliveira[12] António Teixeira[1] João Paulo Silva Cunha[1]

[1] Instituto de Engenharia Electrónica e Telemática de Aveiro
Departamento de Electrónica, Telecomunicações e Informática
Universidade de Aveiro
3810-193 Aveiro, Portugal

[2] Hospital Infante D. Pedro
Avenida Artur Ravara
3814-501 Aveiro, Portugal
{lsferreira, ctelmo, ajst, jcunha}@ua.pt

Resumo

A utilização, cada vez mais frequente nos serviços de saúde nacionais, de sistemas de Registo Clínico Electrónico tem levado a um aumento significativo da informação disponível em formato electrónico. Embora muita desta informação exista, actualmente, numa forma estruturada, uma parte significativa encontra-se sob a forma de texto livre não estruturado. A necessidade de processar e gerir estas grandes quantidades de texto tem motivado o recente interesse em aproximações semânticas. Este artigo descreve o trabalho desenvolvido no âmbito do projecto MedAlert para a criação de um corpus anotado semanticamente e no desenvolvimento de um sistema de extracção automática de informação capaz de identificar entidades clínicas relevantes, bem como os seus relacionamentos. Para tal, o MedAlert possui actualmente um corpus de cerca de 48 000 textos médicos relativos a episódios de internamento ocorridos no Hospital Infante D. Pedro, em Aveiro. Um subconjunto do corpus foi seleccionado para a criação das directivas de anotação e anotação semântica manual e automática. O sistema de reconhecimento de entidades mencionadas REMMA foi usado numa primeira avaliação. Os primeiros resultados são apresentados indicando a necessidade de desenvolver directivas precisas para a anotação de textos médicos, de modo a melhorar a concordância entre anotadores.

1 Introdução

O acesso a informação clínica em instituições de saúde nacionais é feito, cada vez mais, através de variados sistemas de Registo Clínico Electrónico (RCE). Embora alguns relatórios médicos existam actualmente, nestes sistemas, numa forma estruturada, uma parte significativa é guardada ainda como texto livre não estruturado. Este é o caso dos relatórios relativos a episódios de internamento. Estes documentos contêm informação importante, não só para a manutenção do cuidado de saúde do doente, mas também de uso potencial em investigação. Descrevem, por exemplo, qual a medicação usada em cada tratamento, porque foi interrompida, quais os resultados de exames físicos e quais os problemas considerados relevantes na discussão com o paciente mas que nem sempre são considerados relevantes na codificação interna.

A necessidade de gerir este tipo de informação está a motivar aproximações semânticas, cujos principais objectivos são a redução de erros clínicos, a melhoria da eficiência, da segurança e da satisfação no serviço médico. Por exemplo, a informação contida nestes documentos poderia ser usada para assistir o clínico na formação de hipóteses, caso este pudesse obter respostas a questões relevantes, como por exemplo *Quantos pacientes com AVC isquémico agudo foram tratados com Enoxaparina e permaneceram sem outras complicações?* O tratamento individual de pacientes beneficiaria também, caso pudessem ser obtidos sumários concisos da história clínica do paciente ou se existisse acesso a histórias clínicas de pacientes com manifestações semelhantes reportadas em diversas ocasiões e localizações.

O MedAlert usa a tecnologia de extracção automática de informação nos dados disponibilizados no sistema de RCE em utilização no Hospital Infante D. Pedro em Aveiro, a Rede Telemática de Saúde (RTS) (Cunha et al., 2006).

Este artigo reporta a construção de uma colecção dourada para o projecto MedAlert, na qual os documentos clínicos são anotados com as suas múltiplas entidades e relacionamentos. Uma primeira avaliação do sistema de extracção au-

tomática de informação REMMA - *Reconhecimento de Entidades Mencionadas do MedAlert* é também apresentada.

A secção seguinte apresenta o projecto MedAlert e a sua motivação. A Secção 1.2 sumaria algum trabalho relacionado apresentado na literatura. Os recursos utilizados no MedAlert são apresentados na secção 2, onde é descrito o processo de selecção de documentos para a colecção dourada, o método de anotação usado e as respectivas entidades e relacionamentos. As fontes de conhecimento usadas na extracção automática de informação são descritas na secção 2.2. A secção 3 descreve o sistema REMMA e os primeiros resultados obtidos são discutidos na secção 4. O artigo termina na secção 5 com as conclusões e algumas sugestões de trabalho futuro.

1.1 MedAlert

Nos últimos anos tem sido realizado um investimento significativo em sistemas que permitam o acesso electrónico a informação clínica. Este tipo de acesso é cada vez mais uma realidade através de numerosos sistemas de RCE. No entanto, pouco tem sido feito na criação de sistemas que permitam a comunicação entre diferentes instituições médicas (Cunha et al., 2006). A Rede Telemática de Saúde (RTS)[1] tenta colmatar esta dificuldade através de uma infra-estrutura que permite a comunicação clínica entre os múltiplos serviços de saúde regional. Esta rede promove, assim, o acesso seguro a informação existente em vários serviços de saúde, a todos os profissionais credenciados. A RTS implementa um *Processo Clínico Electrónico Regional* resumido, que combina diversos documentos electrónicos existentes em todas as instituições que pertencem à rede, permitindo, assim, o acesso dos profissionais de saúde a informação como cartas de alta, resultados de exames e boletins de vacinação.

O MedAlert usa a informação disponibilizada pela RTS, em utilização no Hospital Infante D. Pedro e na região de Aveiro e tem como principal objectivo a utilização de técnicas de extracção automática de informação de textos médicos, de modo a inferir, de uma forma automática, irregularidades/dúvidas suscitadas pelas decisões tomadas pelos profissionais de saúde. O MedAlert, que deverá tomar a forma dum módulo escalável e adaptável a diferentes configurações de sistemas de informação hospitalares, pretende usar técnicas de Processamento de Linguagem Natural (PLN) para extrair informação de um amplo conjunto de textos médicos, particularmente cartas de alta e textos contendo directivas médicas. Esta informação, bem como a proveniente de recursos externos como ontologias e outras fontes de conhecimento médico, deverá ser utilizada no suporte e validação de decisões, melhorando, assim, o cuidado médico, com a redução de erros, melhoria de segurança e satisfação.

1.2 Trabalho relacionado

Várias aplicações de suporte à decisão clínica têm sido desenvolvidas recentemente, fazendo uso de técnicas de PLN e fontes de conhecimento como ontologias. Consequentemente, uma grande variedade de *corpora* anotados semanticamente e outras fontes de conhecimento médico foram desenvolvidas tendo em vista a investigação em extracção de informação biomédica. O *thesaurus Medical Subject Headings* (MeSH)[2] e o *Unified Medical Language System* (UMLS) (NLM, 2008), com as suas vertentes de *metathesaurus* e de rede semântica, são exemplos do esforço feito no sentido de facilitar o desenvolvimento de sistemas computacionais capazes de processar linguagem médica. Ambos são actualmente utilizados numa grande variedade de sistemas na catalogação, indexação e recolha de informação biomédica e de saúde.

Um esforço semelhante foi realizado no desenvolvimento do vocabulário trilingue DeCS - Descritores em Ciências da Saúde[3]. O DeCS foi desenvolvido a partir do MeSH com o objectivo de permitir o uso de terminologia comum para a pesquisa em três línguas, inglês, espanhol e português, proporcionando uma forma consistente e única para a recolha de informação médica. Os conceitos que compõem o DeCS são organizados numa estrutura hierárquica permitindo a execução de pesquisa em termos mais amplos ou mais específicos ou de todos os termos que pertençam a uma dada estrutura hierárquica.

2 Recursos

No desenvolvimento do sistema MedAlert são utilizados vários recursos, desde o *corpus* usado no desenvolvimento da colecção dourada MedAlert, até às várias fontes de conhecimento externo usadas na extracção automática de informação. Esta secção apresenta em mais detalhe estes recursos, começando por apresentar na Secção 2.1 o *corpus* MedAlert e o método usado na anotação semântica manual. A Secção 2.2 apresenta as fontes de conhecimento usadas no reconhecimento automático das entidades e relacionamentos definidos na anotação manual.

2.1 O *corpus* MedAlert

O *corpus* MedAlert é actualmente constituído por 48 229 textos relativos a episódios de internamento

[1] http://www.rtsaude.org

[2] http://www.nlm.nih.gov/mesh/

[3] http://decs.bvs.br/

ocorridos no Hospital Infante D.Pedro, em Aveiro. Estes relatórios incluem informação relativa a:

- Motivo de internamento;
- História clinica;
- Exame físico;
- Evolução;
- Terapêutica;
- Destino.

A Tabela 1 apresenta a distribuição de informação no *corpus*, em particular, a quantidade de documentos, frases e tokens existente para cada estrutura.

Os relatórios provêm do *Processo Clínico Electrónico Regional* implementado pela RTS, onde toda a informação confidencial relativa aos doentes e profissionais de saúde está já de uma forma estruturada e separada. Assim, os relatórios usados neste trabalho não contêm qualquer informação confidencial ou passível de identificação dos intervenientes no processo.

2.1.1 Colecção dourada MedAlert

A construção de uma colecção dourada MedAlert tem como objectivo servir três propósitos principais:

1. focar e clarificar os requisitos do sistema através da análise de dados anotados manualmente por peritos da área;
2. o desenvolvimento de um *gold standard* contra o qual os resultados da extracção automática de informação serão calculados;
3. o fornecimento de dados para o desenvolvimento do sistema: as regras de extracção podem deste modo ser criadas automaticamente ou manualmente, bem como podem ser desenvolvidos modelos estatísticos dos dados para a utilização de algoritmos de *machine learning*.

Dado o elevado custo da anotação manual, a ser realizada, neste caso, por pessoal médico especializado, a percentagem de relatórios a anotar teve de ser reduzida a um subconjunto relativamente pequeno de todo o corpus de 48 229 relatórios. Nesta fase inicial do processo e de modo a facilitar a introdução das directivas aos peritos, optou-se por focar nas estruturas Motivo de Internamento e História Clínica e num conjunto reduzido de documentos, embora no alcance dos objectivos finais do projecto seja necessária a existência de mais dados anotados manualmente e relativos a todas as estruturas dos relatórios.

Assim, optou-se pela utilização de um subconjunto de 120 relatórios, 20 para cada estrutura, tendo destes, 10 documentos sido usados no desenvolvimento das directivas de anotação e 10 na anotação manual.

Deste modo, a colecção dourada é constituída actualmente por 20 documentos anotados manualmente, relativos às estruturas Motivo de Internamento e História Clínica.

O restante artigo foca na anotação semântica e extracção automática de informação relativa aos relatórios de Motivo de Internamento.

2.1.2 Método de anotação

A construção de uma colecção dourada para o projecto MedAlert pressupõe a existência de um *corpus* de documentos médicos anotados semanticamente, quer com múltiplas entidades, quer com as suas relações.

De modo a garantir a qualidade da colecção dourada todos os documentos foram anotados pelo mesmo *standard* e foram desenvolvidas directivas especificas de modo a que as várias questões que surjam ao anotar os relatórios estejam devidamente esclarecidas. As directivas desenvolvidas pretendem, assim, garantir a consistência, descrevendo em detalhe o que deve e o que não deve ser anotado, respondendo a questões relevantes tais como, decidir se duas entidades estão relacionadas ou como lidar com correferência. As directivas apresentam também uma sequência de passos, uma receita, que os anotadores deverão seguir quando trabalham com os documentos, de modo a minimizar os erros de omissão. Deste modo, o desenvolvimento das directivas de anotação foi realizado através de um processo rigoroso e iterativo, criado de modo a garantir consistência (Roberts et al., 2007).

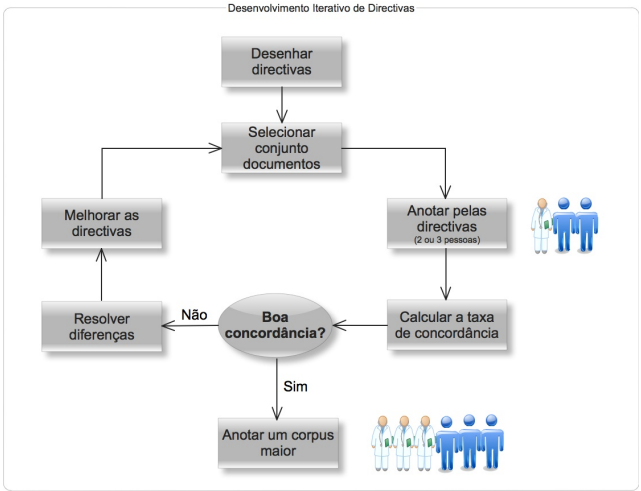


Figura 1: Processo Iterativo de anotação de relatórios.

Em detalhe o processo incluiu vários passos, apresentados na Figura 1, entre os quais se des-

Tabela 1: Relatórios MedAlert

Documento	Tokens	Frases	Textos
Motivo Internamento	104 833	11851	8 563
História Clinica	1 179 960	56 202	9 775
Exame Físico	414 558	37 499	7 071
Evolução	474 303	26 663	8 106
Terapêutica Efectuada	332 017	11 569	8 363
Destino	219 189	13 834	6 351
Total	**2 724 860**	**157 618**	**48 229**

taca:

1. Dupla anotação: um documento anotado por uma única pessoa pode reflectir vários problemas, como os valores ou erros frequentemente efectuados por um único anotador. A anotação dupla é uma forma comum de minimizar estes problemas, na qual cada documento é anotado independentemente por dois ou mais anotadores, e o conjunto de anotações comparado de modo a determinar a concordância.
2. Métricas de Concordância: o nível de concordância entre anotadores foi medido através do *índice de concordância inter-anotadores* (IAA):

$$IAA = \frac{\text{concordância}}{\text{concordância} + \text{não concordância}} \quad (1)$$

O índice de concordância foi calculado segundo um processo "relaxado", no qual as concordâncias parciais são contabilizadas como meia concordância. Os relacionamentos também foram avaliados usando IAA, tendo sido convencionado que apenas os relacionamentos envolvendo as entidades que todos os anotadores encontraram são contabilizados, permitindo, assim, isolar melhor a avaliação dos relacionamentos em relação à avaliação das entidades.

2.1.3 Entidades e Relacionamentos

Na definição da informação a anotar começou por definir-se os conceitos de entidade e relacionamento no contexto médico. Assim, *entidade* foi definida como algo real referido no texto, como por exemplo, a medicação mencionada, os exames realizados, etc. Os *relacionamentos* são então ligações entre entidades como o resultado de um exame ou a medicação indicada para uma patologia. A anotação também contemplou palavras que modificam marcações, tais como negação e caracterização. Duas ou mais marcações podem referir-se à mesma *entidade* real, e foram, neste caso, marcadas como *correferências*.

A Figura 2 apresenta alguns aspectos relevantes da anotação, tais como a marcação das entidades e dos seus relacionamentos.

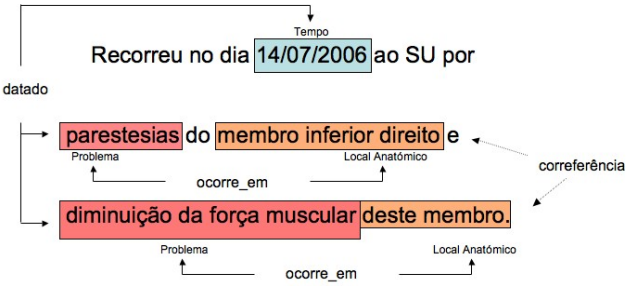


Figura 2: Exemplo ilustrativo de anotação.

A cada entidade e relacionamento foi atribuída uma *categoria*, tendo algumas sido classificadas também com o atributo *tipo*. No caso dos relatórios relativos ao Motivo de Internamento foram definidas as seguintes categorias:

- Problema - Sintomas, diagnósticos, complicações, condições e restantes problemas manifestados pelo doente;
- Local Anatómico - Estrutura ou localização anatómica, substância corporal ou função fisiológica, tipicamente a localização de um *Problema* ou *Exame*;
- Tempo - Expressões temporais, incluindo datas e tempos (absolutos e relativos), durações e frequências;
- Exame - Interacção entre o profissional de saúde e o doente com o objectivo de medir ou estudar algum aspecto do *Problema*;
- Resultado - Observação numérica ou qualitativa de um exame, excluindo referências a *Problemas*;
- Valor - Quantidades absolutas, relativas ou classificações;

- Caracterização - expressões que caracterizam outras entidades, como as pertencentes às categorias *Problema* e *Local Anatómico*;
- Negação - expressões que modificam outras entidades, neste caso negam, como por exemplo as entidades pertencentes à categoria *Problema* e *Resultado*.

Foram também definidos os seguintes relacionamentos:

- `inclui` - relação de *inclusão* entre entidades da mesma categoria, em particular aplicável às entidades das categorias *Problema, Local Anatómico* e *Exame*;
- `ocorre_em` - relação de *localização* entre um *Problema* ou *Exame* e o *Local Anatómico* em que é verificado;
- `datado` - relaciona as entidades *Exame*, *Problema* e *Resultado* com a sua indicação temporal (*Tempo*);
- `quantificado` - relaciona entidades quantificáveis, como as pertencentes às categorias *Resultado* ou *Problema* e o *Valor* que as caracteriza.
- `resulta` - relaciona um *Resultado* com o *Exame* que o produziu;
- `indica` - relaciona um *Problema* com o *Exame* que demonstrou a sua presença;
- `modificado` - relaciona um *Problema* ou *Resultado* com uma *Negação* ou *Caracterização*, bem como o *Local Anatómico* com a sua *Caracterização*, tal como a lateralidade: *direita, esquerda, bilateral* e sub-localização: *alto, baixo, extra, etc.*.

Alguns exemplos para cada uma das entidades e relacionamentos definidos, bem como os tipos atribuídos, são apresentados nas Tabelas 2 e 3.

De modo a facilitar o processo de anotação manual por parte dos especialistas, foram desenvolvidos esquemas de anotação para cada uma das estruturas dos documentos. O esquema de anotação relativo ao Motivo de Internamento é apresentado na figura 3, onde é possível visualizar cada uma das entidades definidas e a forma como estas se relacionam entre si.

2.1.4 Ferramentas de Anotação

De modo a realizar a anotação de uma forma consistente os esquemas de anotação foram modelados como ontologias Protégé-Frames[4] (Gennari et al., 2002). A anotação foi realizada usando o *plugin* Knowtator (Ogren, 2006) para Protégé. Este foi escolhido pelo facto de lidar com relacionamentos, após uma avaliação de outras ferramentas disponíveis (MMAX2[5], Wordfreak[6], Callisto[7]) e de arquitecturas de software de PLN como o GATE(Cunningham et al., 2002).

