Volume 2, Número 2 - Junho 2010

# linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 2, Número 2 – Junho 2010

# LinguaMÁTICA

ISSN: 1647–0818

# Conteúdo

# Editorial

*Neste cuarto número de Linguamática, queremos salientar a incorporación dunha nova sección á revista, dedicada á presentación de proxectos de investigación en curso. Agardamos que esta nova posibilidade constitúa un estímulo á colaboración e á difusión do moito traballo de fondo que se está a facer nas nosas universidades e centros de investigación.*

*Desta vez contamos cun artigo convidado a cargo de Maria das Graças Volpe Nunes, da Universidade de São Paulo, coordinadora do Núcleo Interinstitucional de Lingüística Computacional (NILC). O NILC, fundado en 1993, é nestes momentos o grupo de investigación en Procesamento de Linguaxe Natural da lingua portuguesa brasileira máis importante do Brasil. Na súa importante contribución a Linguamática, a profesora Nunes fai un exame polo miúdo das actividades e liñas desenvolvidas polo seu grupo nestes 17 anos de labor constante e produtivo.*

*Queremos agradecer máis unha vez a todas as persoas que enviaron traballos de investigación orixinais para este número de Linguamática, publicados finalmente ou non, e aos membros do Comité Científico que participaron no seu comentario e avaliación. Sen a vosa colaboración e apoio, non existiría a revista.*

*Xavier Gómez Guinovart*
*José João Almeida*
*Alberto Simões*

# Comissão Científica

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

# Dossier

# Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade

Maria das Graças V. Nunes, Sandra M. Aluisio, Thiago A. S. Pardo
NILC – ICMC – Universidade de São Paulo
São Carlos – SP, Brasil
{gracan, sandra, taspardo}@icmc.usp.br

**Resumo**

Este artigo faz uma breve apresentação do Núcleo Interinstitucional de Linguística Computacional (NILC), que é um dos principais grupos brasileiros dedicado a pesquisas na área de Processamento de Línguas Naturais, particularmente do português brasileiro. Após apresentar um breve histórico de sua formação, mostramos como as atuais áreas de pesquisa do grupo foram consolidando-se ao longo dos anos. Para cada uma dessas áreas de atuação do NILC, fazemos um breve resumo dos resultados mais importantes e do estado atual das pesquisas no grupo.

## *1. Introdução*

O Núcleo Interinstitucional de Linguística Computacional (NILC)[1] é hoje composto por mais de 30 pesquisadores da área de Processamento de Línguas Naturais (PLN), incluindo professores universitários e alunos de graduação e pós-graduação, com formação principalmente em ciências da computação e linguística. Esse grupo foi criado em 1993, na Universidade de São Paulo, em São Carlos, com o objetivo de formar recursos humanos e desenvolver pesquisa e sistemas de PLN especialmente para o português do Brasil (PB). A criação do NILC foi especialmente motivada pelo convite recebido da empresa de informática Itautec, para implementar, como *plug-in* do Office da Microsoft, um sistema de revisão gramatical do português. O desafio era enorme, tendo em vista que àquela época não existiam recursos disponíveis para essa tarefa. Era necessário construir um léxico computacional, um analisador sintático robusto, voltado à detecção de erros sintáticos, e *corpora* de referência e de testes. Também era grande o desafio de compor e gerenciar uma equipe de pesquisa e desenvolvimento interdisciplinar (computação e linguística), com culturas tão distintas. Tudo isso fez com que o grupo já nascesse grande, com o compromisso de gerar um produto comercial e com a responsabilidade de criar tudo de que precisava. Apesar disso, uma primeira versão do revisor, sem análise sintática automática, foi lançada já em 1994. Outras versões se seguiram até que em 1999, por meio de uma licença que vigora até hoje, a Microsoft adquiriu direito de uso do revisor no Office 2000.

Com os recursos linguístico-computacionais construídos no projeto do revisor gramatical, até então inéditos para o PB, e já estendido com colaboradores de outras instituições – Universidade Federal de São Carlos (UFSCar) e Universidade Estadual de São Paulo (UNESP) – o grupo tornou-se referência na área de PLN e passou a ser convidado para desenvolver outros projetos, como o da Universal Networking Language (UNL). Em 1997, o NILC passou a representar o Brasil no grupo de países que integravam o Projeto UNL, patrocinado pelo Instituto de Estudos Avançados da Universidade das Nações Unidas (UNU/IAS). Mais tarde essa associação deu origem à UNDL Foundation[2], com sede em Genebra. A meta do projeto é criar ferramentas de tradução, dentro do paradigma de interlíngua, em um primeiro momento para as línguas oficiais da ONU e outras línguas de muitos falantes, para a comunicação na internet. Ao grupo brasileiro, cabia criar os recursos para a tradução entre o português e a interlíngua UNL. O projeto continua ativo na UNDL, porém, o NILC não participa mais como membro institucional. A participação, por cerca de 4 anos, no projeto UNL abriu no NILC uma importante área de pesquisa, a da tradução automática (TA). Tratava-se, à época, de uma área de pesquisa com muito pouca expressão no país. Vários outros projetos e importantes publicações têm sido gerados pelo grupo, que encontram-se em (Martins *et al*., 2004a) e na Seção 4 deste artigo.

A partir do envolvimento nesses dois grandes projetos, o grupo ganhou expressão no país e no exterior e passou a agregar novos membros. Sua atuação se estendeu a áreas mais teóricas e à

[1] http://www.nilc.icmc.usp.br/nilc/index.html

[2] http://www.undlfoundation.org/undlfoundation/

construção de recursos robustos para outras aplicações de PLN, o que acabou por aproximá-lo a outros grupos brasileiros de PLN.

O grupo se destaca também na organização e promoção da área de pesquisa em PLN no Brasil. Junto com outros grupos nacionais de expressão, como os da Universidade Católica do Rio Grande do Sul (PUC-RS) e do Rio de Janeiro (PUC-Rio), tem sido responsável por projetos de cooperação nacional e pelos principais eventos científicos dessa área. Esses grupos de pesquisa criaram, em 2003, o que é hoje o principal evento científico nacional dessa área, o STIL: Simpósio de Tecnologia da Informação e da Linguagem Humana[3], que está na sua 8ª edição. Da mesma forma, participam ativamente, e em conjunto com vários pesquisadores de Portugal, da organização do PROPOR[4], a conferência internacional e bianual sobre processamento do português brasileiro, hoje na sua 9ª edição.

Os projetos de parceria com colegas do Brasil e do exterior têm possibilitado a geração de recursos e ferramentas de interesse de toda a comunidade e que representam avanços significativos para o processamento do PB. Podemos destacar, e detalharemos nas próximas seções, os *corpora* compilados e anotados; os diferentes léxicos computacionais; as bases e redes lexicais; ferramentas avançadas, como as que fazem análise discursiva e simplificação sintática; ferramentas aplicadas à tradução automática; aplicações como a sumarização mono e multidocumento e os ambientes de auxílio à escrita e à leitura; novos métodos de avaliação de sistemas de PLN; etc.

Atualmente o NILC conta com 14 pesquisadores seniores, de quatro diferentes instituições brasileiras, e cerca de 20 estudantes de graduação e pós-graduação associados. Sob uma perspectiva histórica, este artigo procura mostrar algumas das principais contribuições do NILC para a área de PLN no Brasil, às vésperas de completar sua maioridade, bem como apresenta um cenário das áreas atuais de atuação dos autores signatários. Na Seção 2 descrevemos brevemente os principais recursos linguístico-computacionais criados no NILC e que servem de apoio a todas as demais pesquisas. A Seção 3 apresenta os principais resultados das pesquisas do grupo na área de sistemas de auxílio à escrita e à leitura, uma das áreas de pesquisa pioneiras do NILC. A experiência do grupo em TA é relatada brevemente na Seção 4. Na Seção 5 apresenta-se a trajetória das pesquisas do grupo em sumarização automática e análise discursiva. Finalmente, na Seção 6, concluímos o artigo arriscando fazer algumas projeções para o futuro próximo.

[3] http://www.nilc.icmc.usp.br/til/index.htm
[4] http://www.nilc.icmc.usp.br/cgpropor/

## *2. Ferramentas e recursos básicos para o processamento do PT brasileiro*

O primeiro recurso importante criado no NILC foi o léxico computacional (do NILC) que faz parte dos revisores ortográfico e gramatical do MS-Office. Do ponto de vista linguístico, a versão atual do léxico é capaz de gerar cerca de 1.500 mil lexemas a partir de cerca de 100 mil lemas. Cada lexema pode pertencer a uma ou mais de 13 classes, cada uma com atributos distintos. Do ponto de vista tecnológico, o léxico é implementado como um autômato finito minimizado, ocupando um espaço mínimo de memória e com desempenho otimizado (Jesus *and* Nunes, 2000). A partir do léxico, vários outros recursos lexicais foram produzidos no NILC: um tesauro eletrônico, a base Diadorim, o Unitex-BR, e finalmente a WordNet.Br.

O tesauro eletrônico TEP é resultado da primeira tentativa de se estender o léxico do NILC com informações semânticas de sinonímia e antonímia. Esse tesauro também é usado pelas ferramentas de revisão do Office para a tarefa de sugestão de alternativas. A base Diadorim é a versão do TEP disponível para a consulta na internet, na forma (ineficiente) de uma base de dados[5]. Já o Unitex-BR[6], criado segundo os formatos da ferramenta de *corpus* INTEX, é sua versão em código aberto, veiculada pela rede RELEX na web[7]. O conjunto de palavras simples no padrão DELA fez com que o número de ocorrências crescesse 93.28% em relação à fonte original. No entanto, o número de entradas do dicionário de palavras compostas, assim como o número de regras de remoção de ambiguidades, ainda é bastante tímido.

A evolução mais ambiciosa quanto à semântica lexical é a construção, em andamento, da WordNet.Br (Di Felippo *and* Dias-da-Silva, 2007; Dias-da-Silva *et al*., 2008), que segue os mesmos pressupostos da Wordnet de Princeton (Fellbaum, 1998). A versão preliminar, sob o nome TeP 2.0 (Maziero *et al*., 2008), tem interface disponível na

[5] http://www.nilc.icmc.usp.br/nilc/tools/intermed.htm
[6] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html
[7] http://infolingu.univ-mlv.fr/brasil/

web[8]. Atualmente, o TeP 2.0 contém 19.888 conjuntos de sinônimos e 44.678 unidades lexicais, tendo a média de 2,5 unidades por conjunto de sinônimos. Quanto à antonímia, há 4.276 relações entre os synsets da base, ou seja, aproximadamente 22% da base está reLácionada por meio dessa relação. Além disso, para 253 unidades lexicais pertencentes à categoria dos verbos, o TeP 2.0 armazena uma frase-exemplo distinta para cada uma das unidades. A frase-exemplo fornece o contexto de uso mínimo do item lexical. O recurso armazena também uma glosa (ou seja, uma definição informal do conceito) para 6.648 synsets, todos eles constituídos por unidades da categoria dos verbos.

Toda a evolução dos recursos lexicais, bem como o desenvolvimento de ferramentas e aplicações para o processamento do português, foi acompanhada pela construção sucessiva e progressiva de diferentes *corpora*. O primeiro grande *corpus*, chamado de *corpus* NILC[9], com cerca de 40 milhões de palavras, foi compilado para subsidiar as pesquisas do revisor gramatical. Para tanto, deveria ser representativo dos desvios da língua escrita por usuários "médios" de editores de texto digital. Era preciso identificar e modelar os principais desvios gramaticais. Ao mesmo tempo, o *corpus* também deveria servir como referência para a construção de uma gramática normativa, já que a função do revisor é detectar desvios e sugerir correções. Essa dupla finalidade criou as três divisões do *corpus* NILC conhecidas como *corpus* corrigido (obras literárias, livros didáticos, textos jornalísticos, etc.), *corpus* não corrigido (redações de vestibulares) e *corpus* semi-corrigido (teses acadêmicas, cartas comerciais, etc.). O *corpus* NILC está disponível para consulta na Linguateca, no âmbito do projeto AC/DC[10].

O *corpus* NILC foi, durante muito tempo, a fonte de informação sobre o PB contemporâneo escrito para as pesquisas no grupo. A partir de 2002, com o apoio do CNPq, e em parceria com o IME (Instituto de Matemática e Estatística) e a FFLCH (Faculdade de Filosofia, Letras e Ciências Humanas), da USP-São Paulo, o projeto Lácio-Web[11], de construção de *corpora*, teve início no NILC. O objetivo deste projeto era divulgar e disponibilizar livremente na Web vários *corpora* do PB escrito contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados em padrão de fácil intercâmbio, navegação e análise. Além disso, disponibilizar ferramentas linguístico-computacionais, tais como contadores de freqüência, etiquetadores morfossintáticos e concordanciadores. A idéia era prover recursos para um público heterogêneo: de um lado linguistas, cientistas da computação, lexicógrafos, entre outros, e, de outro, não especialistas em geral. Formado por quatro grandes *corpora*, o Lácio-Web contém 10,5 milhões de palavras de textos dos gêneros informativo, jurídico, científico, literário e instrucional.

Após o Lácio-Web, outros importantes *corpora* foram construídos pelo grupo. Destacamos a participação do NILC no Projeto do Dicionário Histórico do Português Brasileiro dos séculos 16 até o início do século 19 (HDBP)[12], tratou de várias características inerentes a textos históricos, tais como: ausência de uma ortografia, uso extensivo de abreviações e suas variações de grafia, falta de espaço entre as palavras, uso irregular da hifenização e símbolos tipográficos que caíram em desuso (Candido Jr. *et al*., 2009).

Mais recentemente, no âmbito do projeto de cooperação multi-institucional (USP, UFSCar, Unisinos, PUC-RS, PUC-Rio, Mackenzie, UNESP), o grupo coordenou a criação do Portal de *Corpus*[13] (Muniz *et al*., 2007), formado por 3 *corpora*:
(a) PLN-BR FULL, que contém 103.080 mil textos da Folha de São Paulo e 29.014.089 *tokens*; está formatado segundo etiquetas do Unitex;
(b) PLN-BR CATEG, que tem 30 mil textos e 9.780.220 *tokens*, originalmente criado para compor um *benchmark* para avaliação de métodos de classificação textual;
(c) PLN-BR GOLD, que possui 1024 textos e 338.441 *tokens* e pode ser acessado livremente via Web. O tamanho deste *corpus* é tal que representa 1% do *corpus* PLN-BR FULL de forma a conservar, proporcionalmente, a distribuição deste *corpus* maior. Trata-se de uma amostra aleatória estratificada e proporcional à distribuição do *corpus* PLN-BR FULL com relação aos textos dos cadernos do jornal. Foi criado para exemplificar e tornar pública a proposta de anotação de *corpora* da Língua Portuguesa, considerando vários níveis linguísticos (Bruckschen *et al*., 2008).

Vários outros *corpora*, de uso mais restrito a determinadas pesquisas e aplicações, têm sido compilados no NILC, como o TeMário, de sumários feitos manualmente; o *Corpus*TCC, de teses acadêmicas, e o RHETALHO, de textos acadêmicos

[8] http://www.nilc.icmc.usp.br/tep2/index.htm
[9] http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm
[10] http://acdc.linguateca.pt/acesso/
[11] http://www.nilc.icmc.usp.br/Lácioweb/index.htm
[12] http://www.nilc.icmc.usp.br/nilc/projects/hpc/
[13] http://www.nilc.icmc.usp.br:8180/portal/

e jornalísticos, ambos anotados pela ferramenta de análise discursiva RSTTool; o *Corpus* Paralelo, de textos alinhados português e inglês, o *corpus* paralelo de textos originais e simplificados léxica e sintaticamente, entre outros. Vários destes *corpora* são detalhados nas próximas seções.

Entre as ferramentas de PLN desenvolvidas no NILC, destacamos os etiquetadores POS construídos no âmbito do projeto Lácio-Web[14] (Aluísio *et al*., 2003), e o *parser* Curupira, que é derivado do revisor gramatical, e provê o conjunto de todas as análises sintáticas possíveis para uma dada sentença em PB (Martins *et al*. 2003).

## *3. Sistemas de Auxílio à Escrita e à Leitura*

A necessidade de escrever artigos científicos em inglês é um dos grandes problemas de pesquisadores de vários países cuja língua nativa não é o inglês. Os trabalhos do NILC nesta área têm explorado uma estratégia de escrita baseada no reuso de trechos de textos escritos por pesquisadores nativos do inglês, indexados pelos componentes da estrutura esquemática da seção na qual aparecem.

Embora grande parte dos problemas enfrentados por escritores nativos se apresente no nível estrutural, problemas nos níveis lexical e sentencial também ocorrem. De fato, esses escritores têm o conhecimento da língua no seu uso geral, mas podem não dominar o seu uso em um gênero específico, tendo problemas na escolha de itens lexicais e estruturas sintáticas apropriadas.

Ferramentas de suporte à escrita científica em inglês com base na abordagem baseada em casos (*case-based reasoning*) e em sistemas de críticas *(expert/computer-aided critiquing systems*), largamente usados na grande área de Inteligência Artificial, foram desenvolvidas no projeto AMADEUS (AMiable Article DEvelopment for User Support) (Fontana *et al*., 1993; Aluísio *and* Oliveira, 1995; Aluísio *and* Oliveira Jr, 1996; Aluísio *and* Gantenbein, 1997; Aluísio *et al*., 2001). Estas ferramentas foram portadas para o ambiente Web, seguindo a tendência atual para facilitar o acesso de sistemas (por exemplo, o sistema SciPo-Farmácia[15] (Aluísio *et al*., 2005)), e também uma delas, chamada SciPo[16] (Feltrim, 2004; Feltrim *et al*., 2004; Feltrim *et al*., 2006), foi disponibilizada para a língua portuguesa para ser usada por escritores nativos do português escrevendo teses e dissertações.

Experiências realizadas com as ferramentas de auxílio à escrita científica têm demonstrado que a boa aceitação das mesmas por parte de seus usuários se deve fortemente ao fato de possuírem *corpora* específicos da área de pesquisa do usuário-escritor. Assim, uma questão que se coloca é o custo de se estender esse auxílio computacional a pesquisadores de diferentes áreas do conhecimento, pois o gargalo da construção das ferramentas é a anotação dos textos com os componentes da estrutura esquemática de um artigo, tese ou dissertação. A solução proposta no NILC foi a utilização de detecção automática dos elementos estruturais de textos científicos, dado que esta proposta se apresenta também como um desafio científico, pois trata da automatização de uma tarefa que é problemática mesmo quando realizada por humanos. Alguns sistemas têm sido propostos na literatura para a realização dessa tarefa (Burstein *et al*., 2003; Antony *and* Lashkia, 2003; Teufel *and* Moens, 2002). No NILC foram desenvolvidos dois sistemas de detecção automática de estrutura esquemática de resumos, o AZPort (Feltrim *et al*., 2004) e o AZEA[17] (Genoves *et al*., 2007a). O primeiro é voltado para resumos em português e o segundo para resumos em inglês (*abstracts*). Ambos se baseiam no método AZ (*Argumentative Zonning*) (Teufel *and* Moens, 2002)

O SciPo (*Scientific Portuguese*), inspirado no projeto AMADEUS, é um ambiente Web voltado para escritores cuja língua mãe é o português, em especial aqueles que estão iniciando sua carreira acadêmica e ainda não estão familiarizados com as convenções do gênero científico. Ele baseia-se em teses e dissertações da área de Computação.

O SciPo apóia a estruturação e a realização linguística de textos científicos de forma flexível, deixando o usuário livre para escolher entre dois modos de trabalho, a saber: (i) um processo *top-down*, que parte do planejamento estrutural para a escrita propriamente dita, incluindo ciclos de críticas e refinamentos da estrutura, herdado do projeto AMADEUS; ou (ii) um processo *bottom-up*, em que se submete um texto já escrito à análise (detecção e crítica) automática da estrutura. Na verdade, trata-se de pontos de partida distintos para um mesmo processo cíclico de refinamento, já que a estrutura detectada e criticada em (ii) pode ser

[14] http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html
[15] http://www.nilc.icmc.usp.br/scipo-farmacia/
[16] http://www.nilc.icmc.usp.br/~scipo/
[17] http://www.nilc.icmc.usp.br/azea-web/

aprimorada por meio dos recursos disponíveis em (i).

O SciPo-Farmácia é um conjunto de ferramentas computacionais desenvolvido para ajudar os usuários a escreverem artigos científicos em inglês. Possui a mesma interface do SciPo, porém um número menor de funcionalidades e baseia-se em artigos científicos da área de Ciências Farmacêuticas. Este sistema foi desenvolvido com o intuito de ajudar estudantes e pesquisadores que não têm o inglês como língua materna e necessitam escrever artigos científicos nessa língua e/ou também não estão familiarizados com a estrutura e as peculiaridades do gênero científico. O desenvolvimento do SciPo-Farmácia resultou de uma parceria entre pesquisadores da Faculdade de Ciências Farmacêuticas da USP de São Paulo e o NILC.

Outras linhas recentes de pesquisa para apoio computacional para a escrita e a leitura, no NILC, incluem:
(i) a diversificação de gênero, com o desenvolvimento de uma ferramenta Web inteligente de auxílio à escrita de planos de negócios em português (Ferraz Jr *et al.*, 2007; Raymundo *et al.*, 2007);
(ii) a implementação de uma rubrica baseada no gênero científico para analisar resumos de artigos (Aluísio *et al.*, 2005; Schuster *et al.*, 2005; Genoves *et al*, 2007a; Genoves *et al.*, 2007b);
(iii) e, mais recentemente, o desenvolvimento de tecnologias para facilitar o acesso de informação por pessoas com baixo nível de letramento ou outros problemas de leitura, no escopo do projeto PorSimples[18] (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) (Aluisio *et al*, 2008).

O grande objetivo do projeto PorSimples é poder ajudar pessoas com problemas de leitura a compreender documentos do gênero informativo disponíveis na Web brasileira, por exemplo, informações do governo e notícias de jornais de grande circulação.

No Brasil, o Indicador de Alfabetismo Funcional (INAF) tem sido computado desde 2001 para medir os níveis de letramento da população brasileira. O relatório mais atual, de 2009, apresenta um cenário ainda desanimador: 7% das pessoas são analfabetas; 21% são alfabetizadas no nível rudimentar; 47% são alfabetizadas no nível básico; e somente 25% são totalmente alfabetizadas (INAF, 2009). O número de pessoas com alfabetização nos níveis rudimentar e básico totaliza 68% da população do Brasil e estas podem somente achar informação explícita em textos curtos (rudimentares), ler e entender textos um pouco maiores, além de serem capazes de fazer inferências simples (básicas). Estes dois níveis são o alvo do projeto PorSimples, e para isso foram desenvolvidos três sistemas destinados a públicos alvos diferentes:

- um sistema de autoria, chamado SIMPLIFICA[19], para ajudar autores a produzirem textos simplificados destinados aos alfabetizados rudimentares e básicos (Candido Jr *et al.*, 2009);
- sistemas facilitadores para ajudar o mesmo público acima a ler um dado conteúdo da Web. Estes incluem tarefas de sumarização textual e simplificação sintática (sistema FACILITA[20]) (Watanabe *et al.*, 2009) e elaboração léxica, apresentação do texto salientando as relações retóricas entre as idéias do texto, explicitação das Entidades Mencionadas e dos argumentos dos verbos (sistema FACILITA EDUCATIVO[21]) (Watanabe *et al.*, 2010).

O sistema SIMPLIFICA (Figura 1) é um editor WYSIWYG baseado no editor WEB TinyMCE[22].


**Figura 1:** Tela principal do SIMPLIFICA que dá acesso às 3 funcionalidades do editor: simplificação léxica e sintática (no topo, acima do texto) e verificador da inteligibilidade (na barra de status)

O usuário insere um texto no editor e realiza: (i) as escolhas para a simplificação relacionadas ao tipo de público alvo, podendo ser: simplificação forte (para alfabetizados rudimentares) em que todos os fenômenos sintáticos complexos de uma sentença são tratados; simplificação natural (para alfabetizados básicos) em que somente as sentenças apontadas por um classificador treinado em um *corpus* anotado manualmente serão tratadas; e

[18] http://caravelas.icmc.usp.br/wiki/

[19] http://www.nilc.icmc.usp.br/porsimples/simplifica/

[20] http://vinho.intermidia.icmc.usp.br:3001/facilita/

[21] http://vinho.intermidia.icmc.usp.br/watinha/Educational-Facilita/

[22] http://tinymce.moxiecode.com/

simplificação customizada em que o usuário escolhe o fenômeno alvo de simplificação, e (ii) um ou mais tesauros a serem utilizados no processo de simplificação léxica.

Após as escolhas acima, o usuário pode ativar o verificador de inteligibilidade (Aluisio *et al*., 2010). Este módulo mapeia o texto em um dos 3 níveis de letramento definidos pelo INAF: rudimentar, básico, avançado. De acordo com o resultado do verificador, o usuário pode ativar simplificações léxicas e sintáticas, revisar as simplificações e iniciar novamente o ciclo, via nova checagem da inteligibilidade do texto simplificado.

O sistema FACILITA (Figura 2) é um plug-in destinado a facilitar a leitura de um documento da Web por alfabetizados dos níveis rudimentar e básico.

FACILITA inclui módulos separados de sumarização textual e simplificação sintática.

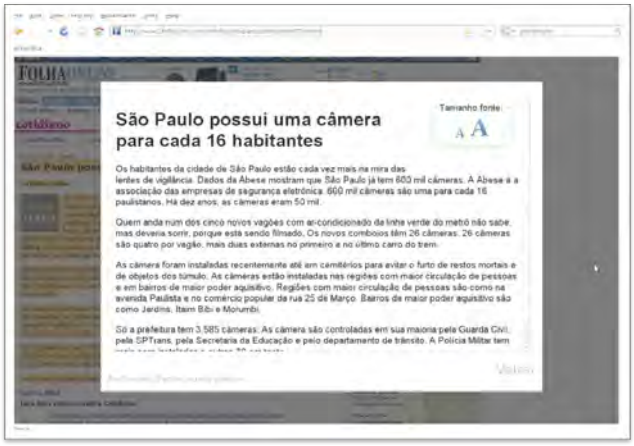



**Figura 2:** Janela *popup* mostrando o conteúdo facilitado de uma página Web cujo trecho em foco foi selecionado pelo usuário (ao fundo)

O usuário pode selecionar um texto de um site da Web e ativar FACILITA para obter o conteúdo facilitado. O módulo de sumarização é baseado na técnica EPC-P (extração de palavras-chaves por padrão) que verifica a presença de palavras-chaves nas sentenças do texto; aquelas que possuem palavras-chaves são retidas para o sumário final. O módulo de simplificação é melhor descrito em (Candido Jr *et al*., 2009).

O sistema Educational FACILITA[23] (Figura 3) é uma aplicação Web destinada a ajudar pessoas com baixo letramento a entenderem o conteúdo de documentos. As entidades nomeadas são marcadas e, ao serem selecionadas, definições curtas são apresentadas, vindas da Wikipédia. Também marca palavras complexas para as quais apresenta sinônimos simples.

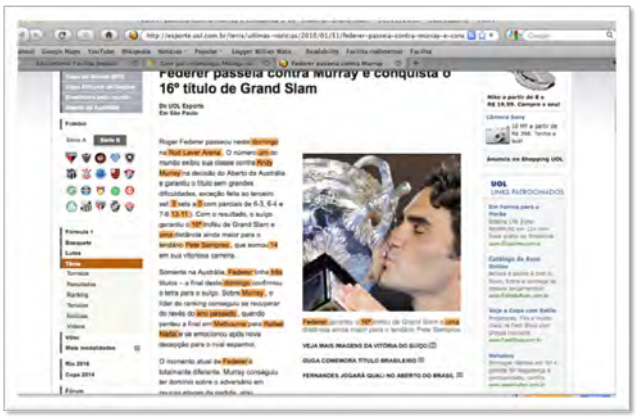



**Figura 3:** Resultado do sistema de elaboração textual FACILITA EDUCACIONAL ao ser acionado de uma página Web.

Mais detalhes dos recursos, métodos, sistemas e ferramentas de suporte disponibilizados pelo PorSimples podem ser vistos em Aluisio *and* Gasperin (2010).

