Volume 5, Número 2- Dezembro 2013

linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 5, Número 2 – Dezembro 2013

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Recentemente fomos contemplados com alguns estudos que, supostamente de forma inocente, levaram a uma descridibilização das revistas de acesso aberto, e em especial, daquelas que apenas publicam na rede. Como editores ficamos algo preocupados com qual seria o efeito desta notícia no número de propostas para a Linguamática.

Felizmente tivemos um número de propostas bastante elevado, com artigos relevantes e interessantes, alguns dos quais não foram aceites inicialmente para publicação não pela falta de qualidade do trabalho, mas por necessitarem de algum trabalho adicional.

O facto dos autores continuarem a enviar propostas para a Linguamática demonstra um voto de confiança no nosso trabalho e dos revisores, confiança essa que muito agradecemos. Também acreditamos que o facto de a Linguamática estar indexada em novos índices (como o da plataforma e-Revistas do CSIC espanhol) seja um factor relevante.

Em relação a este ponto, relativo à indexação da Linguamática, cumpre-nos informar que está neste momento a decorrer a avaliação por parte da SCOPUS para a possível inclusão da Linguamática. Supomos que com a adição de títulos, resumo e palavras chave em inglês será mais fácil que outros índices acolham e incluam a nossa revista nos seus índices.

Ao terminar este quinto ano de vida da Linguamática não podemos de deixar de agradecer a todos os membros da comissão científica, bem como a todos os investigadores que nos fazem chegar os seus trabalhos para publicação.

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,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Universidade Beira Interior

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Henrique Barroso (convidado),
Universidade do Minho

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José Carlos Medeiros,
Porto Editora

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Lluís Padró,
Universitat Politècnica de Catalunya

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Mikel Forcada,
Universitat d'Alacant

Patrícia Cunha França,
Universidade do Minho

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Rui Pedro Marques,
Universidade de Lisboa

Salvador Climent Roca,
Universitat Oberta de Catalunya

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Dossier

imaxin|software
16 anos desenvolvendo aplicações no campo do processamento da linguagem natural multilingue

16 years developing applications for multilingual natural language processing

José Ramom Pichel Campos
imaxin|software
jramompichel@imaxin.com

Diego Vázquez Rey
imaxin|software
diegovazquez@imaxin.com

Antonio Fernández Cabezas
imaxin|software
afernandez@imaxin.com

Luz Castro Pena
imaxin|software
luzcastro@imaxin.com

Resumo

imaxin|software é uma empresa criada em 1997 por quatro titulados em engenharia informática com o objetivo de desenvolver videojogos multimédia educativos e processamento da linguagem natural.

16 anos depois tem desenvolvido recursos, ferramentas, aplicações multilingues para diferentes línguas: Português (Galiza, Portugal, Brasil, etc.), Espanhol (Espanha, Argentina, México, etc.), Inglês, Catalão, Francês.

Neste artigo redigido em português-galego faremos uma descrição daqueles principais fitos deste percurso tecnológico e humano.

Palavras chave

processamento da linguagem natural, correçao ortográfica e linguística, correçao gramatical e de estilo, dicionários eletrônicos, corpus, traduçao automática, opinion mining, biotext mining, sentiment analysis, Microsoft, open source

Abstract

imaxin|software is a company created in 1997 by four computer engineers with the aim of developing educational multimedia games and natural language processing tools.

After 16 years **imaxin**|software has developed resources, tools and applications for different languages, specially for Portuguese (Galiza, Portugal, Brazil, etc.), Spanish (Spain, Argentina, México, etc.), English, Catalan, French.

In this article we will describe the main highlights of this technological and human challenge.

Keywords

natural language processing, spell-checkers, grammar checkers, electornic dictionaries, corpus, machine translation, opinion mining, biotext mining, sentiment analysis, Microsoft, open source.

1 Introdução

imaxin|software é uma empresa dedicada ao desenvolvimento de serviços e soluções avançadas de software e multimédia desde o ano 1997, especializada em engenharia linguística e videojogos multimédia educativos e formativos.

Dentro da primeira linha de desenvolvimentos, **imaxin**|software é desde o ano 2000 fornecedor de tecnologia linguística para Microsoft. Além disso, podemos destacar entre os principais desenvolvimentos em PLN os sistemas de correção ortográfica, gramatical, estilística; sumarizadores de textos, sistemas de opinion mining, pesquisa semântica, sistemas de codificação médica de histórias clínicas, detecção automática de entidades (NER), bem como a plataforma líder europeia em tradução automática de código aberto: Opentrad[1].

No campo dos multimédia, **imaxin**|software desenvolve desde o ano 1999 sistemas de aprendizagem para meninos e adultos mediante o uso dos videojogos, o que na atualidade é conhecido por serious games. Dentro destes desenvolvimentos multilíngues podemos destacar: Vetas (jogo formativo de riscos laborais para as empresas do granito do Porrinho), Climântica (um jogo simulador de cidades sustentáveis), Keco e a Ecopanda (videojogo educativo de educação meio ambiental), minijogos com a plataforma wii e aplicações móveis (Apps).

No campo da localização de software, **ima-**

[1] http://www.opentrad.com

xin|software encarregou-se de traduções de software especialmente a idiomas como (galego/português, espanhol, inglês, francês) de grande volume como Microsoft Windows XP, Vista ou 7, Office XP ou Office 2013, OpenOffice.org, etc.

2 Linhas de trabalho

imaxin|context

Realização de soluções multilíngues de processamento linguístico e documentário. **imaxin**|software aplica a tecnologia linguística através de ferramentas informáticas na gestão de conteúdos, gestão documentário e gestão do conhecimento, melhorando a produtividade e optimizando a exploração dos recursos intangíveis da empresa:

- **Processamento linguístico:** corretores ortográficos, sintácticos e de estilo, corretores ortográficos em rede, dicionários eletrónicos, tradutores automáticos (Opentrad).
- **Processamento documentário:** reputação online, pesquisadores documentários, pesquisadores semânticos, sumarizadores de documentos, classificadores automáticos documentários, extractores de informação.

in|gaming

Área multimédia de **imaxin**|software especializada em soluções tecnológicas para a educação, formação e ócio. Contamos com uma equipa multidisciplinar perito na transformação de conteúdos a formato digital e na criação de aplicações inovadoras. Pioneiros no desenvolvimento de jogos educativos multimédia na Galiza, hoje contamos com uma carteira de produtos que englobam serious games, realidade aumentada, aplicações multitácteis, etc.

imaxin|localiza

Criada fruto da experiência na localização de Windows e Office a galego para Microsoft. **imaxin**|localiza presta serviços integrais de localização de software para português europeu (galego, português), espanhol e inglês contando com uma equipa altamente qualificada de linguistas e informáticos. O objetivo deste área é prestar serviço a grandes e medianas empresas de software que precisem traduzir as suas aplicações para chegar a um maior número de utentes e a mercados internacionais.

3 Certificação CMMI-3

imaxin|software conseguiu a certificação CMMI de nível 3 para duas das três áreas de desenvolvimento de software (Context e ingaming). Resultado desta certificação definiu um conjunto de processos regulares da organização que regem esta atividade e garantem a obtenção de produtos com a qualidade requerida e a satisfação do cliente.

CMMI é um modelo para a melhoria dos processos de uma organização muito complexo de atingir no mundo do software. Aplica-se em áreas de processos que garantem as práticas a seguir pola organização para a consecução dos seus objetivos e se representa por níveis de maturidade atingidos no desenvolvimento da atividade produtiva.

Entre os principais benefícios que proporciona CMMI a **imaxin**|software, estão:

- Alinha o processo Software com a estratégia de negócio da organização.
- Melhora a Qualidade e Produtividade da empresa.
- Antecipa-se os problemas mediante técnicas proativas de gestão.
- Melhora a comunicação na organização, conseguindo uma linguagem comum na mesma.
- Permite que o conhecimento fique na organização.
- Proporciona uns mecanismos de melhoria contínua de processos através de análises de medições.

4 Principais projetos PLN

4.1 Ferramentas de conversão a XML de dicionários e corpus

2013 Projeto: Dicionário da RAG

Este projeto consistiu no desenvolvimento de ferramentas de conversão de um dicionário monolíngue em Word a formato TLEX utilizando a DTD específica para dicionários. Também se desenhou a DTD, realizado a conversão semiautomática de referências cruzadas (Sinónimos, Antónimos, Equivalente). TshwaneLex é uma suíte específica para a elaboração e gestão de dicionários monolíngues, dicionários bilíngues

ou multilíngues em XML. A partir do dicionário convertido a XML-TLEX pode-se optar pola publicação de dicionários em papel, meios eletrónicos ou online. O armazenamento e o manejo de dados lexicográficos utilizando standards da indústria como XML e Unicode, aumenta a produtividade e a qualidade na criação, gestão, revisão e publicação.

4.2 Jogos e Lexicografia

2013 Projeto: Portal das Palavras

O Portal dás palavras é um site educativo que põe em valor o dicionário da Real Academia Galega mediante jogos relacionados com as palavras. Com o Portal das Palavras não só melhoraremos a nossa concorrência em idioma galego senão que também aprenderemos jogando. Inclui também o dicionário da RAG com buscas de lemas e sinônimos, vídeos explicativos e guias didáticas para a língua.

4.3 Dicionários monolíngues

1999-2002 Projeto: Dicionário de dicionários contemporâneos

Um dos primeiros Dicionários de dicionários eletrónico da Europa em SGML que permite consultas associadas. Contém 25 dicionários históricos galegos desde o século XVIII ao século XXI, desde Sarmento, até o de Elixio Rivas de 2001 na versão inicial. Sobre este se podem realizar Consultas simples, Consultas complexas, Visualização de entradas, Procuras por refrões, Buscas por poemas, Histórico, Cesta, Hipertexto, Impressão, etc.

Aplicação que trabalha com dicionários previamente convertidos a formato SGML com ferramentas automáticas desenvolvidas adhoc e sobre o que se podem realizar as funções acima indicadas.

2000 Projeto: Dicionário Eletrónico Cumio da Língua Galega

Dicionário eletrónico com perto de 40.000 verbetes. Podem-se realizar as seguintes funções sobre estas: Visualização hipertextual, Seleção de texto, Pesquisas complexas, Impressão, Histórico, Ajuda on-line, Pesquisas por sinônimos, por antônimos etc. Esta aplicação trabalha sobre um dicionário previamente convertido a formato SGML/XML e sobre o que se podem realizar as funções acima indicadas.

2001 Projeto: Dicionário Morris inglês-euskera, euskera-inglês

imaxin|software desenvolveu para o Governo basco o Dicionário de Morris inglês-euskera, euskera-inglês em formato site utilizando diretamente pola primeira vez na Península ibéria a tecnologia XML. Este dicionário eletrónico consta de 40.000 entradas para os dois idiomas presentes no Dicionário regular em Euskadi de inglês-euskera e euskera-inglês. O utente pode fazer consultas por lemas, introduzindo nas caixas de texto correspondentes a palavra, tanto em inglês como em euskera, acedendo à definição do dicionário. Para um determinado lema mostra todos os diferentes verbetes mostrando em ecrã em um formato de saída cômodo para o utente de Internet. Este dicionário site permite a consulta de lemas em inglês e euskera, sendo uma ferramenta fundamental para a internacionalização do euskera.

4.4 Dicionários medievais

2006 Projeto: Dicionário de dicionários medieval

Dicionário eletrónico que contém 13 dicionários construídos a partir de corpus medievais galegos (em estudo da incorporação dos corpus portugueses).

Sobre este se podem realizar Consultas simples, Consultas complexas, Visualização de entradas, Buscas por provérbios, Buscas por poemas, Historial, Cesta, Hipertexto, Impressão, etc. De especial relevância é a construção de tipos de letra próprios medievais.

4.5 Tesouros informatizados contemporâneos

2003 Projeto: Tesouro informatizado da língua galega (Tilga)

Um site que contém textos do galego moderno, desde o ano 1612 à atualidade. Tem 11.409.358 registos e ao redor de 90 mil lemas. Está pensada como corpus de referência onde se podem realizar diferentes consultas: consultas por lema, por palavra, por ano de publicação, por intervalo de anos de publicação, por autor, por obra, etc. É o maior corpus de referência de galego num site e foi realizado para o Instituto da Língua Galega dirigido na altura por D. Antón Santamarina Fernández.

4.6 Tesouros informatizados medievais

2004 Projeto: Tesouro informatizado Medieval dá língua galega (TMILG) (Site que contém textos do galego medieval, desde o século XII até o XVII)

Contém um total de 140.000 palavras indexadas em 80 obras, ao redor de 23.000 páginas. Está pensada como um corpus de referência onde se podem realizar diferentes consultas: por palavra, gênero, subgênero, obra específica, intervalo de datas, consultas complexas. As consultas apresentam relatórios estatísticos com informação que abarca desde intervalos de séculos até tipos de documentos nos que aparece em texto consultado. É o maior corpus de referência do galego-português medieval em site.

4.7 Optimizadores semânticos de pesquisas

2008 Projeto: "Optimizador de pesquisas em bibliotecas mediante tesaurus"

O objectivo do módulo Optimizador é sugerir sinônimos nas buscas efetuadas polos utentes nos sistemas de consulta bibliográfica do CSBG (Centro Superior Bibliográfico da Galiza). Estes sistemas permitem a consulta de bancos de dados bibliográficas, mostrando resultados sobre os termos procurados. A função do Optimizador é sugerir a possibilidade de alargar essa consulta realizada atendendo a sinônimos do termo consultado. Inicialmente o sistema apresenta os resultados em dois idiomas: galego e castelhano. foi desenvolvido para que em um futuro se possam incorporar outras variantes e idiomas, como o português, inglês, francês etc. A opção de realizar a implementação através de um serviço site é para facilitar a integração do módulo em diferentes tipologias de sites, tanto quanto a linguagens como a plataformas. Está integrado com o software de gestão de bibliotecas de código aberto Koha.