A Figura 4 apresenta a interface gráfica do Knowtator. No lado esquerdo da figura é possível visualizar o esquema de anotação criado para a anotação dos documentos do Motivo de Internamento. O quadro central e direito da figura apresenta um excerto de um relatório destacando a anotação da palavra `DPOC` como pertencente à classe `Diagnóstico` e o seu relacionamento de *inclusão* e *caracterização* com as palavras `insuficiência` e `agudizada`, respectivamente.

2.2 Fontes de conhecimento

O REMMA, sistema de Reconhecimento de Entidades Mencionadas do MedAlert, usa uma aproximação baseada em conhecimento de modo a detectar e classificar as expressões pertencentes às diversas categorias. Assim, várias fontes de conhecimento foram necessárias para a realização desta tarefa. Este é o caso da lista com cerca de 3 400 actos médicos e 1500 análises realizados no Hospital Infante D. Pedro, bem como da lista dos vários medicamentos disponíveis e comercializados em Portugal, com cerca de 12 800 entradas. Uma pequena lista com os nomes de problemas clínicos mais comuns, cerca de 200, foi também utilizada.

Apesar dos esforços realizados no sentido de obter o vocabulário biomédico DeCS - Descritores em Ciências da Saúde, tal não foi, até à data, possível. Assim, de modo a colmatar a falta de uma fonte de conhecimento especializada de grande abrangência, foi necessário recorrer a outras fontes de conhecimento não especializado como é o caso da Wikipédia. A secção seguinte faz uma pequena introdução à Wikipédia e à sua utilização em PLN.

2.2.1 Wikipedia

Recentemente, assistiu-se a um crescimento rápido e bem-sucedido da Wikipédia[8], uma enciclopédia electrónica livre e que está a ser construída por milhares de colaboradores em todo mundo. A Wikipédia tinha em Janeiro de 2009 mais de 2 700 000 artigos na versão inglesa e cerca de 454 000 artigos na sua versão portuguesa. Uma vez que a Wikipédia pretende ser uma enciclopédia, a maior parte dos artigos são sobre entidades mencionadas e mais estruturados que texto livre. A Wikipédia é actualizada diariamente, ou seja, novas entidades

[4] http://protege.standford.edu

[5] http://mmax.eml-research.de

[6] http://wordfreak.sourceforge.net

[7] http://callisto.mitre.org

[8] http://www.wikipedia.org

Tabela 2: Entidades MedAlert.

Categorias	Tipos	Exemplos
Problema	Sinal	*Prostração* marcada
	Sintoma	*Poliartralgias* MIs
	Diagnóstico	*Dpoc* agudizada
	Patologia	*Bronquite* Aguda
Local Anatómico		Hemorragia *digestiva* alta
Tempo	Tempo Calendário	Recorreu no dia *14/07/2006*...
	Duração	*Durante o internamento* ...
	Frequência	...a repetir a *cada meia hora*...
Exame	Físico	*Auscultação* pulmonar
	Analítico	...tendo sido realizada *biópsia*
	Imagiológico	cuja *EDA* revelou...
Resultado		Abdómen *sem alterações evidentes*
Valor		Lexotan *1,5mg*
Caracterização		Abdómen *sem* alterações evidentes
Negação		Acidente Vascular Cerebral *isquémico*

Tabela 3: Relacionamentos MedAlert.

Relacionamentos	Exemplos
inclui	[*arg1* dores] de garganta com [*arg2* tosse] e [*arg2* expectoração]
ocorre_em	[*arg1* dores] de [*arg2* garganta]
caracterizado_por	[*arg1* bronquite] [*arg2* aguda]
negado_por	[*arg2* sem] episódios prévios de [*arg1* convulsões]
datado_de	sem episódios [*arg2* prévios] de [*arg1* convulsões]
quantificado_por	[*arg1* febre] [*arg2* 40°C]
indica	realizou [*arg1* ecografia] abdominal que mostrou [*arg2* hepatoesplenomegalia] e [*arg2* esteatose] hepática
resulta	[*arg2* sem alterações] à [*arg1* auscultação]

são adicionadas e revistas constantemente (Voss, 2005). Deste modo, a extracção de conhecimento a partir da Wikipédia para o PLN é uma forma promissora de permitir a criação de aplicações em grande escala, aplicáveis em situações da vida real. De facto, vários estudos surgiram recentemente em que a Wikipédia é explorada como fonte de conhecimento ((Auer et al., 2007); (Ruiz-Casado, Alfonseca, and Castells, 2006); (Santos et al., 2008); (Wu and Weld, 2007); (Zesch, Müller, and Gurevych, 2008)). A maior parte destes estudos concentram-se na extracção automática de almanaques da Wikipédia (Toral and Munoz, 2006) e na utilização da estrutura interna da Wikipédia para a desambigação de entidades mencionadas (Bunescu and Pasca, 2006). O REMMA baseia-se no método apresentado em (Kazama and Torisawa, 2007), onde se utiliza o sintagma nominal da primeira frase de um artigo Wikipédia para a extracção da categoria semântica. No REMMA, optou-se por identificar na primeira frase do artigo um conjunto de palavras indicativas da categoria e tipo de uma dada entidade. Por exemplo, o artigo Wikipédia sobre o *Acidente Vascular Cerebral* começa com a seguinte frase:

O Acidente Vascular Cerebral (AVC), ou Acidente Vascular Encefálico (AVE), vulgarmente chamado de"derrame cerebral", é caracterizado pela perda rápida de função neurológica, decorrente do entupimento ou rompimento de vasos sanguíneos cerebrais; é uma doença de início súbito, que pode ocorrer por dois motivos: isquemia ou hemorragia.

A extracção da palavra `doença` desta frase permite inferir a classificação a atribuir à entidade *Acidente Vascular Cerebral*. O método utilizado na obtenção destas classificações é descrito em detalhe na secção 3.

A Wikipédia disponibiliza o todo conteúdo para cada uma das diferentes línguas, em formato XML, bem como as ferramentas necessárias para a sua conversão para SQL, formato utilizado pelo REMMA na tarefa de classificação de entidades[9].

[9] O esquema completo da base de dados pode ser consultado em `http://www.mediawiki.org/wiki/Manual:Database_layout`

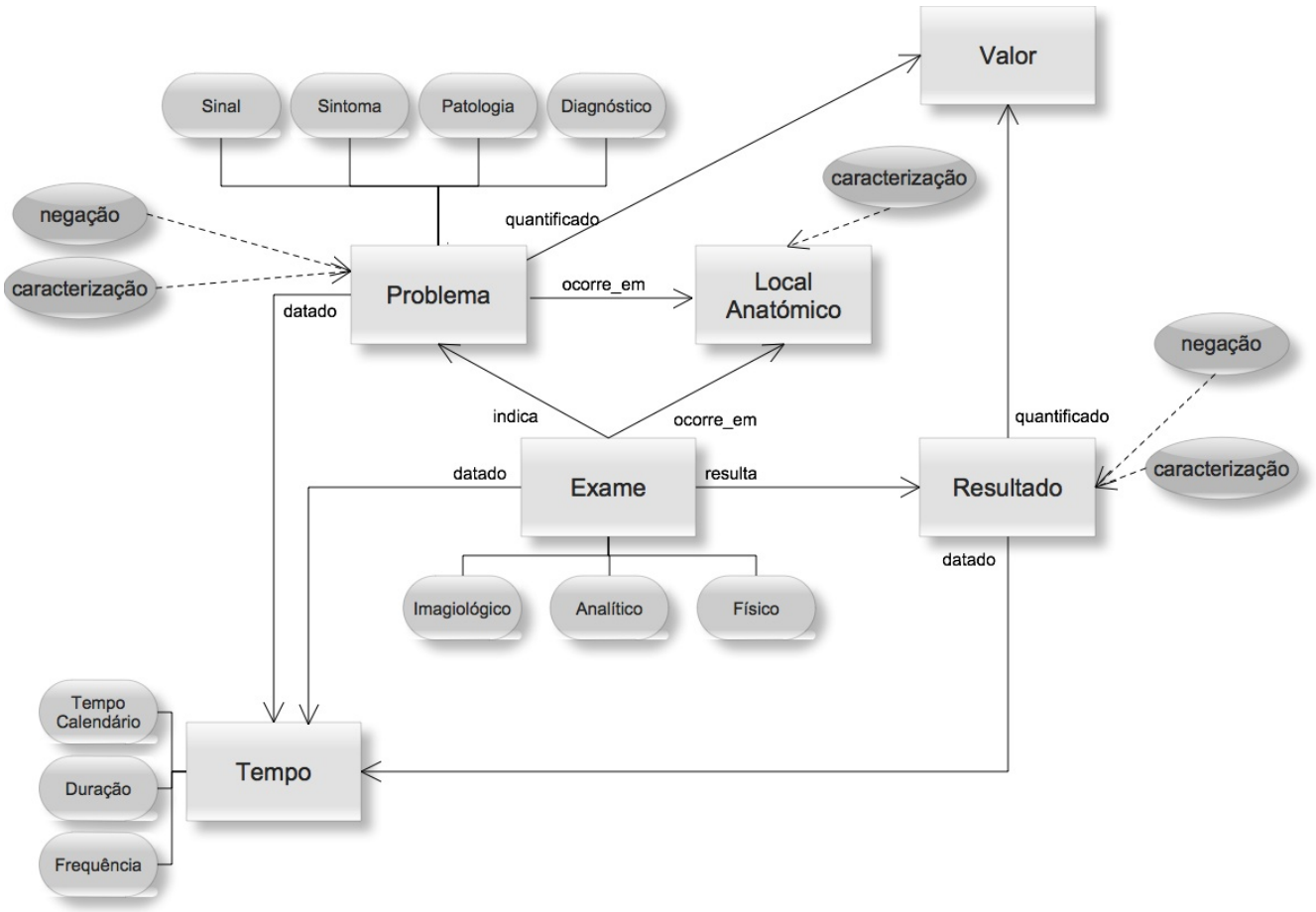


Figura 3: Esquema de anotação do Motivo de Internamento.

3 O sistema REMMA

O sistema REMMA foi inicialmente desenvolvido tendo em vista a participação no Segundo HAREM (Mota and Santos, 2008), uma avaliação conjunta na área do reconhecimento de entidades mencionadas em português, realizada em Abril de 2008. Para este evento o REMMA tinha como objectivo o reconhecimento de entidades mencionadas em textos de domínio geral, principalmente noticiosos (Ferreira, Teixeira, and Cunha, 2008). Para a extracção de informação de textos médicos, especificamente relativos a motivos de episódios de internamento hospitalar, várias adaptações foram realizadas. A secção seguinte descreve a arquitectura e os métodos usados para a identificação e classificação semântica das entidades e relacionamentos destes relatórios.

3.1 Arquitectura

Uma característica do sistema é a sua integração na plataforma UIMA. O UIMA, *Unstructured Information Management Architecture* (Ferrucci and Lally, 2004), é uma plataforma livre, escalável e extensível, para a criação, integração e desenvolvimento de sistemas de gestão de informação não estruturada. Embora seja uma arquitectura com um certo grau de complexidade, tem diversas vantagens, como por exemplo:

- Disponibiliza algumas ferramentas de pré-processamento, tais como leitores e finalizadores genéricos, atomizador, separador em frases e outros anotadores simples;
- Uniformiza a estrutura dos resultados;
- Foca na modelação em vez de na programação.

O UIMA usa uma Estrutura de Análise Comum (CAS, *Common Analysis Structure*) que permite aos anotadores acesso de leitura ao objecto a ser processado (por exemplo, um documento) e acesso de leitura/escrita aos resultados da análise ou às anotações associadas às diferentes regiões dos objectos. Estas regiões podem corresponder a palavras, frases ou parágrafos no texto. O CAS é partilhado entre os diversos anotadores que processam a colecção de objectos, passando de um anotador para seguinte no processo.

A arquitectura do REMMA está apresentada na Figura 5.

O REMMA começa por ler os documentos, um por um, e guardar os respectivos metadados.

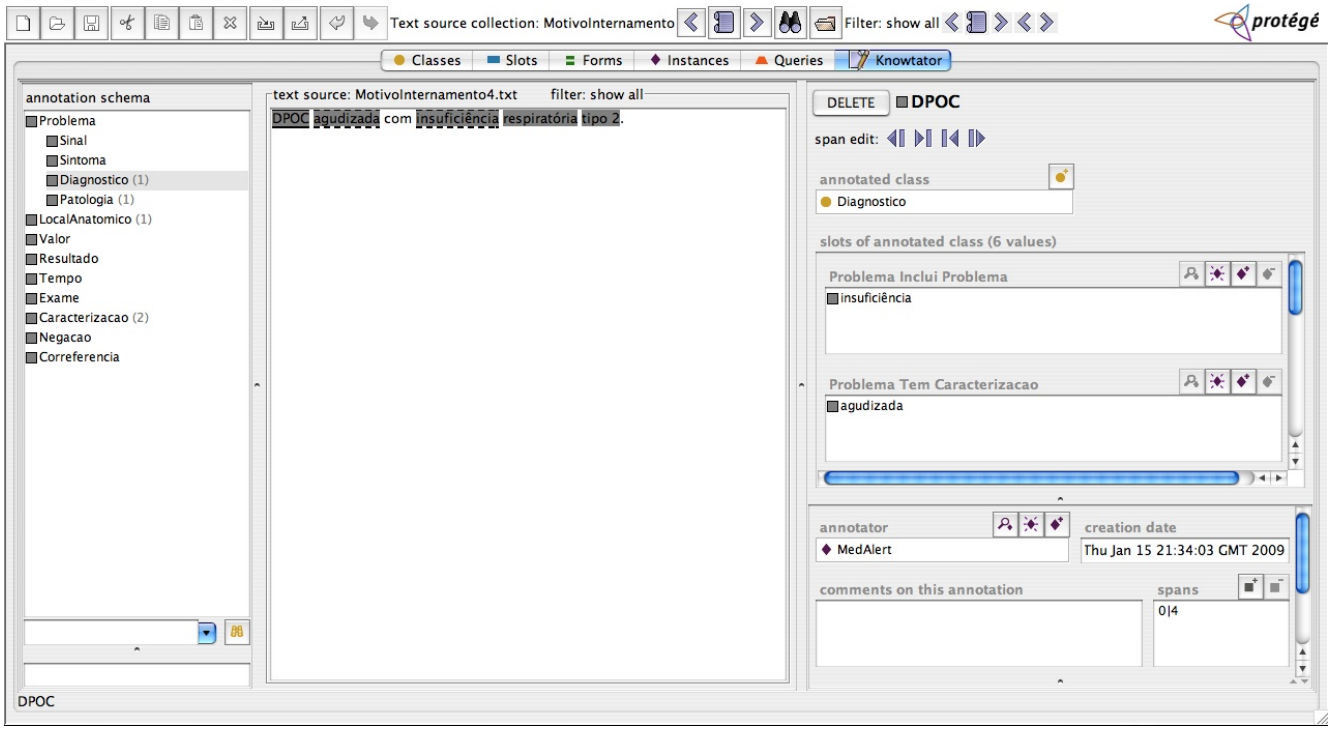


Figura 4: Motivo de Internamento no knowtator.

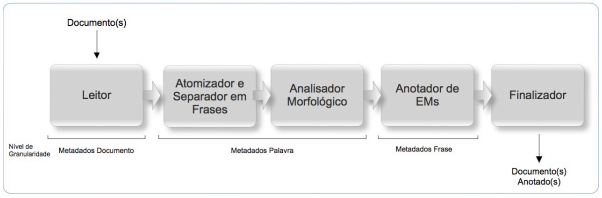


Figura 5: Arquitectura do REMMA

Os textos são posteriormente divididos em frases e tokens com a ajuda das ferramentas de pré-processamento disponíveis no UIMA. O analisador TreeTagger (Schmid, 1995) foi usado na obtenção das categorias morfossintáticas.

As anotações geradas por estas ferramentas são armazenadas no CAS e usadas nos diversos anotadores que constituem o módulo de extracção de informação. A Figura 6 apresenta a sequência de anotadores utilizados na identificação e classificação das entidades e relacionamentos. Estes anotadores são apresentados em mais detalhe nas secções seguintes.

O primeiro anotador a ser invocado é o Anotador de Candidatos que identifica excertos de frases com mais possibilidade de conterem entidades. As expressões candidatas são todos os conjuntos de palavras separadas por termos de ligação como as preposições *com* ou *por*, ou por pontuação. Estas expressões candidatas são posteriormente analisadas pelos anotadores de classificação.

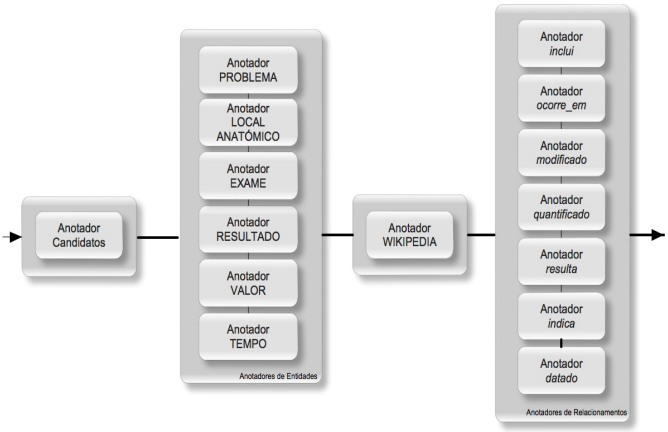


Figura 6: Anotadores do REMMA

O REMMA foi desenvolvido de modo a contemplar duas abordagens de classificação distintas. A primeira baseia-se em almanaques e regras muito simples, apresentada na secção 3.1.1 e a segunda é realizada com base na informação extraída da Wikipédia. Esta ultima é descrita em mais detalhe na secção 3.1.2. Os anotadores desenvolvidos para a identificação e classificação de relacionamentos são descritos na secção 3.1.3

3.1.1 Classificação semântica com base em regras e almanaques

Esta primeira abordagem baseou-se numa utilização combinada de um conjunto de regras de

análise de contexto com a consulta das fontes de conhecimento externas descritas na Secção 2.2.
As regras utilizadas foram criadas manualmente e baseiam-se, não só no contexto em que a expressão é referida, mas também na existência de palavras com prefixos ou sufixos indicativos de classificação semântica. Por exemplo, na identificação e classificação de termos pertencentes à classe semântica *Problema* foram procuradas expressões começadas por `Síndrome de` ou `Insuficiência`, bem como palavras começadas por `hiper, hipo, hemo` ou terminadas em `patia, algia, ismo, ose, oma`.