## *4. Tradução Automática*

O NILC possui trabalhos em TA de diferentes paradigmas. Com o projeto UNL, tiveram origem alguns trabalhos em TA por interlíngua (Martins *et al*., 2004a). O projeto Retratos, por sua vez, investiga a tradução entre português, espanhol e inglês, por meio de regras de tradução aprendidas automaticamente de *corpus*. Já no paradigma estatístico, alguns trabalhos procuram criar os primeiros recursos para PB. Outros trabalhos relacionados a TA dizem respeito à desambiguação lexical e à avaliação de traduções automáticas. Comentamos a seguir sobre essas linhas de pesquisa.

O projeto EPT-Web[24] propôs um sistema de tradução por interlíngua de *headlintes* de notícias do The NewYork Times para o português. Para a criação do protótipo foi necessário criar um conjunto interessante de recursos: um dicionário trilíngue inglês-UNL-português (Antiqueira *et al*., 2002 ), um sistema de tradução inglês-UNL (Martins *et al*., 2004b) e um sistema de tradução UNL-português, este derivado dos trabalhos desenvolvidos no Projeto UNL.

[23] http://vinho.intermidia.icmc.usp.br/watinha/Educational-Facilita/

[24] http://www.nilc.icmc.usp.br/nilc/projects/ept-web.htm

Outra experiência com a UNL ocorreu no projeto LIBRAS[25], que visava a tradução de PB para a língua brasileira de sinais, Libras. Resultados preliminares evidenciaram a complexidade de se relacionar semanticamente 3 línguas de naturezas distintas: uma língua natural (PB), uma língua gestual-visual (Libras), e uma interlíngua (UNL) cuja função é representar a semântica comum entre as outras duas (Nunes *et al*., 2003).

Na linha da tradução baseada em regras, o projeto Retratos (Caseli, 2007) desenvolveu ferramentas de alinhamento sentencial e lexical para as línguas portuguesa, inglesa e espanhola, criou *corpora* paralelos e sistemas de aprendizagem automática de léxicos bilíngues e de regras de tradução (Nunes *et al*., 2008). Este projeto compartilhou alguns recursos do sistema de tradução de código aberto, Apertium[26]. Os recursos criados têm servido para apoiar outras pesquisas, como a de reconhecimento de multipalavras a partir de *corpus* paralelo bilíngue (Caseli *et al*., 2009).

Pelo que se sabe, os primeiros trabalhos no Brasil na linha da TA estatística foram realizados no NILC. Aziz *et al*. (2008) desenvolveram um tradutor estatístico entre o PB e o espanhol. Com base em um *corpus* paralelo relativamente pequeno de notícias de divulgação científica da Revista Pesquisa FAPESP, treinaram-se alguns modelos estatísticos clássicos baseados em palavras (Brown *et al*., 1993). Os resultados obtidos foram levemente inferiores ao tradutor Apertium. Dando continuidade a este trabalho, Aziz *et al*. (2009a) treinaram modelos estatísticos mais sofisticados baseados em *phrases* (que, nesse contexto, significam sequências quaisquer de palavras) (Koehn *et al*., 2003), incluindo, além das línguas anteriores, o inglês americano. Utilizando-se o mesmo *corpus*, os resultados obtidos foram superiores aos obtidos pelo Apertium para o par de línguas português-espanhol. Por meio de um experimento preliminar, constatou-se que os resultados são comparáveis ao Google Translate[27] para o par de línguas português-inglês.

Além dos trabalhos anteriores, trabalhos complementares de Caseli e Nunes (2009), Nunes e Caseli (2009) e Aziz *et al*. (2009b) investigaram como alguns parâmetros e simples escolhas de modelagem podem interferir na qualidade da tradução produzida pelos métodos de TA estatística. Por exemplo, investigaram-se as questões de uniformização de fonte, uso de pontuação no texto, e aplicação de otimização dos valores de parâmetros estatísticos, dentre outros, demonstrando-se que algumas pequenas alterações podem influenciar positivamente os resultados.

O projeto LeAR investigou a desambiguação lexical de sentido (WSD) para a TA. Propôs uma nova abordagem de WSD voltada especificamente para a tradução automática, que segue uma metodologia híbrida - baseada em conhecimento e em *corpus* - e utiliza um formalismo relacional para a representação de vários tipos de conhecimento e de exemplos de desambiguação, por meio da técnica de Programação Lógica Indutiva (ILP). Experimentos diversos mostraram que a abordagem proposta supera abordagens alternativas para a desambiguação multilíngue e apresenta desempenho superior ou comparável ao do estado da arte em desambiguação monolíngüe. Adicionalmente, tal abordagem se mostrou efetiva como mecanismo auxiliar para a escolha lexical na tradução automática estatística (Specia *et al*., 2009a). Este trabalho também mostrou como a ILP, juntamente com vários tipos de conhecimento de fundo, pode melhorar consideravelmente o desempenho de sistemas de desambiguação lexical de sentido (Specia *et al*., 2009b)

Outra linha relacionada à TA é a que investiga métodos alternativos para avaliação automática de traduções automáticas. O estabelecimento de métricas para avaliação automática da qualidade dos sistemas de tradução automática é crucial devido ao amplo uso da TA na web, e isto pode ser feito representando-se textos como redes complexas. Os conceitos e metodologias de redes complexas vêm sendo usados numa enorme variedade de áreas (Costa *et al*., 2008), incluindo a análise automática de textos em PLN. O potencial uso de redes complexas para esse tipo de análise foi demonstrado em várias oportunidades, a partir da comprovação de que um texto pode ser representado por uma rede livre de escala (Cancho *and* Sole, 2001), isto é, uma rede com poucos vértices fortemente conectados e muitos vértices fracamente conectados. Resultados consolidados no grupo incluem a determinação de autoria (Antiqueira et. al., 2007), a avaliação da qualidade de sumários automáticos (Antiqueira *et al*., 2009), e de tradução automática (Amancio *et al*., 2008). Neste último cenário, métricas de redes complexas foram aplicadas e os resultados foram utilizados como entrada para métodos de aprendizado de máquina, e permitiram que textos traduzidos automaticamente e manualmente fossem distinguidos. Tal método foi aplicado para o par de

[25] http://www.nilc.icmc.usp.br/nilc/projects/LIBRAS2.htm
[26] http://www.apertium.org/
[27] http://translate.google.com/

línguas inglês-português e espanhol-português. Os resultados mostram que é possível capturar um contexto mais amplo com a utilização de níveis hierárquicos mais profundos em conjunto com os métodos de aprendizado de máquina.

## *5. Sumarização mono e multidocumento*

Há tradição no NILC em trabalhos de sumarização automática, principalmente em sumarização monodocumento. Há trabalhos tanto da abordagem profunda baseados em teorias discursivas como baseados em aprendizado de máquina e métodos empíricos. Pelo que se sabe, o NILC é o único grupo de pesquisa no Brasil que desenvolve pesquisas nesse assunto.

O primeiro trabalho de sumarização foi teórico e com base em conhecimento discursivo, realizado por Rino (1996) e validado na forma de um gerador automático de sumários por Pardo e Rino (2002). Estes trabalhos foram baseados na combinação de 3 modelos discursivos: RST (Mann e Thompson, 1987), o modelo intencional de Grosz e Sidner (1986) e o modelo Problema-Solução (Jordan, 1980). Combinando-se o conhecimento fornecido por esses três modelos, gerava-se o sumário de textos científicos. Essa abordagem produziu bons resultados, apesar de ser altamente custosa devido à demanda por conhecimento muito especializado.

Outros trabalhos baseados somente em RST seguiram os trabalhos anteriores. O melhor representante desta linha talvez seja o trabalho de Uzêda *et al*. (2008), onde se analisaram diversos métodos de sumarização com base na RST e se demonstrou que todos eles têm desempenho comparável. Mostrou-se também que os métodos baseados em RST são melhores do que métodos superficiais clássicos.

Ainda nesta linha, Seno e Rino (2005) e Carbonel *et al*. (2007) investigaram o uso da Teoria das Veias (Cristea *et al*., 1998) para lidar com correferências em sumários, já que a ocorrência de anáforas não resolvidas em sumários provoca sérios problemas de coesão e coerência. A Teoria das Veias é um modelo que permite que se identifiquem os segmentos textuais possíveis em que antecedentes de anáforas ocorram, o que possibilitaria a inclusão destes segmentos no sumário, resolvendo a anáfora e melhorando sua qualidade, portanto. Esse modelo, entretanto, trabalha sobre estruturas RST, demandando novamente conhecimento especializado. Como a Teoria das Veias indica várias possibilidades para a ocorrência do antecedente de uma anáfora, Tomazela e Rino (2009) investigaram como informação semântica superficial (de nível lexical) pode ajudar neste processo. Sua hipótese principal foi que o antecedente deve apresentar os mesmos traços semânticos da anáfora, o que permitira descartar algumas possibilidades de segmentos fornecidas pela Teoria das Veias.

Em outra linha, mas ainda na abordagem profunda, Martins e Rino (2002) usaram uma interlíngua para representar o conteúdo textual e manualmente desenvolveram regras para produzir suas versões comprimidas. A interlíngua utilizada foi a UNL, já citada na seção anterior.

É importante notar que muitos dos sistemas citados anteriormente se baseiam na RST. Diante desta demanda, foi produzido para o PB um analisador automático chamado DiZer (Pardo e Nunes, 2008). Esse analisador, de natureza simbólica (com regras de análise produzidas manualmente a partir de estudo de *corpus*) produz as estruturas RST possíveis para um texto-fonte de entrada. Como esse analisador foi desenvolvido para textos científicos em português e era de difícil adaptação para outros tipos textuais e línguas, novos trabalhos foram iniciados e se desenvolveu o DiZer 2.0[28], que está *online* e consiste em uma solução web de mais fácil portabilidade para outras línguas e tipos textuais. Esta versão do analisador permite que um usuário de forma relativamente simples adicione os recursos necessários e personalize seu próprio analisador.

Na abordagem superficial, Pardo *et al*. (2003a) desenvolveram um sistema de sumarização baseado principalmente em freqüência de palavras. Esse sistema é provavelmente um dos sistemas superficiais mais usados no Brasil e, apesar dos sumários gerados apresentarem diversos problemas de coesão e coerência, seus resultados são interessantes. Trabalhando sobre esses resultados, Gonçalves *et al*. (2008) usaram regras de pós-processamento para resolver anáforas e demonstraram que muitos dos problemas anteriores eram resolvidos.

Pardo *et al*. (2003b) usaram uma rede neural de Kohonen e atributos superficiais para modelar o processo de sumarização. O princípio deste trabalho consistia em agrupar sentenças de igual importância por meio da rede treinada, de forma que fosse possível descartar sentenças menos importantes para a produção do sumário.

[28] http://www.nilc.icmc.usp.br/dizer2

Leite *et al*. (2007, 2008) usaram, para seu sumarizador, um método de aprendizado de máquina bayesiano para combinar atributos superficiais simples e complexos, produzindo os melhores resultados até o momento para a língua portuguesa. Um dos pontos interessantes deste trabalho é que seus atributos complexos codificam métodos completos de sumarização automática, atribuindo, desta forma, grande informatividade ao processo como um todo.

Antiqueira *et al*. (2009) modelaram textos como redes complexas e usaram métricas das redes para selecionar informação relevante para compor o sumário, produzindo resultados muito bons. Sua modelagem de texto como rede é muito simples e elegante, demonstrando que não é necessária grande quantidade de conhecimento linguístico para se gerar bons sumários.

Trabalhos mais antigos do grupo incluem as propostas de Souza e Nunes (2001) e Pereira *et al*. (2002), as quais também usaram atributos textuais superficiais para sumarização. Outra questão relacionada investigada foi a compressão sentencial, ou seja, a tarefa de se produzir uma versão mais curta de uma sentença (Kawamoto e Pardo, 2010), utilizando-se aprendizado de máquina. Tal abordagem investigou o aprendizado automático de regras simbólicas para detecção de palavras de uma sentença que poderiam ser excluídas, observando-se critérios de gramaticalidade, informatividade e foco textual.

Recentemente, sistemas de sumarização multidocumento começaram a ser investigados no Brasil. O primeiro sistema foi proposto por Pardo (2005) e era trivial: o sistema simplesmente justapõe todos os textos e aplica métodos de seleção de sentenças com base na freqüência das palavras. Desde 2009, um grande projeto de sumarização multidocumento da abordagem profunda foi iniciado. Com base no modelo CST (*Cross-document Structure Theory*) (Radev, 2000), diversas estratégias de sumarização estão sendo investigadas, com alguns resultados promissores já produzidos. A CST, inspirada na RST, modela o relacionamento entre diversos textos sobre um mesmo assunto, permitindo que se lide adequadamente com os fenômenos multidocumento, como a presença de informação redundante, contraditória e complementar, a ordenação das informações textuais no sumário, e a própria questão de coerência e coesão.

Os primeiros trabalhos nesta linha de sumarização multidocumento (Jorge e Pardo, 2009, 2010) relacionaram preferências de sumarização do usuário com os relacionamentos previstos na CST, produzindo operadores de sumarização que, quando aplicados ao conteúdo textual, produzem um ranque de informações a partir do qual se devem selecionar as que serão incluídas no sumário.

Novamente, devido à demanda por análise CST, investiga-se atualmente a questão da análise automática multidocumento segundo este modelo. Os primeiros resultados obtidos (usando aprendizado de máquina e atributos superficiais) são promissores e avançam significativamente o estado da arte (Maziero *et al*., 2010).

Durante a investigação da sumarização automática no NILC, diversos recursos e ferramentas dedicados ao assunto foram produzidos. Dentre os *corpora*, os de mais destaque são o TeMário (Pardo e Rino, 2003; Maziero *et al*., 2007), o CSTNews (Aleixo e Pardo, 2008), o Summ-it (Collovini *et al*., 2007) e o Rhetalho (Pardo e Seno, 2005). Em termos de ferramentas, valem citar a RST Toolkit e a CSTTool, que são ferramentas de suporte à análise RST e CST, respectivamente.

## *6. Conclusões*

Nos últimos 17 anos, o NILC tem se dedicado à pesquisa e ao desenvolvimento de recursos e sistemas de PLN, especialmente para o PB escrito. Ao contrário do cenário inicial, hoje já é possível desenvolver pesquisa em qualquer área de PLN para o português em condições competitivas com outras línguas. Recursos básicos como léxicos, *corpora*, *parsers* e modelos de língua estão ao alcance dos pesquisadores, e o NILC se orgulha de ter contribuído significativamente para isto. Os desafios, no entanto, continuam grandes. É necessário fazer crescer a comunidade de PLN no país, que atualmente encontra dificuldades decorrentes do modelo de educação superior formal. Um linguista encontra barreiras para complementar sua formação em Computação, da mesma forma que um cientista da computação as encontra para complementar a sua em Linguística. Essa formação híbrida tem acontecido de maneira quase *ad hoc*, o que impede uma formação continuada. Para alterar o modelo, no entanto, é preciso fortalecer a área, inicialmente dentro dos limites de ambas as comunidades, e posteriormente além deles. Esse fortalecimento decorre de pesquisas de boa qualidade e reconhecidas internacionalmente, bem como de uma comunidade local unida e com objetivos comuns. Nesse sentido, ações como a organização dessa comunidade em comissões especiais (como a Comissão Especial de PLN na

Sociedade Brasileira de Computação[29]), e a promoção de eventos científicos para atrair novos pesquisadores (como a Escola Brasileira de Linguística Computacional[30]), a aproximação a sociedades internacionais, como a ACL e a NAACL, são muito relevantes.

Do ponto de vista das pesquisas do NILC, o momento atual é de consolidação de trabalhos iniciados há bastante tempo, como os de Ferramentas de Auxílio à Escrita e à Leitura e os de Sumarização Automática.

Na linha de trabalhos sobre ferramentas de suporte à escrita científica, o foco de pesquisa para os próximos anos será estender a rubrica baseada em gênero científico para uso em outras seções além do resumo. Quando totalmente automatizada, esta rubrica possibilitará que uma ferramenta de suporte à escrita detecte erros e ofereça sugestões para melhorias.

Quanto aos trabalhos dentro do escopo do projeto PorSimples, trabalhos futuros focarão na avaliação das ferramentas com usuários reais. Também pretendemos melhorar o desempenho da nossa abordagem de simplificação sintática via experimento com *parsers* sintáticos de abordagens diferentes do atual utilizado no projeto.

Sobre os trabalhos de sumarização automática, é interessante notar sua evolução natural. No início, o grupo investia pesadamente em abordagens profundas, necessitando de ferramentas de análise sofisticadas. Atualmente, tais ferramentas já existem (mesmo que ainda longe de produzirem dados ideais) e a transição entre as investigações monodocumento para multidocumento foi iniciada. Na linha superficial, resultados do estado da arte foram atingidos, incentivando a continuidade das investigações nesta direção.

Os trabalhos em tradução automática têm se concentrado cada vez mais na linha estatística, mas não abandonando o uso de conhecimento linguístico. Investigações recentes procuram saber como o conhecimento sintático-semântico pode auxiliar nesse processo. Acredita-se que, como na maior parte das aplicações de PLN, a combinação das abordagens pode produzir resultados melhores.

[29] http://www.nilc.icmc.usp.br/cepln/

[30] http://www.corpuslg.org/ebralc/Inicial.html

## *Agradecimentos*

Agradecemos a todos os colaboradores do NILC, desde sua criação, que têm tornado possível o desenvolvimento de todos os trabalhos - entre muitos outros - descritos neste artigo. Agradecemos também o apoio das agências brasileiras de pesquisa – CNPq, FAPESP, CAPES e FINEP –, da UNU/IAS e da Itautec S.A.

## *Referências*

Aleixo, P. e Pardo, T.A.S. 2008. *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, Maio, 12p.

Aluísio, S. M.; Pelizzoni, J. M.; Marchi, A. R.; Oliveira, L. H.; Manenti, R.; Marquivafável, V. 2003. An account of the challenge of tagging a reference *corpus* of Brazilian Portuguese In: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003, Proceedings. *Lecture Notes in Computer Science 2721* Springer 2003

Aluísio, S. M., Barcelos, I., Sampaio, J., Oliveira Jr, O. N. 2001. How to Learn the Many Unwritten ´Rules of the Game´ of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases to Support Scientific Writing In: *IEEE International Conference on Advanced Learning Technologies,* Madison, Wisconsin. 2001. v.1. p.257 – 260.

Aluísio, S. M., Fontana, N., Oliveira JR., O. N., Oliveira, M. C. F. 1993. Computer Assisted Writing - Applications to English as a Foreign Language. *Computer Assisted Language Learning Journal. v*.6, p.145 - 161, 1993.

Aluísio, S. M., Gantenbein, R. E. 1997. Towards the Application of Systemic Functional Linguistics in Writing Tools In*: Proceedings of International Conference on Computers and their Applications,* 1997. v.1. p.181 - 185

Aluísio, S. M., Oliveira JR, O. N. 1995. A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users In: *Proceedings of the First International Conference - ICCBR-95. Lecture Notes in Artificial Intelligence.* Berlin: Springer-Verlag, v.1010. p. 121 – 132

Aluísio, S. M., Oliveira JR., O. N. 1996. Detailed Schematic Structure of Research Papers Introductions: An Application in Support-Writing Tools. *Revista de La Sociedad Espanyola Para El*

*Procesamiento Del Lenguage Natura.* v.1, p.141 – 147.

Aluísio; S. M.; Schuster; E.; Feltrim; V.D.; Pessoa Jr; A.; Oliveira JR, O. N. 2005. Evaluating scientific abstracts with a genre-specific rubric. In: *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005).* Amsterdam: v.1, p. 738-740.

Aluísio, S. M., Specia, L., Pardo, T.A.S., Maziero, E. G. *and* Fortes, R. 2008. Towards Brazilian Portuguese Automatic Text Simplification Systems. In the *Proceedings of the 8th ACM Symposium on Document Engineering*, pp. 240-248.

Aluisio, S., Specia, L., Gasperin, C. *and* Scarton, C. 2010. Readability Assessment for Text Simplification. To be published in the Proceedings of the The *5th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2010.*

Aluísio, S.M. *and* Gasperin, C. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. To be published in the *Proceedings of The Young Investigators in the Americas Workshop, NAACL-HLT 2010.*

Amancio, D.R.; Antiqueira, L.; Pardo, T.A.S.; Costa, L.F.; Oliveira Jr. O.N.; Nunes, M.G.V. 2008. Complex networks analysis of manual *and* machine translations. *International Journal of Modern Physics C - IJMPC*, V. 19, N. 4, pp. 583-598.

Anthony, L., Lashkia, G.V. 2003. Mover: A machine learning tool to assist in the reading *and* writing of technical papers. *IEEE Transactions on Professional Communication* 46 (2003) 185-193

Antiqueira, L ; Oliveira Jr, Osvaldo N ; Costa, Luciano F ; Nunes, M. G. V. 2009. A complex network approach to text summarization. *Information Sciences,* v. 179, p. 584-599.

Antiqueira, L.; Pardo, T. A. S. ; Nunes, M. G. V.; Oliveira Jr., O. N. 2007. Some issues on complex networks for author characterization. *Inteligencia Artificial*, v. 11, p. 51-58.

Antiqueira L.; Fossey, M.F.; Pedrolongo, T.; Greghi, J.G.; Martins, R.T.; Nunes, M.G.V. 2002. *A construção do corpus e dos dicionários Inglês-UNL e UNL-português para o projeto EPT-Web* - Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - NILC-TR-02-24.

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. 2009a. Statistical Phrase-based Machine Translation: Experiments with Brazilian Portuguese. In *Anais do VII Encontro Nacional de Inteligência Artificial* - ENIA, pp. 769-778. July 20-24, Bento Gonçalves/RS, Brazil.

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. 2009b. Fine-tuning in Portuguese-English Statistical Machine Translation. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-4. September 8-10, São Carlos/SP, Brazil.

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. 2008. An Experiment in Spanish-Portuguese Statistical Machine Translation. In the Proceedings of the 19th Brazilian Symposium on Artificial Intelligence - SBIA *(Lecture Notes in Computer Science 5249*), pp. 248-257. Salvador-BA, Brazil. October, 26-30.

Brown, P.E.; Pietra, S.A.D.; Pietra, V.J.D.; Mercer, R.L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 16, N. 2, pp. 79-85.

Bruckschen, M.; Muniz, F.; Souza, J. G. C.; Fuchs, J. T.; Infante, K.; Muniz, M.; Gonçalves, P. N.; Vieira, R.; Aluísio, S. M. 2008. *Anotação Linguística em XML do Corpus PLN-BR*. Série de Relatórios do NILC (NILC-TR-09-08). São Carlos - SP, Junho 2008, 39 p.

Burstein, J.; Marcu, D.; Knight, K. 2003. Finding the WRITE Stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1):32–39.

Candido Jr., A., Aluísio, S. M. 2009. Building a *Corpus*-based Historical Portuguese Dictionary: Challenges *and* Opportunities. *Traitement Automatique des Langues (TAL),* [S.l.], v.50, p.73 – 102. ISSN: 1965-0906

Carbonel, T.I.; Pelizzoni, J.; Rino, L.H.M. 2007. Validação Preliminar da Teoria das Veias para o Português e Lições Aprendidas. In the *Proceedings of the V Workshop on Information and Human Language Technology.* Rio de Janeiro-RJ.

Caseli, H.M. *and* Nunes, I.A. 2009. Statistical Machine Translation: little changes big impacts. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL. September 8-10, São Carlos/SP, Brazil.

Caseli, H.M.; Ramisch C.; Nunes, M.G.V.; Villavicencio, A. 2009. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, p. 1-19.

Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. 2008. Automatic induction of bilingual resources from

aligned parallel *corpora*: application to shallow-transfer machine translation. *Machine Translation.* v. 1, p. 227-245.

Caseli, H.M. 2007. *Indução de léxicos bilíngues e regras para a tradução automática*. Tese de Doutorado. ICMC-USP, Abril, 2007. 158 p.

Collovini, S.; Carbonel, T.I.; Fuchs, J.T.; Coelho, J.C.B.; Rino, L.H.M.; Vieira, R. 2007. Summ-it: Um *corpus* anotado com informações discursivas visando à sumarização automática. In the *Proceedings of the V Workshop on Information and Human Language Technology*. Rio de Janeiro/RJ.

Costa, L. F.; Oliveira Jr., O. N.; Travieso, G.; Rodrigues, F. A.; Villas Boas, P. R.; Antiqueira, L.; Viana, M. P.; Rocha, L. E. C. 2008. Analyzing *and* Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. *Physics and Society*.

Cristea, D.; Ide, N.; Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion *and* Coherence. In the *Proceedings of the Coling-ACL*, pp. 281-285. Montreal, Canadá.

Di Felippo, A. and Dias-da-Silva, B.C. 2007. Towards na automatic strategy for acquiring the WordNet.Br hierarchical relations. In *Proceedings of the 5th Workshop in Information and Human Language Technology.* Rio de Janeiro, Brasil.

Dias-da-Silva, B.C.; Di Felippo, A. and Nunes, M.G.V. 2008. The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco.

Fellbaum, C. 1998. *WordNet: an electronic lexical database*. Ca., MA: MIT Press.

Feltrim, V., Aluisio, S.M., Nunes, M.G.V.2003. Analysis of the rhetorical structure of computer science abstracts in portuguese. In Archer, D., Rayson, P., Wilson, A., McEnery, T., eds.: *Proceedings of Corpus Linguistics 2003*, *UCREL Technical Papers,* Vol. 16, Part 1, Special Issue. (2003) 212-218

Feltrim, V. D. 2004. *Uma Abordagem baseada em Corpus e em Sistemas de Crítica para a onstrução de Ambientes Web de Auxílio à Escrita Acadêmica em Português*. Tese de Doutorado. ICMC – USP, São Carlos, 181p.

Feltrim, V. D., Pelizzoni, J. M., Teufel, S., Nunes, M. G. V., Aluisio, S.M. 2004. Applying Argumentative Zoning in an automatic critiquer of academic writing. In Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004), *Lecture Notes in Artificial Intelligence*, 3171, Springer, p. 214-223.

Feltrim, V., Teufel, S., Nunes, M.G.V., Aluísio, S. M. 2006. Argumentative Zoning Applied to Critiquing Novices'ScientificAbstracts In: *Computing Attitude and Affect in Text: Theory and Applications*. Ed. Dordrecht, The Netherlands : Springer, 2006 v.1, p. 159-170.

Ferraz Jr, C.C.P., Boas, E.V.B., Dornelas, J., Amancio, M.A., Raymundo, E., Aluísio, S. M., Feltrim, Valéria D. 2007. PlaNInt!: Uma ferramenta Web inteligente de auxílio à escrita de planos de negócios em português. *Locus Científico.* v.1, p.48 - 57.

Fontana, N.; Aluísio, S.M.; Oliveira, M.C.F.; Oliveira JR., O.N. 1993. Computer assisted writing - applications to English as a foreign language. 145-161. *CALL (Computer Assisted Language Learning Journal),* 6, 145-161.

Genoves JR, Luiz Carlos, Feltrim, Valéria D., Dayrell, C., Aluísio, S. M. 2007a. Automatically detecting schematic structure components of English abstracts: building a high accuracy classifier for the task. In: *International Workshop on Natural Language Processing for Educational Resources in conjunction with the International Conference RANLP´2007*, 2007, Borovets, v.1. p.23 – 29.

Genoves JR, L.C., Lizotte, R., Schuster, E., Dayrell, C., Aluísio, S. M. 2007b. A two-tiered approach to detecting English article usage: an application in scientific paper writing tools In: *Proceedings of the RANLP-2007,* Sofia: Bulgarian Academy of Sciences, 2007. v.1. p.225 – 229.

Gonçalves, P.N.; Vieira, R.; Rino, L.H.M. 2008. CorrefSum: Referencial Cohesion Recovery in Extractive Summaries. *Lecture Notes in Artificial Intelligence* (Proc. of the 8th International Conference on Computational Processing of Portuguese Language, Propor2008). Berlin : Springer, 2008. v. 5190. p. 224-227.