4.8 Hemerotecas e Bibliotecas digitais

2009 Projeto: "Bibliotecas digitais"

O projeto consistiu na digitalização dos boletins históricos da Real Academia Galega, seguindo as adendas A e B das "Diretrizes para projetos de digitalização de coleções e fundos de domínio público, designadamente para aqueles custodiados em bibliotecas e arquivos", na sua última versão publicada pola Subdirección Geral de Coordenação Bibliotecária do Ministério de Cultura. Ademais atribuíram-se dados e metadatos que codificam as suas descrições e permitem a sua carga em um repositório OAI definido segundo as especificações "The Open Arquives Initiative Protocol for Metadata Harvesting" e utilizando a norma ISSO 15836 (Dublin Core). Integração em Galiciana-Hispana-Europeana. Criação de uma aplicação de base de dados em formato site que permita a consulta e visualização dos boletins em formato eletrónico, de forma singela e eficiente.

4.9 Tradução automática

2005-2013 Projeto: "Opentrad: plataforma de serviços de tradução de código aberto"

Clientes: Instituto Cervantes, El Pais, Ministério de Administrações Públicas, Xunta de Galicia, La Voz de Galicia, Faro de Vigo, El Progreso de Lugo, Universidade de Santiago de Compostela, Universidade de Vigo, Universidade da Corunha, CHUAC, Eroski, etc.

Opentrad é a plataforma de tradução automática em código aberto líder no mercado espanhol. Opentrad melhora a comunicação multilíngue, permite publicar informação em diferentes idiomas, reduz custos e os tempos de tradução humana e permite contar com versões multilíngues das aplicações empresariais.

Opentrad está presente em administrações, empresas e portais de Internet traduzindo milhões de palavras diariamente. Ministério de Fazenda e Administrações Públicas, Xunta de Galicia, Instituto Cervantes, La Voz de Galicia, Parlamento da Galiza, NovaGaliciaBanco ou a Kutxa são alguns dos clientes que confiam a sua comunicação multilíngue a Opentrad.

A melhoria contínua do sistema permite-nos oferecer melhor qualidade entre línguas próximas (Espanhol-Francês Espanhol-Português, Espanhol-Português do Brasil, Espanhol-Catalão, Espanhol-Galego, etc.) que outros tradutores automáticos (Google Translate, Systran, Babelfish).

4.10 Serviços

O nosso modelo de negocio está baseado em diferentes serviços ao redor do tradutor automático.

Serviços de tradução automática em aluguer

Opentrad Server (Personalização do Servidor de tradução Opentrad com dicionário personalizado em 150 termos): Ferramentas online: Tradução de textos, Tradução de documentos, Tradução de URL, API de integração.

Opentrad Server Premium

Personalização a medida do cliente (1.000 termos), Ferramentas online: Tradução de textos, Tradução de documentos, Tradução de URL, Ferramentas servidor: API de integração, Pastas dinâmicas, Tradução via e-mail. Integração em Microsoft Word ou Navegadores.

Opentrad na Aplicateca

Desde o passado 2012 está presente na Loja Cloud Aplicateca de Telefónica o que nos permite chegar a todos os clientes de Movistar. Este 2013 está planificado integrar Opentrad em dous novos operadores internacionais.

4.11 Correção ortográfica e gramatical

1998-2012 Projeto: "Galgo e Galgo 2.0"

O corretor imaxin Galgo é o primeiro corretor do mercado que soluciona não só erros de carácter tipográfico e ortográfico, senão também léxicos, que se podem encontrar em textos escritos em galego. Esta nova versão de Imaxin Galgo, aplicação pioneira para a correção de textos em galego, atualiza e melhora esta ferramenta pensada para a análise dos problemas de carácter ortográfico e léxico intrínsecos a todo texto escrito no nosso idioma.

Assim, é possível que se misturem num mesmo documento desde castelhanismos, verbos mau conjugados, vulgarismos, erros propriamente ortográficos ou até intersecções destas tipologias em uma mesma palavra. Entre as atualizações, destaca a modificação da base de dados segundo a norma aprovada no ano 2003 pola Real Academia Galega e o Instituto da Língua Galega.

4.12 Correção de linguagem não sexista

2008 Projeto: "Exeria: corretor de linguagem não sexista" [em parceria com Tagen Ata]

O corretor de linguagem não sexista é uma ferramenta informática integrável no pacote ofimático livre OpenOffice que tem como finalidade ajudar a realizar documentos com uma linguagem não-sexista. Exeria oferece uma ajuda interativa para a edição de textos que reflitam um tratamento igualitário da linguagem. Exeria nasceu dentro de um plano do Governo galego para a igualdade entre mulheres e homens 2007-2010. Exeria pretende ajudar a construir uma linguagem que visibilize a presença das mulheres no colectivo e integrar a sua consideração enquanto agentes participativos nos diferentes âmbitos sociais. Com esta finalidade, Exeria facilita a construção de um discurso mais inclusivo oferecendo alternativas e soluções para aqueles termos que podem, em alguns contextos, implicar um uso discriminatório no que atinge à linguagem.

5 Principais projetos I+D

CELTIC (Conhecimento estratégico liderado por tecnologias para a Inteligência Competitiva)

[Em desenvolvimento]: O projeto está orientado no campo da vigilância tecnológica e o Social Média Marketing. Participantes no projeto: INDRA, Elogia, **imaxin**|software, Saec-data, USC-GE, USC-CA, Gradiant. Financiado polo programa FEDER-INNTERCONECTA, através do CDTI.

Coruxa Biomedical Text Mining: Extrator e codificador automática de informação médica relevante mediante uso-o da engenharia linguística em código aberto

Direção-geral de I+D+i. Xunta de Galicia. Pesquisador principal: **imaxin**|software, USC-GE, IXA Taldea, Doutor QSolutions. Transferência ao sector produtivo: talento de codificação SNOMED-CT para histórias clínicas. Financiado polo Programa I+D+i Galego.

Coati Opinion mining: Pesquisa avançada multilíngue em blogues para a recuperação de opiniões e tendências para ou âmbito empresarial e dá administração pública. Direção Xeral de I+D+i. Xunta de Galicia. Pesquisador principal: **imaxin**|software, USC-GSI, USC-GE, UDC- IR-Lab. Transferência ao sector produtivo: desenvolvimento de ETLs, talento de recuperação de opiniões em blogues para a administração pública e o sector empresarial. Financiado polo Programa I+D+i Galego.

EurOpenTrad "Traducción automática avanzada de código abierto para la integración europea de las lenguas del Estado español"

(PROFIT-350401-2006-5), 2006-2007-2008. Investigador principal: **imaxin**|software, Eleka, Elhuyar, IXA Taldea, TALP (UPC), Transducens- UA, SLI. Transferência ao sector produtivo: adaptação do tradutor automático OpenTrad a UNICODE, incorporação de detectores de idiomas, detectores de nomes próprios automático. Financiado polo Programa AVANZA-PROFIT do Ministério de Indústria.

OpenTrad "Traducción Automática de Código Abierto para las Lenguas del Estado Español"

(PROFIT-340101-2004-0003, PROFIT-340001-2005-2), 2005-2006. Investigador principal: Eleka, Elhuyar, IXA Taldea, TALP (UPC), Transducens-UA, **imaxin**|software, SLI-Universidade de Vigo. Transferência ao sector produtivo: tradutor automático de espanhol-catalão, espanhol-galego e espanhol-euskera desenvolvido em código aberto com transferência de recursos linguísticos galego e português e motor de tradução adaptado a estes dois idiomas. Financiado polo Programa AVANZA-PROFIT do Ministério de Indústria.

EixOpenTrad "Tradução automática avançada de código aberto entre as variantes do portugués de Portugal e do português da Galiza"

2006-2007. Investigador principal: **imaxin**|software, Universidad de Santiago de Compostela y SLI. Transferência ao sector produtivo: tradutor automático de galego-português e português-galego desenvolvido em código aberto com transferência de recursos linguísticos galego e português e motor de tradução adaptado a estes dois idiomas.

RecursOpenTrad "Recursopentrad: recursos lingüístico-computacionais de tradución automática avanzada em código aberto para a integración europea da língua galega"

Pesquisador principal: **imaxin**|software, TALP (UPC), Transducens-UA, SLI. Transferência ao sector produtivo: melhoria do tradutor inglês-galego RBMT, criação de um protótipo estatístico SMT de tradução automática inglês-galego/português, desenvolvimento de ETLs, integração de detectores de idiomas, detectores de entidades no motor apertium.

Extração da informação: "Estudio de necesidades e xerazón de recursos e ferramentas intelixentes em xestión da información e enxeñaría lingüística para a mellora das empresas exportadoras galegas"

Subvenciona: Xunta de Galicia, ref. PGIDT03TICC22Y. Entidades participantes: **imaxin**|software, CESGA (Centro de Supercomputación da Galiza), SLI, Cidadanía rede de aplicacións sociais. Pesquisador responsável: José Ramom Pichel (**imaxin**|software), Xavier Gómez Guinovart (SLI-Uvigo). Transferência ao sector produtivo: realização de um protótipo de recuperador de informação a partir de um site crawler libertado com licença GPL.

Etiquetagem de textos e desambiguação automática: "Estudo e adquisición de recursos básicos de lingüística computacional do galego para a elaboración e mellora de aplicacións informáticas de tecnoloxía lingüística"

Subvenciona: Secretaria Xeral de Investigação e Desenvolvimento, Xunta de Galicia, 2001-2004 (ref. PGIDT01TICC06E). Pesquisadores principais: José Ramom Pichel (**imaxin**|software), Xavier Gómez Guinovart (SLI-Uvigo). Equipa de projeto: Elena Sacau (SLI), Ángel López (**imaxin**|software). Transferência ao sector produtivo: o lexicón gerado neste projeto foi reutilizado posteriormente para a realização do corretor ortográfico de OpenOffice.org que foi liberto com licença GPL.

"Estudo do erro gramatical para o galego"

Subvenciona: **imaxin**|software, Projeto de I+D (Universidad - Empresa), 2002-2003. Investigador principal: Xavier Gómez Guinovart (SLI-Uvigo). Transferência ao sector produtivo: este projeto inicial foi necessário para desenvolver entre o ano 2006 e 2007 o corretor gramatical Golfiño.

6 Projetos de I+D+i em desenvolvimento

CELTIC: Conocimiento Estratégico Liderado por Tecnologías para la Inteligencia Competitiva (FEDER-INNTERCONECTA)

Os FEDER INNTERCONECTA são projetos Integrados de desenvolvimento experimental altamente competitivos, com carácter estratégico, de grande dimensão e que tenham como objectivo o desenvolvimento de tecnologias novas em áreas tecnológicas de futuro com projeção econômica e comercial a nível internacional, supondo ao mesmo tempo um avanço tecnológico e industrial relevante para as autonomias destinatárias das ajudas do "Programa Operativo de I+D+i por e para o benefício das empresas - Fundo Tecnológico," como é o caso da Galiza.

imaxin|software conseguiu no ano 2012 um projeto FEDER-INNTERCONECTA com um consórcio formado polas seguintes empresas e Universidades: Indra, Elogia, SaecData, Gradiant, USC-PRONAT-L (USC), ComputationalArchitecture Group (USC).

Objectivo principal do projeto

Desenvolvimento de tecnologias capacitadoras que facilitem ao tecido empresarial a tomada de decisões estratégicas em tempo quase-real, a partir do conhecimento tanto do meio científico-tecnológico como dos impactos econômicos presentes e futuros. Ou o que é o mesmo, o desenvolvimento de tecnologias capacitadoras para a Inteligência Competitiva nas organizações.

As tecnologias a desenvolver durante o projeto cobrirão o processo completo da Inteligência Competitiva, nas suas respectivas fases: agregação de informação, análise da informação extraindo dela o conhecimento necessário, e a distribuição mediante mecanismos de visualização e iteração avançados para facilitar a tomada de decisões estratégicas.

Aplicações do projeto

Marketing: A competitividade atual imprime a necessidade de dispor de sistemas de monitorização inteligente e em tempo real de redes sociais e análises do impacto dos produtos de uma marca determinada no consumidor, mediante tecnologias avançadas de processamento da linguagem natural e tecnologias semânticas.

Vigilância tecnológica: os desenvolvimentos a realizar neste projeto permitirão o acesso e gestão em tempo real dos conhecimentos científicos e técnicos às empresas, bem como a informação mais relevante sobre o seu contexto, junto ao entendimento a tempo do significado e envolvimentos das mudanças e novidades no meio. Isto é indispensável na tomada de decisões das empresas para o desenvolvimento de um novo produto, serviço ou processo para uma organização.

7 Prémios

- Prémios Eganet 2006: Prémio especial a 10 anos.
- Prémios AETIC 2007: Galardoado Opentrad como a melhor aplicação TIC do 2007.
- Prémios Eganet 2008: Prémio à melhor iniciativa de Comunicação Site Institucional polo jogo Keco e procura da Ecopanda
- Prémios Eganet 2008: Finalista na categoria Software Livre pola aplicação Opentrad.
- Prémios Eganet 2009: Prémio Melhor ambiente Laboral
- Prémios Eganet 2009: Finalista na categoria Cultura Digital polo jogo "Climántica"
- Prémio Eganet 2010: Prémio Educação Digital polo projeto "O Valor de IGU" programa interativo de educação em valores através de videojogos.
- Prémio Leixa-prem 2012 polo uso habitual do galego na empresa
- Prémio Ada Lovelace a Luz Castro Pena polo Colexio de Enxeñeiros/as en informática.

Agradecimentos

A todas as pessoas que têm partilhado connosco este caminho de aprendizagem humana e tecnológica.