Os anotadores que usam a informação contida nestes almanaques e regras começam por dividir a expressão candidata nos seus vários termos e atribuem categoria semântica caso algum dos termos da expressão exista nas listas usadas. Quando esta anotação não é conseguida, aplicam na expressão candidata as regras contextuais desenvolvidas para a classe semântica em análise.

3.1.2 Classificação semântica com recurso à Wikipédia

Na tarefa de classificação semântica com base na informação extraída da Wikipédia foi utilizado um subconjunto de todo o conteúdo da Wikipédia, que é disponibilizado em XML para cada uma das diferentes línguas. Foi utilizada a Wikipédia portuguesa de Fevereiro de 2008, que inclui 1 290 836 páginas. Os dados foram posteriormente exportados para uma base de dados SQL, de modo a poderem ser usados neste sistema.

O Anotador Wikipédia foi desenvolvido de modo a encontrar uma entidade na Wikipédia correspondente à identificada nos textos em análise. Deste modo, cada um dos termos existentes nas entidades candidatas identificadas é convertido num nome de entidade Wikipédia através da concatenação dos vários termos da expressão com o caracter "_". Por exemplo, a expressão *Acidente Vascular Cerebral* é convertida em `Acidente_vascular_cerebral` e o artigo correspondente recuperado.

Embora não exista uma regra de formatação estrita, é normal que os artigos Wikipédia comecem com uma pequena frase que define a entidade descrita no artigo. Por exemplo, como foi visto anteriormente o artigo com o título `Acidente_vascular_cerebral` ou `AVC` contém a frase:

O Acidente Vascular Cerebral (AVC) ... é uma doença de início súbito, que pode ocorrer por dois motivos: isquemia ou hemorragia

Tal como neste exemplo a primeira frase, da maioria dos artigos, contém uma expressão que indica a categoria semântica da entidade em análise. Neste caso, a palavra *doença*.

O método seguido concentra-se assim na extracção de tais nomes, a partir da primeira frase do artigo. Para tal foi necessário começar por remover etiquetas desnecessárias, tais como itálicos, negritos e ligações internas. O artigo foi posteriormente dividido em frases de acordo com os padrões `\n, <br>` e regras simples de segmentação para o ponto final (.).

Após obtenção da primeira frase foram aplicadas regras simples, semelhantes às utilizadas no método anterior, ou seja, procuram na primeira frase do artigo Wikipédia palavras-chave indicativas da classe semântica do artigo. Alguns exemplos, bem como a quantidade de palavras utilizadas por este anotador, são listados na tabela 4.

Tabela 4: Exemplos e quantidade de palavras-chave usadas na extracção de uma categoria semântica da primeira frase de um artigo.

Categoria	Exemplos

Problema (N=13)	`doença`
	`trauma`
	`sintoma`
	. . .

Local Anatómico (N=6)	`corpo humano`
	`órgão`
	`sistema`
	. . .

Exame (N=5)	`exame`
	`método de diagnóstico`
	`meio complementar de diagnóstico`
	. . .

3.1.3 Identificação e classificação de relacionamentos entre entidades

O anotador de relacionamentos do REMMA usa ainda um método muito simples e inicial para a detecção de relacionamentos entre entidades. Este usa a informação relativa às várias entidades identificadas nos passos anteriores, em conjunto com os termos de ligação usados pelo anotador de candidatos na identificação dos termos candidatos.

Especificamente, este anotador analisa as entidades identificadas e classificadas em cada uma das expressões candidatas e determina a presença na mesma expressão candidata de entidades pertencentes a categorias relacionáveis, por exemplo, caso uma expressão candidata contenha entidades pertencentes às categorias *Problema* e *Caracterização*, o relacionamento `modificado` é marcado entre estas entidades.

Um método particular é utilizado na identificação dos relacionamento de *inclusão*. Neste

caso, todas as sequências de expressões candidatas ligadas pela preposição *com* são analisadas. Caso ambas contenham pelo menos uma entidade pertencente às categorias *Problema* ou *Exame*, estas são marcadas como relacionadas.

Após a anotação das entidades identificadas pelos vários métodos descritos, um último anotador é chamado, o Finalizador. Este anotador analisa o CAS e cria o(s) documento(s) de saída. É este anotador que produz o documento XML final, através da análise das anotações associadas às diferentes regiões do(s) documento(s). Um exemplo da saída gerada por este anotador é apresentado de seguida. No exemplo, as entidades identificadas são marcadas com a etiqueta equivalente ao nome da entidade, sendo ainda atribuída uma identificação única, `ID`, usada na marcação dos relacionamentos entre entidades.

```
<PROBLEMA ID='p1'
    REL='c6' TIPOREL='caracterizado'
    REL='p20' TIPOREL='inclui'>
      DPOC
</PROBLEMA>
<CARACTERIZACAO ID='c6'>
      agudizada
</CARACTERIZACAO>
       com
<PROBLEMA ID='p20'
   REL='l34' TIPOREL='ocorre_em'
   REL='c47' TIPOREL='caracterizado'>
      insuficiência
</PROBLEMA>
<LOCAL ID='l34'>
      respiratória
</LOCAL>
 <CARACTERIZACAO ID='c47'>
      tipo 2
</CARACTERIZACAO> .
```

4 Resultados

Ao longo do processo de criação da colecção dourada MedAlert diversas avaliações foram efectuadas. A secção 4.1 apresenta os resultados obtidos no processo de definição das directivas de anotação e posterior anotação manual. A secção 4.2 concentra-se nos resultados obtidos na tarefa de reconhecimento automático de entidades e dos relacionamentos entre estas.

4.1 Anotação Manual

Na construção das directivas finais para a anotação da colecção dourada foi obtido um nível de concordância (IAA) de 100%, quer na anotação manual de entidades, quer na anotação de relacionamentos entre estas.

A anotação manual da colecção dourada foi realizada por dois anotadores que seguiram os vários passos e conceitos descritos nas directivas desenvolvidas. Um dos anotadores possui conhecimento médico especializado, mas não tem conhecimentos de processamento de linguagem natural, enquanto que o outro anotador não possui qualquer conhecimento médico especializado, mas tem alguma experiência na anotação de colecções de texto médico.

O nível de concordância inter-anotadores obtido na anotação manual das entidades e seus relacionamentos é apresentado nas tabelas 5 e 6, respectivamente, onde se verifica um IAA de 80% para a anotação de entidades e de 66% na anotação de relacionamentos. Relembramos que apenas os relacionamentos que ambos os anotadores encontraram foram contabilizados.

Estes resultados demonstram claramente a dificuldade, não só na definição de directivas claras em áreas tão especializadas como a medicina, mas também em conseguir que os anotadores sigam as directivas de uma forma consistente. Foram verificados vários problemas como a não concordância em limites de entidades, a inclusão ou não de preposições nas entidades, a dificuldades em separar o conceito de caracterização ou negação, ou mesmo os conceitos de caracterização e local anatómico. Na anotação de relacionamentos verificou-se uma dificuldade acrescida na definição de quais as entidades envolvidas no relacionamento. Por exemplo, qual a entidade caracterizada ou qual a entidade que inclui outra entidade.

4.2 Extracção de Informação

Os resultados obtidos na tarefa de reconhecimento de entidades e relacionamentos são sumariados nas tabelas 7 e 8, respectivamente. As linhas apresentam o número de entidades e relacionamentos correctamente anotados pelo sistema, parcialmente correctos, falsos positivos e as entidades e relacionamentos que o sistema não foi capaz de identificar. Os resultados em termos de Precisão, Abrangência e Medida F estão nas linhas finais da tabela.

Uma precisão de 100% foi obtida para as entidades `LOCAL_ANATOMICO` e `TEMPO`, bem como para os diversos relacionamentos definidos, excepto para o relacionamento `datado`. Estes resultados permitem afirmar que o REMMA, embora esteja ainda numa fase inicial de adaptação à área médica e usando ainda métodos muito simples, é um sistema bastante preciso. Note-se que no contexto da extracção de informação na área da medicina, importa a existência de um sistema preciso, capaz de anotar a informação existente, em oposição a um sistema que extraia muita informação incor-

Tabela 5: Índice de concordância inter-anotadores na anotação manual das entidades

	PROBLEMA	LOCAL_ANATOMICO	CARACTERIZACAO	TEMPO	**Total**
Concordância	20	9	9	0	**38**
Concordância parcial	4	2	0	1	**7**
Não concordância	1	3	4	2	**10**
IAA	**0,96**	**0,89**	**0,69**	**0,20**	**0,80**

Tabela 6: Índice de concordância inter-anotadores na anotação manual dos relacionamentos

	inclui	ocorre_em	caracterizado	datado	**Total**
Concordância	3	11	7	0	**21**
Concordância parcial	0	0	0	0	**0**
Não concordância	2	3	5	1	**11**
IAA	**0,60**	**0,78**	**0,58**	**0,00**	**0,66**

recta ou com ruído.

É de notar a presença em alguns relatórios de problemas na escrita de algumas palavras, situação comum na escrita deste tipo de relatórios descritivos realizados em simultâneo ou imediatamente após a observação do paciente. Um exemplo comum deste tipo de problema é a escrita da palavra `disgestiva` em vez de `digestiva` dificultando a procura do seu significado nas fontes de conhecimento usadas pelo REMMA.

5 Conclusões

Para a extracção automática de informação de relatórios médicos é indispensável a existência de um *corpus* anotado semanticamente, quer com múltiplas entidades, quer com as suas relações. Para tal, foi apresentada uma metodologia para a anotação manual de uma colecção dourada de relatórios médicos de episódios de internamento hospitalar. Esta colecção dourada pretende auxiliar o processo de extracção de informação e sua avaliação. Os resultados iniciais mostram a importância da criação de directivas claras e precisas de modo a atingir bons valores de concordância entre anotadores, bem como a necessidade de coordenação e motivação entre anotadores.

Para a extracção de informação foi utilizado o sistema REMMA, um sistema composto por um conjunto de anotadores UIMA, capaz de usufruir de vários tipos de recursos, sejam estes almanaques especializados, ou, categorias semânticas extraídas a partir da análise da primeira frase de um artigo da Wikipédia. Apesar estar ainda em fase inicial, o REMMA apresenta resultados consistentes com um sistema bastante preciso, característica importante em sistemas de apoio à decisão médica.

5.1 Trabalho Futuro

O projecto MedAlert pretende actuar como um sistema de apoio à decisão clínica, capaz por exemplo, de inferir de uma forma automática dúvidas suscitadas pelas decisões médicas através da análise de relatórios médicos e de textos contendo directivas médicas. Assim, o aumento do conjunto de textos anotados semanticamente, textos estes pertencentes a todas as fases relativas ao processo de internamento hospital, é crucial no desenvolvimento de um sistema útil. De modo a melhorar a qualidade da anotação, é também essencial o aumento do leque de anotadores especializados.

A utilização da Wikipédia no REMMA foi útil para a melhoria da classificação das entidades mencionadas, dando uma indicação clara da utilidade deste tipo de fontes de conhecimento. Existem actualmente diversas wikis públicas e relativas a vários domínios. O futuro do sistema REMMA poderá passar, assim, pela utilização de recursos semelhantes relativos à área da medicina, de modo a melhorar a tarefa de extracção de informação. No entanto, o acesso e utilização de fontes de conhecimento especializadas, em particular o acesso ao vocabulário biomédico DeCS, é uma das tarefas prementes no âmbito do projecto MedAlert. Este tipo de informação segue uma estrutura bem definida e aceite internacionalmente, pelo que permite a estandardização das regras a serem aplicadas em sistemas como o MedAlert.

A natureza descritiva e expontânea dos relatórios médicos analisados, escritos em contexto de consulta hospitalar, leva à existência de vários erros ortográficos. Esta situação é mais grave quando se utilizam de sistemas de extracção de informação baseados em fontes de conhecimento, como é o caso do REMMA. Este problema ficou demonstrado nos resultados obtidos. Assim, a utilização e adaptação de um sistema de correcção ortográfica à área da medicina é um dos próximos passos do projecto MedAlert.

Agradecimentos

O projecto RTS foi financiado pelo programa "Aveiro Digital" da iniciativa "Portugal Digital"

Tabela 7: Resultados na tarefa de reconhecimento de entidades

	PROBLEMA	LOCAL_ANATOMICO	CARACTERIZACAO	TEMPO	**Total**
Saídas correctas	22	13	9	1	45
Parcialmente correctas	3	0	1	0	4
Falsos positivos	0	0	0	0	0
Em falta	3	3	2	0	8
Total	**28**	**16**	**12**	**1**	**57**
Precisão	0,88	1,00	0,90	1,00	0,92
Abrangência	0,89	0,81	0,83	1,00	0,86
Medida F	**0,88**	**0,89**	**0,86**	**1,00**	**0,89**

Tabela 8: Resultados na tarefa de reconhecimento de relacionamentos

	inclui	ocorre_em	caracterizado	datado	**Total**
Saídas correctas	4	11	7	0	22
Parcialmente correctas	0	0	0	0	0
Falsos positivos	0	0	0	0	0
Em falta	1	3	3	2	9
Total	**5**	**14**	**10**	**2**	**31**
Precisão	1,00	1,00	1,00	0,00	1,00
Abrangência	0,80	0,78	0,70	0,00	0,71
Medida F	**0,89**	**0,88**	**0,82**	**0**	**0,82**

e pelo programa POSI do Governo Português.

References

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *In 6th Int'l Semantic Web Conference, Busan, Korea*, pages 11–15. Springer.

Bunescu, Razvan and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Abril.

Cunha, João Paulo Silva, Isabel Cruz, Ilídio Oliveira, António Sousa Pereira, César Telmo Costa, Ana Margarida Oliveira, and Amândio Pereira. 2006. The RTS project: Promoting secure and effective clinical telematic communication within the Aveiro region. In *Em eHealth 2006 High Level Conference*, pages 1–10, Maio.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Julho.

Ferreira, Liliana, António Teixeira, and João Paulo Silva Cunha. 2008. REMMA- Reconhecimento de Entidades Mencionadas do MedAlert. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, Aveiro, Portugal, 7 de Setembro.

Ferrucci, David and Adam Lally. 2004. UIMA an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Gennari, John H., Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, and Samson W. Tu. 2002. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. *International Journal of Human-Computer Studies*, 58:89–123.

Kazama, Jun'ichi and Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, June.

Mota, Cristina and Diana Santos, editors. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, Aveiro, Portugal, 7 de Setembro.

NLM, editor. 2008. *UMLS Knowledge Sources*. National Library of Medicine, Novembro.

Ogren, Philip. 2006. knowtator: A plug-in for creating training and evaluation data sets for biomedical natural language systems. In *Proceedings of the 9th INternational Protégé Conference*, pages 73–76, Stanford, California.

Roberts, A., R. Gaizauskas, M. Hepple, N. Davis, G. Demetriou, Y. Guo, J. Kola, I. Roberts, A. Setzer, A. Tapuria, and B. Wheeldin. 2007. The CLEF Corpus: Semantic Annotation of Clinical Text. In J. M. Teich, J. Suermondt, and G. Hripcsak, editors, *American Medical Informatics Association 2007 Proceedings. Biomedical and Health Informatics: From Foundations to Applications to Policy*, pages 625–629, Chicago, IL, USA, November. American Medical Informatics Association.

Ruiz-Casado, Maria, Enrique Alfonseca, and Pablo Castells. 2006. From wikipedia to semantic relationships: a semi-automated annotation approach. In *1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006)*, Junho.

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. 2008. Getting geographical answers from Wikipedia: the GIKIP pilot at CLEF. In *Working notes for the Cross Language Evaluation Forum, CLEF'2008*, 17–19 Setembro.

Schmid, Helmut. 1995. TreeTagger, a language independent part-of-speech tagger. *Institut fur Maschinelle Sprachverarbeitung, Universidade de Estugarda.*

Toral, Antonio and Rafael Munoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition using wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Abril.

Voss, Jakob. 2005. Measuring Wikipedia. In *10th International Conference of the International Society for Scientometrics and Informatics*, pages 221–231, Julho.

Wu, Fei and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50, New York, NY, USA. ACM.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Maio.

Novas Perspectivas

Conceitos, classes **e/ou** universais: com o que é que se constrói uma ontologia?

Patrícia Cunha França
Mestranda em Ciências da Linguagem
(Área de Especialização em Língua e Tecnologias de Informação)
Universidade do Minho
pg10122@alunos.uminho.pt

Resumo

O termo 'ontologia' é frequentemente usado no âmbito da Ciência da Computação para referir-se a uma "especificação [...] de uma conceptualização" (Gruber, 1993: 2). Mas será de conceitos que a ontologia trata? Smith, por exemplo, que tem vindo a desenvolver o seu trabalho sobre ontologias no âmbito da biomedicina argumenta que as ontologias, pelo menos as ontologias científicas de domínio específico e científico, não tratam de conceitos mas de universais (Smith, s.d. [b.])

Este artigo tem por objectivo analisar os termos usados por diferentes autores, que têm vindo a contribuir para o estudo das ontologias, de forma a encontrar um denominador comum.

1. Introdução

Não obstante o termo 'ontologia' ter surgido no âmbito da Filosofia ele tem vindo a ganhar uma nova dimensão no seio da comunidade das Ciências da Computação e Informação pelo menos desde a década de 90 do século passado. As vantagens da criação, uso e aplicação de ontologias têm sido largamente defendidas e demonstradas dentro dessa comunidade (Abecker and van Elst, 2004; Mika et al., 2004), nomeadamente no que concerne à organização e partilha do conhecimento, pressupostos essenciais no que concerne a políticas de acesso livre.

Actualmente o interesse pelas ontologias tem vindo a estender-se a outras áreas e domínios específicos: às Ciências Sociais (Lawson, 2004), às Ciências Naturais, especificamente à Biomedicina (Smith, 2008; Heuer and Hennig, 2008), à Bioética (Cohnitz and Smith, s.d.; Smith and Brogaard, 2003) e à informação geográfica (Smith and Mark, 2001), bem como às Ciências da Linguagem (Schalley and Zaefferer, 2008). As ontologias têm vindo a afirmar-se como instrumentos eficazes de disseminação de conhecimento, de partilha e de diálogo.

Ora, este alargamento de interesse levou a que diferentes pessoas de diferentes áreas, que trouxeram consigo a terminologia específica das suas disciplinas, começassem a trabalhar em conjunto. Na construção de uma ontologia cooperam – ou idealmente deveriam cooperar – filósofos, linguistas, engenheiros informáticos e especialistas de um domínio específico (no caso específico das ontologias de domínio). Esta interdisciplinaridade, desejável e inevitável, trouxe consigo alguns desafios, nomeadamente a determinação de um consenso terminológico.