Grosz, B. *and* Sidner, C. 1986. Attention, Intentions, *and* the Structure of Discourse. *Computational Linguistics*, Vol. 12, No. 3.

INAF 2009. Instituto P. Montenegro e Ação Educativa. INAF Brasil - Indicador de Alfabetismo Funcional - 2009. Disponível em: ibope.com.br/ipm/relatorios/relatorio_inaf_2009.pdf

Jesus, M.A.C.; Nunes, M.G.V. 2000. Autômatos Finitos e Representação de Grandes Léxicos:

Aplicação a um Léxico de Português Brasileiro. In *Anais do V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000),* v.1, p.29-42.

Jordan, M.P. 1980. Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science*, Vol. 9, pp. 221-252.

Jorge, M.L.C. and Pardo, T.A.S. 2009. Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-8. September 8-10, São Carlos/SP, Brazil.

Jorge, M.L.C. *and* Pardo, T.A.S. 2010. Formalizing CST-based Content Selection Operations. In the *Proceedings of the International Conference on Computational Processing of Portuguese Language* - PROPOR. April, 27-30, Porto Alegre/RS, Brazil.

Junior, A. C., Maziero, E., Gasperin, C., Pardo, T., Specia, L.*;* Aluisio, S. M. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In the *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado.

Kawamoto, D. and Pardo, T.A.S. 2010. Learning Sentence Reduction Rules for Brazilian Portuguese. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* – NLPCS. Funchal, Madeira, Portugal.

Koehn, P.; Och, F.J.; Marcu, D. 2003. Statistical phrase-based translation. In the *Proceedings of the HLT-NAACL*, pp. 48-54.

Leite, D.S.; Rino, L.H.M.; Pardo, T.A.S.; Nunes, M.G.V. 2007. Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In C. Biemann, I. Matveeva, R. Mihalcea, *and* D. Radev (eds.), *Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pp.17-24. Rochester, NY, USA.

Leite, D.S. and Rino, L.H.M. 2008. Combining Multiple Features for Automatic Text Summarization through Machine Learning. In *Lecture Notes in Artificial Intelligence* (Proc. of the 8th International Conference on Computational Processing of Portuguese Language, Propor2008), 2008. v. 5190. p. 122-132.

Mann, W.C. *and* Thompson, S.A. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Martins, C.B. and Rino, L.H.M. 2002. Revisiting UNLSumm: Improvement through a case study. In the *Proceedings of the Workshop on Multilingual Information Access and Natural Language Processing*, Vol. 1. pp. 71-79. Sevilha, Espanha.

Martins, R.T.; Pelizzoni, J.M.; Hasegawa, R.; Nunes, M. G. V. 2004a. Da tradução automática para a língua portuguesa: apontamentos de três experiências baseadas em interlíngua. *Palavra (PUCRJ),* Rio de Janeiro, v. 12, n. 1, p. 37-55.

Martins, R.T., Hasegawa, R., Nunes, M. G. V. 2004b. HERMETO: A Natural Language Analysis Environment In: TIL- Workshop em Tecnologia da Informação e da Linguagem Humana, 2004, Salvador. *Anais do SBC 2004*.

Martins, R. T.; Hasegawa, R.; Nunes, M.G.V. 2003. Curupira: a functional *parser* for Brazilian Portuguese. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, Maria das Graças Volpe Nunes (Eds.): Proceedings of the Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. *Lecture Notes in Computer Science 2721* Springer 2003, ISBN 3-540-40436-8.

Maziero, E.G.; Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. 2007. *TeMário 2006: Estendendo o Córpus TeMário*. Série de Relatórios do NILC. NILC-TR-07-06. São Carlos-SP, Agosto, 8p.

Maziero E.G.; Jorge, M.L.C.; Pardo, T.A.S. 2010. Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* – NLPCS. Funchal, Madeira, Portugal.

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.

Muniz, M.; Paulovich, F. V.; Minghim, R.; Infante, K.; Muniz, F.; Vieira, R.; Aluísio, S. 2007. Taming the tiger topic: an XCES compliant *corpus* Portal to generate sub*corpus* based on automatic text topic identification. In: *Proceedings of the Corpus Linguistics 2007 Conference*.

Nunes, M.G.V., Pelizzoni, J. M., Greghi, J. G., Hasegawa, R., Martins, R. T. 2003. *Projeto PULO*. NILC Project Report, Jun. 2003

Nunes, I.A. e Caseli, H.M. 2009. Primeiros Experimentos na Investigação e Avaliação da Tradução Automática Estatística Inglês-Português. Em *Anais do Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana* – TILic. São Carlos, Brasil.

Pardo, T.A.S. and Rino, L.H.M. 2002. DMSumm: Review *and* Assessment. In E. Ranchhod and N. J. Mamede (eds.), 3rd International Conference: Portugal for Natural Language Processing – PorTAL (*Lecture Notes in Artificial Intelligence 2389*), pp. 263-273. Faro, Portugal. June 23-26.

Pardo, T.A.S. e Rino, L.H.M. 2003. *TeMário: Um Corpus para Sumarização Automática de Textos*. Série de Relatórios do NILC. NILC-TR-03-09. São Carlos-SP, Outubro, 13p.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. 2003a. GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*. Faro, Portugal.

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. 2003b. NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. In *Anais do IV Encontro Nacional de Inteligência Artificial* – ENIA, pp. 1-10. Campinas-SP, Brazil.

Pardo, T.A.S. 2005. *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP, Fevereiro, 8p.

Pardo, T.A.S. e Seno, E.R.M. 2005. Rhetalho: um *corpus* de referência anotado retoricamente. In *Anais do V Encontro de Corpora*. São Carlos-SP, Brasil. 25 a 26 de Novembro.

Pardo, T.A.S. *and* Nunes, M.G.V. 2008. On the Development *and* Evaluation of a Brazilian Portuguese Discourse *Parser*. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43-64.

Pereira, M.B.; Souza, C.F.R.; Nunes, M.G.V. 2002. Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. *Revista Eletrônica de Iniciação Científica*. Ano II, Vol. 2, N. 1.

Radev, D.R. 2000. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Raymundo, E., Amancio, M.A., Feltrim, Valéria D., Aluísio, S. M. 2007. Análise da Estrutura Retórica da Seção Sumário Executivo de Plano de Negócios In: *Anais do VI Encontro de Linguística de Corpus*, p.1 – 18.

Rino, L.H.M. 1996. *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-USP. São Carlos - SP.

Seno, E.R.M. *and* Rino, L.H.M. 2005. Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In the *Proceedings of the RANLP 2005 Workshop on Crossing Barriers in Text Summarization Research*, pp. 70-75.

Souza, C.F.R. and Nunes, M.G.V. 2001. *Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português*. Technnical Report NILC-TR-01-09.

Specia, L.; Nunes, M.G.V.; Stevenson, M. 2009a. Assessing the contribution of shallow *and* deep knowledge sources for word sense disambiguation. *Language Resources and Evaluation*, Springer. DOI 10.1007/s10579-009-9107-y.

Specia, L.; Srinivasan, A.; Ramakrishnan, G.; Joshi, S.; Nunes, M.G.V 2009b. An Investigation into Feature Construction to Assist Word Sense Disambiguation. *Machine Learning,* 76(1):109-136, Springer.

Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge applied linguistics series.

Teufel, S; Moens, M. 2002: Summarizing Scientific Articles -- Experiments with Relevance and Rhetorical Status. In *Computational Linguistics, 28* (4), Dec. 2002.

Tomazela, E.K. e Rino, L.H.M. 2009. O uso de informações semânticas para tratar a informatividade de sumários automáticos com foco na clareza referencial. Em *Anais do VII Encontro Nacional de Inteligência Artificial*, pp. 799-808. Bento Gonçalves/RS, Brasil.

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. 2008. Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory. In the *IEEE Proceedings of the 8th International Conference on Intelligent Systems Design and Applications* - ISDA, pp. 389-394. Taiwan. November, 26-28.

Watanabe, W. Candido Jr. A., Uzêda, V. Fortes, R., Pardo, T. and Aluísio, S. 2009. Facilita: reading assistance for low-literacy readers. In: *Proceedings of the 27th ACM International Conference on*

*Design of Communication. SIGDOC '09*. ACM, New York, NY, 29-36.

Watanabe, W. M.; Candido Jr. A.; Amancio, M. A.; Oliveira, M.; Pardo, T. A. S.; Fortes, R. P. M.; Aluísio, S. M. 2010. Adapting web content for low-literacy readers by using lexical elaboration *and* named entities labeling. *Accepted for publication at W4A* 2010 (http://www.w4a.info/).

Weissberg, R.; Buker, S. 1990. *Writing up Research: Experimental Research Report Writing for Students of English.* Prentice Hall.

# Artigos de Investigação

# Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português.

Paulo Malvar Fernández
Area of Language Technology, **imaxin**|software
paulomal@gmail.com

José Ramom Pichel Campos
Area of Language Technology, **imaxin**|software
jrampichel@imaxin.com

Óscar Senra Gómez
Area of Language Technology, **imaxin**|software
oscar@imaxin.com

Pablo Gamallo Otero
Universidade de Santiago de Compostela
pablogam@usc.es

Alberto García
Igalia Free Software Company
agarcia@igalia.com

**Resumo**

À hora de desenvolver muitas ferramentas estatísticas de Processamento da Linguagem Natural torna-se essencial a utilização de grandes quantidades de dados. Para salvar a limitação da escassez de recursos computacionais para línguas minorizadas como o galego é necessário desenhar novas estratégias. No caso do galego, importantes romanistas têm teorizado que galego e português são variantes do português europeu. De um ponto de vista pragmático, esta hipótese poderia abrir uma nova linha de investigação para fornecer ao galego ricos recursos computacionais. Partindo do corpus paralelo inglês–português Europarl, **imaxin**|software compilou um corpus paralelo inglês–galego que utilizamos para criar um protótipo de tradutor automático estatístico inglês–galego, cuja performance é comparável a Google Translate. Mantemos que é possível implementar esta estratégia para desenvolver uma grande variedade de ferramentas computacionais para línguas, como o galego, intimamente relacionadas com línguas que já contam com um grande repertório de recursos computacionais.

## *1 Prefácio*

Do ponto de vista da teoria lingüística sistémico-funcional hallidiana, as línguas funcionam, de acordo com Gee (1999, 1) "*tanto como uma ferramenta para a acção quanto como um andaime para as relações humanas dentro das culturas e grupos sociais e instituições*"[1]. Noutras palavras, a linguagem funciona como uma ferramenta não só para a comunicação mas para negociar as relações e as estruturas sociais da própria sociedade. É precisamente, mercê a esta dimensão social que a linguagem joga um papel simbólico crucial. Ao desenvolverem ferramentas computacionais para línguas concretas, os linguistas computacionais, sejam principalmente informáticos ou linguistas, são responsáveis para com as línguas com que trabalham. É possível que no caso de línguas prestigiadas esta responsabilidade não pareça óbvia. Nestes casos, as decisões a respeito de que fenómenos linguísticos se estudam e (mais importante do ponto de vista deste artigo) que ferramentas se desenvolvem; podem parecer triviais, pois semelham não implicar nenhum posicionamento ideológico. Porém, aqueles cientistas que decidiram trabalhar com e para línguas minorizadas, especialmente se são falantes dessas línguas, as suas decisões não são nunca inócuas.

É com esta responsabilidade como investigadores linguísticos e falantes que foi levado a cabo o projecto sobre o qual se debruça este artigo.

## *2 Introdução*

Em 2008 e 2009, em **imaxin**|software levamos a cabo um projecto, subsidiado pola Dirección Xeral de I+D+i da Xunta de Galicia, chamado "RecursOpentrad: Recursos lingüístico-

[1] Tradução dos autores

computacionais para a tradución automática avanzada de código aberto para a integración europea da lingua galega". Dentro deste projecto, além de construirmos um sistema inglês–galego de Tradução Automática (TA) baseada em regras , pensamos que, dados os progressos[2] na actualidade atingidos no campo da Tradução Automática Estatística (TAE), era um excelente momento para dar mais um passo no desenvolvimento de ferramentas de Processamento da Linguagem Natural (PLN) para o galego.

Quando decidimos desenvolver um protótipo de um sistema de TAE inglês–galego, sabíamos que "*quanto maior [fosse] o corpus de treino disponível, mellor [seria] o desempenho [do] sistema de tradução*"[3] (Popović e Ney, 2006, 25) que poderíamos conseguir. Contudo, enquanto compilávamos os recursos necessários para o desenvolvimento de um protótipo para o citado par de línguas, chegamos à seguinte conclusão, absolutamente coincidente com uma das afirmações com que Popović e Ney (2006) começam a sua comunicação em Language Resources and Evaluation (LREC) em 2006:

"*Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality parallel text for the desired domain and language pair requires a lot of time and effort, and for some languages is not even possible.*"(Popović e Ney, 2006, 25)

Convém termos em conta que não é impossível encontrar corpora paralelos inglês–galego na Internet.[4] De facto, o grupo de investigação de Xavier Gómez Guinovart na Faculdade de Tradução e Interpretação da Universidade de Vigo dispõe de uma colecção de corpora parlelos[5] dentro da qual o par inglês–galego está representado com um subcorpus de aproximadamente 9 milhões de palavras. Um corpus deste tamanho, porém, é a todos os efeitos insuficiente para o propósito de construír um sistema de TAE.

Chegados a este ponto, tornava-se, na nossa opinião, necessário tomar um rumo diferente para conseguirmos o nosso objectivo. É, neste sentido, conhecido na comunidade linguística que importantes romanistas, como por exemplo Coseriu (1987), Cunha e Cintra (2002) e Aracil (1985), têm teorizado que, de um ponto de vista linguístico, o galego deve ser considerado uma variante do português junto com o português europeu, brasileiro, africano e asiático. Isto é exactamente o que Coseriu (1987) e Rei (1991) apontam:

"*los romanistas e hispanistas están en general de acuerdo en que el gallego es una forma particular del conjunto dialectal gallego-portugués, en cuanto opuesto al conjunto dialectal español (no "castellano", sino: astur-leonés, castellano, en sus muchas formas, y navarro-aragonés) y al conjunto catalán (o catalán-valenciano)*" (Coseriu, 1987, 795)

"*Na actualidade, desde o punto de vista estrictamente lingüísitico, ás dúas marxes do Miño fálase o mesmo idioma, pois os dialectos miñotos e transmontanos son unha continuación dos falares galegos, cos que comparten trazos comúns que os diferencian dos do centro e sur de Portugal; pero no plano da lingua común, e desde unha perspectiva sociolingüísitca, hai no actual occidente peninsular dúas linguas modernas, con diferencias fonéticas, morfosintácticas e léxicas, que poden non impedi-la intercomprensión ó existir un bilingüismo inherente entre o galego e o portugués, semellante ó existente entre o catalán e o occitano, o danés e o noruegués, o eslovaco e o checo, o feroés e o islandés.*" (Rei, 1991, 17–18)

Deste modo, partindo da suposição de que galego e português são variantes linguísticas intimamente relacionadas e tentando aproveitar a posição privilegiada do português como língua computacionalmente desenvolvida –isto é, uma língua para a qual muitas ferramentas e recursos de PLN foram desenvolvidos–, em **imaxin**|software investigámos a possibilidade de utilizar corpora paralelos inglês–português de livre acesso para criar um corpus paralelo inglês–galego que utilizaríamos para desenvolver um protótipo de tradutor automático estatístico inglês–galego.

## *3 Compilação e processamento do corpus*

### 3.1 O corpus de origem

Já que o nosso projecto estava claramente guiado pola filosofia do movimento do *Open Source*,

[2]Sirva de exemplo a grande popularidade da Tradução Automática Estatística (TAE) de alta qualidade atingida com a implementação feita por Google do seu sistema de TAE, Google Translate (disponível para consulta on-line em `http://translate.google.com/`).

[3]Tradução dos autores

[4]Graças à localização de projectos de ferramentas e sistemas operativos de código aberto levados a cabo pola comunidade galega de usuários de código aberto é possível compilar manualmente corpora paralelos inglês–galego do domínio da localização de software publicados baixo a General Public License (GPL). Contudo, estes corpora, ao serem traduzidos de maneira voluntária por grupos de pessoas não coordenados, não têm uniformidade e o seu tamanho resulta insuficiente para o propósito de crear um sistema de TAE.

[5]Esta colecção pode-se consultar em `http://sli.uvigo.es/CLUVI/`.

queríamos que tantos componentes do sistema como for possível fossem de código aberto, ou polo menos de livre acesso para uso não comercial.

Devido ao seu grande tamanho e liberal licença de *copyright*[6] escolhemos o corpus paralelo Europarl v3[7] inglês–português como corpus de origem do nosso projecto.

O corpus Europarl é um corpus paralelo extraído das Actas do Parlamento Europeu que inclui versões, desde 1996, do seu contido em onze línguas europeias: línguas romances (francês, italiano, espanhol e português), línguas xermánicas (inglês, neerlandês, alemão, danés e sueco), grego e finlandês.

Após um processo inicial de limpeza das etiquetas XML que marcam a estrutura discursiva das elocuções contidas no corpus, obtivemos um corpus paralelo inglês–português não-tokenizado que contém quase 65 milhões de palavras em total. Este corpus foi realinhado oração-a-oração[8] após o citado processo de limpeza empregando a ferramenta *sentence aligner*[9], incluída entre as ferramentas do Europarl v3.

### 3.2 Conversão de inglês–português a inglês–galego

A conversão do corpus paralelo de origem num corpus paralelo inglês–galego que desenhamos em **imaxin**|software é um processo semi-automatizado que envolveu o uso de duas peças de software principais: um sistema de tradução automática baseada em regras e um conversor ortográfico –isto é, um motor de transliteração.[10]

Deste modo, o fluxo de trabalho desenhado foi o seguinte:

- Tradução automática para galego do *lado* português do corpus paralelo de origem utilizando EixOpentrad.[11]
- Identificação dos erros de tradução devidos a erros de codificação de EixOpentrad. Quando em EixOpentrad existe uma regra de transferência ou uma entrada de dicionário mal formulada, o tradutor falha e marca a existência deste tipo de erros imprimindo os caracteres @ ou #, dependendo do tipo de erro, junto às palavras motivadoras dos erros.
- Revisão e correcção manual dos erros de tradução marcados com @ e #. As palavras marcadas com @ são palavras deficientemente codificadas no dicionário bilingue do tradutor. Os erros marcados com # corresponde-se, por sua vez, bem com erros de codificação nos dicionários monolingues, bem com erros de construção das regras de transferência do tradutor.
- Identificação das palavras desconhecidas, e portanto, não traduzidas por EixOpentrad. EixOpentrad marca as palavras não traduzidas com *, de modo que a sua identificação pode ser totalmente automatizada.
- Transliteração para galego das palavras desconhecidas, marcadas com *, utilizando um script de transliteração português–galego chamado port2gal.[12] As palavras que se transliteram no corpus são também armazenadas numa lista com a sua correspondente versão original não transliterada para a sua posterior revisão.
- Revisão e correcção manual dos erros de transliteração identificados na lista de palavras transliteradas obtidas no processamento anterior. Este processo de correcção, que não pode ser automatizado, é o passo que mais demora em se completar devido o tamanho limitado dos dicionários de EixOpentrad. É também, dada a sua extensão em número de palavras afectadas, um passo que convém realizar exaustivamente para assegurar a qualidade do corpus galego que se deseja obter.

[6] "The European Parliament web site states: "Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged."" (Koehn, 2005)(2)

[7] De livre acesso em `http://www.statmt.org/europarl/archives.html`

[8] Isto é, *sentence-to-sentece*.

[9] Esta ferramenta pode ser descarregada no site `http://www.statmt.org/europarl/v3/tools.tgz`

[10] Os conversores ortográficos utilizam-se normalmente para escrever o mesmo código de duas maneiras diferentes. Este tipo de conversores não fazem mais do que substituir padrões de sequeências de caracteres da língua de origem nos seus correspondentes padrões de sequeências de caracteres na língua de chegada. Esta estratégia não envolve informação morfológica, sintáctica nem semântica.

[11] EixOpenTrad é uma versão posterior de OpenTrad, uma plataforma de serviços de tradução automática (`http://www.opentrad.com`). EixOpenTrad é um protótipo de tradução automática galego-português e português–galego que contém 8.500 palavras em ambas as direcções. Este sistema está baseado no motor de tradução de Apertium espanhol–português, (Armentao-Oller et al., 2006).

[12] port2gal, que é um simples script de Perl, foi inicialmente desenvolvido por Alberto García (Igalia Free Software Company) e posteriormente melhorado por Pablo Gamallo (Departmento de Língua Espanhola da Universidade de Santiago de Compostela). Este script simplesmente converte a ortografia do portugês europeu para a ortografia actual do galego. port2gal está disponível baixo GPL em `http://gramatica.usc.es/~gamallo/port2gal.htm`.

Todo este processo conversão demorou três meses de trabalho de uma só pessoa a tempo completo (isto é, à volta de 3.600 horas) em total em se finalizar. Este é, sem dúvida, um período de tempo insignificante se comparado com o esforço em tempo e custos que suporia a compilação manual de um corpus inglês–galego deste tamanho.

### 3.3 O corpus final

Após finalizar o processo de conversão do corpus inglês–português obtivemos um corpus tokenizado inglês–galego composto de 34.715.016 tokens em inglês e 34.688.010 tokens em galego. Isto é, de aproximadamente 69 milhões de palavras, tamanho que é significativamente maior do que o tamanho do corpus citado na secção 2.

## *4 Tradução Automática Estatística*

É comummente aceite por investigadores e profissionais da tradução que o principal desafio de todo o processo de tradução de uma língua para outra é basicamente encontrar um equilibrio entre a fidelidade com o significado expressado na língua de origem e a fluidez do texto equivalente na língua de chegada. De acordo com Jurafsky e Martin (2008, 875), "*Statistical MT is the name for a class of approaches that do just this by building probabilistic models of faithfulness and fluency and then combining these models to choose the most probable translation*". Assim, a melhor tradução $\hat{T}$ de uma frase de origem concreta $S$ pode-se formalizar do seguinte modo:

$$\hat{T} = \text{argmax}_{\text{T}}\text{fidelidade}(T, S)\text{fluidez}(T) \quad (1)$$

Esta intuitiva definição informal da melhor tradução $\hat{T}$ pode ser matematicamente redefinida como a probabilidade condicional de uma possível tradução dada uma frase concreta da língua de origem:

$$\hat{T} = \text{argmax}_{\text{T}} P(T|S) \quad (2)$$

Utilizando a Regra de Bayes esta probabilidade condicional pode ser reescrita como:

$$\hat{T} = \text{argmax}_{\text{T}} \frac{P(S|T)P(T)}{P(S)} \quad (3)$$

Já que $P(S)$ não varia pois permanece constante para qualquer provável tradução $T$, $P(S)$ pode-se ignorar:

$$\hat{T} = \text{argmax}_{\text{T}} P(S|T)P(T) \quad (4)$$

Após a aplicação da Regra de Bayes podemos ver que, embora a nossa formalização intuitiva fazia a tradução $T$ condicional na frase de origem $S$, a equação 4 fai a $S$ de origem condicional na tradução $T$. Este modo inverso de formalizar problemas estatísticos, que é normal nos modelos conhecidos como *Noisy Channel*, tem a vantagem de que a a equação resultante pode ser perfeitamente paralelizada com a definição informal do problema de encontrar a melhor tradução $\hat{T}$:

$$P(S|T) = \text{fidelidade}(\text{T}, \text{S}) \quad (5)$$

$$P(T) = \text{fluidez}(\text{T}) \quad (6)$$

### 4.1 Alinhamentos Palavra-a-Palavra

Nos anos 90 o grupo de investigação de IBM em Yorktown Heights (NY) começou a publicar algoritmos, Brown et al. (1990) and Brown et al. (1993), que, com relativo sucesso, utilizavam uma derivação bayesiana do modelo do *Noisy Channel* para construir tradutores automáticos estatísticos. A aproximação de IBM começava por estabelecer alinhamentos palavra-a-palavra entre frases alinhadas num corpus paralelo. Os alinhamentos palavra-a-palavra simplesmente formalizam a ideia de que existe um mapeamento explícito, embora não perfeito, entre as palavras das frases de origem e de chegada dos corpora paralelos. Seguindo a mesma aproximação do modelo do *Noisy Channel*, os algoritmos de alinhamento palavra-a-palavra modelam a probabilidade condicional de uma frase de origem $S$ dada uma tradução $T$, alinhando palavra-a-palavra estas frases $S$ e $T$:

$$P(S|T) = \sum_{A} P(S, A|T) \quad (7)$$

Noutras palavras, para um par concreto de frases alinhadas, $S$ e $T$, a probabilidade condicional de $S$ dada $T$ encontra-se sumando todos os possíveis alinhamentos palavra-a-palavra $A$ entre $S$ e $T$.

Já que normalmente não há disponíveis corpora parlelos etiquetados à mão[13], é necessário utilizar um algoritmo para calcular as probabilidades de correspondências palavra-a-palavra utilizando a informação dada pola co-ocorrência de palavras num conjunto de frases paralelas. Para a realização desta tarefa normalmente utiliza-se o algoritmo conhecido como *Expectation Maximization* (EM).[14]

[13] De facto, seria muito caro em termos económicos e de recursos humanos etiquetar à mão as correspondências palavra-a-palavra em corpora paralelos do tamanho necessário para obter tradutores automátivos estatísticos de qualidade.

[14] Para uma explicação detalhada do funcionamento

### 4.2 TAE baseada em frases

Embora na TAE baseada em frases, em inglês *Phrase-based Statistical Machine Translation*, como qualquer outro sistema de TAE, a tradução se formalize com mesma equação 4 básica, os sistema de TAE baseada em frases são diferentes em termos daquilo que constitui a unidade de tradução básica. Assim, a principal intuição por trás deste tipo de TAE é que as palavras nem sempre são a melhor unidade de tradução pois a correspondência entre línguas normalmente não é 1 : 1. Poder-se-ia argumentar que esta limitação foi superada polos sistemas de TAE baseada em palavras desde que o algoritmo de tradução de Brown et al. (1993) apresentasse um modelo de tradução conceitualmente preparado para tratar os alinhamentos $1 : n$. Porém, os sistemas de TAE baseada em frases, dão mais um passo simplificando o problema ao converterem os alinhamentos de palavras em unidades de maior ordem, conhecidos como *frases*.[15] Assim, os sistemas de TAE baseada em frases, não realizam mapeamentos entre várias unidades, mas antes entre uma unidade e outra, embora de maior ordem que as palavras.

O modelo de TAE baseada em frases que se seguiu no desenvolvimento do nosso protótipo de TAE inglês–galego é o descrito em Koehn, Och e Marcu (2003).

## *5 Carvalho: sistema de TAE inglês–galego*

Tal e como foi mencionado na secção 2, Carvalho é um protótipo de tradução automática estatística para o par de línguas inglês–galego. Carvalho foi treinado seguindo o paradigma da mencionada TAE beasada em frases. Para o seu treino três peças principais de software foram utilizadas:

- GIZA++[16]: GIZA++, originalmente desenvolvido durante o *John Hopkins University 1999 Summer Workshop*, é uma implementação de Och e Ney (2000) de todos os algoritmos de alinhamento palavra-a-palvra de IBM assim como do algoritmo HMM, acrónimo de Hidden Markov Models.[17]
- Moses[18]: Moses é a implementação de Koehn et al. (2007) da sua proposta de TAE baseada em frases feita em 2003, Koehn, Och e Marcu (2003). Moses utiliza os alinhamentos palavra-a-palavra aprendidos por GIZA++ para criar um modelo de tradução baseada em frases utilizado para determinar a melhor tradução $\hat{T}$ dada uma frase de origem $S$.
- SRILM[19]: SRILM, que pode ser utilizado livremente com fins não comerciais, é um modelizador de língua, isto é, uma ferramenta que aprende sequências de n-gramas, que servem para determinar a fluidez das traduções saintes de Moses e, deste modo, reordenar o ranking de traduções para finalmente determinar a tradução mais provável $\hat{T}$. SRILM foi treinado utilizando o texto completo do "lado" inglês ou galego, dependendo da direcção de traducção, do corpus de treino de GIZA++ e Moses.