Bibliografía

Aguirre Moreno, José Luis, Alberto Álvarez Lugrís, Luz Castro Pena, Xavier Gómez Guinovart, Angel López López, José Ramom Pichel Campos, Elena Sacau Fontenla, e Lara Santos Suárez. 2003a. Adquisición de recursos básicos de lingüística computacional del gallego para aplicaciones informáticas de tecnología lingüística. *Procesamiento del Lenguaje Natural*, 31:303–304.

Aguirre Moreno, José Luis, Alberto Álvarez Lugrís, Iago Bragado Trigo, Luz Castro Pena, Xavier Gómez Guinovart, Santiago González Lopo, Angel López López, José Ramom Pichel Campos, Elena Sacau Fontenla, e Lara Santos Suárez. 2003b. Alinhamento e etiquetagem de corpora paralelos no CLUVI (Corpus Linguístico da Universidade de Vigo). Em José João Almeida, editor, *CP3A 2003, Corpora Paralelos: Aplicações e Algoritmos Associados*, pp. 33–47, Universidade do Minho, Braga, Portugal.

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

de Moura Barros, António Carlos, Angel López López, e José Ramom Pichel Campos. 2008. TMILG: tesouro medieval informatizado da lingua galega. *Procesamiento del Lenguaje Natural*, 41:303–304.

Gamallo, Pablo e Jose Ramom Pichel. 2008. Learning Spanish-Galician translation equivalents using a comparable corpus and a bilingual dictionary. *Lecture Notes in Computer Science*, 4919:423–433.

Gamallo, Pablo e José Ramom Pichel. 2005. An approach to acquire word translations from non-parallel text. *Progress in Artificial Intelligence, LNAI*, 3808.

Gamallo, Pablo e José Ramom Pichel. 2007. Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural*, 39:241–248.

Gamallo, Pablo e José Ramom Pichel. 2010. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. Em *Cycling2010*.

Gamallo Otero, Pablo, Marcos Garcia, e José Ramom Pichel Campos. 2013. A method to lexical normalisation of tweets. Em *Tweet Normalization Workshop at SEPLN*.

Malvar, Paulo e José Ramom Pichel Campos. 2010. Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. the Galician and Portuguese case. Em *II Congreso Internacional de Lingüística de Corpus (CILC10)*, Universidade da Coruña.

Malvar, Paulo e José Ramom Pichel Campos. 2011a. Generación semiautomática de recursos de opinion mining para el gallego a partir del portugués y el español. Em *ICL11: Workshop on Iberian Cross-Language NLP Tasks*.

Malvar, Paulo e José Ramom Pichel Campos. 2011b. Métodos semiautomáticos de generación de recursos de opinion mining para el gallego a partir del portugués y el español. *Novática: Revista de la Asociación de Técnicos de Informática*.

Malvar, Paulo, José Ramom Pichel, Óscar Senra, Pablo Gamallo, e Alberto García. 2010. Vencendo a escassez de recursos computacionais. Carvalho: Tradutor automático estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. *Linguamática*, 2(2):31–38.

Pichel, José Ramom, Paulo Malvar López, Oscar Senra Gómez, Pablo Gamallo Otero, e Alberto García. 2009. Carvalho: English-Galician smt system from english-Portuguese parallel corpus. *Procesamiento del Lenguaje Natural*, 41.

Pichel Campos, José Ramom. 1996. Problemas e solucións para a terminoloxía galega. Em *I Congreso Internacional da lingua galega (ed. ILG)*.

Pichel Campos, José Ramom. 1997. Funciones terminológicas en la lengua gallega: problemas metodológicos y soluciones al respecto. *Uzei, Centro Vasco de Terminología y Lexicografía*.

Pichel Campos, José Ramom e Antonio Fernández Cabezas. 1999. "imaxin Galgo v.1.0. *Procesamiento del Lenguaje Natural*, 25:225–226.

Pichel Campos, José Ramom e Antonio Fernández Cabezas. 2002. Diccionario de diccionarios da lingua galega. *Procesamiento del Lenguaje Natural*, 26:99–100.

Artigos de Investigação

Desenvolvimento de um recurso léxico com papéis semânticos para o português

Developing a lexical resource annotated with semantic roles for Portuguese

Leonardo Zilio
Universidade Federal do Rio Grande do Sul
ziliotradutor@gmail.com

Carlos Ramisch
Laboratoire d'Informatique Fondamentale de Marseille
carlos.ramisch@lif.univ-mrs.fr

Maria José Bocorny Finatto
Universidade Federal do Rio Grande do Sul
mariafinatto@gmail.com

Resumo

Os objetivos deste estudo são os seguintes: apresentar uma metodologia para desenvolver um recurso léxico com informações semânticas; comparar papéis semânticos de verbos em linguagem especializada e não especializada; e observar a anotação de papéis semânticos por vários anotadores.

Foram desenvolvidos dois experimentos relacionados à anotação de papéis semânticos em português: comparação de um *corpus* de linguagem especializada com um *corpus* de linguagem não especializada; e teste da concordância entre diversos anotadores na atribuição de papéis semânticos.

Quanto aos resultados, observaram-se diferenças qualitativas entre os *corpora* estudados, sendo o apagamento de agentes um traço marcante do *corpus* especializado. A não concordância averiguada entre vários anotadores indica que a tarefa é complexa, requerendo mais treinamento ou uma maior simplificação da tarefa, o que não parece ser possível no atual estágio de desenvolvimento.

Palavras chave

Linguística Computacional, Processamento de Linguagem Natural, anotação de papéis semânticos, recursos léxicos

Abstract

The objectives of this study are as follows: to present a methodology for the development of a lexical resource with semantic information; to compare semantic roles in specialized and non-specialized language; and to observe the semantic role labeling (SRL) made by a group of annotators.

Two experiments revolving around SRL in Portuguese were developed: a comparison between data in specialized and non-specialized language corpora; and an annotation evaluation for verifying the agreement among multiple annotators for the task of SRL.

As for results, a qualitative difference between the corpora was observed, and the most prominent feature was the omission of agents in specialized texts. There was little agreement among annotators, which points toward the necessity of more training, or a simplification of the task, which does not seem to be possible at this stage of development.

Keywords

Computational Linguistics, Natural Language Processing, semantic role labeling, lexical resources

1 Introdução

A área de Processamento de Linguagem Natural (PLN) tem por objetivo facilitar a interação entre o computador e as pessoas, de modo que essa interação seja o mais natural possível, por meio do uso de línguas naturais. Nesse âmbito, a tecnologia da linguagem se concretiza como um grande desenvolvimento na história do ser humano, sendo comparada por Branco et al. (2012) "com a invenção da imprensa por Gutenberg". Tendo isso em vista, é importante desenvolver um esforço colaborativo entre várias áreas do conhecimento, incluindo a Ciência da Computação e a Linguística.

Nesse contexto, a Linguística pode se constituir numa fonte de conhecimento e recursos que, somados ao trabalho do PLN, contribuem para a interação homem-máquina, seja para a redação de um texto, seja para a interpretação de um comando de voz etc. Ao lado dos estudos de léxico, morfologia, sintaxe e texto, a semântica também tem uma função a desempenhar, pois em seu âmbito se encontra o estudo dos significados. Existem vários tipos de abordagens semânticas, desde as que observam o léxico e o seu valor na língua (por vezes, sem observar os contextos de uso de uma palavra), até as que tentam reconhecer os significados nos textos ou na interação com o mundo. A abordagem neste artigo parte principalmente da sintaxe e do léxico, e enfoca o significado de verbos em termos de seus papéis semânticos. Para isso, observamos o léxico em contexto e levamos em consideração a sintaxe em

torno dos verbos. Discutiremos os papéis semânticos de forma mais detalhada ao longo do artigo, porém, cabe aqui apresentar um breve exemplo. Observe-se a seguinte sentença:

1. [O homem] bateu [no cachorro].

Na sentença 1, o sujeito *O homem* desempenha um papel de AGENTE (ou ARG0), ou seja, de participante no evento que executa a ação, e o objeto indireto *no cachorro* tem o papel de PACIENTE (ou ARG1), isto é, ele é o participante no evento afetado pela ação. Assim, a semântica dos papéis se configura como uma abstração do significado da oração.

Apesar de o português ser atualmente a quinta língua mais utilizada na Internet[1], a quantidade de recursos semânticos disponíveis para o seu processamento automático ainda é pequena. Estamos distantes de outras línguas que recebem mais investimento no desenvolvimento de recursos e ferramentas para o processamento da linguagem, como é o caso, por exemplo, do inglês (Branco et al., 2012).

Neste artigo, trata-se da criação de um recurso semântico para o português brasileiro que possa ser utilizado para o processamento semântico de verbos do português. Assim, os objetivos deste artigo são:

- Apresentar os métodos para o desenvolvimento de um recurso léxico com informações semânticas[2].
- Comparar papéis semânticos de verbos em linguagem especializada com aqueles dos mesmos verbos em linguagem não especializada.
- Observar a concordância entre anotadores em uma tarefa de anotação de papéis semânticos.

Este artigo está dividido da seguinte maneira: na Seção 2, apresentamos brevemente alguns conceitos e trabalhos relacionados a este estudo; na Seção 3, detalhamos e discutimos o método utilizado para o desenvolvimento de um recurso léxico com informações sobre papéis semânticos; na Seção 4, apresentamos os resultados da anotação de papéis semânticos e realizamos a comparação entre linguagem comum e especializada à luz dos papéis semânticos; na Seção 5, relatamos e discutimos um experimento realizado com múltiplos anotadores; por fim, na Seção 6, expomos nossas considerações finais.

2 Conceitos e trabalhos relacionados

Nesta seção, procuramos apresentar brevemente a base teórica deste artigo. Na Seção 2.1, discorremos sobre papéis semânticos e estruturas de subcategorização; na Seção 2.2, discutimos trabalhos como o de Levin (1993) e recursos como a VerbNet, o PropBank e a FrameNet.

2.1 Papéis semânticos e estruturas de subcategorização

2.1.1 Papéis semânticos

Os papéis semânticos foram introduzidos na teoria linguística há milhares de anos, sendo o seu precursor o gramático indiano Panini (Dowty, 1991; Gildea e Jurafsky, 2002; Levin e Rappaport-Hovav, 2005). Como se comentou rapidamente na Seção 1, os papéis semânticos representam uma forma abstrata de semântica: "os papéis semânticos distinguem [...] as facetas do significado que são gramaticalmente relevantes" (Levin e Rappaport-Hovav, 2005). Essas facetas do significado podem ser identificadas a partir da observação do léxico e da sintaxe, porém, elas não são nem tão específicas quanto uma semântica lexical (por exemplo, acepções em dicionários), nem tão abstratas quanto uma semântica puramente sintática (por exemplo, a utilização de categorias sintáticas como sujeito e objeto direto como indícios de diferenciação semântica). Em outras palavras, os papéis semânticos nem são tão semânticos para delimitar definições para cada palavra, mas também não são tão sintáticos a ponto de atribuir um mesmo papel para todos os sujeitos e objetos. Esse território intermediário entre semântica e sintaxe em que os papéis semânticos se encontram serve seu propósito para o processamento automático da linguagem.

Para exemplificar o que são os papéis semânticos, tomemos como exemplo as sentenças a seguir[3]:

2a. [João] abriu [a porta] [com a chave].
2b. [A porta] abriu [com a chave].
2c. [A chave] abriu [a porta].

[1] Dados de 2010, retirados a 10 de setembro de 2013 do site http://www.internetworldstats.com/stats7.htm.

[2] Por *recurso léxico com informações semânticas*, estamos nos referindo a um banco de dados que contenha sentenças do português anotadas com papéis semânticos.

[3] Os exemplos são inventados. Não provêm dos corpora envolvidos no estudo. Opta-se aqui por usar frases fictícias para simplificar o exemplo e permitir que o foco recaia sobre a explicação do que são papéis semânticos, sem envolver outras questões que poderiam surgir a partir de exemplos reais de uso.

Nas três sentenças acima, o verbo é sempre o mesmo (*abrir*), os sujeitos se alternam, mas sempre há um sujeito, e os demais elementos variam conforme a estrutura sintática do verbo permite. Os elementos a que chamamos atenção aqui, porém, não são os sintáticos, mas sim os semânticos. Em 2a, *João* está executando uma ação, o que lhe confere o papel de AGENTE (ou ARG0); *a porta* está sofrendo os efeitos dessa ação (está passando por uma modificação de fechada para aberta), o que caracteriza o papel de PACIENTE (ou ARG1); já *a chave* é o INSTRUMENTO (ou ARG2) utilizado pelo AGENTE para realizar a modificação no PACIENTE. Em 2b, por mais que o sujeito agora seja *a porta*, ela não passa para uma função de AGENTE (ou ARG0), pois ela não está em condições de **executar** a ação de *abrir*; assim, ela permanece como PACIENTE (ou ARG1), porque a ação está sendo executada por um elemento não divulgado na sentença. Na sentença 2c, o sujeito é *a chave*, mas, novamente, esta não é a executora da ação, ela permanece sendo apenas o INSTRUMENTO (ou ARG2) utilizado por um AGENTE implícito.

A partir desse exemplo, pode-se perceber que os elementos sintáticos (sujeitos, objetos etc.) nem sempre têm uma semântica óbvia. Desse modo, discriminar os papéis desempenhados pelos elementos sintáticos em diversos contextos pode ajudar no processamento automático de textos. Por exemplo, em um sistema de extração de informações hipotético, deseja-se conhecer o nome de todas as empresas compradas pela Google nos últimos 10 anos. Para isso, não é suficiente detectar apenas verbos de compra das quais Google seja o sujeito, pois seriam ignoradas frases como esta: *[Android Inc.] foi comprada [pela Google] [em 2005]*.