Em nome de uma compreensão mútua por parte dos intervenientes no processo de construção de ontologias, é desejável que, se não for possível encontrar uma terminologia comum, pelo menos a terminologia usada pelas várias partes seja compreendida reciprocamente. Assim, questões como o que é exactamente uma ontologia, como se constrói e o que faz parte dela tornam-se questões cujas respostas exigem um acordo prévio.

O maior problema das terminologias propostas não é, como creio, a sua incompatibilidade nem mesmo as questões em torno de posições epistemológicas opostas, mas a quantidade de termos usados indiscriminadamente sem uma definição clara e compreensível.

Neste artigo proponho analisar algumas propostas terminológicas que têm vindo a contribuir para os estudos no âmbito da ontologia, nomeadamente no que ajudam a esclarecer o objecto específico que lhe dá corpo.

Este artigo trata especialmente das denominadas ontologias genéricas[1] (de "*top-level*") (Guarino,

1 As ontologias genéricas - *"top-level ontologies"*- são definidas por Guarino (1998) como as ontologias que "descrevem conceitos gerais como espaço, tempo, matéria, objecto, evento, acção, etc., independentes de um domínio ou problema particulares". São usualmente referidas como exemplo deste tipo de ontologias a Wordnet (http://wordnet.princeton.edu) e a Cyc (http://opencyc.org). As ontologias genéricas distinguem-se das ontologias de domínio, das ontologias de tarefa e das ontologias de aplicação (Guarino, 1998).

1998), muito embora considere que a distinção entre os tipos de ontologias propostos por Guarino seja, em certa medida, irrelevante aqui. E isto é assim porque este artigo incide sobre a fase inicial do processo de construção de uma ontologia, fase esta que, à partida, fará parte de todas as ontologias. Como referem Degen e Herre (s.d.) "toda a ontologia de domínio específico terá de usar como base de trabalho alguma ontologia de nível superior que descreva as categorias da realidade mais gerais e independentes de domínio".

No ponto 2 serão analisadas algumas propostas de definição do termo 'ontologia' que partem de conceitos. Partirei da origem da noção de ontologia para a origem da palavra no seio da Filosofia para chegar à definição de Gruber, no seio da Ciência da Computação e à noção de Ontolinguística, no âmbito das Ciências da Linguagem.

O ponto 3 trará para discussão algumas objecções à definição que liga ontologia a conceitos, explorando a relação entre ontologia e realidade e, por consequência, a relação entre termo, conceito e realidade.

No ponto 4 será exposto e estudado o quadro conceptual de análise das várias propostas para a noção de ontologia proposto por Nickles et al. (2003). Partindo deste quadro as várias propostas de definição de ontologia serão comparadas tendo em conta a sua posição ali.

O ponto 5 será deixado para as conclusões.

2. Ontologia e conceptualização

> "*D. So you define dress by referring to what people think dresses are?*
>
> *A. Yes. […] What I try to define is the concept 'dress' that people have, not actual dresses*"
>
> Geeraerts, 2006: 425

2.1 Das origens

É um lugar-comum começar um trabalho sobre ontologias com a definição que liga o termo à Filosofia. Uma ontologia é definida a partir do seu estatuto etimológico: do grego ón, *óntos* - 'ser' - e *logos* - 'palavra', 'discurso', 'razão'.

Da última vez que procurei a palavra num comum dicionário de língua o único sentido existente era retirado do domínio da Filosofia: estudo do ser, do que existe.

Não obstante o facto de parecer consensual reportar a origem do termo 'ontologia' a Aristóteles, foi com o cunho de Jacob Lorhard que a palavra ganhou existência em 1606 no seu livro *Ogdoas Scholastica*[2], um volume composto por oito livros referentes a matérias como gramática latina e grega, lógica, retórica, astronomia, ética, física e metafísica (ou ontologia) (Øhrstrøm et al., 2007: 3). Lorhard define a sua 'ontologia' como

> the science of the intelligible as intelligible insofar as it is intelligible by man by means of the natural light of reason without any concept of matter
>
> Lorhard, J., 1606: Livro 8, p. 1[3]

De sublinhar aqui que Lorhard define a ontologia como a ciência do que é inteligível pelo homem através da razão, sem influência da matéria. Esta definição vai contra a proposta de ontologia, ou filosofia primeira[4], de Aristóteles, tal como é entendida nas suas duas obras mais relevantes sobre o tema: os escritos que mais tarde foram compilados sob o título *Metafísica* e as *Categorias*[5]. A ênfase de Lorhard na razão em detrimento da matéria é determinante para a construção da mais recente noção de ontologia no âmbito da Ciência da Computação. Ela é também determinante para a noção de ontologia em alguns estudos recentes no âmbito da Linguística.

2.2 Ontologia e Ciência da Computação

Segundo pretende Smith (Smith, s.d.[a.]: 22-23) o termo 'ontologia' é usado pela primeira vez dentro da comunidade da Ciência da Computação em 1967, num trabalho de S. H.

2 Uma tradução em inglês, feita por Sara L. Uckelman (Institute for Logic, Language, and Computation da Universiteit van Amsterdam), do capítulo 8 desta obra de Lorhard está disponível em http://www.illc.uva.nl/Publications/ResearchReports/X-2008-04.text.pdf [cons. 19-09-2008].

3 Lorhard, J. (1606). *Ogdoas scholastica*. Sangalli, Livro 8, p. 1, apud Øhrstrøm et al. (2007: 4).

4 O conceito de filosofia primeira de Aristóteles pode ser considerado o embrião do conceito do que mais tarde, já no século XVII, como vimos, viria a chamar-se ontologia.

5 Aristóteles vê a matéria como algo enganoso, mutável, da qual nada se pode dizer com verdade e entende a forma como a essência dos seres. Não obstante, Aristóteles entende a forma como inseparável da matéria. Para Aristóteles, "a substância [entendida como a forma e a matéria] deverá ser qualquer coisa, «um sujeito real e determinado»" (Ricoeur, 1992: 904).

Mealy sobre processamento de *data*. Não obstante, é uma definição de Gruber do termo 'ontologia' que aparece citada com maior frequência nos trabalhos sobre ontologias daquela comunidade (Uschold and Gruninger, 1996; Almeida e Bax, 2003; Staab and Studer (ed.s), 2004; Mika, s.d.; Pisanelli et al., s.d.; Morais, s.d.). Diz Gruber que uma ontologia é "uma especificação explícita de uma conceptualização", sendo que o termo 'conceptualização' é definido como "uma visão do mundo abstracta e simplificada que desejamos representar para um propósito qualquer"[6] (Gruber, 1993: 1). E um pouco antes, no mesmo artigo, Gruber toma de Genesereth & Nilsson (1987) a definição de conceptualização como "os objectos, conceitos, e outras entidades que se assumem existir dentro de uma área de interesse e as relações que existem entre eles"[7] (Gruber, 1993: 1).

Então, numa conceptualização cabem ao mesmo tempo conceitos, objectos e as relações que se assumem existir entre esses objectos e conceitos dentro de uma área de interesse? Cabem todas as entidades[8], tudo, independentemente de serem consideradas materiais, imateriais, processuais, enfim...?

Antes de respondermos a esta questão, tomemos para análise o exemplo que Nickles et al. (2007: 27) usam para interpretar a definição de Gruber. Se a nossa área de interesse for, por exemplo, a nossa secretária, e se presumirmos que existem ali em cima objectos - uma caneta, papéis, lápis, um livro, etc. - será que esses objectos cabem numa conceptualização? Uma caneta, um lápis, um livro, sendo objectos, fazem parte de uma conceptualização? À partida, dificilmente responderíamos afirmativamente a esta questão. Como referem Nickles et al., o que faz parte de uma conceptualização são os conceitos desses objectos: o conceito de caneta, o conceito de livro, etc..

Independentemente de considerarmos que são os objectos ou os conceitos o material de trabalho de um ontologista, colocar objectos e conceitos no mesmo nível é partir do pressuposto errado, pelo menos do ponto de vista teórico.

Sendo assim, o erro de Gruber foi a sua definição do termo 'conceptualização' e não a sua definição do termo 'ontologia'. E, se olharmos para a definição de ontologia de Lorhard, verificamos que a definição de Gruber, ao reportar-se a uma "especificação [...] de uma conceptualização" assenta no mesmo princípio do pedagogo do século XVII.

A definição de Gruber do termo 'ontologia' foi já analisada por Guarino (1996) e Guarino e Giaretta (1995). Nestes dois artigos o alvo da crítica não é tanto a definição de ontologia mas, precisamente, a definição que Gruber adopta para o termo 'conceptualização'. Guarino e Giaretta começam por propor que uma conceptualização seja entendida como "uma estrutura semântica intensional que codifica as regras implícitas que determinam a estrutura de uma porção da realidade" (Guarino e Giaretta, 1995). Uma conceptualização deve ser distinguida de uma ontologia que, por sua vez, deve ser definida, em sentido restrito, como "uma teoria lógica que fornece uma proposta[9] explícita e parcial de uma conceptualização" (Guarino e Giaretta, 1995).

Uma ontologia, então, é uma teoria que fornece uma linguagem para uma outra teoria que, por sua vez, também fornece uma linguagem que dá conta de um pedaço da realidade?

A distinção que Guarino faz mais tarde, num artigo de 1998, parece-me mais esclarecedora. Neste artigo, uma conceptualização é entendida com a leitura feita do termo 'ontologia' no seio da Filosofia, *i.e.*, "um sistema particular de categorias que dão conta de uma certa visão do mundo" (Guarino, 1998). Uma ontologia, por sua vez, é definida com a leitura feita do mesmo termo no seio da Inteligência artificial (IA), *i.e.*, "um artefacto de engenharia, constituído por um vocabulário específico usado para descrever uma certa realidade, mais uma série de pressupostos explícitos acerca do significado que se atribui a esse vocabulário" (Guarino, 1998).

6 Tradução livre.

7 Tradução livre.

8 Neste artigo, o termo 'entidade' será usado no seu sentido mais alargado, como tudo aquilo que se supõe existir (ou que existe), incluindo coisas, estados, processos, funções, qualidades, crenças, acções, documentos,.... Tudo o que pode ser inserido nos níveis 1, 2 e 3 (Smith, 2006) a que farei referência no ponto 4 deste artigo.

9 O termo '*account*' foi traduzido por 'proposta'.

Esta reformulação da definição do termo 'ontologia' trazida por Guarino é um pouco mais compreensível. É de extrema relevância, penso, a distinção que o autor introduz entre conceptualização e ontologia a partir da linguagem. Uma conceptualização é entendida como uma visão do mundo independentemente da linguagem usada para a representar, enquanto que uma ontologia é dependente de um vocabulário[10] (Guarino (1998).

Isto significa que duas ontologias podem usar diferentes vocabulários e partilhar, ao mesmo tempo, a mesma conceptualização[11]. Aqui está uma das questões deixadas de fora do quadro conceptual da noção de ontologia que será adiantado mais à frente no ponto 4[12].

A questão que aqui se coloca é a de saber onde pertencem os termos que estarão dispostos numa ontologia: à conceptualização ou à ontologia? Nas palavras de Guarino, as categorias pertencem à conceptualização, que é, como diz, independente de uma linguagem. Então, sendo assim, não é necessário usar nenhum vocabulário específico para que essas categorias tenham existência? Estará Guarino a referir-se a conceitos quando fala de "categorias"? À partida, se Guarino considera que uma conceptualização é independente da linguagem, deveria pelo menos especificar a que é que ele se refere quando usa o termo 'categorias'. Como se fazem categorizações sem recorrer a termos, a uma linguagem?

Porque a questão que Guarino desencadeia é a mesma que tem vindo a ser discutida por filósofos e linguistas há mais de dois milénios e prende-se com a questão do significado[13] e pelas disputas acerca dos elementos que fazem parte daquilo que ficou conhecido pelo triângulo de Ogden & Richards (1985, 11): (i) o símbolo, (ii) o pensamento ou a referência e (iii) o referente.

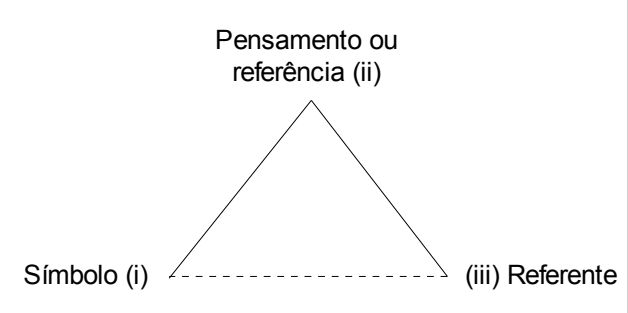


Ilustração 1: Triângulo semiótico de Ogden & Richards (adaptado de Ogden & Richards (1985: 11)).

Pertencerão estas categorias de Guarino à parte (ii) do triângulo de Ogden & Richards? Estas "categorias" a que Guarino se refere correspondem a conceitos, ou àquilo que Ogden & Richards denominam 'pensamento' ou referência?

Qualquer que seja o lado do triângulo onde Guarino desejasse colocar as suas "categorias", seria necessário uma consequente justificação. Não cremos que fosse necessário explorar as teorias semânticas acerca do significado, mas cremos que seria importante saber, para bem do entendimento mútuo, a que se refere Guarino quando usa o termo 'categorias'.

Talvez estas questões percam importância no exacto acto de escrever/inserir os termos numa ontologia, mas ganham uma outra importância quando se tenta definir o termo.

Questões como a que surgiram neste ponto têm vindo a ser trabalhadas numa área que tem contribuído com alguns estudos importantes para aqueles que se dedicam à pesquisa e elaboração de ontologias. Refiro-me à Ontolinguística e dela tratarei no ponto a seguir.

2.3 A Ontolinguística

É precisamente sobre a noção de conceptualização, numa aceitação da definição de Gruber, que os trabalhos mais recentes no domínio da Ontolinguística[14] assentam. Em

10 Guarino usa os termos 'linguagem' e 'vocabulário' como sinónimos.

11 Guarino dá o exemplo do uso de palavras inglesas ou italianas. E aqui cremos que poderiam ser usadas outras formas de convenções que não a linguagem natural.

12 Não obstante, podemos afirmar que na própria definição de ontologia de Gruber que vimos acima, nomeadamente quando se fala de "especificação", está já subjacente a ideia de um vocabulário ou linguagem.

13 Com questões muito pertinentes acerca de saber se a construção de uma conceptualização pode partir de categorias pré-linguísticas ou extra-linguísticas, seja qual for a linguagem a que nos estejamos a referir. Ou se existem categorias de referência que sejam independentes da língua.

14 Tanto quanto sei, o termo 'Ontolinguística' foi usado pela primeira vez como título de um livro editado por Schalley e Zaefferer datado de 2007. Neste livro foram reunidos vários estudos em torno da contribuição do

termos sucintos, a Ontolinguística pode inserir-se no âmbito da Linguística Cognitiva e apresenta-se como uma área de estudo onde se procura encontrar uma ponte entre os mecanismos linguísticos que usamos no nosso dia-a-dia e o nosso conhecimento ontológico. Na verdade, Schalley e Zaefferer acreditam mesmo que o conhecimento linguístico é um tipo especial de conhecimento ontológico (2003:10). A Ontolinguística assenta no pressuposto que existem universais mentais[15] e que as opções que as línguas fornecem para expressar um conceito estão intimamente dependentes da posição que esse conceito ocupa dentro de um sistema conceptual, i.e., dependem do estatuto ontológico desse conceito, das relações que esse conceito estabelece com outros conceitos dentro de um mesmo sistema. Nas palavras de Schalley e Zaefferer a Ontolinguística entende a ontologia como "uma sistema de conceptualizações", ou, para ser mais completo, "uma rede de conceptualizações interconectadas do fenómeno que constitui o mundo" (Schalley e Zaefferer (eds.), 2007: 3). Ora, esta definição, como referi acima, não está muito longe da definição de Gruber exposta no ponto anterior.

Segundo Schalley e Zaefferer (2007: 8-10), o conhecimento ontológico pode ser caracterizado por conhecimento definicional ou analítico[16], mas deve ser distinguido do conhecimento enciclopédico ou do conhecimento do mundo[17]. As relações ontológicas são, para estes autores, relações interconceptuais.

2.3.1. As relações na Ontolinguística

Shalley e Zaefferer distinguem cinco relações taxonómicas e cinco relações meronímicas[18]. As relações taxonómicas dividem-se em (i) subordinação conceptual, em que o conceito A é *c-subordinado* ao conceito B se e só se toda a instância de A for também uma instância de B (por exemplo, PÉ HUMANO é *c-subordinado* ao conceito PARTE DO CORPO HUMANO por que é inconcebível que uma instância do primeiro não seja uma instância do último); (ii) superordenação conceptual, em que se dá o inverso; (iii) equivalência conceptual, em que o conceito A é *c-equivalente* ao conceito B se e só se toda a instância de A for também uma instância de B e vice-versa (por exemplo, PÉ HUMANO é *c-equivalente* ao conceito PÉ HUMANO ESQUERDO OU DIREITO porque é inconcebível que uma entidade instancie apenas um destes dois conceitos; (iv) compatibilidade

conhecimento linguístico para o conhecimento ontológico. Não obstante, o termo 'ontolingua' tinha já sido usado por T. Gruber em 1992 (Gruber, 1992a:5). Ver também Gruber (1992b).

15 Estes universais mentais vêm sendo estudados por exemplo por Wierzbicka (1996; 1992), pela Linguística Cognitiva e pela Linguística Generativa, e são também um dos fundamentos da Ontolinguística – ou pelo menos de uma parte muito considerável de estudos neste domínio. É de notar, no entanto, algumas divergências no seio da Linguística Cognitiva, em relação a este assunto, nomeadamente no que concerne à metodologia. Uma interessante discussão, em forma de diálogo ficional, em torno dos métodos usados nos estudos na Semântica Cognitiva foi elaborada por Dirk Geeraerts (2006).

16 Ao referirem-se ao conhecimento definicional ou analítico Schalley e Zaefferer estão a referir-se ao significado intensional, e ligam-no, precisamente, a conceitos. Não é por acaso que, para estes autores, as relações ontológicas são relações interconceptuais.

17 A única diferença apontada na distinção entre conhecimento ontológico, ou analítico, e o conhecimento enciclopédico, ou conhecimento do mundo, é que o primeiro constitui conhecimento acerca de como o mundo deverá ser, dada a forma como o conceptualizamos, enquanto que conhecimento enciclopédico diz respeito ao conhecimento do mundo como ele é (Schalley and Zaefferer, 2007: 8-9). Não obstante, é de notar que os autores defendem que as linhas que separam os diferentes tipos de conhecimento não são fáceis de traçar (Schalley and Zaefferer, 2007: 10).