### 5.1 Carvalho *vs.* Google Translate

Para exemplificar visualmente o sucesso que supôs a utilização do corpus paralelo inglês–galego obtido após o processamento descrito na secção 3.2 gostaríamos de mostrar dous exemplos de tradução; um realizado por Carvalho e outro por Google Translate[20], da seguinte frase, tirada da entrada da Wikipedia *Art*[21]:

*Art is the process or product of deliberately arranging elements in a way that appeals to the senses or emotions. It encompasses a diverse range of human activities, creations, and modes of expression, including music, literature, film, sculpture, and paintings. The meaning of art is explored in a branch of philosophy known as aesthetics.*

A tradução realizada por Carvalho é a seguinte:

*Arte é o proceso ou produto de arranxar deliberadamente elementos dunha forma que apela á sentidos ou emocións. Engloba un diversificado abano de actividades humanas, creacións e modos de expresión, inclusive da música, da literatura, filmes, escultura e pinturas. O significado de arte é explotada en un ramo da filosofía coñecida como aesthetics.*

À continuação mostra-se a tradução realizada por Google Translate a dia 2 de Março de 2010:

deste algoritmo ver Jurafsky e Martin (2008, 886–888).

[15] As *frases* na TAE baseada em frases não estão em absoluto linguisticamente motivadas, pois nada têm a ver com o conceito linguístico de frase derivado da teoria sintáctica de constituintes. Mesmo assim, empregaremos esta denominação pois é a mais estendida no campo da TAE.

[16] Disponível em `http://fjoch.com/GIZA++.html`.

[17] Para uma descrição detalhada do funcionamento dos algoritmos de IBM e HMM ver Och e Ney (2003).

[18] Disponível em `http://www.statmt.org/moses/`.

[19] Disponível em `http://www.speech.sri.com/projects/srilm/`

[20] O serviço de tradução de Google, Google Translate, incorporou em 2008 o galego entre o seu catálogo de línguas com ferramentas de PLN.

[21] `http://en.wikipedia.org/wiki/Art`.

| | Inglês–Galego | Galego–inglês |
|---|---|---|
| Carvalho | 0,1559 | 0,1895 |
| GT | 0,2559 | 0,3591 |

Tabela 1: Comparativa do *BLEU score* de Carvalho *vs.* Google Translate (GT).

*A arte é o proceso ou produto de deliberadamente organizar elementos de un modo que pide aos sentidos ou emocións. Engloba unha variada gama de actividades humanas, creacións, e modos de expresión, incluíndo a música, literatura, cine, escultura e pintura. O significado da arte é explotado desde unha rama da filosofía coñecido como estética.*

Embora resulte interessante poder comparar visualmente as traduções realizadas por estes dous sistemas de TAE, é obrigado empregar uma medida numérica objectiva para pôr em perspectiva a performance de ambos os sistemas. A medida escolhida foi *BLEU score*[22], (Papineni et al., 2001), calculada pola versão 11b do *National Institute of Standars and Technology* (NIST) dos Estados Unidos da América. A obtenção de um valor numérico de *BLEU score* realizou-se mediante a tradução de um pequeno corpus de refeência, *goldstandard*, de 11.500 palavras que compilámos em **imaxin**|software manualmente traduzindo uma colecção de 500 frases em inglês extraídas da versão online do jornal inglês *The Guardian*.

Em **imaxin**|software somos conscienteces de que as medidas obtidas têm as suas limitações. Por um lado, entre as críticas mais importantes vertidas a respeito de *BLEU score* está que esta medida em muito diversos contextos correlaciona-se deficientemente com as percepções humanas à hora de avaliar uma mesma tradução automática (ver Ananthakrishnan et al. (2007) ou Callison-Burch, Osborne e Koehn (2006)). Deste modo, se compararmos as traduções de exemplo de Carvalho e Google Translate, as diferenças entre dous sistemas parecem não ser tão dramáticas como sugerem os resultados obtidos mediante a tradução do nosso *goldstandard*, que apresentamos na tabela 5.1. Por outro lado, somos também conscientes de que, para a availiação de uma tradução automática, a utilização de uma só tradução de referência com *BLEU score* é insuficiente, já que se-lhe atribui a uma só tradução demasiado peso e valor, o qual não reflecte a realidade de que não existe uma *tradução perfeita* e que um mesmo texto de origem pode e deve ser traduzido de modos diferentes dependendo do contexto sociocultural, histórico, etc.

Tendo todas estas críticas em conta, a obtenção de uma medida numérica objectiva não deixa de ser útil quanto peça de informação de referência para comparar estes dous sistemas de TAE.

## 6 Conclusões

Neste artigo mostrou-se, por um lado, uma sólida estratégia de dramática redução do tempo de compilação de um corpus paralelo inglês–galego do tamanho necessário para o desenvolvimento de um protótipo de TAE para o citado par de línguas mediante o processo de conversão semi-automatizado descrito na secção 3.2. E demonstou-se, por outro lado, a alta qualidade dos resultados que podem ser obtidos seguindo esta estratégia (ver sec. 5.1). Estatégia que cremos foi também a seguida por Google na incorporação do galego no seu serviço Google Translate, tal e como sugere a seguinte tradução que em Abril de 2009 realizámos com este serviço durante os primeiros testes de avaliação de Carvalho:

*A arte é o proceso ou produto de deliberadamente organizar elementos dun modo que apelido aos sentidos ou* **emoções***. Engloba un conxunto diversificado de actividades humanas,* **criações***, e modos de expresión, incluíndo a música e a literatura. O significado da arte é explorador no ramo da filosofía coñecido como estética.*[23]

Em **imaxin**|software cremos firmemente que de não ser pola minimização do tempo de desenvolvimento e a alta qualidade dos resultados obtidos, Google muito provavelmente teria demorado muito mais tempo em incorporar o galego entre o leque de línguas das suas ferramentas de

[22] *BLEU score* é uma medida de avaliação de TA que mede a proximidade de uma tradução automática de uma tradução profissional humana, assumindo que quanto mais próxima esteja a tradução automática da tradução humana melhor é a primeira. Assim, o que *BLEU score*, a *grosso modo*, faz é contar o número de n-gramas da tradução automática que se sobrepõem aos da tradução humana, que se utiliza como tradução de referência. Na prática, *BLEU score* funciona combinando n-gramas sobrepostos ponderados de diferentes tamanhos –quatrogramas, trigramas, bigramas e unigramas. Além deste modelo de *backoff* de n-gramas sobrepostos, *BLEU score* também implementa um factor de penalização de brevidade que impede que as traduções sejam demasiado curtas com respeito à tradução humana de referência.

[23] Tal e como indica esta tradução, Google Translate foi muito provavelmente treinado utilizando corpora paralelos inglês–portugês parcialmente convertidos à ortografia galega. Contudo, à diferença da estrategia de **imaxin**|software, Google não parecia utilizar conversores ortográficos. Deste modo, as palavras portuguesas que não se encontravam nos seus dicionários permaneciam na sua ortografia original.

PLN.

É por tudo isto que podemos concluír com confiança que a estratégia de criar ferramentas de PLN para o galego partindo de recursos computacionais do portugês não é simplesmente justificável do poto de vista lingu ístico, mas absolutamente legítima.

Não é, do nosso ponto de vista, aventurado concluír que a utilização de recursos de uma língua intimamente relacionada, especialmente se esta é uma língua computacionalmente desenvolvida, é extremadamente útil para variedades linguísticas, como o galego, que carecem de ferramentas de PLN devido à sua posição de minorização.

## *Agradecimentos*

A todos os investigadores/as que reconheceram que o galego tinha uma dimensão internacional e que tínhamos que nos aproveitar disso: Carvalho Calero, Manuel Rodrigues Lapa, Eugene Coseriu, etc.

Ao Parlamento Europeu por ter libertado as suas actas no domínio público.

À Dirección Xeral de I+D+i da Xunta de Galicia que financiou parte deste projecto RecursOpentrad.

## *Referências*

Ananthakrishnan, R., P. Bhattacharya, M. Sasikumar, e R. M. Shah. 2007. Some issues in automatic evaluation of English–Hindi MT: More Blues for BLEU. Em *International Conference On Natural Language Processing (ICON)*.

Aracil, Ll. 1985. *Lingüística e sócio-lingüística galaico-portuguesa: reintegracionismo e conflito lingüístico na Galiza.* Asociaçom Socio-Pedagógica Galaico-Portuguesa.

Armentao-Oller, C., R. C. Carrasco, A. M. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, Felipe Sánchez-Martínez, e M. A. Scalco. 2006. Open-source Portuguese–Spanish machine translation. Em *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)*, pp. 50–59. (c) Springer-Verlag.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, e P. Roosin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Brown, P., S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, e R. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C., M. Osborne, e P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. Em *Proceedings of the European Association for Computational Linguistics (EACL)*, pp. 249–256.

Coseriu, E. 1987. El gallego en la historia y en la actualidad. Em *Actas do II Congresso Internacional da Língua Galego-Portuguesa*, pp. 793–800.

Cunha, C. e L. Cintra. 2002. *Nova Gramática do Português Comtemporâneo.* Edições João Sá da Costa.

Gee, J. P. 1999. *An Introduction to Discourse Analysis: Theory and Method.* Routledge.

Jurafsky, D. e J. H. Martin, 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 25, pp. 859–908. Pearson, 2 edition.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. Em *MT Summit 2005.*

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, M. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, e E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. Em *Annual Meeting of the Association for Computational Linguistics (ACL).*

Koehn, P., F. J. Och, e D. Marcu. 2003. Statistical phrase-based translation. Em *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 127–133.

Och, F. J. e H. Ney. 2000. Improved statistical alignment models. Em *Proceedings of 38th Annual meeting of the ACL*, pp. 400–447.

Och, F. J. e H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K. A., T. Roukos, T. Ward, e W. J. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Relatório técnico, IBM Research Division, Thomas J. Watson Research Center.

Popović, M. e H. Ney. 2006. Statistical machine translation with a small amount of bilingual training data. Em *Language Resources and Evaluation (LREC) 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority language"*, pp. 25–29.

Rei, F. Fernández. 1991. *Dialectoloxía da lingua galega*. Edicións Xerais de Galicia.

# Inducción de constituyentes sintácticos en español con técnicas de *clustering* y filtrado por información mutua

Fernando Balbachan
Facultad de Filosofía y Letras
Universidad de Buenos Aires
fernando_balbachan@yahoo.com.ar

Diego Dell'Era
Facultad de Filosofía y Letras
Universidad de Buenos Aires
diego.dellera@gmail.com

**Resumen**

El Argumento de la Pobreza de los Estímulos (*Argument from the Poverty of Stimulus*, *APS*) se presenta como el gran campo de debate epistemológico entre el paradigma simbólico y el paradigma estadístico en lingüística computacional (Pullum y Scholz 2002). Desde 2000 en adelante aparecieron algunos trabajos dentro del paradigma estadístico que se propusieron atacar el Argumento de la Pobreza de los Estímulos a partir de la postulación de algún algoritmo general no supervisado de adquisición integral del lenguaje. Entre los aportes más importantes, la tesis de doctorado de Clark (2001) recurre a diversas técnicas estadísticas para dar con un algoritmo general no supervisado de inducción del lenguaje, y en particular, de una gramática independiente de contexto para el inglés.

Clark (2001) trabaja con distintas técnicas de inducción para cada fenómeno lingüístico modelizado: morfología mediante modelos markovianos, categorización (*POS-tagging*) mediante *clustering*, etc. Puntualmente, en este trabajo estamos interesados en la inducción de constituyentes sintácticos, dado un *corpus* etiquetado por clase de palabras (*POS-tagged*), como paso previo al procedimiento de inducción de una gramática independiente de contexto. En su propia tesis, el autor reconoce que es necesaria una mayor evidencia translingüística que apoye la plausibilidad psicolingüística de un enfoque como el suyo. Actualmente, no existen trabajos que se hayan propuesto probar el enfoque de Clark (2001) para la inducción de sintaxis en lenguas flexivas y con orden libre de constituyentes, como el español. Así pues, nuestro trabajo se propone contribuir con dicha evidencia translingüística, estudiando la factibilidad de aplicación del algoritmo de inducción de constituyentes de Clark (2001) para el español.

El algoritmo de Clark (2001) que nos ocupa consiste en aplicar técnicas de *clustering K-means* para agrupar secuencias de etiquetas de clase de palabra, según su información distribucional. Luego, se procede a filtrar los resultados para encontrar *clusters* que efectivamente se correspondan con grupos de constituyentes, recurriendo a un criterio de información mutua entre los símbolos inmediatamente anteriores y posteriores a dichas secuencias. Este criterio de filtrado evita el sesgo de un *corpus* escaso, al tiempo que logra distinguir la dependencia buscada entre los límites de las secuencias candidatas a constituyentes por sobre el umbral de la entropía natural de símbolos que co-ocurren a una cierta distancia en el lenguaje (Li 1990).

Nuestra implementación del algoritmo ha sido evaluada en un *corpus* de dimensiones prototípicas, con resultados prometedores. Se obtuvo una cobertura de 74%, una precisión de 58% y una medida F de 65%, en la etapa prototípica. Estos resultados alientan la continuidad del trabajo de investigación a largo plazo, con la meta de lograr un robusto algoritmo de adquisición integral del lenguaje para el español.

## *1. Introducción*

### 1.1 El debate epistemológico acerca de la adquisición del lenguaje

En el campo transdisciplinario de la lingüística computacional, los enfoques estadísticos, que surgieron principalmente a mediados de la década del '90, vinieron a desplazar al paradigma simbólico, dominante hasta entonces (Moreno Sandoval 1998; Manning y Schütze 1999). El paradigma simbólico, que evolucionó bajo la égida chomskyana de las gramáticas generativas desde la década del '50 (Chomsky 1957, 1965; Moreno Sandoval 1998), se propone manipular categorías sintácticas dadas *a priori* (gramática formal) para deducir o derivar el conjunto de oraciones que constituye una lengua, a partir de la aplicación de reglas, parámetros o principios. El paradigma estadístico, en cambio, echa mano de diversas técnicas probabilísticas, aplicadas a grandes *corpora* de entrenamiento, con vistas a inducir categorías y fenómenos específicos del lenguaje natural a partir de la

detección de patrones estadísticamente significativos en la *tabula rasa* que constituyen los *corpora*. Sin embargo, el paradigma estadístico es más que una mera aplicación de técnicas y modelización matemática: estos enfoques aportan evidencia de plausibilidad psicolingüística a un renovado debate acerca de la naturaleza misma del lenguaje. En efecto, entre el paradigma simbólico y el paradigma estadístico se ha entablado un manifiesto contrapunto de concepciones epistemológicas opuestas en torno al atávico problema de la adquisición del lenguaje, a partir del encolumnamiento de las obras fundacionales del campo detrás de teorías innatistas o teorías empiristas, respectivamente (Piatelli-Palmarini 1980, Cowie 1999, Pullum y Scholz 2002).

> "Probabilistic methods are providing new explanatory approaches to fundamental cognitive science questions of how humans structure, process and acquire language […] Probabilistic models can account for the learning and processing of language, while maintaining the sophistication of symbolic models." [Chater y Manning 2008:335]

Aunque algunos entusiastas de la polémica aseguran que el debate acerca de la adquisición del lenguaje bien podría remontarse al siglo XVII con las posturas filosóficas de Descartes y de Locke (Clark 2001), más recientemente podemos empezar a rastrear esta confrontación en obras primordiales de la lingüística teórica (Chomsky 1957, 1965, 1986) y la psicolingüística (Fodor 1983; Pinker 1994) de la segunda mitad del siglo XX, las cuales defienden un innatismo a ultranza; mientras que las posturas empiristas son esgrimidas por la lingüística cognitiva prototípica de Lakoff y Langacker (Lakoff 1987; Langacker 2000) y filósofos del lenguaje como Quine y otros. Con respecto al problema de la adquisición del lenguaje en el campo transdisciplinario de la lingüística computacional, mientras el paradigma simbólico adscribe a sistemas deductivos que hipotetizan como condición necesaria un estado inicial de conocimiento innato y ricamente estructurado frente a la pobreza de los datos lingüísticos primarios de que dispondrían los niños, los enfoques estadísticos postulan, más bien, sistemas inductivos a partir del aprendizaje de patrones de ocurrencia de eventos en un *corpus* masivo no estructurado, mediante algún algoritmo de aprendizaje de propósitos generales –es decir, no específico de dominio (Clark 2001).

Justamente, el *Argumento de la Pobreza de los Estímulos* (*Argument from the Poverty of Stimulus* o *APS*) se presenta como el gran campo de debate epistemológico entre el paradigma simbólico y el paradigma estadístico. Así pues, el APS se empezó a perfilar como el más robusto adalid de la hipótesis innatista, aunque como bien señalan Pullum y Scholz (2002), ninguna teoría que avale tácita o taxativamente dicha hipótesis deja en claro las propiedades y la estructura de ese conocimiento innato de que dispondríamos durante el proceso de adquisición del lenguaje:

> "The one thing that is clear about the argument from the poverty of the stimulus is what its conclusion is supposed to be: it is supposed to show that human infants are equipped with innate mental mechanisms specifically for assisting in the language acquisition process – in short that the facts about human language acquisition support 'nativist' rather than 'empiricist' epistemological views. What is not clear at all is the structure of the reasoning that is supposed to support this conclusion. Instead of clarifying the reasoning, each successive writer on this topic shakes together an idiosyncratic cocktail of claims about children's learning of language and claims that nativism is thereby supported." [Pullum y Scholz 2002:12]

Por supuesto, desde la otra orilla, las teorías empiristas no deben ser confundidas con un trasnochado conductismo y su concepción del lenguaje como una mera asociación de esquemas estímulo-respuesta. Los empiristas no refutan la existencia de algún mecanismo inicial como condición necesaria para adquirir el lenguaje; simplemente postulan que este mecanismo se trataría de un aspecto más de la inteligencia humana (Piatelli-Palmarini 1980, Clark 2009), un algoritmo de aprendizaje de propósitos generales y no de una habilidad que presupone *a priori* conocimiento de dominio específico (*cf.* concepto de *gramática universal* en Chomsky 1957, 1965, 1986 y concepto de *facultad vertical* en Fodor 1983). Más aún, algunos empiristas no reniegan completamente del procesamiento encapsulado de dominio

específico (Fodor 1983), pero rechazan la idea de que la adquisición del lenguaje sea un proceso llevado a cabo *íntegramente* por este tipo de capacidades cognitivas:

> "There may well be domain-general parts of cognition that are applied to the task of language-acquisition even though the core of it is domain-specific. This sort of research could fruitfully focus the attention of researchers on particular aspects of language where the domain-specificity is more essential; moreover, I think it is clear that at some points in the language acquisition process, even nativists must propose some sort of statistical learning, albeit just for low-level tasks such as word segmentation." [Clark 2001:20]

## 1.2 Técnicas estadísticas para inducción de sintaxis

Así pues, la confrontación entre el paradigma simbólico y el paradigma estadístico se desató en varios frentes. Por un lado, la supuesta imposibilidad de aprendizaje del lenguaje ante la falta empírica de evidencia negativa (Gold 1967), argumento refutado en Clark (2001). Por otro lado, la renuencia de Chomsky y sus seguidores a dar crédito a las nociones estadísticas de la época como herramienta de análisis:

> "It seems to have been demonstrated beyond all reasonable doubt that, quite apart from any question of feasibility, methods of the sort that have been studied in taxonomic linguistics are intrinsically incapable of yielding the systems of grammatical knowledge that must be attributed to the speaker of a language." [Chomsky 1965:54]

> "Dixon speaks freely throughout about the 'probability of a sentence' as though this were an empirically meaningful notion. [...] We might take 'probability' to be an estimate of relative frequency […]. This has the advantages of clarity and objectivity, and the compensating disadvantage that almost no 'normal' sentence can be shown empirically to have a probability distinct from zero. That is, as the size of a real corpus (e.g. the set of sentences in the New York Public Library, or the Congressional Record, or a person's total experience, etc.) grows, the relative frequency of any given sentence diminishes, presumably without limit." [Chomsky 1966:34-35]

Sin embargo, a partir de la denominada revolución bayesiana en lingüística computacional (Manning y Schütze 1999; Clark 2001), las técnicas estadísticas, otrora ineficaces para lidiar con la aceptabilidad de oraciones que requerían los *corpora* reales, se renuevan incorporando la noción de probabilidad en términos de *grado subjetivo* de incertidumbre (*subjective degree of uncertainty*) para ser utilizadas en procesos masivos de inducción que modelan efectivamente distintas áreas del procesamiento del lenguaje natural, demoliendo así el otro bastión de la polémica contra el paradigma estadístico, con lo que queda en pie un último refugio del reinado del innatismo: el Argumento de la Pobreza de los Estímulos. Parafraseando a Klein y Manning (2004), los estímulos no parecen ser tan pobres como se creería:

> "We make no claims as to the cognitive plausibility of the induction mechanisms we present here; however, the ability of these systems to recover substantial linguistic patterns from surface yields alone does speak to the strength of support for these patterns in the data, and hence undermines arguments based on 'the poverty of the stimulus'." [Klein y Manning 2004:478]

Desde 2000 en adelante aparecieron algunos trabajos dentro del paradigma estadístico que se propusieron atacar el Argumento de la Pobreza de los Estímulos –y consecuentemente, la hipótesis innatista– a partir de la postulación de algún algoritmo general no supervisado de adquisición integral del lenguaje. Pese a que se proponen confrontar con el APS, estos trabajos, enmarcados en el paradigma estadístico, abordan el problema desde la misma perspectiva inicial que Chomsky: la sintaxis como punto de partida para la adquisición del lenguaje y el isomorfismo entre lenguajes formales y naturales (Chomsky 1957).

Entre los trabajos que concitan mayor interés, la tesis de doctorado de Clark (2001) recurre a diversas técnicas estadísticas para dar con un algoritmo general no supervisado de inducción

de sintaxis y, en particular, de una gramática independiente de contexto para el inglés como un modelo formal para la adquisición del lenguaje (Pinker 1979):

> "This question is in one sense thoroughly Chomskyan: I fully accept his characterization of linguistics as, ultimately, a branch of psychology, though for the moment it relies on very different sorts of evidence; I fully accept his argument for complete formality in linguistics, a formality that computer modeling both requires and enforces; I fully accept the idea that one of the central problems of linguistics is how to explain the fact that children manage to learn language in the circumstances that they do. On the other hand, there are many areas in which this work is not so congenial to followers of the Chomskyan paradigm. First, the work here is fully empirical; it is concerned with authentic language, rather than artificial examples. Secondly, it eschews the use of unnecessary hidden entities; far from considering this as the hallmark of a good scientific theory, the unnecessary proliferation of unobservable variables renders the link between theory and surface tenuous and unstable." [Clark 2001:3]

Clark (2001) hace uso de distintas técnicas de inducción para cada fenómeno lingüístico modelizado: morfología mediante modelos markovianos, categorización (*POS-tagging*) mediante *clustering* distribucional, etc. Puntualmente, en este artículo estamos interesados en la inducción de constituyentes sintácticos, dado un *corpus* etiquetado por clase de palabras (*POS-tagged*), como paso previo al procedimiento de inducción de una gramática probabilística independiente de contexto (*Stochastic Context-Free Grammar* o *SCFG* o *Probabilistic Context-Free Grammar* o *PCFG*), que es el fin último de su tesis. Clark mismo reconoce que es necesaria una mayor evidencia translingüística que apoye la plausibilidad psicolingüística de su investigación:

> "There are a number of possible avenues for future research. The most important, in my opinion, is to experiment with other languages, particularly languages with very free word order. There is some evidence that these techniques will work with Chinese (Redington et al. 1995), which has quite fixed word order, but no work has been done in highly inflected languages with free word order." [Clark 2001:148]

Actualmente no existen trabajos que se hayan propuesto probar dicho enfoque para la inducción de sintaxis en español, una lengua flexiva y con orden libre de constituyentes. Así pues, nuestro trabajo se propone contribuir con dicha evidencia translingüística, estudiando, en principio, la factibilidad de aplicación del algoritmo de inducción de constituyentes de Clark (2001) para el español.

## 1.3 Trabajos previos

La tarea de inducción de constituyentes sintácticos, como primer paso para la inducción de gramáticas, ha venido atrayendo la atención temprana de investigadores:

> "Early work on grammar induction emphasized heuristic structure search, where the primary induction is done by incrementally adding new productions to an initially empty grammar (Olivier 1968; Wolff 1988). In the early 1990s, attempts were made to do grammar induction by parameter search, where the broad structure of the grammar is fixed in advance and only parameters are induced (Lari and Young, 1990; Carroll and Charniak 1992). However, this appeared unpromising and most recent work has returned to using structure search." [Klein y Manning 2002:128]

Más recientemente, con el advenimiento del paradigma estadístico, la tarea adquirió un nuevo vigor científico. Ya sea para rebatir empíricamente el argumento de la pobreza de los estímulos (Clark 2001), para construir bancos masivos de árboles sintácticos –*treebanks*– (van Zaanen 2000) o como parte de modelos de lenguaje (Chen 1995), la inducción no supervisada de estructura sintáctica básica bajo la forma de constituyentes ha probado ser uno de los campos más fértiles de investigación básica en lingüística computacional, a partir de la diversidad de enfoques de los algoritmos de trabajos previos. Clark (2001) releva en detalle los distintos enfoques con que se ha encarado la tarea y agrupa los trabajos más recientes en 3

diferentes aunque en alguna medida superpuestas categorías:

1) Enfoques basados en la probabilidad (Carroll y Charniak 1992 y otros trabajos): Se trata de experimentos basados en el algoritmo *Expectation-Maximization* o *EM* –también conocido como algoritmo *inside-outside* o *IO* (Manning y Schütze 1999)– o en algoritmos genéticos, que han obtenido resultados no muy exitosos:

> "[…] [This approach] produced some rather discouraging research that seemed to indicate that the fact that the IO algorithm converged to a local optimum meant that it would almost always converge to a linguistically implausible grammar." [Clark 2001:124]

2) Enfoques basados en la compresión (Wolff 1988 y otros trabajos): Básicamente recurren a una heurística de compresión de gramáticas inicialmente extensas, a partir del algoritmo *Minimum Description Length* o *MDL* (Manning y Schütze 1999). Si bien estos experimentos obtuvieron cierto éxito en lenguajes artificiales, Clark (2001) observa que este tipo de enfoques fallan al no dar cuenta de las dependencias de larga distancia que caracterizan a los constituyentes sintácticos del lenguaje natural.

3) Enfoques basados en la información distribucional: En este tipo de enfoques se encuadran los trabajos de Finch et al. (1995), Clark (2001) y nuestro propio experimento. La idea subyacente es que las secuencias de palabras o etiquetas morfosintácticas que componen un constituyente sintáctico –símbolo no terminal como, por ejemplo, Sintagma Nominal SN o Sintagma Preposicional SP– aparecerán en similares contextos distribucionales –a izquierda y a derecha– a lo largo de *corpora* masivos.

   Entre los enfoques basados en la información distribucional, encontramos uno de los primeros trabajos de inducción de gramáticas para el español: el algoritmo de inducción de gramática del español de Juárez Gambino y Calvo (2007). Basándose en la noción de sustituibilidad de Harris (2000) para hallar regularidades estructurales, estos investigadores desarrollaron un algoritmo no supervisado para entrenar al sistema de inducción de gramática ABL (*Alignment-Based Learning*) (van Zaanen 2000) con un *corpus* (CAST-3LB) de español de características similares a las de nuestro *corpus* (véanse *Tabla 1* y *Tabla 2*), reportando una medida F de 32,45% en la tarea de inducción de constituyentes sintácticos. No obstante, cabe aclarar que la evaluación del experimento de Juárez Gambino y Calvo (2007) apuntó al *parsing* de oraciones con los constituyentes inducidos, meta más ambiciosa que la evaluación manual de un listado de constituyentes inducidos, como en nuestro caso.