Nos exemplos fornecidos até aqui, apresentamos duas possibilidades de anotar os papéis semânticos: a forma descritiva (AGENTE, PACIENTE, INSTRUMENTO etc.) ou numerada (ARG0, ARG1, ARG2 etc.). As formas descritivas são a base para a VerbNet (Kipper-Schuler, 2005) e também para os vários projetos baseados na FrameNet (Baker, Fillmore e Lowe, 1998). Já a forma numerada foi proposta por Palmer, Kingsbury e Gildea (2005) ao desenvolverem o PropBank. Esses trabalhos serão discutidos na Seção 2.2.

Na linguística moderna, os papéis semânticos ressurgiram com os trabalhos de Gruber (1965) e Fillmore (1967), posteriormente se desenvolvendo em trabalhos como os de Jackendoff (1990), Dowty (1991) e Levin e Happaport-Hovav (2005). Para o português, na teoria de papéis semânticos, podemos citar estudos de Franchi e Cançado (2003), Perini (2008), Cançado (2009; 2010); Cançado, Godoy e Amaral (2012).

As principais discussões concernentes aos papéis semânticos giram em torno de questões como a quantidade de papéis necessários para representar uma linguagem natural e a subjetividade envolvida na atribuição dos papéis semânticos. Em particular, essas questões são discutidas com bastante propriedade por Levin e Rappaport-Hovav (2005). Em síntese, as autoras evidenciam a dificuldade de se estabelecer uma lista de papéis semânticos que não seja nem genérica demais a ponto de não apresentar diferenças suficientes entre os papéis, nem específica demais a ponto de que não se possam depreender generalizações.

A subjetividade é um fator que está constantemente presente nas discussões sobre semântica. Isso ocorre porque, em última instância, cada pessoa identifica um significado diferente (ainda que muitas vezes coincidente ou quase coincidente com o significado atribuído por outras pessoas) para cada texto com que se depara. Assim, existem discussões, por exemplo, sobre como as seguintes frases, retiradas de Kasper (2008), deveriam ser interpretadas:

3a. The cardinal loaded bottles on the wagon.
(O cardeal colocou garrafas na carroça.)
3b. The cardinal loaded the wagon with bottles.
(O cardeal carregou a carroça com garrafas.)[4]

A interpretação, conforme indicada por Jackendoff (1990), é de que em 3a as garrafas não preenchem a carroça, enquanto em 3b a carroça está completamente cheia. Porém, Fillmore (1968, *apud* Kasper, 2008) considerava que ambas eram sinônimas. Do ponto de vista dos papéis semânticos, se ambas veiculam o mesmo significado, então os papéis utilizados para os substantivos *wagon* e *bottles* serão os mesmos nas duas sentenças (assim como foi apresentado no primeiro exemplo, em que *a porta* e *a chave* não mudam de papel semântico). Porém, se seus significados forem diferentes, então os papéis também vão diferir.

Em português, temos um exemplo parecido com o que foi apresentado para o inglês, porém, com o verbo *encontrar*:

4a. O estudo encontrou a doença em 15 pacientes.
4b. O estudo encontrou 15 pacientes com a doença.

[4] A tradução em português, infelizmente, não faz jus à ambiguidade existente no inglês, pois não há um verbo que se aplique ao contexto para as duas sentenças com duas estruturas sintáticas.

Assim como nos exemplos 3a e 3b do inglês, a estrutura sintática das sentenças 4a e 4b apresentam diferenças claras devido ao emprego de diferentes preposições; porém, as duas sentenças podem ser consideradas paráfrases. Por um lado, as duas sentenças podem indicar que os pesquisadores encontraram a doença nos pacientes. Desse modo, o objeto encontrado, nas duas sentenças, é a doença, pois ela está sendo procurada, e não os pacientes (os pesquisadores sabem onde os pacientes estão). Por outro lado, a sentença 4b pode indicar que, em uma busca, foram encontrados 15 pacientes que sofriam de uma determinada doença, de modo que o objeto encontrado, de fato, são os pacientes, pois eles estavam sendo procurados, e não a doença. A doença é apenas um atributo dos pacientes.

Do nosso ponto de vista, esse tipo de diferença parece só poder ser realmente averiguado a partir da observação do referente no mundo real. Partindo apenas dessas frases escritas, uma pessoa pode interpretar o sentido das duas formas. Desse modo, há uma ambiguidade que só pode ser desfeita pela observação direta da realidade. Como a atribuição de papéis semânticos não entra no domínio da Pragmática, torna-se inviável esse tipo de atribuição.

2.1.2 Estruturas de subcategorização

As estruturas de subcategorização, mais amplamente conhecidas por seu nome em inglês, *subcategorization frames*, são estruturas sintáticas mais abstratas do que as descrições normais de sujeitos, objetos e complementos. Segundo Messiant, Korhonnen e Poibeau (2008), as "estruturas de subcategorização de predicados capturam as diferentes combinações de argumentos que um predicado pode ter no nível sintático", ou, como aponta Manning (1993), "uma estrutura de subcategorização é uma ratificação dos tipos de argumentos sintáticos que um verbo (ou adjetivo) apresenta". Apesar de as definições fazerem menção ao nível sintático, as estruturas de subcategorização não descrevem, em geral, funções de elementos sintáticos, mas sim sua morfologia básica. Por exemplo, na seguinte sentença:

5. João viu Maria.

A classificação sintática da sentença-exemplo 5 seria: *João* = sujeito; *viu* = verbo/predicado; *Maria* = objeto direto. Porém, na classificação de estrutura de subcategorização, essa mesma sentença teria a seguinte análise: *João* = NP (do inglês, *nominal phrase*) [5] ou SN (sintagma nominal); *viu* = V (verbo); *Maria* = NP ou SN. Se tivéssemos um caso com um objeto indireto ou um adjunto preposicionado, ele seria marcado como PP (*prepositional phrase*) ou SP (sintagma preposicional). Assim, as estruturas de subcategorização se apresentam em formatos como NP_V_NP e NP_V_NP_PP, ou, simplesmente, NP_PP (sem indicação da posição do verbo e, às vezes, também do sujeito). Com base nessas estruturas, é possível se obter uma boa indicação da estrutura sintática e do número de argumentos que um verbo admite.

O trabalho de Beth Levin (1993), que será discutido mais adiante, partiu do pressuposto de que verbos com uma semântica próxima compartilham estruturas sintáticas, sendo possível agrupá-los em classes semânticas com base apenas em seu comportamento sintático. Dado que as estruturas de subcategorização são um bom indicador da sintaxe das sentenças (pode se dizer que elas indicam a sintaxe de forma implícita), os estudos de PLN as têm usado para a classificação de verbos. Por serem relativamente fáceis de observar em grandes *corpora* analisados sintaticamente, as estruturas de subcategorização acabam servindo como substitutos de classificações sintáticas que identificam explicitamente sujeitos, objetos etc.

As estruturas de subcategorização já foram utilizadas para o agrupamento de verbos em diversas línguas, como alemão (Schulte im Walde, 2002), francês (Messiant, 2008; Messiant, Korhonen e Poibeau, 2008), inglês (Preiss, Briscoe e Korhonen, 2007) e italiano (Ienco, Villata e Bosco, 2008). No Brasil, um trabalho pioneiro no reconhecimento automático de estruturas de subcategorização foi o de Zanette (2010), o qual será descrito na Seção 3. Um trabalho que usou essas estruturas para agrupar verbos automaticamente foi a dissertação de mestrado de Scarton (2013), cujos resultados estão expostos de modo resumido em Zanette, Scarton e Zilio (2012) e Zilio, Zanette e Scarton (2012).

2.2 Trabalhos relacionados

Começamos esta seção com o trabalho de Levin (1993), para depois prosseguirmos com a VerbNet, o PropBank, a FrameNet e a WordNet.

O trabalho de Levin (1993) é importante não só para o inglês, a língua utilizada, mas para a Linguística como um todo, pois mostrou que é possível agrupar verbos semanticamente próximos a partir de suas estruturas sintáticas. Apesar de

[5] Em nosso estudo, privilegiamos o uso das siglas em inglês, por ser a forma utilizada durante o trabalho de anotação.

haver várias críticas ao trabalho desenvolvido[6], Levin (1993) foi pioneira na área, principalmente pela magnitude do trabalho, de modo que merece destaque e consideração em estudos que abordem sintaxe e semântica associada a verbos.

Levin (1993) observou que, quando os verbos admitem as mesmas (ou quase as mesmas) alternâncias sintáticas, eles podem ser agrupados em categorias semânticas. Por exemplo, a partir da observação dos verbos *break*, *cut*, *hit* e *touch* e das suas possibilidades de alternâncias mediais, conativas e que envolvem partes do corpo, é possível chegar à Tabela 1.

	Break	Cut	Hit	Touch
Medial	X	X		
Conativa		X	X	
Parte do corpo		X	X	X

Tabela 1: Comportamento dos verbos *break*, *cut*, *hit* e *touch*.

Assim, percebe-se que, apesar de os quatro verbos serem transitivos, eles não autorizam os mesmos tipos de alternâncias sintáticas e, por isso, pertencem a quatro classes diferentes de verbos. O verbo *break*, por exemplo, compartilha as mesmas alternâncias de verbos como *crack* (rachar), *rip* (rasgar) e *shatter* (despedaçar), já o verbo *hit* está na mesma classe de *kick* (chutar), *whack* (bater), *bash* (espancar), e assim por diante. Além de perceber essa diferença na sintaxe, Levin também apontou que esses verbos apresentam diferenças em seus traços semânticos: o verbo *cut* envolve movimento, contato e mudança de estado; o verbo *hit* envolve contato e movimento; o verbo *break* envolve apenas mudança de estado; e o verbo *touch* envolve apenas contato.

Com base nessas observações de alternâncias sintáticas e de traços semânticos, Levin organizou mais de quatro mil verbos do inglês em um total de 193 classes e subclasses. Ao apresentar as classes, Levin contribuiu em muito para os estudos sobre verbos do inglês, pois determinados fenômenos aplicáveis a um verbo geralmente se aplicam também a toda uma classe.

Para o português, ainda não foi publicado um trabalho como o de Levin (1993)[7], porém, Cançado, Godoy e Amaral (2012) já apresentaram um projeto que intenta levar a cabo essa empreitada.

Partindo das classes de Levin (1993), Kipper-Schuler (2005) desenvolveu a VerbNet. O recurso apresentado na VerbNet contém as classes de Levin associadas aos papéis semânticos que podem ser apresentados pelos verbos de cada uma das classes. No estágio atual da VerbNet (versão 3.2), foram utilizados efetivamente 30 papéis semânticos, partindo de uma lista com 36 papéis.

Por partir das classes de Levin, a anotação de apenas 191 classes (na versão 1.0) já dava cobertura para 4.173 verbos. Atualmente, com o acréscimo de outras classes de verbos, já existe anotação para cerca de 5.800 verbos, divididos em 272 classes.

Para o português, além do nosso trabalho, sabemos da existência do estudo de Scarton (2013), que se propôs a transpor as anotações do inglês para o português aproveitando-se das conexões que existem entre a VerbNet, a WordNet e a WordNet.Br. Desse modo, para as classes sinônimas entre a WordNet (Fellbaum, 1998) e a WordNet.Br (Dias-da-Silva, 2005; Dias-da-Silva, Di Felippo e Nunes, 2008), os papéis foram importados diretamente do inglês para os verbos em português. Desse modo, já existe uma VerbNet.Br, porém, ela foi construída de modo semiautomático, podendo conter erros, e apresenta apenas aquelas classes que são sinônimas entre o português e o inglês.

A principal diferença que se deve ressaltar em relação ao trabalho de Scarton (2013) e este estudo é o fato de que Scarton usou o inglês como base e importou semiautomaticamente os dados que apresentam sinonímia entre as WordNets do inglês e do português. O trabalho aqui apresentado parte do português e se baseia em uma anotação manual dos dados por um linguista. Assim, apesar de nosso estudo ser menos abrangente, ele apresenta uma menor propensão a erros.

Como mencionamos, o trabalho de Scarton (2013) usou como base os alinhamentos entre a WordNet de Princeton (Fellbaum, 1998) e a WordNet.Br (Dias-da-Silva, 2005; Dias-da-Silva, Di Felippo e Nunes, 2008). As WordNets são recursos que apresentam *synsets* (conjuntos de sinônimos) e as relações entre eles. As relações podem ser de hiperonímia, antonímia, holonímia etc. Além disso, existem definições formais para os possíveis significados de cada um dos *synsets*. Por tomarem *synsets* como base, e não palavras soltas, as relações construídas entre apenas dois *synsets*

[6] Para uma amostra das críticas feitas ao trabalho de Levin (1993), consulte Perini (2008). O estudo de Lima (2007) também mostra como verbos de um mesmo grupo semântico não necessariamente apresentam as mesmas estruturas sintáticas.

[7] Scarton (2013) realizou o agrupamento de verbos em classes, porém, partindo das classes em inglês e usando métodos semiautomáticos. Foram também publicados trabalhos isolados para uma ou algumas classes de verbos, como o trabalho de Lima (2007), mas desconhecemos a existência de um trabalho para o português que tenha a abrangência do trabalho de Levin (1993).

cobrem várias palavras, o que amplia muito a abrangência desse recurso.

Voltando para a anotação de papéis semânticos, além de um recurso mais dicionarístico como a VerbNet, que apresenta classes de verbos e seus possíveis papéis, existe também o PropBank (Palmer, Kingsbury e Gildea, 2005), que apresenta sentenças de um *corpus* anotadas com papéis numerados.

Apesar de esse tipo de opção representar uma facilidade para o anotador, que não precisa fazer distinções entre AGENTES e EXPERIMENTADORES, PACIENTES e TEMAS, entre outras, o resultado diminui muito a informação que se pode adquirir a partir da anotação. Como apontam Zapiran, Agirre e Màrquez (2008), "a interpretação dos papéis do PropBank são dependentes do verbo". Por exemplo, na sentença *João joga bola*, o sujeito do verbo *jogar* não é anotado como AGENTE, mas sim como ARG0, devendo ser interpretado como o papel semântico JOGADOR. Uma das vantagens do PropBank é que, por apresentar vários exemplos de cada um dos verbos anotados (por ser um *corpus* anotado), ele pode ser usado para treinar *softwares* de anotação automática de papéis semânticos, algo que a VerbNet, por ter um número restrito de exemplos, não permite.