18 Algumas das relações propostas pela ontolinguística são equivalentes a algumas das relações semânticas tradicionais, nomeadamente a relação de hierarquia, inclusão, equivalência e oposição (ver Campos e Xavier, 1991; ver também "Terminologia Linguística para o Ensino Básico e Secundário". http://www.prof2000.pt/users/primavera/).
A diferença fundamental entre as relações semânticas tradicionais e as propostas pela Ontolinguística é que na semântica tradicional a ênfase é posta nos itens lexicais – de um modo geral, nas palavras e nas relações que se estabelecem entre palavras e sentidos de palavras -, enquanto que na Ontolinguística lida-se com conceitos, muito embora a definição do termo 'conceito' não esteja definida de forma clara, pelo menos no livro a que faço referência aqui: Schalley and Zaefferer (ed.s), 2007.
De notar ainda que na Ontolinguística há um aproveitamento do referente, numa aceitação do lado (iii) do triângulo semiótico de Ogden & Richards a que fiz referência acima, contrariamente ao que acontece na semântica tradicional. É a chamada semântica extensional (ou semântica referencial). Sobre a relação entre a linguística tradicional e a Linguística Cognitiva com o referente ver Teixeira (2001).

conceptual, em que o conceito A é *c-compatível* com o conceito B se e só se alguma entidade instancie ao mesmo tempo os conceitos A e B (por exemplo, PÉ HUMANO é *c-compatível* com o conceito MAGOADO); e (v) incompatibilidade conceptual, quando se verifica o contrário.

As relações meronímicas compreendem (i) a *x-subordinação meronímica*, (ii) a *x-superordenação meronímica*, (iii) a *x-cosubordinação meronímica a C*, (iv) a *x-compatibilidade meronímica sobre C*; e (v) a *x-incompatibilidade meronímica sobre C*.

Nestas últimas relações (meronímicas: *-m*) 'x' corresponde a uma variável para o tipo de relação 'parte-de'. Os exemplos seguintes representam uma relação de inclusão, referenciada por '*i*-':

(i) o conceito PÉ HUMANO é *m-i-subordinado* ao conceito CORPO HUMANO porque toda a instância completa deste último *i*-inclui uma instância do primeiro;

(ii) o conceito PÉ HUMANO é *m-i-superordenado* ao conceito DEDO GRANDE DO PÉ uma vez que toda a instância completa do primeiro *i*-inclui uma instância do último;

(iii) os conceitos PÉ HUMANO e CABEÇA HUMANA são *m-i-cosubordinados* ao conceito CORPO HUMANO porque toda a instância completa do último *i*-inclui uma instância do primeiro e uma instância do segundo;

(iv) os conceitos DEDO GRANDE DO PÉ e SEXTO DEDO HUMANO são *m-i-compatíveis* sobre o conceito PÉ HUMANO porque há instâncias completas deste último conceito que *i*-incluem tanto uma instância do primeiro conceito como do segundo conceito (supostamente sob uma anomalia chamada polidactilia ou polidatilia);

(v) os conceitos DEDO GRANDE DO PÉ e DÍGITO NUMÉRICO são *m-i-incompatíveis* sobre o conceito PÉ HUMANO porque é inconcebível que uma instância completa do último *i*-inclua tanto uma instância do primeiro como uma instância do segundo conceito.

A principal diferença notável entre as relações taxonómicas e as relações meronímicas é que as primeiras caracterizam-se por existirem apenas a um nível conceptual, enquanto que as relações meronímicas se caracterizam por existirem ao nível das instâncias[19], i.e., podem ser instanciadas.

Tomemos para análise um outro exemplo[20] dado no artigo de Schalley e Zaefferer (2007: 7-8). O primeiro caso refere-se a relações taxonómicas, no segundo caso estamos perante relações meronímicas.

> Se considerarmos por exemplo o pé direito de Edward Teller [...] ao nível da instância e compararmos as suas possíveis conceptualizações como O PÉ DIREITO DE TELLER, PÉ DIREITO e PÉ, respectivamente, isto corresponde a diferentes fotografias com um grau crescente de pormenor da mesma entidade, mas não corresponde a diferentes entidades. [...]
>
> Pelo contrário, se considerarmos, juntamente com o pé direito de Edward Teller, a sua perna direita e o seu corpo e os conceitos PÉ DIREITO DE TELLER, PERNA DIREITA DE TELLER e CORPO DE TELLER, respectivamente, isto dá lugar a uma relação conceptual entre o conceito de uma entidade e os conceitos de outras entidades de que esse conceito faz parte, [...]. Se compararmos estes conceitos com diferentes fotografias, elas não são fotografias da mesma entidade, mas de diferentes entidades que mantêm uma relação material que não é de identidade.[21]
>
> Schalley e Zaefferer , 2007: 7

Como bem referem os autores, é sempre importante relativizar as relações meronímicas de subordinação no instante de proceder à instanciação dos conceitos superordenados[22]. Isto

19 Por instâncias (também particulares ou *tokens*) deve entender-se tudo aquilo que tem existência num espaço e tempo determinados, o que existe aqui e agora. Por exemplo, é frequente distinguir-se tipo, classe ou universal de instância, particular ou *token*, onde, por exemplo, Jean-Pierre Proudhon será considerado instância e o termo 'homem' um possível universal dessa instância. De notar, no entanto, que, como veremos mais adiante, a distinção entre instância e universal, nomeadamente aquando da construção de uma ontologia, é muito ténue, especialmente nas denominadas ontologias de domínio onde o grau de pormenor e o próprio objecto tratado nessa ontologia podem determinar se um termo é considerado uma instância ou uma classe.

20 Todos os exemplos dados para as relações taxonómicas e meronímicas foram retirados do artigo de Schalley e Zaefferer (2007).

21 Tradução livre.

22 Os autores falam da relativização da relação meronímica de subordinação mas esta relativização deve ser mantida para todas as relações meronímicas

porque, tomando o exemplo de Schalley e Zaefferer, Edward Teller perdeu o seu pé direito em 1928 quando estudava na Universidade de Munique. Não é que o conceito PÉ DIREITO deixe de estar *m-i-subordinado* ao conceito PERNA HUMANA ou CORPO HUMANO, no caso preciso do pé direito de Edward Teller (no caso específico desta instância). Segundo os autores, o seu pé direito continua a fazer parte da sua perna direita (concebida como uma entidade completa), apenas a sua perna direita deixou de ser completa. É por casos como este que os autores fazem questão de referir-se a *entidades completas*[23].

Não obstante crer que a noção de completude não deixa de ser passível de crítica, mesmo com a salvaguarda da relativização, entendo que estas relações propostas pela Ontolinguística podem ser úteis para a construção de ontologias. Elas trazem novas formas de encarar as relações semânticas que, por exemplo, na Wordnet, se restringem a sinonímia, antonímia, hiponímia e meronímia[24] (Miller, 1995: 40).

3. Ontologia e realidade

> *"Ontologies do not represent concepts in people's heads. They represent types in reality"*
>
> Smith, s.d. [c.]

3.1 Não conceitos mas universais

Até agora as propostas de definição de ontologia parecem unânimes em relacioná-la com conceptualização. Uma ontologia é definida em relação directa com o termo conceptualização, partindo do pressuposto que é de conceitos que uma ontologia trata. De ressalvar apenas a distinção que Guarino faz de ontologia, tornando-a dependente de uma linguagem e distinguindo-a de uma conceptualização. Ainda que com esta diferença, o autor não fornece uma definição adequada sobre o que entende exactamente por 'categorias' dentro de uma conceptualização.

Mas, não obstante este aparente consenso, há uma voz dissonante que insiste em desmistificar drasticamente a noção de que uma ontologia lida com conceitos. Essa voz é representada por Barry Smith.

Para Barry Smith o termo 'conceptualização' deve ser rejeitado na definição de ontologias (Smith et al., 2006).

Smith distingue dois tipos de ontologias – uma ontologia (simples) e uma ontologia de base realista - para nenhuma delas usa o termo 'conceito'. A principal diferença entre as duas ontologias é que a primeira trata de universais[25], classes definíveis[26] e das relações entre eles, enquanto que a segunda trata exclusivamente de universais, universais estes que são definidos a partir dos termos gerais de uma teoria científica aceite. Neste último caso, trata-se de uma ontologia científica, e Smith entende que ela deve ter a mesma importância que um texto científico ou qualquer outro produto decorrente da investigação científica[27].

descritas. E é aqui que se torna extremamente importante a distinção entre relações taxonómicas e relações meronímicas em que as primeiras se ficam pelos conceitos e as segundas podem exigir as instâncias a que os conceitos se referem.

23 Esta noção de completude pode estar directamente relacionada com a noção de prototipicidade, postulado base da Semântica Cognitiva. A teoria dos protótipos baseia-se nas conclusões dos estudos sobre a categorização das cores levados a cabo pela psicóloga Eleanor Rosh e a sua equipa (Rosh, E, 1973. "Natural Categories", *Cognitive Psychology*, Vol.4, No.3, May 1973, p.328. apud Cuenca & Hilferty, 1999). O protótipo é definido como o elemento mais característico dentro de uma determinada categoria e a partir do qual todos os outros elementos se definiam. O "protótipo-objecto" foi, entretanto, substituído pelo "protótipo-entidade cognitiva", e passa a ser entendido como uma imagem mental, uma abstracção.

24 No que concerne à categoria dos nomes, ou, na terminologia de Goddard (2007: 145), ao léxico nominal.

25 Smith define universais, ou tipos, como algo que é partilhado por todos os particulares que são as suas instâncias. Um particular é aquilo que tem existência num dado momento e num dado lugar (Smith, s.d. [b)]).

26 A única diferença que Smith dá para distinguir classes e universais é que as classes referem-se a a conjunto arbitrário de instâncias, enquanto que para os universais não existe essa arbitrariedade. Uma classe é uma colecção de particulares determinada por um termo geral. Podemos pôr todas as instâncias de um universal numa classe (ou *set*) e chamaremos a isso a extensão desse universal, mas podemos também constituir uma classe de uma forma mais arbitrária. Todos os universais têm extensões, mas nem todas as classes são extensões de universais (Smith, s.d. [b.])

27 Smith define uma ontologia como "um artefacto representacional cujas unidades representativas

Smith argumenta que o termo 'conceito' tem sido usado de forma aleatória e confusa. Aqui terei de concordar com Smith e afirmar que, se analisarmos com atenção as propostas onde se defende que uma ontologia lida com conceitos, como aquelas que vimos atrás, teremos de concordar que nenhuma delas define o termo 'conceito'. De resto, já John Lyons (1980: 84-87), na sua obra *Semântica*, ao tentar esclarecer o triângulo de Ogden & Richards, que ficou exposto no ponto anterior, reuniu um conjunto de interpretações possíveis para os três elementos (algumas tomadas de outros autores) e que resumi aqui no seguinte quadro:

(i) símbolo	***(ii) pensamento/ referência***	***(iii) referente***
signo	conceito	significatum
signo	intensão	extensão
palavra / lexema	conceito	coisa
signo	significatum	denotatum
signo	pensamento	objecto

Ilustração 2: Algumas interpretações para o triângulo semiótico de Ogden e Richards dadas por John Lyons (1980: 84-87).

Também Lyons afirmava que 'conceito' é "um termo com uma longa história; e quem quer que defina o significado de uma palavra como o conceito correlacionado com essa palavra, deve aos leitores uma explicação subsequente" (Lyons, 1980: 98)[28].

Barry Smith teria certamente muitas coisas a dizer acerca das propostas da Ontolinguística, muito especificamente no que concerne às relações meronímicas (porque relativamente às relações taxonómicas, tal como são entendidas aqui pela Ontolinguística, seriam certamente desconsideradas, uma vez que se restringem ao nível conceptual).

Uma das críticas que Smith faria a Schalley e Zaefferer é a de que as relações meronímicas a que os autores se referem - por exemplo, às existentes em relação ao pé, perna e corpo de Teller - não se referem a conceitos mas a entidades reais do mundo físico. Mas que diria Smith sobre onde pertence o pé inexistente de Edward Teller um dia depois de ele o ter perdido quando saltou de um carro em movimento[29], sem recorrer a conceitos? Talvez tendo em consideração uma relação espaço-tempo[30].

Mas para Smith, à partida, este problema nem sequer se põe porque para ele uma ontologia (pelo menos uma ontologia científica) não lida com instâncias, mas com universais. Uma ontologia científica não está interessada no pé esquerdo de Edward Teller, nem sequer num qualquer pé esquerdo; ela interessa-se, ou deve interessar-se, pelo universal que dá conta da instância que é designada por 'pé esquerdo de Edward Teller': simplesmente Pé[31].

Terei de precisar aqui que Smith distingue ontologias científicas, ou ontologias em suporte da ciência, de ontologias administrativas (de notar que esta distinção não tem a ver com a distinção feita atrás entre ontologia simples e ontologia de base realista). A principal diferença[32] entre ambas é que as primeiras

(*nodes*) – que podem ser elaborados a partir de uma linguagem natural ou formalizada – pretendem representar:

1. universais na realidade;
2. as relações entre esses universais que obtêm universalidade (= para todas as instâncias)." (Smith, s.d. [b]).

28 Num sentido geral 'conceito' pode significar "uma ideia, pensamento ou construção mental" (Lyons, 1980: 95).

29 Tomo aqui o exemplo de Schalley e Zaefferer (2007:7).

30 E é precisamente por aí que as relações propostas por Smith para as ontologias na área da Biomedicina se vão fundamentar. Ver ponto 3.1.1.

31 Como veremos mais adiante, é o próprio Smith que acaba por admitir que a diferença entre universais e instâncias não é fácil de definir.

32 Smith defende que uma ontologia científica deve ser aberta, passível de ser usado por múltiplas pessoas de diferentes áreas que se interessam por um mesmo objecto, estável, o mais completas possíveis e de longa duração, úteis para o uso da ciência. São exemplos de ontologias científicas a Gene Ontology (http://www.geneontology.org/), a Basic Formal Ontology (http://www.ifomis.org/bfo), o Foundational Model of Anatomy Ontology (http://sig.biostr.washington.edu/projects/fm/AboutFM.html).
As ontologias administrativas não necessitam cumprir estes requisitos; normalmente são elaboradas para uso particular, são parciais e por vezes inúteis para outro uso que não seja o propósito específico para que foram criadas. São exemplos de ontologias administrativas a

restringem o seu âmbito aos universais, enquanto que as segundas vão além deles; elas lidam com classes definíveis[33] (por oposição àquilo que Smith designa por classes naturais) embora sublinhe que devam excluir igualmente os conceitos, entendidos por Smith como aqueles termos para os quais não há instâncias, i.e. são putativos[34].

A imagem seguinte demonstra bem a distinção entre universais, classes e conceitos proposta por Smith (s.d.,[b.]), em que o rectângulo pertencente aos conceitos fica fora do âmbito de uma ontologia:

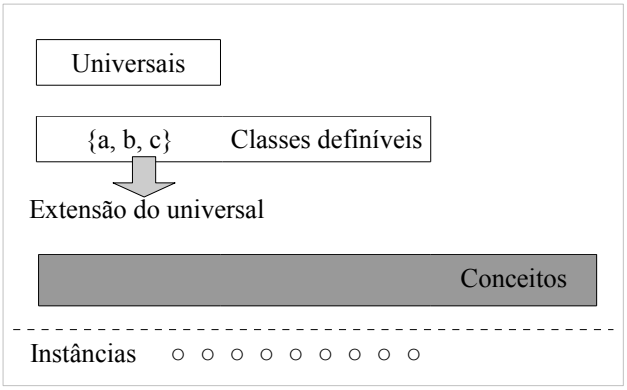


Ilustração 3: Âmbito das noções de universal, classe definível, conceito e instância proposto por Smith (s.d. [b.])

Tendo em conta esta comparação, é fácil entender a importância da distinção entre, por exemplo, o termo 'pé' como universal e o mesmo termo referido como instância. Assim, o mesmo termo 'pé' pode servir para referir-se ao pé de Edward Teller antes do acidente de 1928 ou ao universal/tipo do qual o pé de Teller é uma instância.

Portanto, como vemos, aquilo que Schalley e Zaefferer designam de *entidade completa* corresponde em Smith à noção de universal.

Desta forma, entendemos a posição de Smith quando afirma que conceitos não podem estar nas relações de 'parte_de', 'conectividade', 'causa',... (o que estão nessas relações são entidades, coisas reais).

Apesar de Barry Smith não recorrer a conceitos, e sentir muita relutância em aceitá-los no domínio específico das ciências naturais, e muito concretamente no domínio da biomedicina, é difícil dar uma resposta pronta para o que fazer com aqueles entidades que não podem ser instanciadas.

Por exemplo, Smith advoga que os termos numa ontologia devem ser formulados de forma positiva, i. e., numa ontologia científica não devem constar termos como '*absent nipple*' ou 'cirurgia não praticada por decisão do doente' ou, diria eu, 'pé ausente'. A questão é saber o que fazer com eles, uma vez que, por vezes, é necessário lidar com eles?

Mais, Smith crê ser possível separar epistemologia (aquilo que sabemos/cremos que existe) de ontologia (aquilo que existe) e esse é um dos argumentos essenciais na sua defesa da objectividade na construção de ontologias.

Importa aqui também esclarecer que o termo 'universal' que Smith adopta corresponde àquilo que John Lyons designa por 'conceito objectivo', definido como "entidades extra-mentais postuladas que eram apreendidas pelo espírito no seu conhecimento e percepção do mundo exterior", por oposição a «conceito mental», entendido no sentido que foi descrito na nota 34. Como vemos, Smith também lida com conceitos, mas não no mesmo sentido que Schalley e Zaefferer.

3.1.1 As relações numa ontologia científica de base realista

Smith defende que um dos princípios básicos a ter em conta na construção de uma ontologia de base científica é o uso de definições aristotélicas do tipo

A é um B que é C

em que B representa o *genus* e C representa a diferença específica. Isto pode traduz-se no seguinte exemplo

O ser humano (A) é um animal (B) que é racional (C).

FOAF ontology (http://xmlns.com/foaf/spec/), a Amazon.com (http://www.amazon.com/).

33 Uma classe definível é entendida por Smith como aquela classe que é definida por um termo geral que, obrigatoriamente, não designa um universal (Smith, s.d. b.).

34 Também poderemos designá-los por "conceitos mentais", tomando o termo de John Lyons (1980: 96) (ver mais adiante a distinção entre conceitos mentais e conceitos objectivos).