## 1.4 *Corpora* masivos y *corpora* para implementaciones prototípicas

Los experimentos de aprendizaje de máquina (*machine learning*) nos obligan a reflexionar sobre un aspecto metodológico con profundas incumbencias en el estudio del desarrollo ontogenético del lenguaje. Efectivamente, las técnicas probabilísticas, aplicadas a *corpora* masivos, inducen los fenómenos lingüísticos a partir de la identificación de patrones estadísticamente significativos. De este modo, se busca analogar los datos lingüísticos primarios (*Primary Linguistic Data* o *PLD*) de que dispondría un niño durante el proceso de adquisición del lenguaje a los *corpora* de millones de palabras que son procesados iterativamente por las computadoras.

La preocupación concerniente a la plausibilidad de modelización de los PLD es uno de los requerimientos que detalla Pinker (1979) para una teoría formal que se proponga explicar la adquisición del lenguaje:

> "It is instructive to spell out these conditions one by one and examine the progress that has been made in meeting them. First, since all normal children learn the language of their community, a viable theory will have to posit mechanisms powerful enough to acquire a

> natural language. This criterion is doubly stringent: though the rules of language are beyond doubt highly intricate and abstract, children uniformly succeed at learning them nonetheless, unlike chess, calculus and other complex cognitive skills. Let us say that a theory that can account for the fact that languages can be learned in the first place has met the *Learnability Condition*. Second, the theory should not account for the child's success by positing mechanisms narrowly adapted to the acquisition of a particular language. For example, a theory positing an innate grammar for English would fail to meet this criterion, which can be called the *Equipotentiality Condition*. Third, the mechanisms of a viable theory must allow the child to learn his language within the time span normally taken by children, which is in the order of three years for the basic components of language skill. Fourth, the mechanisms must not require as input types of information or amounts of information that are unavailable to the child. Let us call these the *Time and Input Conditions*, respectively. Fifth, the theory should make predictions about the intermediate stages of acquisition that agree with empirical findings in the study of child language. Sixth, the mechanisms described by the theory should not be wildly inconsistent with what is known about the cognitive faculties of the child, such as the perceptual discriminations he can make, his conceptual abilities, his memory, attention, and so forth. These can be called the *Developmental and Cognitive Conditions*, respectively." [Pinker 1979:219]

Pullum (1996) describe bastante bien esta plausibilidad de modelización entre los PLD y los *corpora* masivos del procesamiento computacional:

> "Ideally, what we need to settle the question is a large machine-readable corpus – some tens of millions of words – containing a transcription of most of the utterances used in the presence of some specific infant (less desirably, a number of infants) over a period of years, including particularly the period from about one year (i.e. several months earlier than the age at which two words utterances start to appear in children's speech) to about 4 years." [Pullum 1996:505]

Sin embargo, como Clark (2001) mismo reconoce, resulta difícil emular por completo los datos lingüísticos primarios, toda vez que los recursos de *corpora* de que dispone la comunidad científica están mayormente basados en lenguaje escrito y manifiestan notables diferencias en el registro y grado de formalidad y complejidad de los enunciados en comparación con los que presumiblemente serían los enunciados a los que se ve expuesto un niño entre el año y los 4 años de vida. Es menester mencionar que existen ciertos *corpora* que ofrecen lenguaje de registro especializado, como el *corpus CHILDES* (2,5 millones de palabras organizadas en interacciones orales madre-niño en inglés norteamericano) o el *corpus Wall Street Journal* o *WSJ* (registro periodístico). Aun así, como el mismo Chomsky (1959) concede, se debe tomar en cuenta que los niños en edad de adquirir el lenguaje no sólo se ven expuestos a los enunciados dirigidos específicamente hacia ellos, sino que los medios audiovisuales de comunicación o incluso las conversaciones entre adultos bien podrían funcionar como otros proveedores de datos lingüísticos primarios.

Para su tesis y en particular, para el experimento de inducción de constituyentes sintácticos, Clark (2001) recurre al *British National Corpus* o *BNC* en su primera edición del año 1994, un *corpus* sincrónico de inglés británico que contiene 100 millones de palabras de registro variado (periódicos, obras literarias, etc.), etiquetadas automáticamente según el estándar *C5* (CLAWS5) –un conjunto de 76 etiquetas morfosintácticas al que Clark agrega un símbolo para indicar el fin de oración. Aunque el BNC abarca registros orales en un 10% de la muestra, Clark recorta el *input* del BNC a 12.000.000 de palabras del registro escrito.

Puesto que nuestro objetivo es el estudio de factibilidad del experimento de Clark (2001) acerca de la inducción de constituyentes sintácticos para el español, nuestro trabajo se propuso adaptar su metodología a una implementación prototípica que probara la viabilidad de este enfoque para una lengua flexiva y con constituyentes sintácticos de orden libre.

Entre los recursos gratuitos de lingüística computacional del español, debemos mencionar el *corpus* CRATER, un *corpus* masivo multilingüístico de alineamiento de oraciones entre el inglés, el francés y el español, anotado morfosintácticamente. No obstante, una primera evaluación de la utilidad de este *corpus* para nuestro experimento resultó poco prometedora, ya que CRATER emplea alrededor de 500 etiquetas morfosintácticas y su interfaz de consulta resulta completamente obsoleta.

Entre los recursos de acceso gratuito sólo con fines académicos, analizamos los 2 corpus morfosintácticamente anotados más conocidos del español: CAST-3LB (Civit 2003) y el *Spanish Treebank* (Moreno Sandoval *et al.* 1999).

| | **CAST-3LB** | **Spanish Treebank** |
|---|---|---|
| Tamaño en palabras | ≈100.000 | ≈45.000 |
| Tamaño en oraciones | ≈3.500 | ≈1.600 |
| Extensión promedio de oraciones | ≈30 palabras | ≈28 palabras |
| Anotación morfosintáctica | ≈350 etiquetas | ≈200 etiquetas |
| Criterio de anotación | Anotación semi-manual sintáctica, semántica y pragmática. En el nivel sintáctico, se sigue anotación por constituyentes, con marcaje adicional de funciones sintácticas (Civit 2003) | Automático: *chunking* y *POS-tagging*<br><br>Validación manual por muestreo aleatorio (Moreno Sandoval *et al.* 1999) |

Tabla 1: Comparación entre *corpora* CAST-3LB y *Spanish Treebank*

Sin embargo, la proliferación de etiquetas morfosintácticas en un corpus de reducidas dimensiones podría presentar un problema de dispersión de datos, escollo que debíamos evitar para adaptar el algoritmo de Clark (2001), que trabaja sobre un corpus masivo y con un listado reducido de etiquetas.

Dada la escasez de *corpora* morfosintácticamente anotados para nuestro idioma, nos vimos obligados a encarar la esforzada tarea de generar un *corpus* propio en español, etiquetado según los lineamientos morfosintácticos adaptados del BNC (Leech *et al.* 1994), que alcanzara dimensiones suficientes para trabajar a escala con un prototipo de la implementación del algoritmo adaptado y optimizado para el idioma español.

Debemos agradecer la colaboración de un equipo de entusiastas estudiantes de Lingüística de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires UBA, gracias al cual logramos organizar un *corpus* de aproximadamente 50.000 palabras de registro periodístico escrito, etiquetado morfosintácticamente mediante un riguroso criterio metodológico (véanse *Anexos I*, *II* y *III*). El *corpus* resultante, el instructivo describiendo la metodología utilizada y la implementación prototípica del algoritmo del experimento se encuentran disponibles para la comunidad científica, bajo los alcances de una licencia *Creative Commons* en el sitio web[1].

## *2. Algoritmo de inducción de constituyentes sintácticos en Clark (2001)*

El algoritmo de Clark (2001) que nos ocupa consiste en aplicar técnicas de *clustering K-means* para agrupar secuencias de etiquetas de clase de palabra, según su información distribucional. Luego, se procede a filtrar los resultados para encontrar *clusters* que efectivamente se correspondan con grupos de constituyentes, recurriendo a un criterio de información mutua entre los símbolos inmediatamente anteriores y posteriores a dichas secuencias.

Este criterio de filtrado evita el sesgo de un *corpus* escaso, al tiempo que logra distinguir la dependencia buscada entre los límites de las secuencias candidatas a constituyentes por sobre el umbral de la entropía natural de símbolos que co-ocurren a una cierta distancia en el lenguaje (Li 1990).

[1] http://campus.filo.uba.ar/course/view.php?id=587

| | **Clark (2001)** | **Balbachan-Dell'Era (2010)** |
|---|---|---|
| Tamaño en palabras | ≈12.000.000 | 49.925 |
| Tamaño en oraciones | ≈700.000 | 2.108 |
| Extensión promedio de oraciones | 16,6 palabras | 23,8 palabras |
| Anotación morfosintáctica | Manual: 77 etiquetas según estándar C5 | Manual: 48 etiquetas adaptadas del estándar C5 |
| Criterio de anotación | BNC (Leech *et al.* 1994) | propio, adaptado del BNC |

Tabla 2: Comparación entre *corpora* de *input* para ambos experimentos

## 2.1 Acerca de la naturaleza de un constituyente

La noción de constituyente, en sentido amplio, se aplica a conjuntos de palabras (o etiquetas) que funcionan como unidades sintácticas en la oración. En el sentido estricto en que Clark la usa, la noción de constituyente está restringida a una secuencia continua de etiquetas que reescribe a un nodo no terminal en una derivación sintáctica. El requerimiento de continuidad se debe a que los constituyentes encontrados se usan en etapas subsiguientes del experimento para inducir gramáticas libres de contexto, y esa clase de gramáticas no admite estructuras discontinuas.

La definición de constituyente para el algoritmo admite, por lo tanto, la imbricación de constituyentes en otros constituyentes de mayor extensión, pero excluye explícitamente la oración entera (esto es, se elimina la secuencia delimitada por dos símbolos de fin de oración). Se trata de una simplificación operativa: aunque en sentido amplio la oración se pueda considerar como un constituyente, en el experimento de Clark se intenta hallar los constituyentes más básicos para construir reglas gramaticales. Nótese que en los casos liminares donde un constituyente es en sí mismo una oración, se lo sigue considerando como un constituyente; a la inversa, una oración breve que coincide con un constituyente no se convierte por ello en un constituyente.

Si bien es razonable pensar que las secuencias que forman un constituyente han de ocurrir frecuentemente en un texto, cabe aclarar que la mera frecuencia no garantiza que una secuencia sea una de las estructuras que este experimento se propone encontrar. Por ejemplo, la secuencia AT1 NN1 PRP (artículo singular-sustantivo singular-preposición) es mucho más frecuente que AT1 AV0 AJ1 NN1 (artículo singular-adverbio-adjetivo singular-sustantivo singular) y sin embargo, la secuencia AT1 NN1 PRP no es un constituyente, así como tampoco lo es ninguna de las altamente frecuentes secuencias terminadas en PRP. El caso extremo es la secuencia formada por una única PRP (preposición), que es la etiqueta más frecuente en la mayoría de los textos, pero no un constituyente. A la inversa, un constituyente extenso no deja de serlo por ser muy infrecuente. Es por ello que el corte por frecuencia es sólo el primer umbral del algoritmo de Clark.

Además, la longitud de la secuencia tampoco define el carácter de constituyente: NN1 (sustantivo singular) tiene la misma extensión que PRP y sin embargo es un constituyente.

Aunque existen limitaciones prácticas y teóricas, idealmente un experimento de este tipo ha de encontrar constituyentes continuos de cualquier extensión, particularmente cuando están compuestos de varios niveles imbricados de constituyentes breves, como en la *Figura 1*.

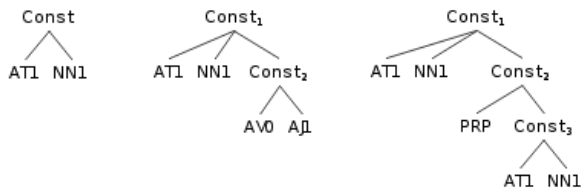



Figura 1: Constituyentes de a) 1 nivel, b) 2 niveles y c) 3 niveles de imbricación

## 2.2 Paso 1: perfil de frecuencias decrecientes de secuencias candidatas a constituyentes

Para el primer paso de su algoritmo, Clark lista las secuencias de etiquetas con una frecuencia mayor que el número de parámetros de la distribución que modela sus contextos. Utiliza 77 etiquetas, lo cual define una distribución de $77^2$ parámetros ≈ 5500, de modo que el piso de frecuencia para las secuencias seleccionadas es de 5000 ocurrencias en su *corpus*.

Si hubiéramos usado este criterio al adaptar el experimento de Clark al español, habríamos debido calcular el siguiente *umbral de ocurrencias u*:

$$u = \left( {tags^2}/{size} \right)^{ext} = \left( {48^2}/{240} \right)^{1,5} = 31 \text{ ocurrencias (i)}$$

donde distribución de símbolos en contexto *tags* = (48 etiquetas: $48^2 \approx 2300$)

tamaño de *corpus size* = 240 veces menor

extensión de oraciones *ext* = 1,5 veces mayor (promedio de longitud de oraciones)

Aunque el cálculo del umbral arrojaba el valor de 31 ocurrencias, decidimos experimentar con diversos escenarios de corte, entre 10 y 110 ocurrencias. De ese modo, podemos afinar a voluntad la base con la que el resto del algoritmo ha de trabajar, a la vez que nos mantenemos en el orden de valores sugeridos por la adaptación del *corpus* de Clark al nuestro. Con todo, consideramos que estos lineamientos en cuanto al umbral de ocurrencias son comparativamente significativos: Clark obtiene 753 secuencias candidatas, y en nuestro caso obtenemos 198 para el escenario más efectivo de 110 ocurrencias (véase *Tabla 4*).

Una idea importante que opera en el experimento de Clark reside en observar que varias secuencias que forman una misma clase de constituyentes (sintagma nominal, sintagma preposicional, etc.) aparecen en contextos similares, de modo que estudiar los contextos puede brindar información distribucional útil para tratar de detectar constituyentes automáticamente. La idea, en esencia, es la misma que subyace a las pruebas de sustitución para determinar si una secuencia de etiquetas conforma un constituyente o no.

Denominaremos *contexto previo* a la etiqueta que precede a la secuencia y *contexto posterior* a la etiqueta que le sigue. Esta información se puede modelar como dos distribuciones que indican cuántas veces aparece cada posible constituyente (compuesto de una secuencia de etiquetas) en cada contexto (compuesto por los pares posibles combinados de etiquetas anteriores y posteriores). Dado que en nuestro experimento hay 48 etiquetas, cada distribución tiene $48^2$ tales pares.

## 2.3 Paso 2: *Clustering* de secuencias candidatas a constituyentes

Una vez obtenida la anterior tabla de información distribucional, el siguiente paso en el algoritmo de Clark consiste en *arracimar* las secuencias de etiquetas en *clusters* o grupos afines. Para ello, considera que la información de la tabla representa la posición de cada secuencia en un espacio vectorial multidimensional, y que la afinidad entre secuencias se puede medir como la distancia que las separa.

> "If two sequences of tags occur mostly forming the same non-terminal, then we would expect the context that those strings occur in to be similar […] If we clustered sequences according to their distributions we would thus expect to find clusters corresponding to various syntactic constituents" [Clark 2001:132]

Clark sugiere como algoritmo de *clustering* el método iterativo *k-means* y usa la distancia euclidiana entre vectores. Como resultado de esta etapa, espera obtener varios grupos de secuencias que contengan constituyentes válidos, por un lado, y el resto de las secuencias, por el otro. Clark sostiene que es posible determinar automáticamente en cuáles de los *clusters* agrupados por información distribucional hay constituyentes válidos y en cuáles no. Este paso se entiende en el experimento de Clark como instancia previa a la inducción de una SCFG, que es el fin último de su investigación en los procesos de inducción de sintaxis, de modo tal que cada cluster válido resulte el germen para una categoría sintagmática mayor (sintagma nominal, sintagma preposicional, etc.).

| | 1 | 2 | 3 | 4 | … | 71 | … | 73 | … | 2203 | 2204 | … | 2303 | 2304 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **AJ0-AJ0** | **AJ0-AJ1** | **AJ0-AJ2** | **AJ0-AJC** | … | **AJ1-NN2** | … | **AJ1-NNP** | … | **VVZ-VVG** | **VVZ-VVI** | … | **$$$-XX0** | **$$$-$$$** |
| AT1 NN1 PRP AT1 NN1 PRP | 0.0 | 0.0 | 0.0 | 0.0 | … | 1.0 | … | 0.0 | … | 1.0 | 1.0 | … | 0.0 | 0.0 |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| NN1 PRP NNP | 0.0 | 1.0 | 0.0 | 0.0 | … | 0.0 | … | 4.0 | … | 0.0 | 0.0 | … | 0.0 | 1.0 |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |

Tabla 3: Tabla de información distribucional (secuencias y contextos)

No obstante, en nuestro caso, sólo estamos interesados en el sub-proceso de inducción de constituyentes (véase sección *Modificaciones al experimento original*).

## 2.4 Paso 3: Criterio de filtrado por información mutua entre etiquetas adyacentes a las secuencias candidatas a constituyentes

Una vez concluido el paso 2, Clark obtiene 100 *clusters*. En nuestro caso, como se ve en el *Anexo IV*, obtuvimos 25 *clusters*. Sin embargo, como Clark observa, esto no significa que todos los *clusters* agrupen constituyentes sintácticos:

> "As expected, the results of the clustering showed clear clusters corresponding to syntactic constituents […] of course, since we are clustering all of the frequent sequences in the corpus, we will also have clusters corresponding to parts of constituents […] we obviously would not want to hypothesize these as constituents: we therefore need some criterion for filtering out these spurious candidates." [Clark 2001:133]

Para determinar cuáles son los grupos de secuencias válidas como constituyentes, Clark propone un filtro basado en el grado de dependencia entre la etiqueta previa y la etiqueta posterior del contexto. La medición de la dependencia entre contextos consiste en estimar su información mutua (*mutual information* o *MI*):

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left( \frac{p(x,y)}{p_1(x)\, p_2(y)} \right) \quad \text{(ii)}$$

*Mutual information* (información mutua)

donde *I* es la información mutua; *X* e *Y* son las distribuciones de contextos previos y posteriores, respectivamente; *p(x,y)* es la probabilidad conjunta de dos etiquetas dadas de dichos contextos; $p_1(x)$ es la probabilidad marginal de ese contexto previo; y $p_2(y)$ es la probabilidad marginal de ese contexto posterior.

Nótese que en esta fórmula no se mide la interdependencia entre las etiquetas que pertenecen a la secuencia, ni la dependencia entre la secuencia y sus contextos. Esta fórmula de información mutua mide cuán dependientes entre sí son los contextos previo y posterior: una MI de 0 refleja total independencia, mientras que valores altos de MI reflejan cuánto disminuye nuestra perplejidad (Manning y Schütze 1999) cuando, conociendo una etiqueta, encontramos la otra.

Sin embargo, aunque la secuencia propiamente dicha no esté presente en la fórmula, influye en el cálculo. Dado que hay una cierta MI 'natural' entre dos símbolos cercanos cualesquiera de un lenguaje (Li 1990), y que esa MI disminuye a medida que la distancia entre los símbolos crece, la longitud de una secuencia determina la distancia entre sus contextos, de modo que se ha de tomar en cuenta en el cálculo de MI. En la *Figura 2* se puede observar la rápida caída en la curva de MI para distancias crecientes, donde una distancia de 2 símbolos corresponde a una secuencia de 1 etiqueta, una distancia de 3 símbolos, a una de 2 etiquetas, y así sucesivamente:

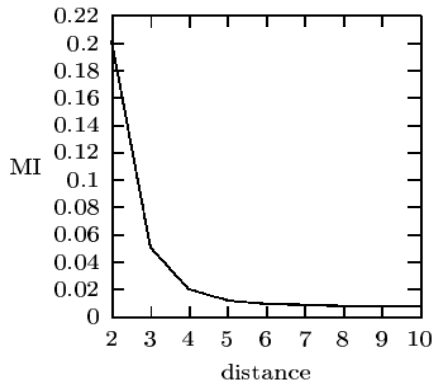



Figura 2: La información mutua entre el contexto previo y el contexto posterior desciende conforme crece la distancia que los separa (medida en símbolos) (Li 1990)

Por esta razón, Clark (2001) toma en cuenta como parámetro la distancia entre las etiquetas del contexto y postula que si en un *cluster* el promedio de la MI de los contextos –ponderado por la longitud de las secuencias– supera el umbral determinado por Li (1990), entonces dicho *cluster* es válido y probablemente agrupe secuencias que son constituyentes.

Cabe aclarar que las conclusiones de Li (1989 y 1990) competen a secuencias de caracteres. Clark (2001) extiende sus conclusiones a secuencias de etiquetas, considerando que una etiqueta funciona como un símbolo; de hecho, Li (1989) mismo ya había contemplado en su artículo la idea de ampliar la unidad de las secuencias de caracteres a palabras.

En la réplica del experimento hasta este punto obtuvimos resultados similares a los reportados por Clark (2001). Mediante una evaluación manual de los *clusters* determinamos una medida F (promedio armónico entre precisión y cobertura) del 65% (véase *Tabla 4*).

## 2.5 Modificaciones al experimento original

En el experimento original, Clark (2001) agrupa las secuencias en *clusters* con el propósito de inducir símbolos no-terminales automáticamente. Una vez detectados, estos símbolos le son de utilidad para inducir reglas gramaticales. Por esta razón, aplica el filtro de MI sobre el *cluster* entero, ya que la estimación de MI resulta más precisa si se hace sobre la variedad de ocurrencias de etiquetas contenidas en ese grupo.

Es legítimo preguntarse qué pasa si en lugar de aplicar el filtro de MI sobre un *cluster* lo aplicamos sobre cada secuencia. Ello implicaría dejar de lado el objetivo de inducir símbolos no-terminales y reglas gramaticales, pero por otro lado brindaría la posibilidad de determinar en forma más o menos inmediata si una secuencia dada y de ocurrencia frecuente es un constituyente, lo cual reviste utilidad práctica. Para ello, es preciso modificar la manera de estimar su MI: en lugar de un promedio sobre todo el *cluster*, usamos la fórmula para MI punto a punto (*pointwise MI*), calculada sobre un promedio entre secuencias de la misma longitud:

$$MI(X;Y) = \log \frac{p(x,y)}{p(x)p(y)} \qquad \text{(iii)}$$

*Pointwise mutual information*
(información mutua punto a punto)

En la *Figura 3* se puede apreciar un diagrama que compara el experimento original de Clark y nuestra adaptación al español. Clark agrupa secuencias en *clusters* y luego determina en cuáles hay constituyentes mediante el cálculo de MI de cada *cluster*, mientras que en nuestro caso calculamos MI de cada secuencia y luego las agrupamos en *clusters* de constituyentes de similar distribución. Nuestro paso de *clustering* no es parte esencial del criterio de definición de constituyentes, sino simplemente una forma de asegurarnos la viabilidad del proceso al demostrar convergencia con los resultados del algoritmo original.

## *3. Evaluación*

Resumiendo la descripción de nuestro experimento:

i. Dividimos las 2108 oraciones en dos grupos: 2000 oraciones para entrenamiento y 108 para la aplicación de los constituyentes inducidos (véase *Anexo VI*).
ii. Definimos un umbral de frecuencia de 110 ocurrencias para el paso 1 del algoritmo: obtuvimos 198 secuencias candidatas a constituyentes.
iii. Aplicamos el criterio de MI punto a punto a las secuencias candidatas de (i.) (paso 3 del algoritmo): obtuvimos 107 secuencias validadas como constituyentes.
iv. Arracimamos las 107 secuencias con *clustering* basado en *k-means*: obtuvimos 25 *clusters* de alta pureza (véase *Anexo IV*).
v. Repetimos el experimento con distintos umbrales en (i.), con los resultados de la *Tabla 4*.
vi. Evaluamos manualmente la salida de (iii.) con nuestros propios juicios de gramaticalidad (véase *Anexo V*), de modo de calcular la medida F para cada escenario de (v.): obtuvimos distintos valores para la columna de constituyentes válidos (*n positivos* en la *Tabla 4*).

Las evaluaciones del algoritmo original de Clark y de nuestra implementación revelaron que ambos métodos convergen a resultados similares. Obtuvimos una medida F de alrededor del 65% para el escenario más efectivo de nuestro experimento (véase *Tabla 4*). En cuanto a la extensión y la calidad de los constituyentes, el listado de constituyentes inducidos abarca no sólo casos obvios con etiquetas triviales, sino que, sorprendentemente, en muchos casos se han inducido correctamente constituyentes con etiquetas poco frecuentes (por ejemplo, ocurrencias de CRD –adjetivo cardinal– bien integradas a sintagmas nominales). En otros casos, la extensión de los constituyentes llega a 5 y 6 etiquetas (véase *Anexo V*), lo que demuestra la viabilidad del enfoque para inducción de complejas estructuras de constituyentes.

Precisión

$$P = \frac{n°\,ConstituyentesPositivos}{n°\,ConstituyentesPositivos + FalsosPositivos} \quad \text{(iv)}$$

Cobertura

$$C = \frac{n°\,ConstituyentesPositivos}{n°\,ConstituyentesPositivos + FalsosNegativos} \quad \text{(v)}$$

Medida $F = \frac{(\beta^2 + 1) * P * C}{\beta^2 * P + C}$ (vi)

(Con $\beta = 1$ para asignar igual peso a $P$ y a $C$)

| umbral de ocurrencias | nº Positivos (Paso3) | Candidatos (Paso1) | Precisión % | Cobertura % | medida F % |
|---|---|---|---|---|---|
| **110** | 62 | 198 | 59 | 74 | 66 |
| **50** | 125 | 464 | 58 | 65 | 61 |
| **30** | 166 | 786 | 53 | 53 | 53 |
| **15** | 242 | 1561 | 49 | 41 | 45 |
| **10** | 291 | 2429 | 51 | 32 | 39 |

Tabla 4: Evaluación de la medida F para distintos escenarios de experimentación según umbral de ocurrencias

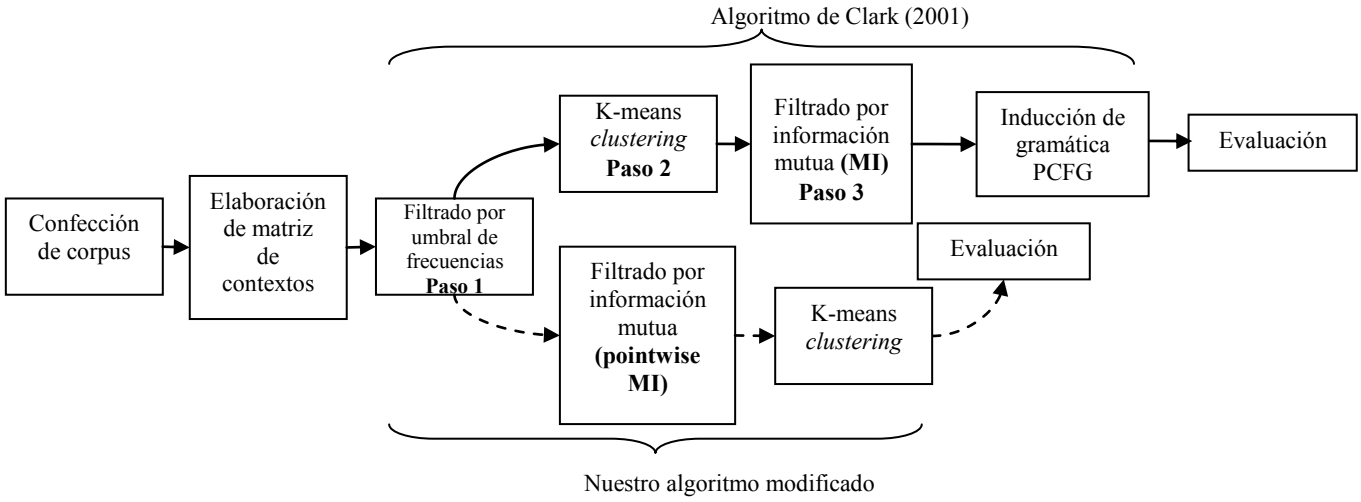



Figura 3: Experimento original de Clark (2001) y nuestra adaptación al español

## *4. Conclusiones y trabajo a futuro*

El presente trabajo verifica la factibilidad de encarar la tarea de inducción de constituyentes sintácticos en español con un enfoque basado en la información distribucional. Además, el experimento ofrece interesantes implicancias acerca de las propiedades formales extrínsecas de los constituyentes sintácticos. Si bien un constituyente es primariamente una secuencia de símbolos más o menos frecuente en la distribución de un *corpus* morfosintácticamente etiquetado, esta condición no es suficiente para definir un constituyente. Más bien, lo que el experimento refleja es que el verdadero filtro entre las secuencias frecuentes de etiquetas, candidatas a ser constituyentes, es la información mutua entre los símbolos que co-ocurren en las adyacencias de dichas secuencias.