O projeto SemLink (Loper, Yi e Palmer, 2007) foi responsável por realizar a vinculação dos papéis semânticos da VerbNet às sentenças do PropBank, de modo que as sentenças estão atualmente anotadas também com papéis descritivos (AGENTE, PACIENTE etc.).

Assim como no caso da VerbNet, também existe para o português um projeto que se encarregou de desenvolver o PropBank.Br. Esse projeto, desenvolvido por Duran e Aluisio (2011; 2012) já se encontra disponível[8] e contém mais de 6 mil instâncias anotadas.

Por fim, existe ainda outro tipo de anotação de papéis semânticos, bastante difundida, que toma como base os cenários comunicativos, chamados de *frames*[9]. É assim que se estrutura a FrameNet (Baker, Fillmore e Lowe, 1998), um projeto que tem por objetivo anotar os papéis semânticos de cada elemento de uma sentença em relação ao seu domínio e ao seu contexto. Por exemplo, os papéis semânticos do *frame* DECISÃO (Copa do Mundo) podem ser VENCEDOR, PERDEDOR, TORNEIO e FINAL[10]. A FrameNet Brasil (Salomão, 2009) utiliza essa mesma abordagem.

As diferenças entre a VerbNet, o PropBank e a FrameNet estão principalmente na granularidade dos papéis. Os papéis da FrameNet são altamente específicos, pois se aplicam apenas a um determinado *frame*. Os papéis da VerbNet são menos específicos, tentando apresentar uma descrição abstrata da semântica que pode ser aplicada para qualquer contexto. Já o PropBank apresenta a solução mais abstrata, pois apenas cinco papéis (ARG0 a ARG4) se aplicam a qualquer contexto, configurando-se como protopapéis.

3 Metodologia de construção do recurso

O cerne deste estudo é o desenvolvimento de um recurso léxico para a língua portuguesa que contenha informações de papéis semânticos. Para efeitos de comparação, esse recurso pode ser entendido como uma mistura entre a VerbNet e o PropBank. Ele contém sentenças extraídas de *corpora* como base para a anotação, assim como o PropBank, porém usa papéis semânticos descritivos, como a VerbNet.

Além disso, queremos comparar os verbos presentes em textos especializados com aqueles presentes em textos não especializados, de modo que, como será visto, utilizamos dois *corpora* de maneira contrastiva.

Para desenvolver esse recurso e a comparação entre os dois tipos de texto, realizamos um estudo-piloto com uma amostra de cinquenta verbos. Nesta seção, serão apresentados os materiais e os métodos empregados nesse estudo-piloto.

3.1 Materiais

3.1.1 *Corpora*

Neste trabalho, foram utilizados dois *corpora*: um composto por textos especializados e outro composto por textos não especializados. Para representar os textos especializados, selecionamos o *corpus* composto por artigos científicos da área da Cardiologia compilado por Zilio (2009). Para representar os textos não especializados, selecionamos o *corpus* de textos do jornal popular Diário Gaúcho, compilado no âmbito do projeto

[8] Disponível no site (acessado em 15/10/2013): http://www.nilc.icmc.usp.br/portlex/index.php/en/projects/propbankbringl.

[9] É importante deixar claro que a palavra *frame* é bastante polissêmica. Neste artigo, trataremos de *subcategorization frames* (estruturas de subcategorização), como vimos anteriormente, e também de *frames* como os da FrameNet, que são compreendidos como domínios semânticos ou estruturas conceptuais (por exemplo, o *frame* dirigir ou o *frame* jogo de futebol). Procuraremos deixar claro pelo contexto qual é o tipo de *frame* a que nos referimos.

[10] Exemplo retirado da FrameNet Brasil (Salomão, 2009). http://200.131.61.179/maestro/index.php/fnbr/report/frames?db=fncopa,

PorPopular [11]. Na Tabela 2, podemos ver a constituição dos *corpora* em relação ao número de palavras. Ambos os *corpora* foram analisados automaticamente pelo *parser* PALAVRAS (Bick, 2000) com árvores de dependências sintáticas.

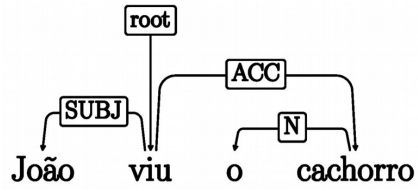
Corpus	Nº de palavras[12]
Cardiologia	1.605.250
Diário Gaúcho	1.049.487

Tabela 2: Tamanho dos *corpora*

Nessa anotação de dependências, o *corpus* anotado apresenta uma hierarquia de ligações entre os elementos sintáticos das sentenças. Um exemplo disso pode ser visto na seguinte sentença analisada com o *parser* PALAVRAS:

João viu o cachorro.

João [João] @SUBJ> ***#1->2***
viu [ver] @FS-STA ***#2->0***
o [o] @>N ***#3->4***
cachorro [cachorro] @<ACC ***#4->2***
$. ***#5->0***
</s>



Na anotação dessa sentença, se observarmos os valores após a cerquilha (#), é possível ver quais elementos estão ligados diretamente aos verbos e, com isso, extrair os argumentos. Isso porque o número antes do sinal "->" é o número da palavra, enquanto o número após o sinal "->" é o número da outra palavra à qual esta se liga. Assim, vemos que as palavras *João* e *cachorro* estão ligadas ao verbo *viu*, e este está ligado a 0, que é a raiz. Com isso, cria-se uma árvore de dependências que tem um verbo ligado à raiz e os demais elementos ligados a ele. Além disso, após a arroba (@), está identificada a categoria sintática à qual pertence cada palavra da sentença. Essa estrutura é então utilizada por um extrator de estruturas de subcategorização, que reconhece automaticamente os argumentos dos verbos e os organiza em um banco de dados.

[11] Para maiores informações sobre o projeto e o *corpus*, acesse: http://www.ufrgs.br/textecc/porlexbras/porpopular/index.php. Os números atuais apresentados no site diferem dos números apresentados neste artigo porque nosso *corpus* não compreende a totalidade de textos.

[12] Os números de palavras foram observados com a ferramenta WordSmith Tools, versão 4.0 (Scott, 2004).

3.1.2 Extrator de estruturas de subcategorização

O extrator de estruturas de subcategorização (Zanette, Scarton e Zilio, 2012; Zilio, Zanette e Scarton, 2012) é um *software* que realiza a preparação dos dados para a anotação. O extrator contém um conjunto de regras de extração, que são aplicadas às frases do *corpus* analisadas pelo *parser* PALAVRAS com árvores de dependências sintáticas. Durante a extração, com base nas informações fornecidas pelo *parser*, o sistema faz a identificação de quais verbos são auxiliares e quais são principais. Estes são utilizados, enquanto aqueles são excluídos e utilizados apenas para que possa ser reconhecido o sujeito da oração. Por exemplo, na sentença:

O cachorro foi visto por João.

O extrator reconhece *ver* como verbo principal. O sujeito *o cachorro* está ligado ao verbo auxiliar *ser*, mas o extrator consegue recuperar essa informação e o associa ao verbo *ver*. Desse modo, são mantidas apenas informações referentes a verbos principais.

Todas as informações extraídas são identificadas por meio de regras. Assim, o extrator busca informações como *@<ACC*, fornecidas pelo *parser*, as extrai e também as traduz em etiquetas mais explícitas para o anotador humano, como *OBJETO DIRETO*.

Após extrair as informações, os dados são armazenados em um banco de dados em formato MySQL. Para facilitar a anotação, existe uma interface de usuário que permite a visualização dos dados extraídos, com a classificação dos argumentos em uma ordem predefinida, assim como a anotação de papéis semânticos. Tal interface pode ser vista na Figura 1.

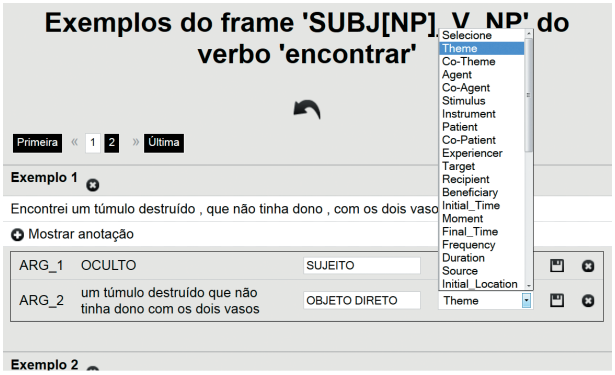


Figura 1: Interface de usuário para anotação

Essa interface mostra a estrutura de subcategorização (chamada de *frame*), o verbo em questão, os exemplos e os argumentos extraídos com a sua respectiva categoria sintática. Essas

informações são extraídas por meio de regras diretamente da anotação já presente nos *corpora*.

Ao anotador humano de papéis semânticos cabe o trabalho de criar uma lista de papéis semânticos, digitá-los em um arquivo de texto separados por vírgulas e selecioná-los a partir da lista de rolagem (que pode ser vista na Figura 1) no momento da anotação. Com essa interface de anotação, o anotador pode se concentrar no que lhe interessa: definir a semântica dos argumentos, sem precisar delimitá-los.

É importante ressaltar que, na estrutura atual, o banco de dados permite apenas a seleção de um papel semântico por argumento. Sendo assim, para teorias que admitem mais de um papel (por exemplo, Gelhausen, 2010), seria necessário modificar a arquitetura do sistema.

Outra característica importante da extração automática é que ela não distingue argumentos de adjuntos. Como aponta Cançado (2009), "a associação do [*status* de] argumento ao complemento de um verbo apresenta dificuldades, e a literatura sobre o assunto não é clara". Messiant (2008) também afirma que "não existem critérios linguísticos relevantes o suficiente para fazer uma distinção entre adjuntos e argumentos, não importando o contexto". Assim, em nossa anotação, não faremos uma distinção *strito sensu* entre adjuntos e argumentos; contudo, existem papéis que são potencialmente atribuídos apenas a adjuntos. Para contornar o problema de distinção entre argumentos e adjuntos, utilizamos a frequência como delimitador. Desse modo, se um elemento ocorre dez vezes ou mais junto ao verbo, ele é anotado, recebendo um *status* de argumento. Não sustentamos que todos os elementos anotados de acordo com esse princípio sejam realmente argumentos, porém, a sua existência tem uma influência sobre o significado de uma sentença, e esse significado deve ser reconhecido para se obter um bom desempenho na interpretação de um texto.

Por fim, apesar de o extrator já deixar os dados prontos para o anotador trabalhar, a análise automática de dependências sintáticas realizada pelo *parser* PALAVRAS nem sempre é correta. Existem erros de análise que vão desde a simples segmentação de sentenças até a delimitação dos argumentos. Além dos possíveis erros decorrentes da análise automática, o extrator de estruturas de subcategorização também organiza os dados de acordo com regras, e estas nem sempre são corretas. Desse modo, existem dados que podem conter ruído no banco de dados. Como veremos mais adiante, na Seção 3.2, esses dados errados são ignorados e não são anotados.

3.1.2 Lista de papéis semânticos

Outro elemento importante para a anotação de papéis semânticos em *corpora* é a definição de uma lista de papéis. Como vimos anteriormente, as listas podem se estender desde alguns poucos papéis (como no caso do PropBank) até dezenas de papéis (como no caso da VerbNet e da FrameNet), dependendo da abordagem escolhida.

Neste estudo, optamos por uma lista de papéis descritivos e genéricos, nos moldes da VerbNet. Optamos por esse tipo de lista tendo em vista que já existe um PropBank.Br que utiliza uma lista de protopapéis e também uma FrameNet.Br que usa papéis específicos para o contexto. A VerbNet.Br, apesar de existir, foi feita a partir de uma importação de dados do inglês, tomando por base o potencial interlinguístico das classes de Levin. No entanto, não houve um estudo linguístico mais profundo que mostrasse o quanto essa importação realmente seria aplicável ao português. Desse modo, decidimos focar nossos esforços no mesmo âmbito da VerbNet.Br, porém partindo de uma anotação manual, que posteriormente poderá ser confrontada com os dados importados do inglês presentes na VerbNet.Br.

Em trabalho anterior (Zilio, Zanette e Scarton, 2012), apresentamos um breve estudo com uma lista de 46 papéis semânticos proposta por Brumm (2008). Ao final do estudo, percebemos que os papéis não eram adequados ao nosso tipo de abordagem, por serem muito específicos em alguns casos, e muito genéricos em outros. Não havia um equilíbrio nos papéis, provavelmente pelo fato de que eles ainda não haviam sido testados em dados reais. A falta de dados anotados com os papéis propostos por Brumm também faz com que não haja exemplos concretos que possam ser observados para melhor compreender a funcionalidade de cada um dos papéis. Por isso, neste estudo, optamos por utilizar uma lista já testada, revisada e com exemplos que podem ser livremente acessados na internet: a lista da VerbNet (Kipper, 2005) em sua versão 3.2. Essa lista, além de permitir posteriormente uma comparação direta com os dados da VerbNet.Br, também nos pareceu ser a melhor para suprir as falhas que observamos na lista de Brumm (2008).

Após um estudo dos papéis semânticos e dos exemplos disponíveis na VerbNet, realizamos algumas pequenas modificações na lista. A principal modificação foi a criação do hiperônimo TARGET[13], que passou a abrigar BENEFICIARY e

[13] Por termos escolhido uma lista em inglês, optamos por não traduzir os nomes dos papéis e por manter a nomenclatura toda em inglês. Assim, quando nos referirmos a papéis genéricos, utilizaremos nomes em português, como AGENTE e PACIENTE,

RECIPIENT, para os casos em que um verbo autoriza ambos. As demais modificações apenas alteraram o entendimento da hierarquia da VerbNet, mas não modificaram os papéis em si.