E que estaria representado pelo esquema seguinte:

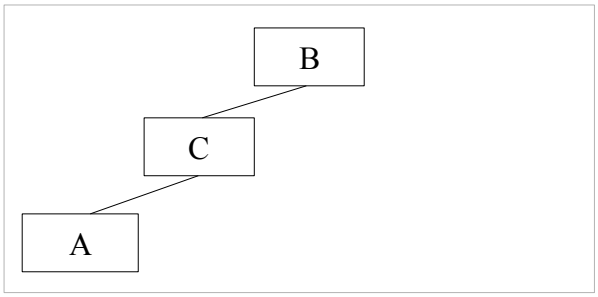


Ilustração 4: Exemplo de uma relação taxonímica de tipo aristotélica.

Este tipo de definições permitem construir uma ontologia com base numa hierarquia, em que cada termo tem apenas uma herança única ou, em outros termos, deve ter apenas um pai. Uma taxonomia, portanto, com relações taxonómicas.

Esta relação hierárquica baseia-se numa relação que é conhecida como "*is_a*". Na verdade, a relação "*is_a*" bem como "*part_of*" são as relações mais básicas nas propostas das ontologias computacionais.

Smith admite ainda outras relações, umas que diferem completamente das relações propostas pela Ontolinguística outras que se assemelham. As semelhanças entre as duas propostas é que aquilo que na Ontolinguística se designa por relações taxonómicas, em Smith apresenta-se como relações entre universais. Mas, ao contrário do que seria de esperar, também Smith admite relações entre universais e instâncias e entre as próprias instâncias. E como foi mencionado no ponto 3, as relações ao nível das instâncias têm a variante tempo em consideração. Isto é assim porque as instâncias, como sabemos, existem num determinado tempo e espaço. Não são universais.

Mas antes de perceber o tipo de relações que Smith propõe é importante definir aquelas que são as três dicotomias básicas da sua proposta. E estas dicotomias baseiam-se nos pares seguintes:

1. **instância** *vs* **universal**
2. ***continuant*** *vs* **ocorrente (processos)**
3. **dependente** *vs* **independente**

A primeira dicotomia foi já definida atrás (ver notas 19 e 25). Em relação à segunda dicotomia, ela assenta no pressuposto de que existem dois tipos de entidades: aquelas que preservam a sua identidade mesmo na mudança e existem continuamente no tempo; e aquelas outras entidades que têm partes temporais, existem apenas nas suas fases e podem desdobrar-se nessas mesmas fases (Grenon and Smith, s.d.: 3-4). As primeiras entidades são designadas '*continuants*' ou '*endurants*', as segundas são designadas 'ocorrentes' ou '*perdurants*'[35]. Por exemplo, eu sou um *continuant* e a minha infância é um ocorrente. Ou, para ser mais precisa, eu, sendo uma substância, sou uma instância do universal de nível superior designado '*continuant*'. A minha infância, sendo um processo, é uma instância do universal de níve superior designado 'ocorrente'.

Para Smith, a melhor forma de distinguir se uma entidade é um *continuant* ou um ocorrente é a partir da metáfora da máquina fotográfica e da câmara de vídeo: nós só podemos fotografar *continuants* enquanto que os ocorrentes só podem ser captados em vídeo (Jansen, 2008: 184).

Exemplos de *continuants* são as substâncias, objectos, coisas, formas, qualidades, planos, papéis, funções. Exemplos de ocorrentes são processos, mudanças, eventos, realizações (Smith, sd. [b.]).

Smith entende que tudo o que existe pertence a uma destas duas categorias. Tudo pode aí ser inserido. E, por esta razão, Smith defende que uma ontologia científica deve conter pelo menos estas duas categorias. Elas correspondem aos dois níveis superiores de uma ontologia, aos todos os outros elementos de uma ontologia se deveriam submeter.

Relativamente à terceira dicotomia, ela existe apenas em relação aos *continuants*, i.e., só os *continuants* podem ser dependentes ou independentes. Porque todos os ocorrentes são, necessariamente, entidades dependentes de um *continuant* dependente[36].

35 Os termos '*continuant*' e 'ocorrente' surgem a partir de de William Johnson, que define '*continuant*' como "o que continua a existir apesar dos seus estados ou relações poderem mudar" (Johnson, 1921: 199. Apud Jansen, 2008: 183).

36 Jansen faz corresponder estes dois termos com os termos 'substância' e 'acidente' de Aristóteles em *Categorias*:

> the dependent categories are called accidents and are placed in opposition to substances. A traditional criterion for the opposition of

E a principal diferença entre eles está contida no seu próprio nome. Enquanto que os *continuants* independentes existem por si mesmos, os *continuants* dependentes necessitam dos *continuants* independentes para existir. Por exemplo, peso, uma doença, altura, cor, são *continuants* dependentes, porque necessitam dos seus portadores para existirem. Ao passo que organismos, células, cadeiras são *continuants* independentes. Como exemplifica Smith, não há corrida sem um corredor e não há doença sem um organismo. Corrida e doença são entidades dependentes, corredor e organismo são entidades independentes (Smith, s.d. [b].). De referir ainda que os *continuants* podem ser materiais (uma célula) ou imateriais (uma cavidade).

Se quisermos pôr num esquema as duas dicotomias de que estive a falar, teríamos algo como o seguinte:

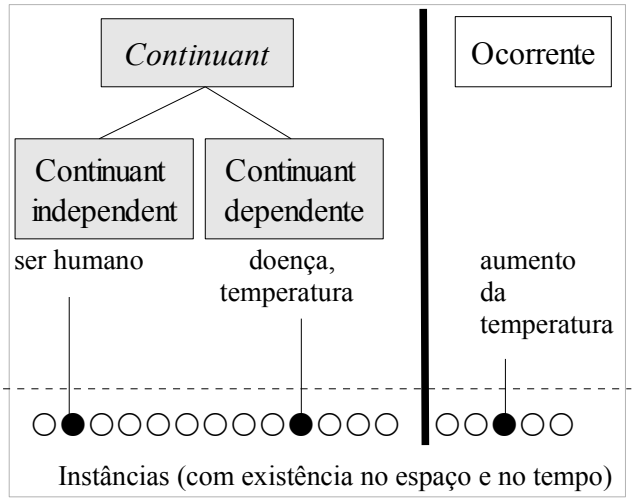


Ilustração 5: Esquema das duas das dicotomias básicas em Smith (adaptado de Smith, s.d.[b.])

Depois de esclarecidas as três dicotomias básicas, é possível agora expor as relações. Como dissemos acima, Smith entende que é possível estabelecer relações entre universais (com relações como *is_a* e *part_of*,...), entre universais e instâncias (a perna de Edward Teller *instance of* universal perna) e entre instâncias (por exemplo, a perna de Teller *part_of* Teller). Devo referir que Schwartz e Smith defendem que uma ontologia científica deve construir-se apenas com universais, no entanto, referem, é necessário definir à partida as relações ao nível das instâncias, na medida em que são essas relações que fornecem as relações para o nível dos universais (Schwartz e Smith, 2008: 221).

Assim, Schwartz e Smith distinguem seis relações primitivas ao nível das instâncias, a saber:

> *c* **instance_of** *C* **at** *t* - a primitive relation between a continuant- instance and a universal which it instantiates at a given point in time […].
> *p* **instance_of** *P* - a primitive relation between a process-instance and a universal which it instantiates independently of time. […].
> *c* **part_of** *c*1 **at** *t* - a primitive part-whole relation between two continuant instances and a time at which the one is part of the other.
> *p* **part_of** *p* - a primitive part-whole relation which, independently of time, obtains between two process-instances (one is a processual part, or segment, of the other).
> *c* **located_in** *r* **at** *t* - a primitive relation between a continuant instance, a 3-dimensional spatial region which this instance occupies, and a time at which this instance occupies this region.
> *p* **has_participant** *c* **at** *t* - a primitive relation between a process, a continuant, and a time at which this instance occupies this region.
> *p* **has_agent** *c* **at** *t* - a primitive relation between a process, a continuant and a point in time (Schwartz e Smith, 2008: 227-228).

De notar ainda que Schwartz e Smith defendem que estas relações devem ser neutras em relação a todos os domínios das ciências. Isto significa que elas devem poder ser aplicadas em todos os domínios. E apesar de todas estas relações primitivas se obterem entre instâncias, elas devem poder ser usadas para definir as relações ao mais alto nível dos universais.

4. Conceitos, classes ou universais num mesmo quadro de análise

Não obstante a convicta afirmação de Smith que nega o termo 'conceito', é o próprio Smith (juntamente com outros autores) que, num artigo acerca das relações nas ontologias biomédicas, faz uma equiparação entre termos, com vista o esclarecimento:

> the term 'class' here is used to refer to what, in the knowledge-representation literature, is

substances and accidents can be found in the second chapter of the *Categories*: qualities and quantities are in a substance, while substances are not in a substance (Jansen, 2008: 181).

Há que precisar no entanto, como refere Jansen, que este 'estar em' não significa, por exemplo o coração estar no corpo. Um *continuant* dependente não existe sem o seu portador; se o seu portador deixa de existir, a entidade dependente deixa também de existir.

> typically (and often somewhat confusingly) referred to under the heading 'concept' and in the literature of philosophical ontology under the headings 'universal', 'type' or 'kind'
>
> Smith et al., 2005

Mas então... estão todos a falar do mesmo? Não importa que me refira a classes, ou conceitos, ou universais, ou tipos? É tudo a mesma coisa?

Não deixa de ser curiosa a afirmação de Smith et al., porque parece que vem tornar irrelevante o que ficou exposto nos pontos anteriores.

À questão de saber se é tudo a mesma coisa, terei de responder sim e não. E explicarei porquê já de seguida.

Os termos 'conceito', 'classe', 'universal' têm em comum o facto de serem o objecto de estudo de um ontologista. A questão terminológica não é irrelevante na medida em que não é o mesmo falar de 'conceitos', 'classes' ou 'universais' indiferentemente fora do seu lugar específico. Com isto eu defendo que a solução para a questão que dá título ao presente artigo não é eliminar nenhum termo, ou dar preferência a um em detrimento de um outro. A solução passa por inserir os termos no seu espaço próprio. E esta tarefa de inserção dos termos no seu lugar específico torna-se mais fácil se estudarmos o quadro conceptual para a noção de ontologia proposto por Nickles et al. (2007).

Pelo que ficou dito atrás, parece-me essencial construir um quadro de análise da noção de ontologia capaz de dar conta de todas as propostas. Foi com esta intenção em mente que Nickles et al. (2007: 23-33) desenvolveram um quadro conceptual capaz de acolher as diferentes definições do termo 'ontologia' quer ao nível interdisciplinar, quer ao nível interno das próprias disciplinas.

Estes autores defendem que, mais do que tentar encontrar um argumento único capaz de dar conta de uma definição universal e totalitária do termo 'ontologia', importa encontrar um espaço de análise das suas diferentes noções para poderem ser comparadas e, com isso, entendidas. Para isso decidem partir de um gráfico, ou espaço, tridimensional onde inserem três eixos ortogonais, que poderíamos designar como a) o eixo da generalidade, b) o eixo da objectividade e, por fim, c) o eixo dos níveis. Isto significa que os autores partem de três dimensões distintas a partir das quais o conceito ou conceitos de ontologia podem ser estudados.

A citação que se segue pode dar-nos um resumo do que os autores entendem por cada uma das três dimensões:

> A dimensão vertical reflecte a generalidade, com os assuntos mais gerais no topo; a dimensão da profundidade reflecte a generalidade com a visão mais objectivista na frente; e a dimensão horizontal que tem três segmentos com o mundo e os seus aspectos e partes à direita, as diferentes visões deste mundo no meio e o(s) campo(s) da Ontologia à esquerda[37].
>
> Nickles et al., 2007: 25

4.1 A dimensão vertical: o eixo da generalidade

O primeiro eixo, que poderíamos designar por eixo da generalidade, diz respeito ao par GERAL vs. ESPECÍFICO. Aqui procura-se determinar se uma ontologia se detém nas propriedades comuns a todas as entidades ou, por outro lado, no lado oposto, nos seus aspectos categoriais. Assim, poderíamos colocar no lado extremo do eixo da generalidade as ontologias definidas por Guarino como ontologias generalistas ou de "*top-level*", no lado oposto, colocaríamos as ontologias de domínio específico. Teríamos qualquer coisa como o seguinte:

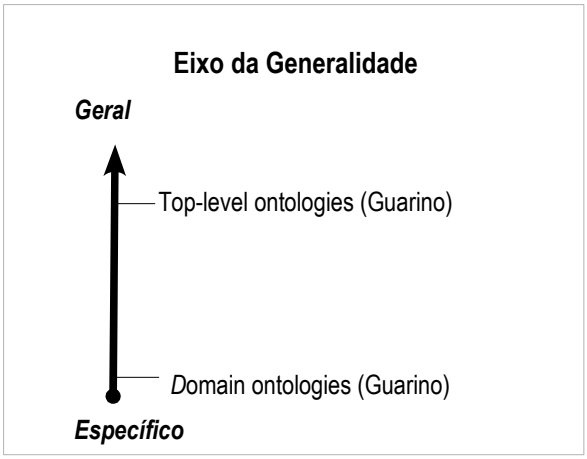


Ilustração 5: Eixo da Generalidade (imagem adaptada de Nickles et al. , 2007: 24).

37 Tradução livre.

4.2 A dimensão da profundidade: o eixo da subjectividade

O segundo eixo, denominado eixo da objectividade, é constituído pelo par SUBJECTIVIDADE vs OBJECTIVIDADE, onde se dá conta das noções de ontologia que, ou assentam no pensamento e na razão ou, pelo contrário, na realidade externa.

Tomando como base de análise esta dimensão, colocaríamos, por exemplo, a definição de ontologia de Barry Smith (*vd.* nota 27) no lado extremo da objectividade e no lado oposto poderíamos inserir, por exemplo, a definição de Lorhard dada acima (*vd.* ponto 2.1). Qualquer outra definição de ontologia teria de ser inserida no nosso eixo tendo em conta estas duas definições já inseridas.

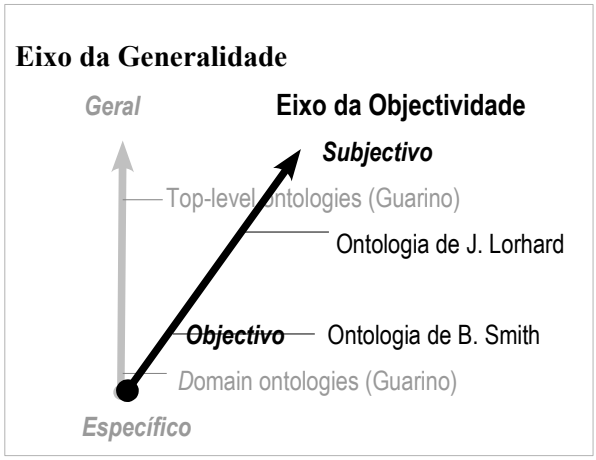


Ilustração 6: Eixos da generalidade e objectividade (imagem adaptada de Nickles et al.,2007: 26).

4.3 A dimensão da horizontalidade: o eixo dos três níveis

Uma terceira dimensão tem a ver com diferentes níveis de distinção da ontologia dentro de um campo disciplinar. Nesta dimensão há uma separação clara entre níveis, e não uma gradação como acontece nos dois níveis anteriores, embora possa haver uma sobreposição dos três níveis, como veremos mais adiante. Esta separação é perfeitamente compreensível se entendermos o critério que lhe subjaz: a ele preside a distinção entre 1) o nível do objecto (*object-level*), 2) o nível da teoria que dá conta desse objecto - o meta-nível (*meta-level*) e 3) o nível que poderíamos traduzir por nível trans-meta (*trans-meta-level*). É precisamente esta distinção que pode agora esclarecer o que existe de errado, à partida, e segundo esta proposta, na definição de conceptualização de Gruber. É que Gruber punha num mesmo nível os objectos e os conceitos que dão conta desses objectos, ou seja os níveis 1 e 2[38].

Para melhor percebermos os três níveis desta terceira dimensão, Nickles *et al.* dão como exemplo o termo 'sintaxe', que pode ser utilizado para referenciar os três níveis propostos:

> Syntax as a mass noun means a field, a certain branch of linguistics; its different outcomes – like say Haider's syntax of German (Haider 1993) – are coded by the corresponding count noun. In fact, in linguistics there is a third use of the term syntax (and a second use of the count noun), one that relates to the subject matter of the second and first use, i.e., that subsystem of a language that constrains the building of phrases from word forms. So there is an object-level use of this term (syntax as language subsystem), a meta-level use (syntax as theoretic account of this subsystem) and in a sense a trans-meta-level use (syntax as subfield or branch of linguistics).
>
> Nickles et al., 2007: 25

Como podemos ver, o termo 'sintaxe' pode ser usado nos três níveis propostos: ao nível do objecto (entendido como o subsistema da língua), a um meta-nível (as várias teorias sobre sintaxe) e a um trans-meta-nível (o ramo da linguística que se ocupa das regras pelas quais se combinam elementos de uma frase).

A questão que os autores colocam é a de saber se também o termo 'ontologia' garante esta polissemia assim especificada, i.e., se é possível garantir esta distinção de três níveis para o termo. Segundo os autores, há duas respostas possíveis.

A primeira resposta é que há, efectivamente, estes três níveis para o termo 'ontologia'. E se aceitarmos uma resposta afirmativa, teremos de colocar no primeiro nível o ser, ou, mais especificamente, o que existe (a realidade), e as suas categorias; num segundo nível, as diferentes teorias que dão conta do primeiro nível e, para o terceiro nível, o espaço de discussão das

38 Esta indiferenciação, como veremos mais adiante, não é assim tão errada quanto Guarino ou Nickles et al. parecem crer.

diferentes teorias dentro de um mesmo campo ou disciplina.

A segunda resposta é negativa na medida em que, como referem Nickles et al.,

> Only the last two levels are properly called ontology, the second one by transparent metonymic extension (and count noun formation) from the name for the third one, whereas the first one requires different means of expression such as *the real world* (as opposed to possible counterparts) or simply *reality* or rather its (ultimate or basic) furniture.
>
> Nikcles et al., 2007: 25

Como vemos aqui, segundo Nickles *et al*., também é possível distinguir na ontologia os três níveis encontrados para 'sintaxe', apenas teremos que advertir que o nível objecto não tem o mesmo nome dos outros dois níveis.