Ahora bien, en nuestro experimento prototípico también tropezamos con algunos obstáculos. Por un lado, nos encontramos con el consabido problema de la sobreestimación del rol de la información mutua (Li 1990) y de la dispersión de datos en los modelos estadísticos (Fong y Berwick 2008), situación que se ve agravada al trabajar con un *corpus* de implementación prototípica que implica dimensiones no tan masivas. Esto explica por qué la medida F decrece tanto cuando el umbral de aceptación de candidatos a constituyentes baja a apenas 15 ocurrencias (véase *Tabla 4*).

Por otro lado, como el mismo Clark (2001) reconoce, el modelo falla en capturar como constituyentes secuencias de etiquetas de muy rara ocurrencia. Esto se condice con la extensión y composición de los constituyentes inducidos (véase *Anexo V*). Los constituyentes extensos (5 o más etiquetas), en general, pueden describirse como constituyentes cortos de etiquetas frecuentes imbricados en otros. Es decir, existe la tendencia a que los constituyentes más extensos se compongan de las etiquetas más frecuentes. Esto se verifica, por ejemplo, en la dificultad que encuentra el experimento para modelar proposiciones subordinadas o constituyentes en los que entran en juego etiquetas menos frecuentes.

El experimento nos revela una importante veta de indagación científica que obliga a replantearse cuestiones tan sensibles para la lingüística como la naturaleza del lenguaje y los mecanismos de adquisición del mismo, a la luz de las promisorias técnicas de aprendizaje de máquina y de los procesos de inducción de gramáticas.

En cuanto al trabajo a futuro, nos trazamos los siguientes ejes de organización. Primero, resulta fundamental continuar con los trabajos de revisión y ampliación del *corpus*, hasta alcanzar dimensiones apropiadas para una experimentación con contundencia científica, más allá de la etapa exploratoria del prototipo. A su vez, debemos refinar y homologar aún más los criterios de anotación para los eventuales colaboradores del proyecto, a fin de alcanzar dicha masividad en el *corpus*.

En segundo término, notamos la necesidad de refinar y validar el algoritmo de inducción, experimentando con otras métricas derivadas de la teoría de la información, tales como la distancia de divergencia Kullback-Leibler (Manning y Schütze 1999). Asimismo, se hace imprescindible considerar otros enfoques propuestos para el problema de la inducción de constituyentes sintácticos y evaluar la factibilidad de los mismos en un *corpus* en español. Los trabajos más conocidos para dicha tarea son los modelos de Klein y Manning (2002 y 2004), como así también trabajos provenientes de otros paradigmas de investigación como el conexionismo (Reali *et al.* 2003). En particular, una muy fructífera línea de investigación podría ser la combinación de modelos de inducción de constituyentes lineales, como el que Clark (2001) propone, con modelos de dependencias sintácticas (Mel'čuk 1988; Paskin 2002; Klein y Manning 2004).

Finalmente, en lo que atañe al objetivo más ambicioso de trabajo a futuro, nos proponemos continuar con una investigación integral para analizar la factibilidad de inducir una gramática independiente de contexto completa del español en relación con la toma de una postura en el debate epistemológico actual entre sesgos fuertes *versus* débiles (*weak bias* y *strong bias*, respectivamente) como componente innato en la adquisición del lenguaje (Lappin y Schieber 2007). Para encarar dicho objetivo, es importante segmentar el proceso total de inducción en tareas parciales, una de las cuales, en principio,

bien podría ser la inducción de constituyentes sintácticos que estamos describiendo en este artículo. De hecho, en investigaciones previas (Balbachan y Dell'Era 2008) probamos la plausibilidad de la inducción de categorías sintácticas en español a partir de un algoritmo no supervisado, tarea cuya salida podría ser aprovechada para que sea automáticamente procesada como los datos de entrada del proceso de inducción de constituyentes sintácticos.

Consideramos que el mérito de la presente implementación prototípica es experimentar con modelos de inducción de fenómenos sintácticos que puedan aportar renovada evidencia al debate acerca de la adquisición del lenguaje; en especial, si consideramos que la investigación de este tipo de enfoques para el español –un idioma particularmente desafiante por el orden libre de sus constituyentes sintácticos– ha venido escaseando durante la última década en el panorama global del estado del arte dentro del paradigma estadístico de la lingüística computacional. En última instancia, la evidencia psicolingüística debería ser refrendada por la neurología o incluso la biolingüística, pero la plausibilidad de dicha evidencia mediante una modelización efectiva es asunto para la agenda actual de la lingüística computacional.

## *Anexo I Listado completo de etiquetas morfosintácticas para anotación de corpus*

El conjunto de etiquetas que usaremos se basa en el llamado *C5 tagset*, un estándar que se aplicó al etiquetado del *British National Corpus* (*BNC*). Como etiquetamos texto en español, prescindimos de algunas etiquetas específicas del idioma inglés y agregamos otras más apropiadas.

| Nro. | *Tag* | Ejemplos |
|---|---|---|
| 1 | AJ0 | adjetivo neutro en número (*bello* en "lo bello") |
| 2 | AJ1 | adjetivo singular (*amable*) |
| 3 | AJ2 | adjetivo plural (*amables*) |
| 4 | AJC | adjetivo comparativo (*peor*) |
| 5 | AJS | adjetivo superlativo (*pésimo*) |
| 6 | AT0 | artículo neutro (*lo*) |
| 7 | AT1 | artículo singular (*la*) |
| 8 | AT2 | artículo plural (*los*) |
| 9 | AV0 | adverbio (*seguidamente*) |
| 10 | AVQ | adverbio interrogativo (*cuándo*) |
| 11 | CJC | conjunción coordinante (*y*, *así que*, *luego*) |
| 12 | CJS | conjunción subordinante (excepto *que*) |
| 13 | CJT | conjunción subordinante *que* (en "Dijo que...") |
| 14 | CRD | adjetivo numeral cardinal (*tres*) |
| 15 | DAT | fecha (*7 de noviembre*) |
| 16 | DPS | determinante posesivo (*su*, *mi*) |
| 17 | DT1 | determinante definido singular (*aquel* hombre) |
| 18 | DT2 | determinante definido plural (*aquellos* hombres, *todos* los hombres) |
| 19 | EX0 | existencial *hay* |
| 20 | ITJ | interjección (*ah*, *ehmm*) |
| 21 | NN0 | sustantivo neutro en número (*virus*) |
| 22 | NN1 | sustantivo singular (*lápiz*) |
| 23 | NN2 | sustantivo plural (*lápices*) |
| 24 | NNP | sustantivo propio (*Buenos_Aires*) |
| 25 | ORD | adjetivo numeral ordinal (*sexto*, *3ro.*, *último*) |
| 26 | PND | pronombre demostrativo (¿Cuál querés? *Éste*, ¿Cuál querés? *Esto*) |
| 27 | PNI | pronombre indefinido (*ninguno*, *todo*) |
| 28 | PNP | pronombre personal (*tú*) |
| 29 | PNQ | pronombre interrogativo (*quién*) |
| 30 | POS | pronombre posesivo (*mío*) |
| 31 | PPE | pronombre personal enclítico (dar\ *lo*, *se* cuasi-reflejo (*morirse*, él *se* cayó) |
| 32 | PRP | preposición (excepto *de*) (*sin*) |
| 33 | REL | pronombre relativo (*quien* en "el presidente, quien avisó...") |
| 34 | SEP | *se* pasivo ("se venden casas") e impersonal ("se reprimió a los manifestantes") |
| 35 | VBG | gerundio de verbo cópula (*siendo*) |
| 36 | VBI | infinitivo de verbo cópula (*ser*) |
| 37 | VBN | participio de verbo cópula (*sido*) |
| 38 | VBZ | verbo cópula conjugado (*es*) |

| | | |
|---|---|---|
| 39 | VM0 | infinitivo de verbo modal (*soler*) |
| 40 | VMZ | verbo modal conjugado (*debía*) |
| 41 | VMG | gerundio de verbo modal (*pudiendo*) |
| 42 | VMN | participio de verbo modal |
| 43 | VVG | gerundio de verbo léxico (*obrando*) |
| 44 | VVI | infinitivo de verbo léxico (*vivir*) |
| 45 | VVN | participio de verbo léxico (*cifrado*) |
| 46 | VVZ | verbo léxico conjugado (*vive*) |
| 47 | XX0 | adverbio de negación (*no*) |
| 48 | $$$ | fin de oración |

## *Anexo II Composición del corpus*

longitud promedio de oraciones: 23,68
cantidad de oraciones: 2.108
cantidad de palabras etiquetadas: 49.925

| etiqueta | n | % |
|---|---|---|
| PRP | 9613 | 19.25 |
| NN1 | 8280 | 16.58 |
| AT1 | 6484 | 12.99 |
| VVZ | 3857 | 7.73 |
| NN2 | 3409 | 6.83 |
| NNP | 2405 | 4.82 |
| AJ1 | 1897 | 3.80 |
| AV0 | 1610 | 3.22 |
| CJC | 1574 | 3.15 |
| AT2 | 1501 | 3.01 |
| CRD | 1056 | 2.12 |
| VVI | 902 | 1.81 |
| AJ2 | 851 | 1.70 |
| VVN | 847 | 1.70 |
| REL | 835 | 1.67 |
| CJT | 730 | 1.46 |
| PPE | 621 | 1.24 |
| VBZ | 571 | 1.14 |
| DPS | 424 | 0.85 |
| SEP | 347 | 0.70 |
| CJS | 250 | 0.50 |
| DT1 | 250 | 0.50 |
| XX0 | 226 | 0.45 |
| ORD | 153 | 0.31 |
| PNP | 135 | 0.27 |
| AJ0 | 127 | 0.25 |
| VMZ | 118 | 0.24 |
| PNI | 118 | 0.24 |
| DAT | 113 | 0.23 |
| VVG | 112 | 0.22 |
| DT2 | 104 | 0.21 |
| PND | 103 | 0.21 |
| AT0 | 64 | 0.13 |
| EX0 | 58 | 0.12 |
| NN0 | 54 | 0.11 |
| VBI | 47 | 0.09 |
| AJC | 46 | 0.09 |
| PNQ | 11 | 0.02 |
| AJS | 8 | 0.02 |
| VBG | 7 | 0.01 |
| AVQ | 5 | 0.01 |
| VMN | 1 | 0.00 |
| ITJ | 0 | 0.00 |
| POS | 0 | 0.00 |
| VBN | 0 | 0.00 |
| VM0 | 0 | 0.00 |
| VMG | 0 | 0.00 |

## *Anexo III Ejemplo de anotación para el corpus*

CONTRA LOS ACUSADOS
AMIA: piden que se amplíe un embargo
El/AT1 fiscal/NN1 Alberto_Nisman/NNP le/PNP pidió/VVZ a/PRP el/AT1 juez/NN1 federal/AJ1 Rodolfo_Canicoba_Corral/NNP ampliar/VVI a/PRP 540/CRD millones/NN2 de/PRP dólares/NN2 el/AT1 embargo/NN1 contra/PRP los/AT2 iraníes/NN2 acusados/VVN de/PRP haber_planeado/VVI y/CJC
ordenado/VVI la/AT1 ejecución/NN1 de/PRP el/AT1 atentado/NN1 terrorista/AJ1 contra/PRP la/AT1 AMIA/PNP ./$
El/AT1 monto/NN1 --expresado en pesos significan 1.843 millones-- surge/VVZ a/PRP el/AT1 considerar/VVI todos/DT2 los/AT2 daños/NN2 provocados/VVN por/PRP el/AT1 atentado/NN1 de/PRP 1994/NN0 , incluido/VVN un/AT1 resarcimiento/NN1 para/PRP los/AT2 familiares/NN2 de/PRP los/AT2 85/CRD muertos/NN1 y/CJC para/PRP los/AT2 más/AV0 de/PRP 200/CRD heridos/NN2 que/REL dejó/VVZ el/AT1 ataque/NN1 ./$

## *Anexo IV Paso 2 del algoritmo de Clark (2001) aplicado a nuestro corpus: clustering de secuencias candidatas*

## Anexo IV: Paso 2 del algoritmo Clark (2001): clustering de secuencias candidatas a constituyentes

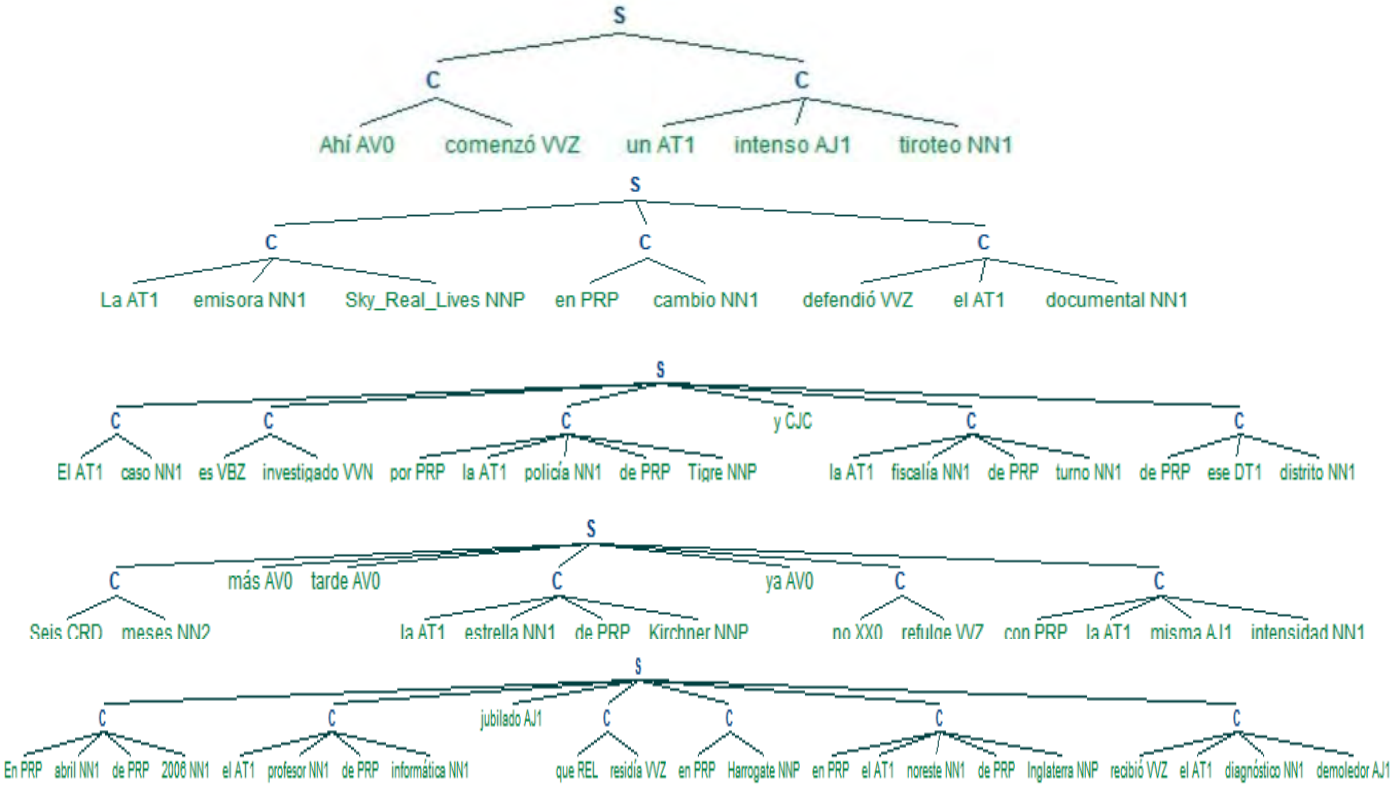
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| PRP DT1 | AV0 AT1NN1 | NNP NNP | PRP AT1NN1PRP AT1NN1 | XX0 | NN2 PRP AT1 | REL VVZ | NN1AT1NN1 | VVZ |
| AV0 PRP AT1 | PRP VVI | AT1NN1CJC | PRP AT1NN1PRP NN1 | SEP | AT1AJ1 | AJ1CJC | AJ1NN1 | VVZ PRP AT1NN1 |
| PRP DPS | AV0 | NNP CJC | VVN | PPE | AT1 | CJC | NN1AV0 | VBZ VVN |
| PRP NN1PRP AT1 | CJT VVZ | | PRP NN2 AJ2 | PNP | CJT AT1 | PRP AT1NN1CJC | NN1 | VVZ AV0 |
| PRP VVI AT1 | CJT AT1NN1 | | PRP AT1NN1NNP | | AT1NN1PRP AT1 | CJC VVZ | NN1PRP AT1NN1AJ1 | VVZ NN2 |
| AV0 AT1 | CJT | | PRP AT1NN1PRP NNP | | DT1 | | NN1AJ1 | VVZ PRP VVI |
| PRP AT1 | | | PRP CRD NN2 | | DPS | | NN1PRP AT1NN1 | PPE VVZ |
| PRP AT1NN1PRP AT1 | | | CJC NN2 | | NNP PRP AT1 | | NN1PRP CRD NN2 | VVZ AT1NN1 |
| PRP AT1AJ1 | | | PRP AT1NN1 | | NNP AT1 | | DAT | VVZ VVN |
| | | | CJC AT1NN1 | | AT2 NN2 PRP AT1 | | NN1PRP NNP | |
| | | | AV0 PRP AT1NN1 | | VVI AT1 | | NN1PRP NN1 | |
| | | | PRP AT1NNP | | | | NN1NNP | |
| | | | PRP AT2 NN2 | | | | | |
| | | | PRP NN1PRP AT1NN1 | | | | | |
| | | | AJ1PRP AT1NN1 | | | | | |
| | | | PRP NN1AJ1 | | | | | |
| | | | PRP AT1NN1AJ1 | | | | | |
| | | | PRP NNP | | | | | |
| | | | VVN PRP AT1NN1 | | | | | |
| | | | PRP NN1 | | | | | |
| | | | PRP AT1AJ1NN1 | | | | | |
| | | | PRP DPS NN1 | | | | | |
| | | | PRP AV0 | | | | | |
| | | | PRP NN2 | | | | | |

| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|
| NN2 PRP | REL VVZ PRP | AT1NN1PRP AT2 | NN2 PRP AT2 | NN2 CJC | NN2 AJ2 | VVZ AT1 | NN1PRP AT2 NN2 | NN1PRP CRD |
| AT1NN1PRP | VBZ VVN PRP | AT2 | | NN2 REL | AJ2 NN2 | CJC AT1 | NN1PRP NN2 | NN1REL VVZ |
| AT2 NN2 PRP | VVZ PRP | | | | NN2 PRP AT1NN1 | VVZ AT1NN1PRP AT1 | NN1VVN | NN1VBZ |
| NNP PRP | AT1NN1VVZ PRP | | | | NN2 | AJ1PRP AT1 | | NN1CJC |
| CRD NN2 PRP | AJ2 PRP | | | | NN2 PRP NN2 | VVN PRP AT1 | | NN1PRP AT2 |
| NN2 PRP AT1NN1PRP | VVZ PRP AT1NN1PRP | | | | NN2 PRP NN1 | VVZ PRP AT1 | | NN1VVZ |
| NN2 AJ2 PRP | NNP VVZ PRP | | | | NN2 AV0 | | | |
| AT1NN1PRP NN1PRP | PRP NN1PRP | | | | NN2 VVZ | | | |
| VVI PRP | PRP NNP PRP | | | | | | | |
| AT1NN1AJ1PRP | CRD PRP | | | | | | | |
| AT1NN1PRP AT1NN1PRP | PRP AT2 NN2 PRP | | | | | | | |
| | PRP NN2 PRP | | | | | | | |
| | PRP AT1NN1PRP | | | | | | | |
| | VVN PRP | | | | | | | |
| | PRP VVI PRP | | | | | | | |
| | PRP | | | | | | | |
| | CJC PRP | | | | | | | |
| | SEP VVZ PRP | | | | | | | |
| | AV0 PRP | | | | | | | |
| | PRP AT1NN1AJ1PRP | | | | | | | |
| | VVZ AT1NN1PRP | | | | | | | |
| | PPE VVZ PRP | | | | | | | |
| | AJ1PRP | | | | | | | |

# *Anexo V Muestra de la salida final del experimento con constituyentes filtrados*

| | secuencia | longitud | MI max | MI argmax | MI promedio |
|---|---|---|---|---|---|
| | AJ1 NN1 | 2 | 8.345 | AT1--CJC | 0.053 |
| error | AJ1 PRP | 2 | 10.122 | NN1--AT1 | 0.050 |
| error | AJ1 PRP AT1 NN1 | 4 | 8.645 | NN1--AJ1 | 0.014 |
| error | AT1 | 1 | 11.372 | PRP--REL | 0.242 |
| | AT1 AJ1 NN1 | 3 | 8.129 | PRP--PRP | 0.040 |
| | AT1 NN1 | 2 | 11.325 | PRP--CJC | 0.188 |
| | AT1 NN1 AJ1 | 3 | 9.972 | PRP--PRP | 0.052 |
| error | AT1 NN1 AJ1 PRP | 4 | 8.536 | PRP--AT2 | 0.072 |
| error | AT1 NN1 PRP | 3 | 10.391 | PRP--NN2 | 0.080 |
| error | AT1 NN1 PRP AT1 | 4 | 10.411 | PRP--NN1 | 0.102 |
| | AT1 NN1 PRP AT1 NN1 | 5 | 8.203 | PRP--AV0 | 0.077 |
| error | AT1 NN1 PRP AT1 NN1 PRP | 6 | 8.009 | PRP--AT1 | 0.022 |
| | AT1 NN1 PRP NN1 | 4 | 9.223 | PRP--AJ1 | 0.055 |
| | AT1 NN1 PRP NN2 | 4 | 6.180 | PRP--VVZ | 0.017 |
| | AT1 NN1 PRP NNP | 4 | 8.917 | PRP--VVZ | 0.047 |
| | AT1 NN1 VVN | 3 | 8.387 | PRP--PRP | 0.021 |
| | AT1 NN1 VVZ | 3 | 7.885 | $$$--PRP | 0.050 |
| error | AT1 NN1 VVZ PRP | 4 | 8.151 | PRP--VVI | 0.064 |
| | AT2 NN2 | 2 | 9.409 | PRP--CJC | 0.060 |
| error | AT2 NN2 PRP | 3 | 8.630 | PRP--ORD | 0.032 |
| error | AT2 NN2 PRP AT1 | 4 | 8.277 | PRP--NN1 | 0.043 |
| | AT2 NN2 PRP AT1 NN1 | 5 | 5.741 | PRP--$$$ | 0.037 |
| | AT2 NN2 VVZ | 3 | 5.964 | PRP--NN2 | 0.029 |
| error | AV0 | 1 | 10.030 | VVZ--PNI | 0.313 |
| error | AV0 AT1 NN1 | 3 | 5.991 | $$$--AJ1 | 0.032 |
| error | AV0 PRP | 2 | 9.209 | VVZ--DPS | 0.081 |
| | AV0 PRP AT1 NN1 | 4 | 6.107 | PRP--AJ1 | 0.032 |
| | AV0 VVZ | 2 | 7.798 | PRP--DPS | 0.049 |
| error | CJC | 1 | 9.969 | NN1--PNI | 0.270 |
| error | CJC AT1 NN1 | 3 | 8.002 | NN2--ORD | 0.036 |
| | CJC VVZ | 2 | 8.537 | NN1--DPS | 0.046 |
| error | CJT AT1 NN1 | 3 | 6.951 | VVZ--PRP | 0.022 |
| | CJT VVZ | 2 | 7.798 | PRP--DPS | 0.045 |
| | CRD NN2 | 2 | 8.979 | PRP--CJC | 0.061 |
| | NN1 | 1 | 12.421 | AT1--REL | 0.344 |
| | NN1 AJ1 | 2 | 11.061 | AT1--CJC | 0.071 |
| error | NN1 AJ1 PRP AT1 | 4 | 8.852 | AT1--NN1 | 0.023 |
| error | NN1 AT1 NN1 | 3 | 8.630 | PRP--ORD | 0.035 |
| error | NN1 PRP | 2 | 11.943 | AT1--NN0 | 0.129 |
| error | NN1 PRP AT1 | 3 | 11.137 | AT1--ORD | 0.049 |
| | NN1 PRP AT1 NN1 | 4 | 10.639 | AT1--AJ1 | 0.082 |
| | NN1 PRP AT1 NN1 AJ1 | 5 | 6.965 | AT1--$$$ | 0.036 |
| error | NN1 PRP AT1 NN1 PRP AT1 | 6 | 5.970 | AT1--REL | 0.010 |
| error | NN1 PRP AT2 | 3 | 10.419 | AT1--NN2 | 0.046 |
| | NN1 PRP AT2 NN2 | 4 | 6.807 | AT1--VVZ | 0.021 |
| | NN1 PRP CRD | 3 | 10.185 | AT1--NN2 | 0.023 |
| | NN1 PRP CRD NN2 | 4 | 7.544 | AT1--VVZ | 0.013 |
| | NN1 PRP NN1 | 3 | 9.554 | AT1--PRP | 0.049 |

## *Anexo VI Muestra de constituyentes inducidos sobre algunas oraciones de prueba*




## *Bibliografía*


Balbachan, Fernando y Diego Dell'Era. 2008. Técnicas de clustering para inducción de categorías sintácticas en un corpus de español. En *Infosur* (2):95-104.

Carroll, Glenn y Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. En C. Weir, S. Abney, R. Grishman y R. Weischedel (eds.). *Working notes of the workshop statistically-based NLP techniques*. AAAI Press.

Chater, Nick y Christopher Manning. 2006. Probabilistic models of language processing and acquisition. En *Trends in Cognitive Sciences* (10):335-344.

Chen, Stanley. 1995. Bayesian grammar induction for language modeling. En *ACL* (33):228–235.

Chomsky, Noam. 1957. *Estructuras sintácticas*. México. Siglo XXI.

Chomsky, Noam. 1959. A review of B. F. Skinner's 'verbal behavior'. En *Language* (35):26–58.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA. MIT Press.

Chomsky, Noam. 1966. *Topics in the Theory of Generative Grammar*. París. Mouton.

Chomsky, Noam. 1986. *El conocimiento del lenguaje*. Madrid. Alianza.

Civit i Torruella, Montserrat. 2003. *Criterios de etiquetación morfosintáctica de corpus en español*. Tesis de doctorado, Universidad de Barcelona.

Clark, Alexander. 2001. Unsupervised language acquisition: theory and practice. Sussex. School of Cognitive and Computing Sciences, University of Sussex Press.

Clark, Eve. 2009. *First Language Acquisition*. Cambridge, Inglaterra. Cambridge University Press.

Cowie, Fiona. 1999. *What's Within? Nativism Reconsidered*. Oxford. Oxford University Press.

Finch, Steve, Nick Chater y Martin Redington. 1995. Acquiring syntactic information from distributional statistics. En J. P. Levy, D. Bairaktaris, J. A. Bullinaria y P. Cairns (eds.). *Connectionist Models of Memory and Language*. Londres. UCL Press.

Fodor, Jerry. 1983. *La modularidad de la mente*. Madrid. Morata.

Fong, Sandiway y Robert Berwick. 2008. Treebank parsing and knowledge of language: a cognitive perspective. En *Proceedings of the 30th annual conference of the Cognitive Science Society*:539-544. Austin, Texas.

Gambino, Omar J. y Hiram Calvo. 2007. On the usage of morphological tags for grammar induction. En Alexander Gelbukh y A. F. Kuri

Morales (eds.). *MICAI 2007, LNAI 4827*: 912–921. Berlín. Springer-Verlag.

Gold, E. Mark. 1967. Language identification in the limit. En *Information and control* (10):447-474.