No total, definiu-se uma lista com 38 papéis semânticos: THEME, CO-THEME, AGENT, CO-AGENT, STIMULUS, INSTRUMENT, PATIENT, CO-PATIENT, EXPERIENCER, TARGET, RECIPIENT, BENEFICIARY, INITIAL TIME, MOMENT, FINAL TIME, FREQUENCY, DURATION, SOURCE, INITIAL LOCATION, MATERIAL, GOAL, DESTINATION, RESULT, PRODUCT, LOCATION, TRAJECTORY, ATTRIBUTE, TOPIC, PIVOT, VALUE, EXTENT, ASSET, CAUSE, REFLEXIVE, PREDICATE, VERB, MANNER e COMPARATIVE.

Pode parecer estranho o uso dos papéis semânticos em inglês, porém, por estarmos utilizando como fonte a VerbNet, acreditamos que essa escolha simplificará uma comparação futura do português com o inglês.

Alguns desses papéis se aplicam potencialmente apenas a adjuntos, como MANNER e COMPARATIVE, outros são papéis auxiliares, como VERB e REFLEXIVE, que se aplicam, respectivamente, a argumentos que formam um significado complexo com o verbo (por exemplo, casos de verbos-suporte) e à partícula reflexiva.

Para manter o artigo sucinto, não explicaremos aqui as funcionalidades específicas de cada um dos papéis. Essas informações podem ser obtidas na documentação da VerbNet[14].

3.2 Método de anotação

Para realizar a anotação de papéis semânticos nas sentenças dos *corpora*, fizemos inicialmente algumas escolhas em relação às quantidades a serem anotadas. Neste estudo, optamos por uma anotação amostral, almejando um teste dos papéis semânticos apresentados pela VerbNet. Optamos por anotar, nos dois *corpora*, primeiro os 25 verbos mais frequentes do *corpus* de Cardiologia e, em seguida, também nos dois *corpora*, os 25 verbos mais frequentes do *corpus* do Diário Gaúcho, excluindo os que já haviam sido anotados na primeira etapa. Foram anotados, dessa forma, 50 verbos ao todo em cada um dos *corpora*[15] — com os seguintes critérios:

- Os seguintes verbos foram excluídos: *ser*, *estar*, *ter* e *haver*.
- Foram anotadas exatamente dez sentenças de cada estrutura de subcategorização.
- Os verbos anotados tinham de estar presentes nos dois *corpora* com frequência suficiente para que pelo menos dez sentenças fossem anotadas dentro de pelo menos uma estrutura de subcategorização.

A exclusão *a priori* de quatro verbos (*ser*, *estar*, *ter* e *haver*) se deu por eles serem extremamente polissêmicos e/ou frequentes nos dois *corpora*. A anotação desses verbos com o método adotado dificilmente refletiria as suas várias facetas, além de consumir muito tempo devido à quantidade de estruturas de subcategorização existentes para cada um deles.

A escolha de dez exemplos, para cada estrutura de subcategorização, foi apenas um incremento em relação ao método usado em Zilio, Zanette e Scarton (2012). Com a modificação apresentada aqui, garantimos que todas as estruturas de subcategorização tenham dez exemplos anotados. Se uma estrutura tem 16 exemplos, mas apenas nove estão corretos, ela é descartada como um todo.

A presença dos verbos nos dois *corpora* foi uma exigência para a sua anotação tendo em vista o objetivo comparativo deste estudo. Não nos adiantava anotar verbos em apenas um dos *corpora*, pois não seria possível comparar os resultados.

Neste estudo-piloto, a anotação foi desenvolvida por apenas um anotador linguista treinado[16], o qual teve acesso a um manual de anotação com a descrição dos papéis semânticos e de alguns exemplos retirados ou da VerbNet ou da anotação realizada em um estudo anterior. Foi realizado também um experimento com múltiplos anotadores, o qual será relatado mais adiante, na Seção 5, porém, para este estudo-piloto, usamos apenas um anotador.

4 Resultados e considerações sobre a anotação de papéis semânticos

Nesta seção, expomos nossas considerações qualitativas sobre o método empregado na anotação de papéis semânticos e, em seguida, apresentamos os resultados da anotação e da comparação entre os dois corpora.

porém, quando nos referirmos à nomenclatura empregada no estudo, usaremos o inglês.

[14] Disponível no site: http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html.

[15] Houve apenas uma exceção a isso. A título de curiosidade, anotamos o verbo *ir* no *corpus* do Diário Gaúcho. Assim, o Diário Gaúcho tem, na verdade, 51 verbos anotados. Esse verbo seria anotado também no *corpus* de Cardiologia, mas a sua frequência não foi suficiente.

[16] O anotador mencionado é o primeiro autor deste artigo.

4.1 Considerações sobre o método

A lista de papéis semânticos da VerbNet se mostrou adequada na maioria dos casos, pois se aplicou bem aos argumentos dos verbos anotados. Os únicos problemas encontrados nesse sentido foram resultantes da união da lista com uma metodologia que não distingue entre argumentos e adjuntos. Como optamos por anotar todos os elementos que se ligassem ao verbo, considerando que a frequência seria o delimitador entre argumento e adjunto, alguns elementos anotados, por serem de natureza adverbial, não tinham um papel semântico condizente, precisando ser anotados com papéis que se adequavam apenas parcialmente. Esse tipo de problema pode ser solucionado se adicionarmos os papéis semânticos específicos para adjuntos que são utilizados no PropBank.

Em geral, a anotação de adjuntos adverbiais foi uma tarefa complexa. Observando as sentenças 6a a 6d a seguir, extraídas dos *corpora*, temos adjuntos adverbiais que contêm *jogo* e *estudo* (destacados em negrito). Esses adjuntos adverbiais apresentam uma dificuldade para a atribuição de um papel semântico. Poderíamos, por exemplo, anotar essas estruturas como, MOMENT, LOCATION ou mesmo INSTRUMENT, dependendo de sua situação na sentença, mas não tínhamos um papel que representasse um significado como SITUATION. Isso ocorreu porque os papéis semânticos da VerbNet foram pensados apenas para argumentos e não para adjuntos. Assim, cremos que seja necessária fazer uma separação entre papéis para argumentos e papéis para adjuntos, como fez o projeto PropBank. Apesar de alguns problemas no que diz respeito aos adjuntos adverbiais, os demais argumentos sempre tinham algum papel semântico na lista que se adequava.

6a. Eles fizeram um jogo largado e nós demos oportunidade ***em um jogo*** *que estava em nossas mãos.*
6b. Teremos de melhorar muito em relação ao que mostramos ***no primeiro jogo****, mas temos todas as condições de reverter.*
6c. ***No presente estudo****, animais adultos restritos apresentaram aumento de todos os parâmetros estereológicos analisados na aorta, sugerindo hiperplasia da túnica média.*
6d. O prognóstico utilizado para o TC6M foi demonstrado ***no estudo SOLVD10****.*

No que diz respeito ao método amostral escolhido, ele se mostrou adequado para a maioria dos verbos, pois representa bom equilíbrio entre tempo utilizado para anotar e representatividade dos dados anotados. Porém, ficou claro que, para verbos muito polissêmicos (por exemplo, *dar*), a amostragem não capta grande parte dos significados do verbo. No entanto, se aumentarmos o número de exemplos anotados a cada estrutura de subcategorização, o esforço necessário para anotar cada um dos verbos também aumentaria, talvez tornando impossível uma anotação de muitos verbos em um tempo aceitável, tendo em vista que temos apenas um anotador. Por mais que sempre exista um problema com o método amostral (afinal, alguns dados são ignorados), depois que a anotação é feita, é possível perceber quais verbos não estão representados adequadamente e, se necessário, é possível dar um tratamento especial a eles.

A ferramenta utilizada para a extração e anotação dos dados desenvolvida por Zanette (2010) é bastante versátil e pode ser adaptada às necessidades do anotador. Por exemplo, para alterar a lista de papéis semânticos, basta modificar um arquivo de texto. Entretanto, com a anotação de mais verbos em relação ao estudo anterior (Zilio, Zanette e Scarton, 2012), percebemos que alguns elementos linguísticos das sentenças são anotados pelo *parser* PALAVRAS (Bick, 2000) de uma forma que não estava sendo levada em consideração pelo sistema. Por exemplo, agentes da passiva são anotados pelo PALAVRAS como PASS, e os objetos indiretos são anotados tanto como PIV quanto como SA; porém, o sistema estava preparado apenas para reconhecer PIVs e ADVLs. Portanto, alguns agentes da passiva acabaram não sendo reconhecidos (pois não apresentavam a marcação ADVL) e o mesmo aconteceu com os objetos indiretos marcados como SA. Para garantir que não haverá mais esse tipo de problema, fizemos uma análise do conjunto completo de etiquetas empregadas pelo PALAVRAS[17] e acrescentamos ao sistema, com a respectiva descrição, as modificações necessárias.

Apesar de termos utilizado a ferramenta desenvolvida por Zanette (2010), existem outras ferramentas que poderiam ser empregadas para a anotação, como a ferramenta SALTO (Burchardt et al., 2006). Entretanto, o sistema de anotação dessa ferramenta é muito mais complexo, deixando ao encargo do anotador a tarefa de delimitar os argumentos. Por um lado, isso garante uma maior precisão na delimitação dos argumentos; por outro lado, aumenta a chance de erros e aumenta o trabalho dispendido na anotação.

[17] As etiquetas com as respectivas explicações de suas funções em http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html.

4.2 Resultados da anotação e comparação entre os *corpora*

Seguindo os métodos apontados na Seção 3.2, realizamos a anotação de 3.400 orações (1.790 orações no corpus de Cardiologia e 1.610 no *corpus* do Diário Gaúcho). Essas orações se encontram atualmente armazenadas em um banco de dados em formato MySQL, o qual foi exportado também para XML[18]. Com os dados em formato XML, o compartilhamento dos dados se torna mais simples, pois o formato XML é mais acessível do que o formato MySQL. No atual estágio, a distribuição do recurso anotado requer apenas o compartilhamento de um arquivo específico para cada *corpus*[19].

No que diz respeito à diferença de frequências entre os *corpora*, temos exemplos bastante discrepantes. Por exemplo, o verbo *considerar*, bastante frequente em Cardiologia, com 60 sentenças anotadas, encontra no *corpus* do Diário Gaúcho uma contraparte de apenas 10 sentenças.

Essas diferenças poderiam ter sido amenizadas na organização do *corpus* através da seleção de sentenças específicas em vez de textos completos. Porém, isso implicaria na construção de um novo recurso a partir do zero, o que demandaria tempo. Além disso, uma organização desse tipo poderia camuflar algumas diferenças existentes entre os dois tipos de linguagem, o que não seria bom para este estudo; afinal, almejamos observar a linguagem em sua forma natural, com diferenças e semelhanças que variam desde as frequências até as estruturas.

Entre as 1790 orações do *corpus* de Cardiologia, observaram-se 304 estruturas sintático-semânticas[20] diferentes, sendo esta, que conta com apenas um argumento, a mais frequente: SUJ<Theme>; no *corpus* do Diário Gaúcho, entre as 1610 orações, encontraram-se 272 estruturas diferentes, sendo mais frequente uma estrutura com dois argumentos: SUJ<Agent>+OBJ.DIR<Theme>.

Em ambos os *corpora*, houve muitas ocorrências de estruturas sintático-semânticas com frequência 1; dentre elas, 117 estavam no *corpus* de Cardiologia e 106 no *corpus* do Diário Gaúcho. Normalmente, frequências baixas são descartadas, por não representarem informações relevantes. Em nosso caso, porém, por serem informações atribuídas manualmente, dificilmente a baixa frequência pode ser desconsiderada sem um bom motivo. Além disso, o fato de que existe apenas uma sentença no *corpus* até então anotada com a estrutura sintático-semântica SUJ<Theme>+ADJ.ADV [em]<Location>+ADJ.ADV [a]<Goal>[21] para o verbo *chegar* não quer dizer que haja apenas uma ocorrência de cada uma das associações sintático-semânticas SUJ<Theme>, ADJ.ADV[a]<Loca-tion> e ADJ.ADV[em] <Goal> para esse mesmo verbo. Durante o aprendizado de máquina de sistemas de anotação de papéis semânticos, não apenas a estrutura como um todo pode ser relevante, mas também cada um de seus elementos individuais.

Nas Tabelas 3 e 4, podemos ver as cinco estruturas mais frequentes nos dois *corpora*. Nessas tabelas, é possível observar que, enquanto o *corpus* de Cardiologia privilegia construções passivas e intransitivas (o que explica a ocorrência de duas estruturas sem objetos), o Diário Gaúcho apresenta estruturas agentivas transitivas diretas no topo, seguidas por passivas e intransitivas.

Quando observamos no banco de dados os verbos e sentenças que se enquadram nas estruturas mais frequentes sem objetos, percebemos que, no caso da Cardiologia, se trata, na maioria dos exemplos, de utilização de voz passiva[22], e nem tanto de intransitividade. Já no Diário Gaúcho ocorre o oposto, com uma maioria de exemplos intransitivos[23]. Isso contraria nossas observações em estudo anterior[24], quando havíamos observado

[18] A ferramenta de exportação foi desenvolvida por Samy Sassi sob orientação de Leonardo Zilio, Carlos Ramisch e Mathieu Mangeot.

[19] Os arquivos encontram-se disponíveis para *download* gratuito em http://cameleon.imag.fr/xwiki/bin/view/Main/Resources.