Ora é também nesta terceira dimensão que Barry Smith se apoia para construir uma terminologia capaz de ser usada para a pesquisa em ontologias (no seu caso particular, Smith reporta-se a ontologias no domínio específico da biomedicina). Também Smith propõe três níveis que devem ser considerados aquando da elaboração ou estudo de uma ontologia no domínio da biomedicina, a saber,

> • Level 1: the objects, processes, qualities, states, etc. in reality (for example on the side of the patient);
>
> • Level 2: cognitive representations of this reality on the part of researchers and others;
>
> • Level 3: concretizations of these cognitive representations (in for example textual or graphical).
>
> Smith, 2006: 2

também aqui, como em Nickles *et al*., distingue-se o nível 1, ou o nível do objecto, ao qual Smith acrescentou os processos, qualidades, estados da realidade e um nível 2, ou o meta-nível, composto pelas representações cognitivas daquela realidade. Em relação ao nível 3, é óbvio que ele não corresponde ao nível 3 de Nickles et al. No caso de Smith, o nível 3 corresponde às concretizações das representações cognitivas, nível este que Nickles et al. não consideram, pelo menos não explicitamente.

Para termos uma imagem global do quadro conceptual de Nickles et al. com as três dimensões sobrepostas, será útil atentar na imagem seguinte:

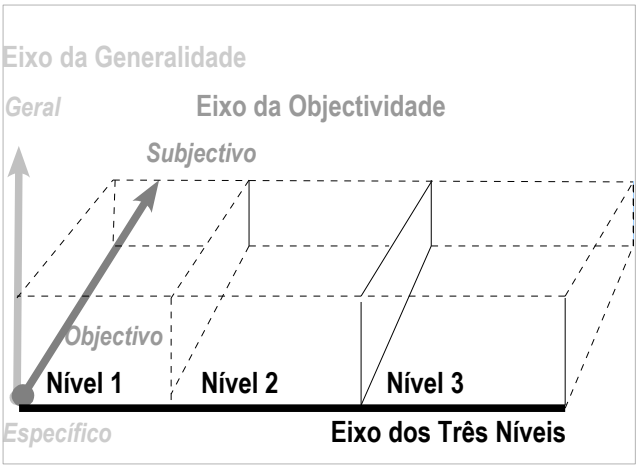


Ilustração 7: A sobreposição das três dimensões de análise das noções de ontologia (imagem adaptada de Nickles et al., 2007: 33).

Como podemos verificar, a separação entre os níveis 1 e 2 neste quadro é feita com uma linha pontilhada e não contínua, ao contrário do que acontece o nível 3, perfeitamente demarcado dos anteriores.

E isto é assim porque, por exemplo, se tomarmos para análise a teoria dos três mundos de Popper, ou mesmo a dos três níveis de Smith, a que me referi acima, entendemos perfeitamente a razão desta diferenciação. Com Karl Popper teremos de rever a forma como são representados os três níveis, nomeadamente a separação que é feita entre os níveis 1 e 2. Na sua teoria dos três mundos acerca do problema mente-corpo[39] Popper diz algo como isto:

> Devo salientar que considero que os produtos da mente humana são reais; não só os que também são físicos – arranha-céus e automóveis, por exemplo, a que toda a gente chamará «reais» - mas também os livros ou as teorias. As teoria em si, a própria coisa abstracta, tenho-a como real porque nos possibilita interagir com ela – podemos

39 Na sua visão pluralista do problema corpo-mente, Popper (1997) distingue 3 mundos que podem ser resumidos em:
a) mundo 1: mundo físico, dos objectos físicos;
b) mundo 2: mundo dos estados mentais, das experiências mentais (conscientes);
c) mundo 3: mundo dos produtos da mente humana (teorias), que pertencem tanto ao mundo 1 como ao mundo 2.

> produzi-la – e porque ela faz o mesmo connosco. Basta isso para considerá-la real.
>
> Popper, 1997: 63

Tanto Smith como Popper fariam Niclkes et al. rever a sua divisão dos três níveis. E teríamos também de rever as considerações que foram tecidas em relação à definição de conceptualização de T. Gruber. Talvez caibam, afinal, numa ontologia e num mesmo nível – na de Popper pelo menos – objectos e teorias sobre esses objectos.

Outra questão que fica em aberto no quadro conceptual de Nickles et al. é o lugar da linguagem. Onde se insere ali a linguagem? No nível 2? Se tomarmos em consideração os três níveis de de Popper e Smith, ela cabe no nível 3, mas onde cabe a linguagem no eixo horizontal dos três níveis do quadro de Nickles et al.?

Para além das três dimensões que Nickles et al. nos propõem, é possível acrescentar outras no momento de analisar diferentes ontologias. Por exemplo, as que dêem conta dos papéis de autor e de usuário; a linguagem utilizada na ontologia (para dar conta do seu grau de formalismo[40]), o fim específico para que foi construída, ou a sua utilidade.

5. Conclusão

Os dois primeiros níveis de Smith e de Nickles *et al*. que apresentei são extremamente relevantes para concluir o presente artigo. É a partir desta dimensão que podemos visualizar um consenso entre as diferentes abordagens sobre a melhor forma de construir uma ontologia capaz de representar informação acerca do mundo (ou de um mundo).

E este consenso existe porque, quer consideremos ou não o nível 1, quer o integremos ou não no nível 2 ou quer consideremos ou não um quarto nível dentro desta dimensão, parce não haver muitas dúvidas que o desenvolvimento e construção de uma ontologia começa no nível 2 de Nickles et al e de Smith e substancia-se no nível 3 de Smith. E isto acontece quer se trate de uma ontologia de base realista ou de uma ontologia de base conceptual.

As ontologias constroem-se com termos, com uma linguagem (natural ou não, formal ou não), que representam ou representa, por sua vez, classes, conceitos, universais ou mesmo instâncias (dependendo da perspectiva adoptada, dependendo do tipo de ontologia que se quer cosntruir e dependendo do grau de pormenor que se quer cobrir).

Optar por uma ontologia conceptual ou por uma ontologia de base realista depende da ontologia que se pretende construir. Numa ontologia de *top-level*, ou de nível superior, não estão representadas instâncias (ou não deveriam estar aí representadas instâncias), por exemplo. Uma ontologia administrativa terá inevitavelmente de ir além dos universais de que fala Smith. Uma ontologia, por exemplo, no domínio das Ciências Naturais construirá uma ontologia de base realista, enquanto que uma ontologia linguística certamente beneficiará de uma perspectiva conceptual.

O objecto de uma ontologia depende de numerosos factores, inclusivamente das diferentes visões epistemológicas ou metodológicas dos participantes no seu processo de construção.

Por isso, talvez, o denominador comum que buscamos não se resolva com uma definição do que é uma ontologia ou o seu objecto específico. Uma definição pode, inclusivamente, surtir o efeito contrário. Como refere Popper,

> a definição constitui um problema lógico em si e que se lhe associa uma grande dose de superstição. As pessoas acham que um termo só tem significado se for definido. […] O que é necessário é fazermo-nos entender e a definição não é por certo o melhor meio para o conseguir.
>
> Popper, 1997: 31-32

Referências

Abecker, Andreas and Ludger van Elst. 2004. "Ontologies for Knowledge Management" in Staab and Studer, 2004. pp. 435-454.

Almeida, Maurício e Marcello Bax. 2003. "Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção" in *Ci. Inf.*, Brasília, n. 3, pp. 7-20. Set./Dez., 2003. http://www.scielo.br/pdf/ci/v32n3/19019.pdf.

40 Que tipo de linguagem deve usar uma ontologia? Terminologia, linguagem comum, linguagem formalizada, números, códigos?

Campos, Mª. Henriqueta e Mª. Francisca Xavier. 1991. "Estrutura semântica do léxico". *Sintaxe e Semântica do Português*. Lisboa: Universidade Aberta. ISBN: 972-674-072-X.

Cohnitz, Daniel and Barry Smith. s.d.. "Assessing Ontologies: The Question of Human Origins and Its Ethical Significance". http://ontology.buffalo.edu/smith/articles/humanorigins.pdf.

Cuenca, Maria Josep & Joseph Hilferty. 1999. *Introducción a la lingüística cognitiva*, Barcelona: Editorial Ariel. ISBN: 84-344-8234-7

Degen, Wolfgang and Heinrich Herre. s.d.. "What is an Upper Level Ontology?". http://www.informatik.uni-leipzig.de/erre/papers/top.ps.

Geeraerts, Dirk. 2006. "Idealist and empiricit tendencies in cognitive semantics" in Geeraerts, Dirk. 2006. *Words and Other Wonders. Papers on Lexical and Semantic Topics*. Berlin/New York: Mouton de Gruyter. pp. 416- 444. ISBN-13: 978-3-11-019042-7.

Goddard, Cliff. 2007. "Semantic primes and conceptual ontology" in Schalley, Andrea C. and Dietmar Zaefferer (ed.s), 2007.

Guarino, Nicola. 1996. "Understanding , Building and Using Ontologies". http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html.

Guarino, Nicola. 1998. "Formal Ontology and Information Systems". http://www.loa-cnr.it/Papers/FOIS98.pdf.

Guarino, Nicola and Pierdaniele Giaretta. 1995. "Ontologies and Knowledge Bases. Towards Terminological Clarification". http://www.loa-cnr.it/Papers/KBKS95.pdf.

Grenon, Pierre and Barry Smith. s.d. "SNAP and SPAN:Towards Dynamic Spatial Ontology". http://ontology.buffalo.edu/smith/articles/SNAP_SPAN.pdf .

Gruber, Thomas. 1992a. "A Translation Approach to

Portable Ontology Specifications". http://ksl.stanford.edu/knowledge-sharing/papers/ontolingua-intro.rtf.

Gruber, Thomas. 1992b. "Ontolingua: a Mecanism to Support Portable Ontologies." http://mas.cs.umass.edu/ãseltine/791S/farquhar.the_ontolingua_server.ps.

Gruber, Thomas. 1993. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing" in *International Journal Human-Computer Studies* Vol. 43, Issues 5-6, Novemer 1995. pp.907-928. http://ksl-web.stanford.edu/KSL_Abstracts/KSL-93-04.html.

Heuer, Peter and Boris Hennig. 2008. "Chapter 9: The Classifications of Living Beings" in Smith and Munn, 2008, pp. 197-217.

Jansen, Ludger. 2008. "Chapter 8: Categories: The Top-Level Ontology" in Smith and Munn, 2008. pp. 173-196.

Johnson, William. 1921. *Logic:Part I.* Cambridge: Cambridge University Press.

Lawson, Tony. 2004. "A Conception of Ontology". http://www.csog.group.cam.ac.uk/A_Conception_of_Ontology.pdf.

Lyons, John.1980. *Semântica I*. Lisboa: Presença.

Miller, George. 1995. "Wordnet: A Lexical Database for English" in *Communications of the ACM*, November 1995/Vol. 38, Nº 11. pp. 39-41.

Mika, Peter. s.d.. "Ontologies are us". http://www.cs.vu.nl/~pmika/research/papers/ISWC-folksonomy.pdf.

Mika, Peter, Victor Iosif, York Sure, Hans Akkermans. 2004. "Ontology-based Content Management in a Virtual Organization" in in Staab and Studer, 2004. pp. 455-476.

Morais, Edison A. M.. s.d.. "O Estado da Arte no Estudo das Ontologias". http://usuarios.cultura.com.br/eds/PDF/fasam.pdf.

Nickles, Mathias, Adam Pease, Andrea Schalley and Dietmar Zaefferer. 2003. "Ontologies across disciplines" in Schalley and Zaefferer (ed.s), 2007. pp 23-67.

Ogden, C. K. and I. A. Richards. 1985. *The Meaning of Meaning*. London: ARK Paperbacks. ISBN: 0-7448-0033-1.

Pisanelli, Domenico M., Aldo Gangemi and Geri Steve. s.d. "Ontologies and Information Systems: the Marriage of the Century?". http://www.loa-cnr.it/Papers/lyee.pdf.

Popper, Karl. 1997. *O Conhecimento e o Problema Corpo-Mente.* Lisboa: Edições 70. ISBN: 972-44-0961-9.

Ricoeur, Paul. 1992. "Ontologie" in *Encyclopedia Universalis, Vol. 16 – Nation-Orchidales*. Paris:*Encyclopedia Universalis France.* pp. 902-910. ISBN: 2-85229-287-4

Schalley, Andrea C. and Dietmar Zaefferer. 2007. "Ontolinguistics - An outline" in Schalley, Andrea C. and Dietmar Zaefferer (ed.s), 2007.

Schalley, Andrea C. and Dietmar Zaefferer (ed.s). 2007. *Ontolinguistics. How Ontological Status*

Shapes the Linguistic Coding of Concepts. Berlin/ New York: Mouton de Gruyter. ISBN: 978-3-11-018997-1.

Schwartz , Ulf e Barry Smith. 2008. "Chapter 10: Ontological Relations" in Smith and Munn, 2008, pp. 219-234.

Smith, Barry. s. d. [a.] "Ontology and Information Systems". http://ontology.buffalo.edu/ontology_long.pdf.

Smith, Barry. s. d. [b.] "Video: How to Build an Ontology". http://ontology.buffalo.edu/smith/articles/ontologies.htm.

Smith, Barry. s. d. [c.] "Towards a Reference Terminology for Talking about Ontologies and Related Artifacts". ontology.buffalo.edu/07/os3/Smith_3_Terminology.ppt.

Smith, Barry. 2006. "Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain". http://ontology.buffalo.edu/bfo/Terminology_for_Ontologies.pdf.

Smith, Barry. 2008. "Chapter 4: New Desiderata for Biomedical Terminologies" in Smith and Munn, 2008. pp. 83-108.

Smith, Barry and David M. Mark. 2001. "Geographical categories: an ontological investigation" in International Journal of Geographical Information Science, 2001, vol. 15, Nº. 7. pp. 591-612. http://www.ncgia.buffalo.edu/ontology/SmithMarkIJGIS2001p591_s.pdf.

Smith, Barry and Berit Brogaard. 2003. "Sixteen Days" in Journal of Medicine and Philosophy, 2003, vol. 28, No. 1. pp. 45-78. http://ontology.buffalo.edu/smith/articles/16Days.pdf.

Smith, Barry and Katherine Munn. 2008. *Applied Ontology. An Introduction*. Frankfurt/Paris/Lancaster/New Brunswick: Ontos Verlag. ISBN 978-3-938793-98-5.

Smith, Barry, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan Rector and Cornelius Rosse. 2005. "Relations in biomedical ontologies". http://genomebiology.com/content/pdf/gb-2005-6-5-r46.pdf.

Staab, Steffen and Rudi Studer (ed.s). 2004. *Handbook on Ontologies*. Berlin/Heidelberg/New York: Springer. ISBN: 3-540-40834-7.

Teixeira, José. 2001. "Referente/Significado: O erro de Saussure", in Revista Portuguesa de Humanidades, Vol. 4-1/2-2000, Faculdade de Filosofia da U.C.P., Braga. pp 125-146. ISSN 0874-0321. http://repositorium.sdum.uminho.pt/bitstream/1822/5365/1/referentSignificSaussur.pdf.

Uschold, Mike and Michael Gruninger. 1996. "Ontologies: Principles, Methods and Applications." in Knowledge Engineering Review, vol. 11, No 2, June 1996. https://eprints.kfupm.edu.sa/55793/.

Wierzbicka, Anna. 1992. *Semantics, Culture and Cognition: Universal human concepts in culture-specific configurations*. New York: Oxford University Press. ISBN 0-19-507325-8/ 0-19-507326-6.

Wierzbicka, Anna. 1996. S*emantics: Primes and Universals*. Oxford: Oxford University Press. ISBN: 0-19-870003-2.

Øhrstrøm, P., S. Uckelman and H. Schärfe. 2007. *Historical and Conceptual Foundation of Diagrammatical Ontology*. UvA-DARE: Digital Academic Repository of the University of Amsterdam. http://www.illc.uva.nl/Publications/ResearchReports/PP-2007-17.text.pdf.

Verificación ortográfica de formas verbais e secuencias de pronomes enclíticos en lingua galega

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

Resumo

Descrición das melloras no comportamento do verificador ortográfico MySpell/Hunspell ante formas verbais en lingua galega, con arquivos de dicionario e de afixos que se elaboraron a partir da versión para o galego dispoñible baixo os termos da licenza GNU GPL, versión 2, aos que se lles modificou a estrutura para que, en cada persoa gramatical da flexión, admita un paradigma diferente de posibles secuencias des pronomes persoais enclíticos á forma verbal consonte a información que se extraeu do *Vocabulario ortográfico da lingua galega* sobre o réxime de construción sintáctica das formas verbais.

1. Introdución

Despois de realizar probas cos ficheiros para a lingua galega dispoñibles baixo os termos da licenza GNU GPL, versión 2, para o verificador ortográfico MySpell/Hunspell (que poden utilizar, entre outros, o paquete de ofimática OpenOffice, o verificador para o navegador Firefox, o xestor de correo electrónico Thunderbird, editores de textos como gedit ou AbiWord, programas de localización de software como o Poedit, Lokalize, WordForge...); calquera se pode decatar de que non verifica adecuadamente certas formas verbais, sobre todo formas rizotónicas de verbos con pouco corpo fónico, e de que non recoñece formas lingüísticas moi expresivas con máis de dous pronomes enclíticos tras a forma verbal. Foron precisamente estas circunstancias as que motivaron o interese por esculcar o arquivo de sufixos dispoñible no sitio web do Centro de Referencia e Servizos de Software Libre Mancomún co fin de examinar a dificultade de engadir máis posibilidades de concorrencias de formas verbais con pronomes enclíticos e ver de mellorar a flexión de certas formas. Despois dalgunhas probas, decatámonos enseguida de que non era difícil ampliar as combinacións sintagmáticas de pronomes enclíticos; no entanto, o deseño do arquivo de afixos ía xerar e admitir moitas máis formas agramaticais das que xa coñecía (formas flexivas de primeira e segunda persoa, singular e plural, con secuencias de enclíticos que comezasen por un pronome reflexivo de terceira persoa, pronomes enclíticos acusativos con verbos que non admiten unha construción transitiva...) de se ampliar o paradigma de pronomes enclíticos sen restricións. Respecto da revisión da flexión verbal, obtivéronse bos resultados mais, en certos casos, resultou insuficiente a revisión do arquivo de afixos e fíxose necesario engadirlle alomorfos con acentos gráficos na raíz ao arquivo de dicionario en certos casos. No momento en que se modificou o arquivo de dicionario, xermolou a idea de impoñer desde este arquivo limitacións no tipo de construción sintáctica de cada lema para restrinxir ou ampliar as secuencias de pronomes enclíticos que se engadisen ás formas flexionadas.