Harris, Zellig S. 2000 (1951). *Structural Linguistics*. University of Chicago Press.

Klein, Dan y Christopher Manning. 2001. Distributional phrase structure induction. En *Proceedings of CoNLL 2001*:113-121.

Klein, Dan y Christopher Manning. 2002. A generative constituent-context model for improved grammar induction. En *Proceedings of ACL 2002*:128-135. Philadelphia.

Klein, Dan y Christopher Manning. 2004. Corpus based induction of syntactic structure: models of dependency and constituency. En *Proceedings of ACL 2004*:478-485. Barcelona.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things*. The University of Chicago Press, Chicago.

Langacker, Ronald. 2000. Estructura de la cláusula en la gramática cognoscitiva. En *Volumen monográfico 2000:*19-65. Universidad de California, San Diego.

Lappin, Shalom y Stuart Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. En *Linguistics* (43):393-427.

Lari, Karim y Steven Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language (4):35–56.

Leech, Geoffrey, Roger Garside y Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. Reporte técnico, Lancaster University.

Li, Wentian. 1989. Mutual information functions of natural language texts. Santa Fe Institute preprint.

Li, Wentian. 1990. Mutual information functions versus correlation functions. En *Journal of statistical physics* (60):823-837.

Manning, Christopher y Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts. MIT Press.

Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany, NY.

Meyer, Charles. 2002. *English Corpus Linguistics*. Cambridge, Inglaterra. Cambridge University Press.

Moreno Sandoval, Antonio. 1998. *Lingüística computacional*. Madrid. Síntesis.

Moreno Sandoval, Antonio, Susana López Ruesga y Fernando Sánchez León. 1999. Spanish Treebank. Reporte técnico, versión 5. Universidad Autónoma de Madrid.

Olivier, Donald. 1968. *Stochastic grammars and language acquisition mechanisms*. Tesis de doctorado, Harvard University.

Paskin, Mark A. 2002. Grammatical bigrams. En T. G. Dietterich, S. Becker y Z. Ghahramani (eds.) *Advances in Neural Information Processing Systems 14*. Cambridge, Massachusetts. MIT Press.

Piattelli-Palmarini, Massimo (ed.). 1980. *Language and Learning: the Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA. Harvard University Press.

Pinker, Steven. 1979. Formal models of language learning. En *Cognition* (7):217–282.

Pinker, Steven. 1994. *El instinto del lenguaje*. Madrid. Alianza.

Pullum, Geoffrey. 1996. Learnability, hyperlearning and the argument from the poverty of the stimulus. En *Parasession on Learnability, 22nd Annual Meeting of the Berkeley Linguistics Society*. Berkeley, California.

Pullum, Geoffrey y Barbara Scholz. 2002. Empirical assessment of stimulus poverty arguments. En *The Linguistic Review* (19):9-50.

Reali, Florencia, Morten Christiansen y Padraic Monaghan. 2003. Phonological and distributional cues in syntax acquisition: scaling-up the connectionist approach to multiple-cue integration. En *Proceedings of the 25th annual conference of the Cognitive Science Society Lawrence Erlbaum Associates, Inc*.:970-975. Mahwah, New Jersey.

Redington, Martin, Nick Chater, Chu-Ren Huang, Li-Pin Chang, Steve Finch y Keh-jiann Chen. 1995. The universality of simple distributional methods: identifying syntactic categories in Chinese. En *Proceedings of the Cognitive Science of Natural Language Processing*. Dublín.

Sánchez León, Fernando. 1994. Spanish tagset for the CRATER project. Reporte técnico, Universidad Autónoma de Madrid.

Santorini, Beatrice. 1991. Part-of-speech tagging guidelines for the Penn treebank project. Reporte técnico MS-CIS-90-47, University of Pennsylvania.

Van Zaanen, Menno. 2000. ABL: Alignment-based learning. En *COLING* (18):961–967.

Wolff, J. Gerard. 1988. Learning syntax and meanings through optimization and distributional analysis. En Y. Levy, I. M. Schlesinger y M. D. S. Braine (eds.). *Categories and Processes in Language Acquisition*. Lawrence Erlbaum, Hillsdale, NJ.

# Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação

Marcos Garcia
Universidade de Santiago de Compostela
marcosgg@gmail.com

Pablo Gamallo
Universidade de Santiago de Compostela
pablo.gamallo@usc.es

**Resumo**

As diferentes tarefas de análise morfossintáctica têm muita importância para posteriores níveis do processamento da linguagem natural. Por isso, estes processos devem ser realizados com ferramentas que garantam bons desempenhos em relação à cobertura, precisão e robustez na análise. FreeLing é uma *suite* com licença GPL desenvolvida pelo Grupo TALP da Universitat Politècnica de Catalunya. Este *software* contém —entre outros— módulos de tokenização, segmentação de orações, reconhecimento de entidades e anotação morfossintáctica. Com o fim de obtermos ferramentas que nos sirvam de base para a análise sintáctica, bem como para disponibilizar *software* livre para o processamento de superfície de Português Europeu e Galego, adaptámos FreeLing para estas variedades. A primeira delas foi desenvolvida com ajuda de recursos linguísticos disponíveis *on-line*, enquanto os ficheiros do Galego tiveram como base a versão anterior de FreeLing (criados pelo Seminario de Lingüística Informática da Universidade de Vigo), que já realizava a análise desta língua. O presente trabalho descreve os principais aspectos da adaptação das ferramentas, com ênfase nos problemas encontrados e nas soluções adoptadas em cada caso. Além disso, são apresentados os resultados de avaliação do módulo PoS-tagger.

## 1 *Introdução*

Os diferentes processos que compõem a análise morfossintáctica constituem uma etapa com enorme importância no processamento da linguagem natural. Tarefas como a recuperação de informação, a análise sintáctica ou a síntese de voz, por exemplo, precisam de um processamento prévio que seja capaz de segmentar orações, reconhecer tokens e inferir os seus lemas ou atribuir uma categoria morfossintáctica (PoS) a cada um deles.

Variedades linguísticas como o Galego (GA) ou o Português Europeu (PE), que apresentam uma flexão verbal complexa, formas homógrafas, ou contracções de tokens ambíguas, precisam ser tratadas com ferramentas desenvolvidas especificamente para elas (Graña, Barcala e Vilares, 2002; Branco e Silva, 2004). Assim, tanto a criação como a adaptação de recursos para estas variedades devem ter em conta os problemas específicos que apresentam, com o fim de evitar erros de análise em etapas posteriores do processamento.

Tanto o Português Europeu como o Galego dispõem de recursos de análise morfossintáctica de alta precisão (Bick, 2000; Marques e Lopes, 2001; Ribeiro, Oliveira e Trancoso, 2003; Branco e Silva, 2004, por exemplo) e (Graña, Barcala e Vilares, 2002), respectivamente, mas até ao momento desconhecíamos *software* com licença livre (salvo os anteriores ficheiros de treino para Galego de FreeLing, desenvolvidos na Universidade de Vigo (Carreras et al., 2004)) para este fim. Neste sentido, a adaptação de FreeLing para Português, bem como a melhoria da versão galega, implicam a disponibilização de ferramentas livres para a análise morfossintáctica destas variedades.

O presente trabalho tem como objectivos principais (i) mostrar os procedimentos de adaptação de FreeLing para Português Europeu e Galego, indicando os casos problemáticos e as soluções adoptadas em cada um deles e (ii) realizar uma avaliação do módulo PoS-tagger nas duas versões desenvolvidas.

A adaptação realizou-se fundamentalmente com recursos linguísticos de livre distribuição disponíveis *on-line*, uma vez que actualmente, este *software* permite ser adaptado de maneira rápida e relativamente simples. As licenças do próprio FreeLing e de alguns dos dicionários e *corpora* utilizados permitem treinar e adaptar os módulos de análise sem um consumo excessivo de recursos, possibilitando ao mesmo tempo correcções, modificações e ampliações posteriores.

A *suite* disponibiliza, desde a versão 2.1, os ficheiros de treino apresentados no presente trabalho. Uma vez que FreeLing contém módulos para diferentes níveis de análise, é preciso referir que até ao momento o desenvolvimento centrou-

se nos seguintes tópicos: (i) Tokenização, (ii) segmentação de orações, (iii) lematização (com base em léxicos) e (iv) PoS-tagging. Outros módulos foram também treinados, mas o seu desenvolvimento ainda está em processo, pelo que não serão apresentados pormenorizadamente: Estes são os (v) reconhecedores de numerais, datas e quantidades e (vii) identificadores de expressões multipalavra de classe fechada.

Actualmente, os módulos de análise do Português Europeu e do Galego incluídos em FreeLing têm desempenhos próximos do estado-da-arte, quer em comparação com outro *software* para as mesmas variedades, quer em relação às avaliações de sistemas para outras línguas.

Para além desta secção introdutória, este artigo apresenta aqueles módulos de FreeLing que foram adaptados (Secção 2), os recursos utilizados (Secção 3), os resultados de diferentes avaliações dos PoS-tagger (Secção 4) bem como as conclusões finais (Secção 5).

## *2 Módulos de FreeLing*

Nesta secção será realizada uma apresentação dos módulos de FreeLing que foram adaptados para o Português Europeu e Galego, destacando os principais problemas encontrados durante o seu desenvolvimento; como foi dito, o *software* fornece mais serviços dos aqui descritos, sendo que não todos eles foram adaptados para as duas variedades referidas.

### 2.1 Tokenizador

O primeiro módulo adaptado foi o tokenizador, que converte, através de regras, um texto plano num vector de palavras. Uma vez que é uma tarefa relativamente simples, o principal aspecto a ter em conta tem a ver com a ordem de aplicação entre o próprio tokenizador e o PoS-tagger, a qual influencia o modo como as contracções ambíguas são tratadas (Graña, Barcala e Vilares, 2002; Branco e Silva, 2003). Assim, formas como `desse` (em PE, ou `dese` em GA), que pode ser um verbo ou uma contracção de preposição e demonstrativo, provocam uma circularidade entre o etiquetador morfossintáctico e o tokenizador. Este último não poderá decidir se separar `desse` em `de+esse` sem conhecer a sua categoria morfossintáctica, mas o PoS-tagger não pode ser aplicado sob um texto não tokenizado. As soluções que FreeLing permite adoptar nestes casos têm a ver com a análise morfossintáctica (dicionário, afixos e PoS-tagger), pelo que neste primeiro processo o tokenizador não separará as contracções. O *output* do tokenizador, portanto, manterá ainda a ambiguidade nas formas contraídas.

Outro aspecto a considerar é a interacção com o segmentador de orações. Na ordem de aplicação proposta (tokenizador > segmentador), o primeiro dos módulos deve reconhecer as abreviaturas (identificando o ponto como parte da abreviatura: `Sr.` e não `Sr .`) para evitar as ambiguidades mais comuns no *input* do segmentador de orações. A diferença entre as configurações do tokenizador de PE e de Galego está, portanto, na lista de abreviaturas.

### 2.2 Segmentador de Orações

O segmentador de orações recebe o *output* do tokenizador e devolve uma nova oração cada vez que detecta uma fronteira. As línguas românicas não apresentam muitas diferenças nos marcadores ortográficos, pelo que a adaptação para Português Europeu e Galego não apresentou grandes dificuldades. Uma vez que o tokenizador eliminou já as ambiguidades mais frequentes entre os pontos finais e os pontos de abreviação, o segmentador não precisa tratar especificamente estes casos.

Entre as duas variedades tratadas, as diferenças de segmentação não são significativas, e dizem respeito a especificidades ortográficas, como a utilização dos pontos de interrogação e exclamação para abrir orações (não utilizados em PE, mas facultativos em GA).

### 2.3 Analisador Morfológico

O módulo de análise morfológica de FreeLing é na verdade um conjunto de módulos que realizam tarefas como a identificação de numerais e de datas, o reconhecimento de entidades nomeadas e de expressões multipalavra, bem como a pesquisa no dicionário e o tratamento dos afixos.

Até ao momento, o maior esforço foi dedicado à adaptação do módulo de pesquisa em dicionário, tranformando o formato dos léxicos disponíveis e criando regras de lematização de afixos verbais e nominais.

Este módulo compõe-se de dois sub-módulos, que actuam em paralelo: Um deles procura no dicionário todas as possibilidadess de análise de cada um dos tokens encontrados no *input*, enquanto o outro aplica as regras de lematização de afixos, que permitem que tokens que não estejam no dicionário sejam analisados pelo sistema.

O dicionário de Português Europeu contém mais de 908.000 entradas (mais de 1.257.000 formas, tendo em conta as entradas com várias análises), enquanto o de Galego supera as 428.000 (577.000 formas). Note-se que o sistema não possui um lematizador próprio: O lema de cada token é procurado no léxico. Isto implica a necessidade de léxicos amplos, com o fim de atingir

níveis altos de precisão nesta tarefa. A avaliação deste processo realizou-se dividindo o número de lemas correctamente atribuídos pelo número total de lemas de um *corpus* de teste. Em PE, o sistema foi avaliado sobre um *corpus* de 50.000 tokens, obtendo uma precisão de 98,583%. Em Galego, o *corpus* de teste foi de 6.200 tokens e o resultado de 99,41%.

O sub-módulo de tratamento de afixos permite criar regras de lematização de formas com prefixos e sufixos. Deste modo não é preciso incluir no dicionário todas as possibilidades de combinação de formas verbais com clíticos, nem diminutivos, aumentativos, advérbios acabados em `mente`, ou formas prefixadas.

A pesquisa em dicionário e o tratamento de afixos permitem que na execução do etiquetador morfossintáctico sejam tratadas as contracções, não divididas pelo tokenizador. O funcionamento é o seguinte:

As contracções não ambíguas (por exemplo `do`: preposição `de` + artigo `o`), estão presentes no dicionário com o formato `do de+o SPS00+DA`, pelo que estas formas serão divididas em dois tokens na saída final.[1] Os casos de ambiguidade (`desse/dese, destes, pelo/polo`, etc.), porém, podem ser tratados —fundamentalmente— de duas maneiras: Incluindo-as no dicionário, ou acrescentando regras de lematização destas formas ao sub-módulo de tratamento de afixos.

As duas soluções referidas permitem evitar a circularidade referida no Ponto 2.1, uma vez que todas as alternativas existentes no módulo de análise morfológica são avaliadas pelo desambiguador morfossintáctico e pelo PoS-tagger que, se for preciso, realizará uma retokenização. A única diferença que encontrámos entre as duas propostas de tratamento diz respeito ao formato do *output*: Enquanto as entradas ambíguas do dicionário são apresentadas sem serem divididas (`desse de+esse SPS00+DD0MS0`), as regras de lematização permitem separar a saída: `de de SPS00 / esse esse DD0MS0` (o formato de saída é `token lema TAG`).[2] Para manter a coerência com o formato das contracções não ambíguas (que são divididas), decidimos utilizar a segunda das opções. Esta solução é similar às adoptadas em (Branco e Silva, 2003) ou em (Graña, Barcala e Vilares, 2002), deixando a decisão de realizar *split* aos módulos de análise morfossintáctica, e não ao tokenizador. Note-se, contudo, que em (Graña, Barcala e Vilares, 2002) a desambiguação de locuções é realizada no mesmo processo.

Dentro do conjunto de módulos de análise morfológica, a seguinte ferramenta de FreeLing (o desambiguador morfossintáctico) assigna uma probabilidade para cada um dos possíveis tags de cada token e, com base na análise das terminações, tenta saber que tags são possíveis nas formas desconhecidas. Esta análise é realizada com base no treino em *corpus*.

Para além destas ferramentas, a análise morfológica contém, como foi dito, outros módulos que realizam tarefas diversas. Até ao momento, com base nas configurações para outras línguas românicas, estão a ser adaptados para Português Europeu e Galego os módulos que têm a ver com o reconhecimento de numerais, de pontuação, de datas, de quantidades, de nomes próprios e de expressões multipalavra. Este último compõe-se de uma lista de expressões multipalavra de classe fechada, extraída dos *corpora* utilizados para o treino do PoS-tagger, e ampliada de modo manual.

## 2.4 PoS-Tagger

FreeLing permite utilizar dois métodos de anotação morfossintáctica: O modelo probabilístico com base nos Hidden Markov Models (Brants, 2000), e um método híbrido que permite combinar informação estatística com informação gerada manualmente (Padró, 1998).

Com este último modelo, um utilizador pode criar um conjunto de restrições para tratar certos tipos de erros produzidos pela abordagem estatística. Estas restrições estabelecem probabilidades de atribuição de uma etiqueta a um token em função do contexto, das propriedades do token, etc.

O modelo híbrido é —apesar de ligeiramente mais lento— de maior precisão do que os HMM, mas requer a criação manual das restrições, pelo que no treino realizado utilizámos o método puramente estatístico.

Para além do *corpus*, um dos aspectos de maior incidência no desempenho de um PoS-tagger deriva do seu tagset; se este for muito complexo, a informação fornecida pelo etiquetador é maior, mas a sua precisão será mais baixa. Pelo contrário, se o tagset for reduzido, a sua precisão

[1] Pode entender-se que a análise de `do` contém ambiguidade relativa à categoria de `o`, que além de artigo, poderia ser pronome nos casos em que o núcleo da frase nominal não está preenchido: `O homem do qual ele falou` (pelo que a entrada do dicionário incluiria SPS00+DA/PD). No caso que nos ocupa, unicamente nos referimos à ambiguidade em que uma única forma pode ser analisada como contraída ou não: `deste` como contracção de preposição+demonstrativo ou como verbo.

[2] A análise das contracções ambíguas não deve ser inserida como a primeira das opções do dicionário, já que desta maneira FreeLing selecciona-a sem aplicar o PoS-tagger, ignorando as restantes hipóteses e, não resolvendo, portanto, a circularidade.

será mais elevada, mas pode correr-se o risco de que a informação obtida não seja suficiente para os objectivos do PoS-tagger.

Um dos propósitos da adaptação de FreeLing para Português Europeu e Galego foi o de utilizá-lo como etiquetador morfossintáctico base para um analisador de dependências multilíngue (Gamallo e González, 2009). Tendo em conta que este sistema requer informação morfológica (género, número, pessoa, tempo e modo verbal, etc.), e com base nos tagsets utilizados nas outras línguas de FreeLing, decidimos usar as recomendações propostas pelo Grupo EAGLES (Leach e Wilson, 1996). O tagset definido para Português Europeu contém 255 tags, enquanto em Galego empregaram-se 277 etiquetas.[3]

O tagset utilizado contém informação morfossintáctica detalhada, mas não todos estes dados são utilizados propriamente pelo PoS-tagger; este usa unicamente os dois primeiros elementos da etiqueta, sendo os restantes extraídos do léxico. O primeiro elemento do tag (`D`, Determinante, `P`, Pronome, `N`, Nome, etc.) indica a categoria morfossintáctica; o segundo (`D` Demonstrativo, `P`, Possessivo, etc., variando em função do primeiro elemento) refere a subclasse da categoria à que pertence. O resto de entradas das etiquetas variam em função da categoria principal, e englobam aspectos como o possuidor (singular ou plural) dos possessivos o grau (aumentativo ou diminutivo) dos nomes, o caso dos pronomes ou informação sobre modo, pessoa e número dos verbos.

## *3 Recursos Utilizados*

Uma vez que alguns dos recursos linguísticos utilizados na adaptação de FreeLing para as duas variedades em causa são de livre distribuição, nesta secção apresentaremos sucintamente a sua origem, bem como as modificações feitas durante o desenvolvimento.

O processo de aprendizagem do módulo de anotação morfossintáctica precisa de um *corpus* etiquetado de alta qualidade. Com este fim, para Português Europeu o *corpus* utilizado foi criado a partir do Bosque 8.0 (Bosque. Uma floresta integralmente revista por linguistas, ), o único que conhecemos disponível livremente com informação morfossintáctica detalhada. Este *corpus* contém aproximadamente os 1.000 primeiros extractos dos *corpora* CETEMPúblico e do CETEMFolha (este último, não empregue, do Português do Brasil), o que faz um total de mais de 138.000 tokens.

O Bosque foi anotado automaticamente e, posteriormente, revisto de forma manual por linguistas. Sendo um *corpus* com informação sintáctica, esta foi eliminada na conversão para o formato requerido por FreeLing.

Para além do *corpus*, o treino de FreeLing requer um dicionário de formas flexionadas (que contenha os lemas e tags possíveis para cada token). Para este fim, utilizou-se o léxico de formas simples LABEL-LEX (SW) (Eleutério et al., 2003), que contém mais de 1.257.000 formas, geradas a partir de perto de 120.000 lemas.

Os dois recursos referidos têm características diferentes em relação à anotação morfossintáctica, pelo que foi preciso fazer uma conversão de cada um deles para o formato utilizado. Neste processo surgiram algumas incoerências que implicaram tomadas de decisão do ponto de vista linguístico. Assim, tags como "pron-indp: pronome independente" (utilizado no Bosque), não tinham correspondente directo nas etiquetas do léxico, pelo que não foi possível uma transferência automática entre os formatos. A conversão destes casos teve de ser incluída no *script* de transformação de maneira individual, e decidir em cada ocorrência dos tokens no *corpus* qual era o tag que lhes correspondia de acordo com o dicionário.

Para além das inconsistências no nível morfossintáctico, a conversão do *corpus* e do dicionário apresenta problemas em termos de lematização nominal. Assim, enquanto o Bosque lematiza os adjectivos superlativos como elementos não derivados (`altíssimo` é o lema de `altíssimo/a/(s)`), o LABEL-LEX (SW) opta por decisões mais coerentes do ponto de vista teórico: `altíssimo/a/(s)` > `alto`. De modo similar, outras diferenças notórias entre as lematizações do *corpus* e do dicionário foram as relacionadas com derivação semântica: O LABEL-LEX (SW) considera que formas como `mulher` (nome) e `melhor` (adjectivo) derivam de `homem` e `bom`, respectivamente, enquanto o Bosque atribui `mulher` e `melhor` como lemas dos mesmos tokens.

Nestes casos, a solução adoptada foi de modo geral aquela que tivesse como base processos morfológicos e não semânticos. Assim, no primeiro dos casos, optou-se por considerar os adjectivos superlativos como derivados do adjectivo simples; no segundo exemplo, a decisão tomada foi consistente com a lematização utilizada no Bosque, que diferencia as formas que não apresentam uma relação morfológica directa.

[3] A estas quantidades são acrescentados 24 tags de símbolos de pontuação (atribuídos não pelo PoS-tagger, mas pelo identificador de pontuação). Na Tabela 4 pode ver-se o formato do tagset. Note-se que, para manter a compatibilidade com outros tagsets de FreeLing, os elementos que não sejam precisos em PE e GA serão marcados com um <0> (veja-se como exemplo os valores semânticos dos nomes, que ocupam os elementos 5 e 6).

No tratamento das locuções e dos nomes próprios compostos por mais de um elemento, o Bosque apresenta algumas inconsistências que, em função dos nossos objectivos, não permitiram avaliar o desempenho do reconhecedor de expressões multipalavra com precisão. Assim, enquanto `Conselho de Administração da PEC-Alimentação` é dividido em quatro tokens, expressões como `director-clínico do Hospital Prisional S. João de Deus` é anotado no *corpus* como um único token/lema. A solução adoptada nestes casos foi a seguinte: Os elementos marcados como locuções no Bosque foram extraídos automaticamente, e adicionados à lista de expressões multipalavra depois de serem revistos manualmente. Nos *corpora* de treino e avaliação, porém, estas formas foram divididas em tokens individuais, pelo que o treino e a avaliação foram realizadas sem locuções.

O desenvolvimento da versão para Galego foi realizado com recursos de diferente procedência, pelo que o processo de adaptação foi diferente ao realizado para PE.

O *corpus* utilizado para treinar o módulo PoS-tagger para Galego foi criado no projecto Gari-Coter (Barcala et al., 2007), e contém mais de 237.000 tokens; o *corpus*, gerado a partir de notícias jornalísticas, é especializado em economia. Uma vez que a anotação morfossintáctica deste recurso seguiu os *standards* do Grupo EAGLES, as únicas adaptações precisas para o treino foram relativas à homogeneização de alguns elementos dos tags: Modificou-se, por exemplo, o caso de alguns pronomes (de nominativo para oblíquo), ou o género dos determinantes indefinidos (de comum para neutro), de acordo com o tagset definido e utilizado no dicionário.

Em relação ao léxico, o trabalho partiu do dicionário criado pelo Seminario de Lingüística Informática da Universidade de Vigo, que fazia parte de anteriores versões de FreeLing. O dicionário foi ampliado com entradas verbais e nominais extraídas de *corpora* e flexionadas automaticamente com ajuda de flexionadores e conjugadores gerados pela equipa de trabalho. Actualmente, o dicionário contém mais de 428.000 entradas, o que se corresponde com mais de 577.000 formas se tivermos em conta aquelas entradas com mais de uma análise.

Assim mesmo, as regras de lematização verbal e nominal (do sub-módulo de tratamento de afixos), foram também ampliadas a partir das publicadas nas anteriores versões de FreeLing.

## 4 Avaliação do PoS-tagger

Para subsequentes tarefas de PLN, a precisão da anotação morfossintáctica é crucial, sobretudo em aquelas formas que contêm ambiguidade, e que maior índice de erros podem provocar em processamentos posteriores.

Nos últimos anos, vários trabalhos têm avaliado diferentes algoritmos de PoS-tagging, tendo em conta variáveis como o tagset utilizado, o *corpus* de treino ou a língua analisada (Megyesi, 2001; Branco e Silva, 2004).

De modo geral, considera-se que a *baseline* para esta tarefa situa-se em 90%, e que o estado-da-arte supera o 97% nos melhores resultados. Contudo, têm surgido algumas críticas à avaliação destas ferramentas, com base no tipo de texto utilizado durante o processo. Comparando diferentes avaliações de PoS-taggers sobre textos de diversas procedências (blogues, jornais digitais e outros *sites*) e tipologias (literário, científico, jornalístico, etc.), e não em texto com "condições artificiais", a precisão desce abaixo de 93%, e apresenta grandes níveis de variação em função do género textual (Giesbrecht e Evert, 2009).

A avaliação do processo de anotação morfossintáctica é realizada dividindo o número de tokens cuja etiqueta foi correctamente atribuída pelo número total de tokens do texto.

Esta tarefa, aparentemente trivial, pode apresentar problemas derivados do alinhamento entre o *gold-standard* e o texto etiquetado automaticamente. Este último pode conter um número diferente de tokens em relação ao primeiro, devido à tokenização ou à identificação de nomes próprios, locuções, etc. Assim, a forma `Presidente Mário Soares`, pode ser analisada como um único nome próprio (`Presidente_Mário_Soares`), pode ser divida em dois elementos (`Presidente / Mário_Soares`), ou em três (`Presidente / Mário / Soares`). Para tratar estes casos, o *script* de avaliação contém três parâmetros de execução, com o funcionamento seguinte:

- *NoTok*: Se são detectados erros de *split* (`Presidente_Mário_Soares NP` vs `Presidente NP / Mário NP / Soares NP`), unicamente é avaliado o tag do primeiro token, pelo que é contabilizado um acerto. Este método considera que os erros de tokenização não devem ser levados em conta na avaliação do PoS-tagger.
- *Tok*: Se houver diferenças de tokenização, são contabilizados todos os erros (no caso anterior, três). Note-se que, no caso de que a tokenização e a atribuição da etiqueta em

palavras com mais de um token sejam correctas, é marcado um único acerto.

- *NoLoc*: Este tipo de avaliação ignora todos os tokens que não estiverem alinhados; no exemplo referido, não seria contabilizado nenhum erro nem acerto.

Como foi dito na Secção 3, o Bosque apresenta alguma inconsistência na anotação das locuções e outras expressões multipalavra, facto que devemos ter em conta na consideração dos resultados destas avaliações.

Por esta razão, foi gerada uma outra versão do *corpus* de teste, na qual se realizou um *split* a todos os elementos que continham mais de um token. Assim, executando o PoS-tagger sem identificação de locuções nem de nomes próprios compostos, o *output* é um texto alinhado perfeitamente com este *gold-standard*, pelo que a avaliação resulta mais simples. Um quarto método (*OnlyTag*) tem em conta unicamente os erros e acertos, evitando diferenças de tokenização entre os *corpora* avaliados. Neste caso, na avaliação das locuções ou dos nomes próprios compostos são contabilizados todos os tokens, pelo que uma locução bem etiquetada sumará mais acertos do que com outros métodos. Esta distorção, contudo, é compensada de alguma maneira pelos casos em que a ferramenta falha, nos quais também são contabilizados um maior número de erros.