[20] Por "estruturas sintático-semânticas", nos referimos às associações entre estruturas sintáticas (sujeito, objeto direto etc.) e papéis semânticos (Agent, Patient etc.) em uma oração. Para simplificar a representação das estruturas sintático-semânticas, utilizaremos as seguintes abreviaturas para a sintaxe:

- SUJEITO = SUJ
- OBJETO DIRETO = OBJ.DIR
- ADJUNTO ADVERBIAL[prep.] = ADJ.ADV[prep.]

[21] A sentença em questão é *Em alguns trechos, a água chegou a 1,5m de altura*.

[22] Alguns exemplos:

- *Foram avaliados os seguintes parâmetros:*
- *Foi observada uma distribuição igual de a população estudada em relação ao sexo.*
- *Não foram demonstradas evidências consistentes de o papel de níveis circulantes de MIF, IL-6 e sCD40L como marcadores de SCA.*

[23] Observou-se uma maioria de verbos como "ocorrer", "existir", "ficar", "acontecer".

[24] Para aquele estudo, foram selecionados quatro verbos com frequências próximas nos dois *corpora*: *levar*, *encontrar*, *usar* e *receber* (Zilio, Zanette e Scarton, 2012). Os resultados sobre o apassivamento, porém, não foram divulgados. Na época, os autores consideraram que seria muito precipitado publicar resultados de um fenômeno tão amplo levando em consideração apenas quatro verbos. Mesmo agora, com nossos 50 verbos, por mais que sejam os mais frequentes presentes nos dois corpora, talvez não tenhamos resultados representativos o suficiente para uma conclusão, devido à

uma tendência maior de apassivamento no Diário Gaúcho, o que provavelmente era um fenômeno pertinente apenas aos verbos estudados.

Cardiologia		
Estrutura	**Freq.**	**Freq. %**
SUJ<Theme>	181	10,11
SUJ<Theme>+ADJ.ADV[em] <Location>	121	6,76
SUJ<Instrument>+OBJ.DIR <Theme>	102	5,70
SUJ<Agent>+OBJ.DIR<Theme>	63	3,52
SUJ<Patient>	40	2,23

Tabela 3: Cinco estruturas mais frequentes no *corpus* de Cardiologia

Diário Gaúcho		
Estrutura	**Freq.**	**Freq. %**
SUJ<Agent>+OBJ.DIR<Theme>	171	10,62
SUJ<Theme>	114	7,08
SUJ<Agent>	92	5,71
SUJ<Theme>+ADJ.ADV [em] <Location>	50	3,11
SUJ<Agent>+OBJ.DIR<Theme> +ADJ.ADV [em]<Location>	45	2,79

Tabela 4: Cinco estruturas mais frequentes no *corpus* do Diário Gaúcho

Tanto o Diário Gaúcho quanto o *corpus* de Cardiologia apresentam estruturas transitivas diretas em posições elevadas na lista, porém, na Cardiologia, há uma tendência para que o sujeito seja um INSTRUMENT, deixando o real agente apagado. O mesmo não se observa no Diário Gaúcho, que apresenta grande quantidade de sujeitos agentes. Esse fenômeno não é algo que se apresenta apenas entre as estruturas sintático-semânticas mais recorrentes, mas ao longo das várias estruturas existentes. A Cardiologia apresentou uma forte tendência a esconder os verdadeiros agentes, colocando em evidência os instrumentos utilizados.

Na comparação, não se pode afirmar que os *corpora* utilizem estruturas sintático-semânticas diferentes, pois quase todas as estruturas ocorrem nos dois tipos de texto. O que se percebe é mais uma tendência diferente no *corpus* especializado, sendo que o principal fator é o apagamento dos agentes. Para sustentar esse resultado com números, observamos que, dentre as 304 estruturas sintático-semânticas anotadas, apenas 31 apresentavam um agente, enquanto no Diário Gaúcho, dentre as 272 estruturas, 121 apresentavam agente. Isso representa um salto de 10,19% para 44,49% entre os *corpora*.

Em termos de exemplos concretos, os sujeitos em Cardiologia tendem a ser expressões como estas (extraídas do banco de dados):

- *Estudos de o perfil lipídico*;
- *a combinação de restrição calórica com exercício físico*; e
- *Análises futuras de feocromocitomas com técnicas de microarray proteômica*;

enquanto o Diário Gaúcho apresenta mais sujeitos como estes:

- *o jogador*;
- *o técnico Abel=Braga*; e
- *Leona=Cavali*.

Como pode ser visto na Tabela 6 (linha 1), os dados sobre a agentividade se mantêm distintos quando olhamos para o número de sentenças. A Cardiologia apresenta 198 sentenças com AGENT em 1790 (11,06%) contra 734 sentenças em um total de 1610 (45,59%) no Diário Gaúcho. Também é possível perceber que a quantidade de sentenças com INSTRUMENT (linha 16 da Tabela 6) é mais de três vezes maior em Cardiologia do que no Diário Gaúcho. Outras diferenças estão no fato de que o papel PIVOT (linha 22 da Tabela 6), que geralmente representa um elemento que contém outro elemento, sem participar em uma ação, ocorre quase seis vezes mais no *corpus* de Cardiologia do que no do Diário Gaúcho, e o papel GOAL (linha 14 da Tabela 6), que geralmente representa um objetivo de uma ação, também é muito mais frequente naquele do que neste.

Utilizando o coeficiente de correlação tau-b de Kendall [25], realizamos três experimentos com diferentes informações.

No experimento 1, avaliamos a correlação entre os *rankings* dos papéis semânticos nos dois *corpora*, considerando também as informações sintáticas e a distribuição nas sentenças. Utilizamos os dados conforme estão representados nas Tabelas 3 e 4. Nesse experimento, os resultados apontaram que há uma correlação inversa entre as amostras, pois encontramos um valor de $\tau = -0{,}394$ ($p < 0{,}001$). Assim, percebe-se que estruturas sintático-semânticas muito frequentes no corpus de cardiologia tendem a ser pouco frequentes no corpus do Diário Gaúcho e vice-versa. Esse

amplitude do fenômeno. Assim, nossos resultados devem ser observados com cautela.

[25] O coeficiente de correlação tau-b de Kendall avalia se existe uma correlação entre os *rankings* de duas amostras. Assim, ele informa se o ranqueamento de uma amostra X é correlacionado ao ranqueamento de uma amostra Y. Os valores possíveis de τ variam entre -1 e 1, sendo 0 uma indicação de que não há correlação. Os cálculos estatísticos foram realizados com a ferramenta IBM SPSS 19.

resultado corrobora algumas tendências observadas na análise qualitativa anterior.

No experimento 2, observamos a correlação entre os dois *corpora* no que diz respeito a papéis semânticos associados às suas respectivas anotações sintáticas. Isto é, em vez de utilizarmos a estrutura sintático-semântica das sentenças (como fizemos no experimento 1), consideramos apenas os argumentos isolados, com suas informações sintáticas e semânticas, da forma como representamos na Tabela 5. Com esse conjunto de dados, não houve correlação entre as duas amostras ($\tau = 0{,}031$; $p = 0{,}608$).

Por fim, no experimento 3, consideramos apenas o *ranking* dos papéis semânticos, sem observar a anotação sintática. Os dados foram utilizados exatamente da forma como estão apresentados na Tabela 6. O valor de τ foi 0,521 ($p < 0{,}001$), indicando uma correlação positiva.

Desse modo, os resultados dos três experimentos mostraram que, quanto mais complexa for a informação analisada, maior é a distância entre as amostras. É importante ressaltar que, para esses experimentos, não consideramos o verbo presente nas sentenças ou ao qual os argumentos estavam associados. Observamos apenas as informações sintáticas e de papéis semânticos de maneira isolada.

Cardiologia	Freq.	Diário Gaúcho	Freq.
SUJEITO<Theme>	659	SUJEITO <Agent>	733
OBJETO DIRETO <Theme>	507	OBJETO DIRETO <Theme>	494
ADJUNTO ADVERBIAL [em] <Location>	356	SUJEITO <Theme>	338
SUJEITO <Instrument>	217	ADJUNTO ADVERBIAL [em] <Location>	259
SUJEITO <Result>	190	SUJEITO <Patient>	171

Tabela 5: Dados sintático-semânticos dos dois *corpora*[26]

N	Papéis Semânticos	Cardiologia	Diário Gaúcho
1	AGENT	198	734
2	ATTRIBUTE	97	46
3	BENEFICIARY	113	109
4	CAUSE	120	71
5	CO-AGENT	0	16
6	COMPARATIVE	19	0
7	CO-PATIENT	19	0
8	DESTINATION	1	91
9	DURATION	38	9
10	EXPERIENCER	41	93
11	EXTENT	29	11
12	FINAL_TIME	0	11
13	FREQUENCY	2	0
14	GOAL	215	84
15	INITIAL_TIME	0	11
16	INSTRUMENT	294	91
17	LOCATION	407	274
18	MATERIAL	0	15
19	MANNER	88	30
20	MOMENT	194	202
21	PATIENT	241	212
22	PIVOT	132	23
23	RECIPIENT	0	12
24	REFLEXIVE	4	20
25	RESULT	269	257
26	SOURCE	57	2
27	STIMULUS	6	11
28	TARGET	8	51
29	THEME	1221	962
30	TOPIC	20	14
31	VALUE	12	0
32	VERB	83	44

Tabela 6: Papéis semânticos e sua frequência nos dois *corpora*

5 Experimento com múltiplos anotadores

Além da anotação, também desenvolvemos um experimento paralelo relacionado ao tema. O experimento foi uma tentativa de passar a anotação a múltiplos anotadores.

Existem estudos que já observaram a tarefa de anotação com mais de um anotador. Hovy et al. (2006) apresentaram uma solução para se obter 90% ou mais de concordância entre anotadores. Para tal, uniram-se os *frames* do PropBank aos significados da WordNet, de modo que o anotador apontava qual era o significado do verbo, e o *frame* era automaticamente selecionado e atribuído.

Atualmente, existe um estudo (Fossati, Giuliano e Tonelli, 2013) que busca levar a anotação da FrameNet para múltiplos anotadores não especialistas. Para tal, foram simplificadas as definições de cada um dos elementos do *frame* e foram conduzidos experimentos em duas etapas: a primeira etapa envolvia apenas a desambiguação do verbo, bastante similar ao experimento de Hovy et al. (2006), porém, que tomou como base o trabalho de Hong e Baker (2011); a segunda etapa era indicar quais argumentos deveriam ser anotados com os papéis semânticos associados ao significado predefinido do verbo. Enquanto a primeira etapa obteve resultados com mais de 90% de acurácia (apesar de o único verbo apresentado ter ficado em 81,9%), a segunda etapa não teve resultados tão positivos.

O elemento em comum nos dois trabalhos apresentados é que já existe um recurso anterior

[26] Essa tabela apresenta apenas os dados mais frequentes, a título de exemplo; porém, para o cálculo do *p*, utilizamos a tabela completa, que contém mais de 140 linhas.

que pode ser utilizado como base. Hovy et al. (2006) tinham o PropBank com milhares de sentenças anotadas e só buscava expandir a anotação para outros *corpora*, e Fossati, Giuliano e Tonelli (2013) têm a FrameNet e, da mesma forma, apenas buscam expandir a anotação para outros *corpora*.

Em nosso caso, não existe ainda um recurso para o português que contenha a anotação de papéis semânticos descritivos. Desse modo, é preciso deixar claro que o ponto de partida para o experimento descrito aqui é diferente dos experimentos já realizados por outros autores.

Nossa intenção com o experimento é observar se, para a criação de um recurso com anotação de papéis semânticos, seria mais útil utilizar a anotação de múltiplos anotadores com pouco treinamento ou de apenas um com muito treinamento (que é o método que está sendo utilizado).

5.1 Procedimento

Para o experimento, foram selecionados dez anotadores linguistas (alunos de pós-graduação em Linguística da UFRGS) e 25 sentenças extraídas dos *corpora* apresentados na Seção 3.1.1. O treinamento foi básico, consistindo apenas em uma explicação sobre a tarefa e o assunto, e no fornecimento de um manual de anotação.

No manual de anotação, cada um dos papéis semânticos que poderiam ser utilizados foi apresentado ao lado de uma descrição, como pode ser visto neste exemplo:

LOCATION	Lugar (físico ou metafórico, real ou fictício) onde uma ação ocorre.

A estrutura das sentenças a serem anotadas foi similar à que apresentamos na Figura 1, com a ressalva de que, por ser uma anotação em papel, não havia uma lista de rolagem para escolher as sentenças (apenas uma lista para consulta no manual de anotação).

Além da anotação dos papéis semânticos, também fazia parte da tarefa a distinção de cada um dos elementos anotados entre argumentos e adjuntos. Para tal, foi apresentada também uma breve explicação sobre a diferença entre argumentos e adjuntos[27].

Eis aqui um exemplo dos dados apresentados para anotação:

O resultado de o exame para investigar vestígios de pólvora em suas mãos , para saber se ele **utilizou** arma , teve resultado negativo .
SUJ = ele _____ ( ) Arg / ( ) Adj
OD = arma _____ ( ) Arg / ( ) Adj
Comentário:

Cada uma das sentenças a ser anotada era apresentada da forma como estava no banco de dados, seguida pelos argumentos (as abreviaturas estavam explicitadas no manual de anotação) com um espaço para escrever o papel semântico e a opção entre argumento ou adjunto. Por fim, acrescentamos um espaço para os comentários do anotador.

5.2 Cálculo de concordância

Após a anotação ter sido realizada, para observar se houve concordância entre os anotadores, utilizamos cálculos com base no coeficiente π, um dos possíveis coeficientes utilizados para a observação de concordância entre anotadores. Em geral, utiliza-se o coeficiente κ para essa tarefa, por isso, discutimos a seguir os motivos que nos levaram a optar por outro coeficiente.