2. Elaboración

2.1 A información sintáctica

A información sobre o tipo de construción, pese a non ser sistemática[1], parece abonda para este propósito e está dispoñible en formato electrónico no portal da Real Academia Galega no apartado dedicado ao *Vocabulario Ortográfico da Lingua Galega* (VOLG). Con este propósito, con data do 14/12/2008 realizouse unha extracción do VOLG mediante buscas por clases de palabra (verbos) que contivesen información sobre o tipo de construción sintáctica (6.738 lemas). Filtráronse as formas non toleradas[2], que se reutilizaron para compoñer regras de substitución que o verificador emprega para ofrecer suxestións nas secuencias que non considera correctas.

2.2 Paradigmas de pronomes persoais enclíticos

A elaboración dos paradigmas de pronomes partiu do paradigma máis extenso (o dunha forma verbal en terceira persoa, singular ou plural, en construción transitiva) con 22 regras para a creación de secuencias de pronomes enclíticos monosílabos e 353 regras para secuencias de pronomes enclíticos polisílabos[3] que chegan a aglutinar ata tres dativos e un acusativo (*trouxéronchemellela*).

1 Vid. Álvarez e Xove, 2002, p. 239.
2 Vid. Estrutura das entradas, no sitio web do VOLG.
3 Os paradigmas de pronomes persoais enclíticos polisílabos tratan de representar, respecto da orde, o que se dispón sobre as secuencias de clíticos en Álvarez e Xove, 2002, p. 570-571.

A partir deste paradigma obtivéronse, por unha banda, o paradigma da terceira persoa da construción intransitiva tras eliminar todas as secuencias que contiñan pronomes enclíticos marcados morfoloxicamente como acusativos e as que contiñan o pronome reflexivo de terceira persoa; e, por outra banda, os paradigmas das demais persoas da construción transitiva tras eliminar de cada paradigma as secuencias que contiñan o reflexivo de terceira persoa ou o pronome reflexivo respectivo da persoa da flexión. A partir do paradigma para unha forma flexionada en terceira persoa en construción intransitiva obtivéronse os paradigmas para as demais persoas na construción intransitiva mediante a eliminación das secuencias que contiñan o pronome reflexivo de terceira persoa ou o pronome reflexivo respectivo da persoa da flexión; e, ademais, os paradigmas da construción pronominal tras eliminar as formas que non contiñan o reflexivo correspondente a cada persoa acompañado ou non de formas de dativo doutras persoas.

2.3 Revisión morfolóxica e asociación cos paradigmas dos pronomes enclíticos

Os paradigmas da flexión parten da revisión morfolóxica dos paradigmas preexistentes para cada conxugación, que se triplicaron (un para cadansúa construción sintáctica) e nos que se relacionou cada regra de creación dun sufixo co seu correspondente paradigma consonte as súas posibilidades sintagmáticas (combinacións con enclíticos monosílabos e polisílabos, acusativos de P3 tras -r, -s ou ditongo decrecente...) e consonte a persoa gramatical da flexión.

2.4 Paradigmas de formas verbais impersoais e defectivas

Os paradigmas das formas impersoais reducíronse á flexión en terceira persoa. Inclúese a terceira persoa de plural para recoller usos metafóricos con suxeito en plural do tipo *choven chuzos, tronaban os canóns.*
Os paradigmas dos verbos defectivos compuxéronse, cando non coincidían cos dos verbos impersoais, de acordo coas limitacións flexivas que precisaban.
Para o verbo *decer* (forma tolerada no VOLG, que remite a *dicir*) creouse un paradigma que inclúe as formas de infinitivo, xerundio, e P4 e P5 do presente de indicativo.

2.5 Elaboración do dicionario

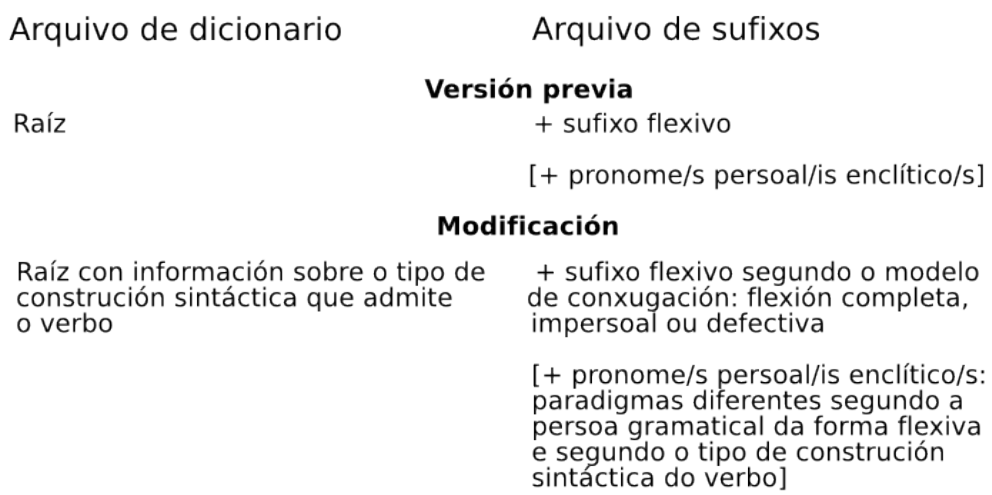
Para os lemas do dicionario empregouse a selección dos termos que resultaron de eliminar os termos non tolerados do extracto do VOLG. Para os alomorfos destes lemas, aproveitáronse, cando existían, os alomorfos da versión anterior. Os demais introducíronse manualmente. A relación cos afixos responde tamén, en parte, coa da versión anterior que se editou, manualmente nalgunhas ocasións e coa axuda dunha folla de cálculo noutras, para adaptalo ás mudanzas que se introduciron durante a revisión morfolóxica, aos lemas novos, ás variacións de paradigmas, á asociación coa súa construción, á anotación morfolóxica...
Cómpre subliñar que os verbos que se consideraron susceptibles de aparecer como auxiliares en perífrases verbais (*acabar, acostumar, andar, botar, cesar, chegar, comezar, continuar, dar, deixar, empezar, estar, levar, parar, pasar, principiar, rematar, terminar, tornar, coller, deber, haber, poder, poñer, pór, ser, ter, volver, ir, seguir* e *vir*), manipuláronse para que, independentemente da información sintáctica que reciban, admitan a posibilidade de engadir enclíticos de calquera dos tres tipos de construcións.
Os lemas do arquivo de dicionario que non se corresponden con formas verbais coinciden coas formas preexistentes.

3. Comportamento

Fronte á versión anterior, prevese un paradigma diferente para que cada persoa gramatical da flexión poida responder a posibilidades combinatorias específicas (deste xeito, evítase, por exemplo, que o reflexivo de terceira persoa poida acompañar formas flexionadas doutras persoas gramaticais: **cómose, *coméchesse, *andádesse, *collémosse...*). As posibles secuencias de pronomes enclíticos varían, ademais, segundo o réxime de construción.

3.1 Paradigma dunha construción transitiva

A forma flexiva non acepta, na secuencia de pronomes enclíticos, o reflexivo da mesma persoa gramatical na primeira posición da secuencia, salvo na terceira persoa, de singular e de plural, para a formación das construcións impersoal activa e impersoal pasiva.
O verificador admite: *retíñanllelas*, *mantíñana*, *mantíñasnola*, *tráiolle*, *produciuse un erro*, *advírteselles ás persoas interesadas...* e rexeita **saltácheste a clase*, **bebinme o leite*, **mercástesvos cadanseu vestido...* A detección deste tipo de castelanismos sintácticos non queda resolta na terceira persoa, de singular e de plural, pois a aparición do pronome reflexivo forma construcións impersoais.



Existen tamén outras limitacións para este comportamento: se a forma verbal ou un termo homógrafo (ou unha forma verbal susceptible de auxiliar a outra nunha perífrase) admiten tamén o réxime de construción pronominal, o verificador non pode diferenciar o uso concreto e vai permitir tamén as secuencias de pronomes enclíticos prevista para esta construción.

3.2 Paradigma dunha construción intransitiva

A forma flexiva non acepta, na secuencia de pronomes enclíticos, o reflexivo da mesma persoa gramatical da forma flexiva na primeira posición da secuencia, salvo na terceira persoa, de singular e de plural, para a formación da construción impersoal intransitiva (agás os verbos impersoais, que tampouco admiten o reflexivo na terceira persoa, pois non precisan marcar sintacticamente a indeterminación do suxeito); tampouco admite na secuencia un pronome enclítico acusativo marcado morfoloxicamente (P2 / P3 / P6).

Deste xeito, o verificador admite usos como *acontecéuchemelles*, *abóndalles*, *concorríase con frecuencia, aquí vívese ben...* e non admite **aconteceuna, *morreuno, *abondádelas, *finóuchemellela, *névase*. Con todo, existen limitacións, coma na construción transitiva, para este comportamento: se a forma verbal ou un termo homógrafo (ou unha forma verbal susceptible de auxiliar a outra nunha perífrase) admiten tamén o réxime de construción pronominal, o verificador non pode diferenciar o uso concreto e vai permitir tamén as secuencias de pronomes enclíticos prevista para esta construción.

Ademais, tanto en construcións transitivas coma en construcións intransitivas, as formas flexivas con sincretismo da persoa gramatical inclúen as posibilidades sintagmáticas de todas as persoas que representan: *cantaba unha canción* pode responder a un uso en primeira ou en terceira persoa de singular do copretérito do verbo cantar, de tal xeito que *cantábame unha canción*, vaise validar independentemente de que responda a un uso indebido en primeira persoa, porque o verificador identificaa cun suxeito en terceira persoa de singular, debido a que a secuencia de enclíticos comeza polo reflexivo de primeira persoa, que non forma parte do paradigma de secuencias de pronomes enclíticos previsto para a forma en primeira persoa.

3.3 Paradigma dunha construción pronominal

A forma flexiva só acepta secuencias de pronomes enclíticos nas que estea presente o reflexivo da mesma persoa gramatical (ou formas marcadas morfoloxicamente como dativo + reflexivo) e non admite na secuencia un pronome enclítico acusativo marcado morfoloxicamente (P2 / P3 / P6)

O verificador admite usos como *desentendéronse, resentíuselle, entrecruzámoschenoslle, ativéstesllesvos..,* pero non admite **desentendéronlle, *resentiulle, *entrecruzámoschelle, *ativéstesslles, *arrepuxémonola...* As limitacións deste comportamento son similares ás que se expuxeron para os paradigmas de construcións transitivas e intransitivas.

Se se toma como exemplo unha forma verbal que admita os tres réximes de construción, obsérvase que, mesmo con todos os posibles paradigmas, o verificador nunca vai admitir unha construción pronominal cunha secuencia de enclíticos que conteña un pronome marcado morfoloxicamente como acusativo: **batémonolos, *botámonola*. Admite usos como *había colonia e botámola*, *había colonia e botamos, había colonia e botámonos*, pero non admite había colonia e *botámonola[4].

4 Álvarez e Xove, 2002, p. 556.

3.4 As suxestións de substitución

Cando se acadou o comportamento que se vén de describir, observouse que as regras para as suxestións de termos cando verificador atopa unha forma que non recoñece melloraran sensiblemente, xa que se eliminaran a formas agramaticais na flexión.
As formas verbais que o VOLG inclúe como formas non toleradas introducíronse como regras de substitución con resultados moi satisfactorios:

```
$ hunspell -m -d gl_ES -i utf-8 -a

@(#) International Ispell Version 3.2.06
(but really Hunspell 1.2.6)1.2.6

olvidouna

& olvidouna 1 0: esqueceuna
```

4. Particularidades de Hunspell

A documentación, en lingua inglesa, e as instrucións de descarga e instalación da programa de verificación ortográfica están dispoñibles no seu sitio web.

4.1 A recursividade

O verificador MySpell/Hunspell permite un envío de relacións desde o arquivo de dicionario (lema e alomorfos) cara ao arquivo de afixos (que xera a súa flexión), e unicamente outro envío desde cada unha destas formas flexivas cara a ese mesmo arquivo de afixos (co que se xera a secuencia de enclíticos).
Esta limitación na recursividade impón que os paradigmas de pronomes persoais enclíticos sexan tan analíticos, debido a que non resulta posible segmentar as secuencias de pronomes enclíticos en unidades palabra, que sería unha descrición máis precisa do comportamento lingüístico.

4.2 A segunda forma do artigo e a guionización

Non se obtiveron resultados satisfactorios nas probas que se fixeron para verificar o alomorfo do artigo determinado (-la, -las, -lo, -los): a documentación de Hunspell xa advirte de que se pode empregar a instrución WORDCHARS - para que o verificador, unicamente nun terminal, non divida as palabras con esta grafía (unha solución deste tipo faise precisa en lingua galega, debido que a sílaba do artigo afecta para a acentuación gráfica na combinación de formas verbais que rematan en -r ou -s co alomorfo do artigo). Cos sufixos que se probaron obtivéronse bos resultados co analizador morfolóxico de Hunspell executándose nun terminal, no entanto, produciron efectos non desexados en todos os programas de edición de texto cos que se experimentou, xa que todos dividen en palabras diferentes a secuencia que haxa antes do guión respecto da que figure a continuación do guión pese a que existan regras de sufixación que o inclúan.
Co exemplo *cóme-lo caldo,* o verificador non vai atopar a secuencia *cóme* na flexión do lema *comer,* porque a súa acentuación responde á ligazón co alomorfo do artigo, que precisa da grafía con guión, mais os editores de texto van interpretar dúas unidades palabra diferenciadas que se segmentan co trazo.
Como froito destas probas e a raíz da documentación que figura wiki de Mancomún comezouse tamén a elaboración dun arquivo de guionización acorde coa silabación en lingua galega, que se inclúe co ficheiro de afixos e co dicionario, e que interpreta xa algunhas características propias do galego, pero que aínda precisa de moitas melloras.
Cómpre ter en conta que este arquivo impón regras de segmentación silábica para o uso do guión ao final de liña e que non o empregan todas as aplicacións que utilizan o verificador, senón que os programas que o xestionan adoitan ter un xeito particular de facelo. As probas realizáronse co motor, propio, do OpenOffice 3.

4. Conclusións

Os arquivos que se obtiveron tras este traballo están dispoñibles, coa mesma licenza que os arquivos orixinarios dos que se partiu, no enlace http://webs.uvigo.es/miguelsolla/gl_ES.zip.
O comportamento que se describiu parece mellorar sensiblemente a eficacia do verificador e abre as portas para establecer novas regras de substitución que aumenten a súa utilidade.
Os comentarios en cada regra de sufixación, que se empregaron inicialmente para identificar o código con maior precisión, reconvertéronse durante o proceso de revisión en anotación que se adaptou, tamén no arquivo de dicionario, para o analizador morfolóxico de Hunspell.

```
$ hunspell -m -d gl_ES -i utf-8

trouxémoschas

trouxémoschas
 st:traer
 is:alomorfo traer transitiva
 ds:pretérito P4 + enclítico
 po:pronome persoal enclítico
  is:monosílabo P4 transitiva dativo P2 +
acusativo P3
```

Deste xeito facilítase moito a depuración de comportamentos inesperados e o verificador fornécese dunha ferramenta que se pode estender no futuro para outras clases de palabras. É preciso ter en conta que para poder empregar o analizador é preciso que estean instalados Hunspell e myspell-gl-es no

sistema e que, na actualidade, myspell-gl-es instala uns arquivos herdeiros doutros verificadores, en normativa de mínimos, que se deben substituír manualmente, para empregalos por omisión, ou indicarlle ao analizador morfolóxico de Hunspell cada vez que se use a localización dos ficheiros que se queiran utilizar.

E alén do obxecto destas liñas, cómpre salientar que aínda quedan por facer diferentes *thesauri* que completen as posibilidades do verificador e amplíen a súa eficacia; pero tamén é certo que xa dispoñemos de ferramentas lingüísticas en código aberto que superan moito a simple verificación da ortografía, coma o Golfiño ou o Exeria, ambos os dous coa análise lingüística do FreeLing.

5. Bibliografía

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

Santamarina, Antón e Manuel González González (coord.). 2004. *Vocabulario ortográfico da lingua galega*, Real Academia Galega / Instituto da Lingua Galega.

Álvarez, Rosario e Xosé Xove. 2002. *Gramática da lingua galega*. Editorial Galaxia, Vigo.

Freixeiro Mato, Xosé Ramón. 2000. *Gramática da lingua galega II. Morfosintaxe.* Edicións A Nosa Terra, Vigo, 1ª edición.

Álvarez, Rosario, X. L. Regueira e H. Monteagudo. 1986. *Gramática galega*, Editorial Galaxia, Vigo.

Hermida Gulías, Carme. 2004. *Gramática práctica (morfosintaxe)*. Sotelo Blanco Edicións, Santiago de Compostela.

Hermida, Avelino. 2006. *Conxugación verbal da lingua galega século21*. Edicións do Cumio / Editorial Galaxia.

González González, Manuel, Carmen García Mateo, Eduardo Rodríguez Banga e Elisa Fernández Rei. 2002. *Diccionario de verbos galegos Laverca*, Edicións Xerais de Galicia, Vigo.

Díaz Regueiro, Manuel. 1992. *Os verbos galegos.* Consellería de Educación e Ordenación Universitaria / Dirección Xeral de Política Lingüística.

Fernández Rei, Francisco. 1991. *Dialectoloxía da lingua galega.* Edicións Xerais de Galicia, Vigo, 2ª edición.

Graña Núñez, Xosé. 1993. *Vacilacións interferencias e outros "pecados" da lingua galega.* Ir Indo Edicións, Vigo.

González Rei, Begoña. 2004. *Ortografía da lingua galega*- Galinova Editorial, A Coruña.

Hermida Gulías, Carme. 2001. *Ortografía práctica.* Sotelo Blanco Edicións, Santiago de Compostela.

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 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 setembro de 2009
- Resultados da selecção até: 31 de sutubro de 2009
- Versão final até: 15 de novembro de 2009
- Publicação da revista: 30 de novembro de 2009

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 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 setembro de 2009
- Resultados da selección até: 31 de outubro de 2009
- Versión final até: 15 de novembro de 2009
- Publicación da revista: 30 de novembro de 2009

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 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 septiembre de 2009
- Resultados de la selección hasta: 31 de octubre de 2009
- Versión final hasta: 15 de noviembre de 2009
- Publicación de la revista: 30 de noviembre de 2009

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 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 de setembre de 2009
- Resultats de la selecció fins a: 31 de octubre de 2009
- Versió final fins a: 15 de novembre de 2009
- Publicació de la revista: 30 de novembre de 2009

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

Design: Alberto Simões

http://www.linguamatica.com/