Os quatro métodos referidos avaliam a precisão do PoS-tagger com o tagset definido no Ponto 2.4. Uma vez que este contém informação muito pormenorizada, e varia notoriamente em relação aos utilizados em outros trabalhos, foi realizada também uma avaliação de cada um dos parâmetros com um tagset mais reduzido (*SingleTags*). Para este fim, unicamente se têm em conta os dois primeiros elementos das etiquetas (categoria e tipo, salvo para os verbos, que se avalia o modo), ignorando assim informação que pode ser inferida por outros meios.[4] Desta maneira, e apesar de os resultados não poderem ser directamente comparáveis (devido a diferenças não apenas no tagset, mas também nos *corpora* de treino e de teste, etc.), temos dados obtidos em condições mais próximas de outras análises, ao mesmo tempo que se verifica a importância do tagset no desempenho de um etiquetador morfossintáctico.

A Tabela 1 mostra os resultados das diferentes avaliações do PoS-tagger para Português Europeu. Os valores são a média de cinco execuções sobre extractos aleatórios de quase 10.000 tokens, com o sistema treinado nos restantes 130.000, salvo para o método *OnlyTag*, treinado sobre 90.000 tokens e avaliado sobre perto de 50.000.

| Avaliação | Tag Completo | SingleTags |
|---|---|---|
| *NoTok* | 94,788% | 96,012% |
| *Tok* | 94,470% | 95,728% |
| *NoLoc* | 95,044% | 96,263% |
| *OnlyTag* | 94,324% | 95,537% |

Tabela 1: Avaliação PoS-tagger PE.

| Avaliação | Tag Completo | SingleTags |
|---|---|---|
| *NoTok* | 97,695% | 98,037% |
| *Tok* | 97,191% | 97,562% |
| *NoLoc* | 97,724% | 98,067% |
| *OnlyTag* | 97,503% | 97,914% |

Tabela 2: Avaliação PoS-tagger GA.

Para Galego (Tabela 2), o treino foi realizado sobre o *corpus* completo (quase 238.000 tokens), e a avaliação sobre um *corpus* extraído de jornais electrónicos e etiquetado manualmente, de 6.200 tokens.[5]

Os resultados das várias avaliações dos dois sistemas indicam que, actualmente, FreeLing consegue realizar análises morfossintácticas de textos de diversa procedência com desempenhos próximos do estado-da-arte.

Entre as duas variedades linguísticas, as diferenças de precisão são notórias, mas os resultados não devem ser directamente cofrontados. A este respeito, devemos notar, por um lado, os *corpora* de treino utilizados; enquanto o PoS-tagger de PE foi treinado sobre extractos de 130.000 tokens, para GA usou-se um texto mais de 100.000 tokens superior, pelo que é esperável que o desempenho seja maior (Banko e Brill, 2001). Em relação a isto, note-se que os resultados mais baixos da avaliação do PE foram com o método *OnlyTag*, treinado sobre um *corpus* menor. Por outro lado, é importante destacar também as características dos *corpora* de teste utilizados para a avaliação. O PE foi avaliado sobre extractos do próprio Bosque 8.0, com mais ruído —e de maior tamanho— do que o *corpus* de avaliação do GA (mais consistente e com menos ruído). Estas diferenças de desempenho sugerem, como (Giesbrecht e Evert, 2009), que as características do texto a etiquetar influenciam decisivamente a qualidade da etiquetagem.

[4] Este tagset pode ver-se na Tabela 3 (também sem os símbolos de pontuação).

[5] A velocidade de etiquetação de FreeLing executado numa máquina com um processador Core 2 Quad a 2.8 GHz num sistema GNU/Linux foi de aproximadamente 12.000 tokens por segundo. O tempo que o *software* precisou para treinar o PoS-tagger na mesma máquina foi de 10 e de 15 segundos sobre os *corpora* de Português Europeu e Galego, respectivamente.

## *5 Conclusões*

Este trabalho apresenta o desenvolvimento de diversos módulos de tratamento morfossintáctico de FreeLing para Português Europeu e Galego. Os primeiros módulos criados, o tokenizador e o segmentador de frases, realizam tarefas relativamente simples, pelo que a adaptação desde outras línguas românicas não apresentou dificuldades. Os sub-módulos de pesquisa em dicionário e de tratamento de afixos foram desenvolvidos fundamentalmente com base em recursos existentes, bem como os ficheiros de treino do PoS-tagger, gerados com base em *corpora* já anotados. Outros módulos de FreeLing, como os de numerais, datas ou expressões multipalavra (ainda em versões não definitivas), são também disponibilizados.

A avaliação do etiquetador morfossintáctico mostrou que, treinado com recursos já disponíveis para Português Europeu e Galego, um PoS-tagger pode atingir valores de precisão próximos do estado-da-arte. Assim, os diferentes testes realizados confirmam a importância do tagset no desempenho do etiquetador (com diferenças de mais de 1% devidas aos dois tagsets utilizados), bem como dos *corpora* de treino e teste.

O principal objectivo do presente trabalho foi mostrar o processo de adaptação de FreeLing para Português Europeu e Galego, indicando as soluções adoptadas nos casos problemáticos, e realizando diferentes avaliações dos módulos de PoS-tagging. A accessibilidade ao *software* e a alguns dos recursos linguísticos permitiu desenvolver ferramentas de alta precisão, bem como disponibilizar recursos de análise morfossintáctica para Português Europeu e Galego com licenças livres.

## *Agradecimentos*

Este trabalho recebeu apoio do Governo Galego através dos projectos com referências PGIDIT07PXIB204015PR e 2008/101.

## *Referências*

Banko, Michele e Eric Brill. 2001. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing. Em *Proceedings of the Conference on Human Language Technology*.

Barcala, Fco. Mario, Eva Mª Domínguez Noya, Pablo Gamallo Otero, Marisol López Martínez, Eduardo Miguel Moscoso Mato, Guillermo Rojo, María Paula Santalla del Río, e Susana Sotelo Docío. 2007. A corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a in a Minority Language. Em Zygmunt Vetulani, editor, *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 3rd Language & Technology Conference*, pp. 359–363, Poznan. Wydawnictwo Poznaskie Sp. z o.o.

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

Bosque. Uma floresta integralmente revista por linguistas. `http://www.linguateca.pt/Floresta/corpus.html#bosque`.

Branco, António e João Silva. 2003. Contractions: breaking the tokenization-tagging circularity. Em *Lecture Notes in Artificial Intelligence*, volume 2721, pp. 167–170, Berlin. Spinger.

Branco, António e João Silva. 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. 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. 507–510, Paris. ELRA.

Brants, Thorsten. 2000. TnT – A Statistical Part-of-Speech Tagger. Em *Proceedings of the 6th Conference on Applied Natural Language Processing, ANLP*. ACL.

Carreras, Xavier, Isaac Chao, Lluís Padró, e Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. Em *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

DepPattern. An Open Source Dependency-Based Analyzer. `http://gramatica.usc.es/pln/tools/deppattern.html`.

Eleutério, Samuel, Elisabete Ranchhod, Cristina Mota, e Paula Carvalho. 2003. Dicionários Electrónicos do Português. Características e Aplicações. Em *Actas del VIII Simposio Internacional de Comunicación Social*, pp. 636–642, Santiago de Cuba.

FreeLing. An Open Source Suite of Language Analyzers. `http://www.lsi.upc.edu/~nlp/freeling/`.

Gamallo, Pablo e Isaac González. 2009. Una gramática de dependencias basada en patrones

de etiquetas. Em *XXV Congresso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Donostia.

Giesbrecht, Eugenie e Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. Em *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, Donostia.

Graña, Jorge, Fco. Mario Barcala, e Jesús Vilares, 2002. *Formal Methods of Tokenization for Part-of-Speech Tagging*, volume 2276/2002, pp. 123–144.

Leach, Geoffrey e Andrew Wilson. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. Relatório técnico, Expert Advisory Group on Language Engineering Standard (EAGLES).

Marques, Nuno e Gabriel Lopes. 2001. Tagging with Small Training Corpora. Em *Proceedings of the International Conference on Intelligent Data Analysis*, volume 2189 of *Lecture Notes on Artificial Intelligente (LNAI)*, pp. 63–72. Springer-Verlag.

Megyesi, Beáta. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. Em *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, pp. 151–158.

Padró, Lluís. 1998. *A Hybrid Environment for Syntax-Semantic Tagging.* Tese de doutoramento, Dept. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.

Ribeiro, Ricardo, Luís C. Oliveira, e Isabel Trancoso. 2003. Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. Em *Proceedings of the 6th Workshop on Computational Processing on the Portuguese Language (PROPOR 2003)*, pp. 143–150, Faro. Springer-Verlag.

| **Tag** | **Valor** |
|---|---|
| AO | Adjectivo Ordinal |
| AQ | Adjectivo Qualificativo |
| CS | Conjunção Subordinativa |
| CC | Conjunção Coordenativa |
| DA | Determinante Artigo |
| DD | Determinante Demonstrativo |
| DI | Determinante Indefinido |
| DP | Determinante Possessivo |
| I | Interjeição |
| NC | Nome Comum |
| NP | Nome Próprio |
| PD | Pronome Demonstrativo |
| PE | Pronome Exclamativo |
| PI | Pronome Indefinido |
| PP | Pronome Pessoal |
| PR | Pronome Relativo |
| PT | Pronome Interrogativo |
| PX | Pronome Possessivo |
| RG | Advérbio Geral |
| RN | Advérbio Negativo |
| SP | Preposição |
| VG | Verbo: Gerúndio |
| VI | Verbo: Modo Indicativo |
| VM | Verbo: Modo Imperativo |
| VN | Verbo: Infinitivo |
| VP | Verbo: Particípio |
| VS | Verbo: Modo Conjuntivo |
| Z | Numeral |

Tabela 3: Tagset Largo (*SingleTags*).

| **Adjectivos** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Adjectivo | A |
| 2 | Tipo | Qualificativo | C |
| | | Ordinal | O |
| 3 | Grau | Aumentativo | A |
| | | Diminutivo | D |
| | | Superlativo | S |
| 4 | Género | Masculino | M |
| | | Feminino | F |
| | | Comum | C |
| 5 | Número | Singular | S |
| | | Plural | P |
| | | Invariável | N |

| **Advérbios** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Advérbio | R |
| 2 | Tipo | Geral | G |
| | | Negativo | N |

| **Determinantes** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Determinante | D |
| 2 | Tipo | Artigo | A |
| | | Demonstrativo | D |
| | | Indefinido | I |
| | | Possessivo | P |
| 3 | Pessoa | 1ª/2ª/3ª | 1/2/3 |
| 4 | Género | Masculino | M |
| | | Feminino | F |
| | | Comum | C |
| | | Neutro | N |
| 5 | Número | Singular | S |
| | | Plural | P |
| | | Invariável | N |
| 6 | Possuidor | Singular | S |
| | | Plural | P |

| **Verbos** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Verbo | V |
| 2 | Tipo | Principal | M |
| 3 | Modo | Gerúndio | G |
| | | Indicativo | I |
| | | Imperativo | M |
| | | Infinitivo | N |
| | | Particípio | P |
| | | Conjuntivo | S |
| 4 | Tempo | Futuro do Pretérito | C |
| | | Mais-que-Perfeito | M |
| | | Futuro | F |
| | | Imperfeito | I |
| | | Presente | P |
| | | Perfeito | S |
| 5 | Pessoa | 1ª/2ª/3ª | 1/2/3 |
| 6 | Número | Singular | S |
| | | Plural | P |

| **Conjunções** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Conjunção | C |
| 2 | Tipo | Coordinativa | C |
| 3 | | Subordinativa | S |

| **Preposições** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Aposição | S |
| 2 | Tipo | Preposição | P |
| 3 | Forma | Simples | S |

| **Nomes** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Nome | N |
| 2 | Tipo | Comum | C |
| | | Próprio | P |
| 3 | Género | Masculino | M |
| | | Feminino | F |
| | | Comum | C |
| 4 | Número | Singular | S |
| | | Plural | P |
| | | Invariável | N |
| 7 | Grau | Aumentativo | A |
| | | Diminutivo | D |

| **Pronomes** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Pronome | P |
| 2 | Tipo | Demonstrativo | D |
| | | Exclamativo | E |
| | | Indefinido | I |
| | | Pessoal | P |
| | | Relativo | R |
| | | Interrogativo | T |
| | | Possessivo | X |
| 3 | Pessoa | 1ª/2ª/3ª | 1/2/3 |
| 4 | Género | Masculino | M |
| | | Feminino | F |
| | | Comum | C |
| | | Neutro | N |
| 5 | Número | Singular | S |
| | | Plural | P |
| | | Invariável | N |
| 6 | Caso | Nominativo | N |
| | | Acusativo | A |
| | | Dativo | D |
| | | Oblíquo | O |
| 7 | Possuidor | Singular | S |
| | | Plural | P |

| **Numerais** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Numeral | Z |

| **Interjeições** | | | |
|---|---|---|---|
| **Elemento** | **Atributo** | **Valor** | **Tag** |
| 1 | Categoria | Interjeição | I |

Tabela 4: Formato do Tagset Estreito.

# Apresentação de Projectos

# Apresentação do projecto Per-Fide: Paralelizando o Português com seis outras línguas

Sílvia Araújo[1], José João Almeida[2], Idalete Dias[1], Alberto Simões[3]
[1]Instituto de Letras e Ciências Humanas, Universidade do Minho
[2]Departamento de Informática, Universidade do Minho
[3]Escola Superior de Estudos Industriais e de Gestão, Instituto Politécnico do Porto
{saraujo@ilch,jj@di,idalete@ilch}.uminho.pt, alberto.simoes@eu.ipp.pt

**Resumo**

Neste documento apresentamos o projecto Per-Fide que tem como principal objectivo a criação de recursos bilingues entre a língua portuguesa e seis outras línguas: espanhol, russo, francês, italiano, alemão e inglês. Este processo iniciar-se-á com a compilação de corpora paralelos em diferentes áreas, nomeadamente a literatura, religião e política (legislativa e jurídica) e técnico-científica. Os corpora serão alinhados à frase e à palavra e serão objecto de extracção automática de dicionários e terminologia bilingues. O documento irá focar inicialmente os principais objectivos do projecto, seguindo-se uma visão geral sobre as tarefas propostas e os resultados que se esperam obter.

## *1 Introdução*

Nos últimos anos a quantidade de recursos paralelos para o processamento de linguagem natural tem vindo a aumentar. Infelizmente este aumento tem sido significativo especialmente para outras línguas que não a portuguesa e, o aumento para esta língua, tem-se verificado essencialmente na área política ou legislativa (graças à adesão de Portugal à União Europeia).

O projecto Per-Fide[1] tem como principal objectivo a compilação de corpora paralelos entre a língua portuguesa e seis outras línguas: espanhol, russo, francês, italiano, alemão e inglês.

A escolha das línguas foi dirigida pela relevância global das línguas em causa, mas também da relevância das línguas dentro do Centro de Estudos Humanísticos da Universidade do Minho (CEH/UM), onde o projecto está sediado.

A construção destes corpora será realizada em diferentes áreas do conhecimento que podem ser divididas em dois grandes blocos:

- **ficção:** neste primeiro bloco inclue-se essencialmente a produção literária (portuguesa e estrangeira) e textos religiosos[2]
- **não ficção:** este segundo bloco focará textos jornalísticos, político ou legislativo, e técnico-científico.

Tentar-se-á, também, a recolha de textos tendo como língua de origem cada uma das sete línguas, incluindo as diferentes variantes da língua portuguesa: português europeu, português brasileiro e português africano (nos seus diferentes dialectos).

O Per-Fide não culminará com a construção dos corpora. Esse será apenas o primeiro passo. O Per-Fide tem um conjunto de objectivos mais vasto. Destes, o mais importante, e ortogonal a tudo o resto, é a disponibilização aberta de todos os recursos recolhidos e produzidos durante o projecto.

## *2 Per-Fide: linha de produção*

Esta secção detalha as principais etapas previstas no projecto, relacionando-as com recursos produzidos, projectos semelhantes já existentes e com trabalho desenvolvido previamente pela equipa Per-Fide. Estas etapas estão esquematizadas na figura 1 que destaca o lado aberto do projecto com a disponibilização de todos os recursos produzidos.

### 2.1 Limpeza, Tratamento e Anotação

No processo previsto pelo Per-Fide, a primeira etapa corresponde à compilação dos recursos (nomeadamente de textos). Esta tarefa vai ser especialmente custosa devido à necessidade de negociação de direitos de autor, o que nem sempre será possível. Assim que negociado o modo

[1]O nome do projecto tem a sua génese nos nomes das sete língua envolvidas: Português, English, Russian, Français, Italiano, Deutch, Español.

[2]A classificação dos textos religiosos não é consensual. A nossa escolha não quer de forma alguma levantar essa discussão já que é irrelevante para o contexto em causa.


Figura 1: Linha de Produção Per-Fide

de disponibilização dos documentos, seguir-se-á a sua limpeza e tratamento. Estas duas tarefas incluem a conversão entre formatos, a remoção de ruído obtido durante as conversões efectuadas, a detecção de documentos duplicados e a avaliação da qualidade dos recursos em causa.

Além disso, os documentos serão classificados e anotados com meta-informação como sejam a língua (e respectiva variante), o género do documento, o título e autor (se aplicável), informação sobre a tradução (caso não esteja na sua língua original[3]), etc. Toda esta informação será armazenada em formato textual usando as estruturas já disponibilizadas pelo TEI — Text Encoding Initiative (Vanhoutte, 2004) — para esta finalidade.

Embora esta etapa não seja completamente automatizável, existe um conjunto de ferramentas já existentes que serão imprescindíveis:

- diversos conversores de formato, como o `antiword` ou o `pdftotext`, que permitem uma conversão de qualidade razoável entre alguns formatos proprietários e documentos de texto;
- identificadores de língua, como seja o módulo Perl `Lingua::Identify`;
- identificadores de variante da língua (ou da grafia), e comparadores de grafia (Almeida, Santos e Simões, 2010);

Nesta etapa serão disponibilizados os corpora monolingues para pesquisa na rede ou, sempre que as licenças o permitam, para serem descarregados e processados localmente.

## 2.2 Etiquetação Morfosintáctica

No Per-Fide não temos como objectivo desenvolver um sistema de etiquetação morfosintáctica, especialmente dado que o teríamos de fazer para as sete línguas envolvidas. Embora neste momento ainda não se tenha procedido à aquisição de nenhum etiquetador morfo-sintáctico está previsto o uso do Palavras (Bick, 2000). Outras hipóteses já se encontram delineadas como seja o uso do TnT (Brants, 2000) ou do TreeTagger (Schmid, 1994). Uma outra alternativa será a etiquetação ambígua usando um simples analisador morfológico como o jSpell (Almeida e Pinto, 1994) ou o FreeLing (Atserias et al., 2006).

Existem outros projectos que disponibilizam corpora etiquetados, como por exemplo a Floresta Sintá(c)tica para a língua portuguesa (Afonso et al., 2001), ou o Penn Treebank para a língua inglesa (Marcus, Marcinkiewicz e Santorini, 1993). No caso concreto do Per-Fide o objectivo não é tanto a disponibilização dos corpora monolingues etiquetados, mas a sua disponibilização como parte integrante de um corpus paralelo.

Os corpora etiquetados serão disponibilizados nos formatos convencionais, e também disponibilizados para pesquisa na rede.

## 2.3 Alinhamento ao Nível da Frase

Existem várias alternativas de ferramentas para o alinhamento ao nível da frase, desde as mais simples como o Vanilla Aligner (Gale e Church, 1991) aos mais usados recentemente como o easy-align, parte do IMS Corpus Workbench (Christ et al., 1999), o HunAlign (Varga et al., 2005) ou o Clue Aligner do PLUG (Tiedemann, 2003).

Em relação à disponibilização de corpora paralelos alinhados para consulta na rede ou para processamento local existem já vários projectos, dos quais salientamos o COMPARA, um corpus paralelo português-inglês literário (Frankenberg-Garcia e Santos, 2003), e o projecto OPUS que disponibiliza vários corpora paralelos de diferentes áreas para todas as línguas europeias (Tiedemann e Nygard, 2004).

Os corpora alinhados serão disponibilizados

[3] Alguns géneros linguísticos, como jornalístico ou religioso, não permitirão uma fácil classificação de língua de origem. Neste caso nenhuma das línguas levará essa classificação.

em TEI[4] (Erjavec, 1999), XCES (Ide, Bonhomme e Romary, 2000) ou TMX (Savourel, 2005)[5].

### 2.4 Extracção de Dicionários de Tradução

Existem duas abordagens diferentes no alinhamento de corpora ao nível da palavra. Uma, que tem como ferramenta preferencial para a sua extracção o Giza++ (Och e Ney, 2003), pretende associar cada instância de uma palavra com a instância correspondente na sua tradução. A outra abordagem, que pode ser obtida usando o NATools (Simões e Almeida, 2003), pretende extrair um dicionário probabilístico de tradução que associa a cada palavra tipo as suas possíveis traduções, devidamente pesadas probabilisticamente.

No Per-Fide serão aplicadas as duas abordagens, sendo que a extracção de dicionários probabilísticos de tradução será imprescindível para a abordagem planeada na extracção de terminologia bilingue. Os dicionários serão também úteis para auxiliar a pesquisa bilingue de concordâncias (Simões, 2008).

### 2.5 Extracção de Terminologia Bilingue

Para além da extracção de dicionários de tradução, o Per-Fide também tem como objectivo a compilação de terminologia bilingue que, sendo originária de corpora devidamente etiquetados, poderá ser facilmente classificada por área de conhecimento. Para esta extracção tenciona-se usar uma abordagem baseada em padrões bilingues (Simões e Almeida, 2008; Guinovart e Simões, 2009). Estes padrões especificam regras morfológicas dos contituintes de cada termo, permitindo a sua extracção automática. Ao usar-se um corpus etiquetado morfo-sintacticamente e não um analisador morfológico levará a um aumento da precisão de extracção.

### 2.6 Disponibilização na Rede

Como foi sendo referido nos itens anteriores, o projecto Per-Fide tem como ponto fulcral a disponibilização dos recursos construídos. Esta disponibilização será feita à medida que os recursos sejam desenvolvidos e não apenas no final do projecto. Deste modo espera-se que o retorno efectuado pelos utilizadores permita melhorar o material disponibilizado.

Além disso, tentar-se-á que os recursos sejam disponibilizados de forma integrada, de modo que o mesmo interface permita a consulta nos corpora (paralelos ou bilingues, com ou sem anotação), nos dicionários de tradução e na terminologia bilingue. Esta integração permitirá também que os recursos extraídos possam ser utilizados para enriquecer as consultas de outros recursos. Por exemplo, a apresentação de concordâncias bilingues poderá beneficiar dos dicionários probabilísticos de tradução para alinhar (ou sublinhar) o resultado da pesquisa em ambas as línguas.

## *3 Conclusões*

O projecto está agora a iniciar e os seus membros têm a perfeita noção de que os objectivos são ousados. Um dos principais problemas será a questão de direitos de autor, essencialmente no que respeita à obtenção de textos literários. Neste sentido está a ser desenvolvido um manual de negociação que permitirá, sucessivamente, aumentar a probabilidade de sucesso.

Um outro problema terá que ver com a língua Russa, e o facto de esta usar um sistema de codificação diferente. Embora tecnologicamente se possa falar do Unicode como solução global, a verdade é que a maioria das ferramentas que prevemos utilizar nunca foi testada com outro tipo de codificação que não seja o ISO-8859-1.

Neste momento o projecto pode ser acompanhado em `http://natura.di.uminho.pt/per-fide`.

## *Agradecimentos*

O Per-Fide, *Português em paralelo com seis línguas (Português, Español, Russian, Français, Italiano, Deutsch, English)*, é parcialmente financiado pelo projecto `PTDC/CLE-LLI/108948/2008` da *Fundação para a Ciência e a Tecnologia*.

## *Referências*

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.

Almeida, José João e Ulisses Pinto. 1994. Jspell – um módulo para análise léxica genérica de linguagem natural. Em *Actas do X Encontro da Associação Portuguesa de Linguística*, pp. 1–15, Évora.

[4]Embora o TEI não seja desenhado para a codificação de corpora paralelos existem alguns projectos que o usam para indicar alinhamentos.

[5]Neste ponto ainda não se decidiu por qual (ou quais) optar. Sendo possível proceder à automatização na conversão de formatos, serão todos disponibilizados.

Almeida, José João, André Santos, e Alberto Simões. 2010. Bigorna – a toolkit for orthography migration challenges. Em *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, May, 2010. forthcomming.

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. Em *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 48–55.

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

Brants, Thorsten. 2000. TnT – a statistical part-of-speech tagger. Em *6th Applied NLP Conference, ANLP-2000*, Seattle, WA, April 29, 2000.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann, e Esther König, 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*. Institute for Natural Language Processing, University of Stuttgart, March, 1999. `http://www.ims.uni\_stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/`.

Erjavec, Tomaz. 1999. A tei encoding of aligned corpora as translation memories. Em *In Proceedings of the EACL Workshop on Linguistically Interpreted Corpora '99*, pp. 49–60.

Frankenberg-Garcia, Ana e Diana Santos. 2003. Introducing COMPARA, the portuguese-english parallel translation corpus. Em Silvia Bernardini Federico Zanettin e Dominic Stewart, editores, *Corpora in Translation Education*. Manchester: St. Jerome Publishing, pp. 71–87.

Gale, William A. e Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. Em *Meeting of the Association for Computational Linguistics*, pp. 177–184.

Guinovart, Xavier Gomez e Alberto Simões. 2009. Parallel corpus-based bilingual terminology extraction. Em *8th International Conference on Terminology and Artificial Intelligence*, Toulouse, France, November, 18–20, 2009.

Ide, Nancy, Patrice Bonhomme, e Laurent Romary. 2000. XCES: an XML-based encoding standard for linguistic corpora. Em *Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association.*

Marcus, Mitchell P., Mary Ann Marcinkiewicz, e Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Och, Franz Josef e Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Savourel, Yves, editor. 2005. *TMX 1.4b Specification*. Localisation Industry Standards Association (LISA), April, 2005. `http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm`.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. Em *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49.

Simões, Alberto e José João Almeida. 2008. Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, 41:281–288, September, 2008.

Simões, Alberto M. e J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September, 2003.

Simões, Alberto Manuel Brandão. 2008. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Tese de doutoramento, Escola de Engenharia, Universidade do Minho, Braga, 19 May, 2008.

Tiedemann, Jörg e Lars Nygard. 2004. The OPUS corpus - parallel and free. Em *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*, Lisbon, Portugal, May, 2004.

Tiedemann, Jörg. 2003. Combining clues for word alignment. Em *10th Conference of the European Chapter of the ACL (EACL03)*, Budapest, Hungary, April 12–17, 2003.

Vanhoutte, Edward. 2004. An introduction to the tei and the tei consortium. *Lit Linguist Computing*, 19(1):9–16, April, 2004.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, e V. Nagy. 2005. Parallel corpora for medium density languages. Em *Recent Advances in Natural Language Processing (RANLP 2005)*, pp. 590–596.

# 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 Outubro de 2010
- Resultados da selecção até: 15 de Novembro de 2010
- Versão final até: 31 de Novembro de 2010
- Publicação da revista: Dezembro de 2010

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 outubro de 2010
- Resultados da selección: 15 de novembro de 2010
- Versión final: 31 de novembro de 2010
- Publicación da revista: 15 de decembro de 2010

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 octubre de 2010
- Resultados de la selección: 15 de noviembre de 2010
- Versión final: 31 de noviembre de 2010
- Publicación de la revista: diciembre de 2010

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 d'octubre de 2010
- Resultats de la selecció: 15 de novembre de 2010
- Versió final: 31 de novembre de 2010
- Publicació de la revista: desembre de 2010

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

http://www.linguamatica.com/