Artstein e Poesio (2008) apresentam uma longa discussão acerca de diversos coeficientes e testes utilizados para avaliar a concordância entre anotadores. Os autores chamam atenção para o fato de que há um problema de terminologia, pois o teste desenvolvido por Fleiss (1971) acabou sendo chamado de multi-κ, apesar de tomar como base o coeficiente π e, portanto, ter um pressuposto diferente. Como existe esse problema de terminologia, Artstein e Poesio (2008) propõem que se utilize κ para o teste de Cohen (1960), multi-π para o teste de Fleiss (1971) e multi-κ para o teste de Davies e Fleiss (1982). Neste estudo, seguiremos a proposta de Arstein e Poesio (2008) em relação à terminologia.

Vejamos as principais diferenças entre os coeficientes. Segundo Artstein e Poesio (2008), os testes que usam π como base partem do pressuposto de que a distribuição das etiquetas não é uniforme, mas que a distribuição entre os anotadores o é. Assim, para um dado conjunto de etiquetas, cada uma das etiquetas tem a mesma probabilidade de ser utilizada por todos os anotadores, mas algumas têm mais chance de serem utilizadas do que outras. No caso dos testes que utilizam κ como base, tanto a distribuição das etiquetas quanto a distribuição das anotações é pressuposta como não uniforme, sendo assim, todas as distribuições são consideradas independentes entre si.

[27] Como já mencionamos anteriormente, sabemos que a distinção entre argumentos e adjuntos é um assunto bastante controverso nas teorias gramaticais, por isso, nos limitamos a mostrar que a diferença se dá em relação ao quanto determinado elemento afeta o significado do verbo.

Por exemplo, dado um conjunto de etiquetas AGENT, THEME e LOCATION e três anotadores A, B e C, um teste com base em π observa a totalidade dos dados e avalia uma distribuição não uniforme para as etiquetas (por exemplo, 50% dos argumentos receberiam a etiqueta AGENT, 30% THEME e 20% LOCATION), essa mesma distribuição será aplicada a todos os anotadores: A, B e C. No caso do κ, para esse mesmo conjunto de etiquetas e anotadores, seria avaliada a distribuição das anotações para cada um dos anotadores; desse modo, teríamos, por exemplo: 40% para AGENT, 35% para THEME e 25% para LOCATION no caso do anotador A; 60% para AGENT, 20% para THEME e 20% para LOCATION no caso do anotador B; e 45% para AGENT, 45% para THEME e 10% para LOCATION no caso do anotador C. Assim, a concordância de κ leva em conta não somente a distribuição das etiquetas, mas também a anotação feita por cada um dos anotadores. Conforme apontam Artstein e Poesio (2008), na teoria, essa diferença é bastante grande, porém, na prática, ela perde um pouco a sua força, pois os coeficientes π e κ resultam em valores muito próximos, e, no caso de multi-π e multi-κ, essa diferença varia muito menos, pois ela tende a se extinguir conforme o número de anotadores aumenta.

Como temos mais de dois anotadores, a diferença entre os coeficientes é muito pequena, mas, ainda assim, é importante que se decida por um ou outro em virtude dos pressupostos assumidos. Neste estudo, assumem-se os pressupostos de π, pois estamos avaliando a confiabilidade dos dados anotados por vários anotadores, de modo que as etiquetas devem ter uma distribuição não uniforme, mas os anotadores deveriam anotar de modo consistente e similar. Sendo assim, para verificar a concordância entre os anotadores e também entre os pares de anotadores, empregamos, respectivamente, os testes multi-π e π. A observação da concordância entre os pares de anotadores serve principalmente para detectar *outliers* (isto é, pessoas que não entenderam a tarefa ou que realizaram a anotação sem prestar muita atenção aos dados) e poder dar mais confiabilidade ao multi-π. Os cálculos foram levados a cabo por meio de uma ferramenta presente no mwetoolkit (Ramisch, Villavicencio e Boitet, 2010a; 2010b) que calcula vários coeficientes de concordância.

5.3 Resultados da anotação com múltiplos anotadores

Primeiramente, observamos a distinção entre argumentos e adjuntos, que consideramos ser uma tarefa mais simples (principalmente por haver apenas duas possibilidades de anotação), para observar se algum dos anotadores se caracterizava como *outlier*. Para essa observação, comparamos os anotadores em pares calculando o π entre eles.

A distinção entre argumentos e adjuntos, apesar de ser bastante controversa no caso de alguns verbos, deveria ser bastante simples na maioria dos casos. Por exemplo, na sentença 7, a seguir, é possível perceber que o sujeito (*O PT*) e o objeto direto (*um projeto de lei*) são argumentos, por serem necessários para que o verbo expresse seu significado completo, enquanto o adjunto adverbial (*no Congresso*) aparece apenas para acrescentar uma informação que não depende do verbo.

7. *O PT* ***apresentou*** *no Congresso um projeto de lei que cria contribuição social sobre fortunas.*

Por isso, esperávamos um alto nível de concordância nessa tarefa. Porém, não foi isso que observamos. Ao analisar os valores de π para os pares de anotadores utilizando apenas dados da distinção entre argumento e adjunto, percebemos que três anotadores apresentaram níveis baixos de concordância com os demais anotadores, a ponto de haver valores negativos entre eles (o que indica discordância). Uma das possíveis explicações para isso é que talvez eles não tenham compreendido a tarefa, ou simplesmente fizeram a anotação com pressa, deixando de ponderar adequadamente cada uma das instâncias a ser anotada. Dado o baixo nível de concordância entre esses anotadores em relação aos demais, o multi-π também foi baixo, com um valor de 0,315020 (multi-κ = 0,320770).

Com a retirada desses três *outliers*, o valor do coeficiente multi-π aumenta para 0,553020, mas continua abaixo dos 0,8, apontados por Neuendorf (2002, *apud* Arstein e Poesio, 2008) como mínimo necessário para que se considere que haja uma boa concordância. Assim, duas conclusões vêm imediatamente à mente: ou a tarefa não estava clara para os anotadores, ou a anotação de argumentos e adjuntos não é tão simples quanto imaginávamos.

Passemos então para a tarefa mais importante, que é a anotação de papéis semânticos. Para o cálculo do multi-π dessa tarefa, também retiramos os mesmos três *outliers*, afinal, se eles não haviam compreendido (ou haviam feito às pressas) a tarefa de distinguir entre argumento e adjunto, cremos que não havia por que confiar nos seus resultados em uma tarefa muito mais complexa, que envolve mais de trinta possíveis anotações, e não apenas duas. Assim, dentre os sete anotadores restantes, obtivemos um multi-π de 0,253407 (multi-κ = 0,256954). Esse valor é extremamente baixo, de

modo que se pode dizer que praticamente não houve concordância entre os anotadores.

Observando-se as anotações individuais, percebe-se que houve alguns pontos de convergência, principalmente na atribuição do papel AGENT, MOMENT, LOCATION e, em alguns casos, THEME. No entanto, quando outros papéis eram requeridos, os anotadores discordaram de modo a ter, em alguns casos, uma anotação diferente para cada anotador. Em mais de um caso, em uma mesma sentença, houve total concordância em um argumento, mas discordância nos demais. Por exemplo, no caso da sentença 7, acima, os 10 anotadores concordaram que o sujeito *O PT* desempenha a função de AGENT, no entanto, o objeto direto *um projeto de lei* teve apenas 5 anotadores concordando com o papel THEME, e o adjunto adverbial *no Congresso* contou com apenas 6 anotadores optando por LOCATION.

Outras sentenças não tiveram concordância em nenhum dos argumentos. Por exemplo, a sentença 8 não teve concordância em nenhum dos argumentos. O sujeito *A versão religiosa* recebeu 4 anotações como AGENT e 3 como THEME, enquanto o objeto indireto *com as mulheres Jaca ou Melancia* foi anotado como THEME por 4 anotadores e como ATTRIBUTE por 3.

8. *A versão religiosa não* ***conta*** *com as mulheres Jaca ou Melancia , mas todas=as velocidades estão lá , em a música .*

Apenas a sentença 9 apresentou uma maior concordância entre os anotadores no que diz respeito aos dois argumentos.

9. *O resultado de o exame para investigar vestígios de pólvora em suas mãos , para saber se ele* ***utilizou*** *arma , teve resultado negativo .*

O sujeito *ele* foi reconhecido como AGENT pelos 10 anotadores, enquanto o objeto direto *arma* foi anotado por 8 anotadores como INSTRUMENT.

Existem vários motivos que podem ter levado a uma concordância tão baixa. É possível, por exemplo, que o material fornecido não tenha sido detalhado o suficiente para a realização da tarefa, ou que os anotadores não tenham entendido claramente o que deveria ser feito. Porém, cremos que o principal fator envolvido é a complexidade da tarefa, que requer um treinamento muito bem desenvolvido para que se possa chegar a níveis maiores de concordância.

Como pode ser visto no trabalho de Hovy et al. (2006), a solução encontrada para se obter alto nível de concordância foi simplificar a tarefa o máximo possível. Para simplificar a tarefa, no entanto, seria necessário que já tivéssemos um recurso existente, do qual pudéssemos tirar insumos para a anotação. Porém, estamos tratando aqui justamente do desenvolvimento de um recurso que ainda não existe para o português, e não da expansão do mesmo.

Algo que poderia aumentar a concordância seria uma interface de anotação mais bem desenvolvida e mais amigável do que uma folha de papel e um manual de anotação. No entanto, não cremos que tal material conseguiria aumentar o valor da concordância (multi-π) de 0,25 para mais de 0,8, que seria um valor aceitável para o desenvolvimento de um recurso.

A baixa concordância averiguada neste experimento faz com que nossa tendência seja por manter a anotação com apenas um anotador, que teve um maior treinamento, com o estudo de outros recursos, como a VerbNet, o PropBank e o PropBank.Br, e com anotações-teste antes de iniciar o trabalho.

6 Conclusões

Realizamos a anotação de uma quantidade amostral de verbos em dois *corpora* com base na metodologia proposta. Desse modo, temos um recurso lexical com informação sobre papéis semânticos disponível em um formato amostral. Após a anotação, foi possível observar semelhanças e diferenças nos papéis semânticos atribuídos para verbos em textos especializados e não especializados.

Quanto à metodologia de anotação, a lista de papéis escolhida foi suficiente para realizar a anotação dos argumentos dos verbos. Porém, a anotação de elementos que podem ser considerados adjuntos se mostrou complexa. Assim, cremos ser necessária uma modificação para incluir papéis específicos de adjuntos, nos mesmos moldes do PropBank.

A opção por uma anotação de algumas sentenças por verbo, e não da totalidade de sentenças, é válida, pois a maioria dos verbos não apresentou uma grande polissemia. Assim, cremos ser prudente manter a metodologia empregada e, se necessário, dar um tratamento especial aos verbos mais polissêmicos. Por exemplo, para os casos de verbos como *dar*, *fazer* e *ir*, realmente seria necessário um olhar mais cuidadoso, pois a quantidade de significados desses verbos é muito grande, e a anotação amostral não dá conta de suas várias facetas.

A ferramenta utilizada para a extração e anotação dos dados apresentou versatilidade e pretendemos continuar com o seu uso. Além disso,

com a possibilidade de exportar os dados para o formato XML, que é mais amigável, a disponibilização dos dados e seu compartilhamento se tornam mais simples.

No que diz respeito à comparação entre linguagem comum e linguagem especializada, os dados estatísticos mostram que, conforme aumenta a complexidade dos dados (anotação de papéis semânticos -> de argumentos -> de sentenças), aumenta também a distância entre as duas amostras.

Essa diferença entre os dados dos *corpora* pode ser visualizada qualitativamente no que diz respeito a alguns elementos específicos, tais como a utilização de INSTRUMENTS na posição de sujeitos, com um apagamento do agente, que é uma marca dos textos da Cardiologia. Além disso, papéis como PIVOT e GOAL também foram mais frequentes na Cardiologia do que no Diário Gaúcho.

A avaliação da concordância entre vários anotadores permitiu que observássemos o quão complexa é a tarefa de atribuição papéis semânticos e de distinção entre argumentos e adjuntos. Os índices multi-π encontrados entre anotadores linguistas ficaram abaixo dos limites apontados na literatura como mínimos para a existência de concordância. A literatura apresenta estudos com múltiplos anotadores com resultados superiores (Hovy et al., 2006; Fossati, Giuliano e Tonelli, 2013), porém, esses estudos se baseiam em recursos já existentes, o que permite simplificar as decisões do anotador. Em nosso caso, seria possível simplificar a anotação (por exemplo, reduzir a lista de papéis semânticos), mas isso modificaria muito o recurso final a ser gerado. Assim, acreditamos que o trabalho de anotação deve ser realizado inicialmente por apenas um anotador treinado. Posteriormente, de posse de um recurso já desenvolvido, será possível simplificar a anotação tomando por base as informações do recurso existente para realizar um novo teste com múltiplos anotadores.

Agradecimentos

Agradecemos ao CNPq e à CAPES, pelo financiamento e também ao Projeto CAMELEON (CAPES-Cofecub 707/11) pelo apoio e oportunidade de intercâmbio no exterior.

Referências

Afonso, Susana, Eckhard Bick, Renato Haber e Diana Santos. 2001. Floresta sintá(c)tica: um treebank para o português. In: *Actas do XVII Encontro da Associação Portuguesa de Linguística*, APL, Lisboa, Outubro de 2001.

Artstein, Ron e Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. In: *Computational Linguistics*, 34(4):555-596. ACL

Baker, Collin F., Charles J. Fillmore e John B. Lowe. 1998. The Berkeley FrameNet project. In: *COLING-ACL '98*: Proceedings of the Conference. Montreal, Canada 1998, pp. 86-90.

Bick, Eckhardt. 2000. *The Parsing System* PALAVRAS*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press. http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf

Branco, António *et al*. 2012. *The Portuguese Language in the Digital Era* / *A língua portuguesa na era digital*. Heidelberg, Nova Iorque: Springer.

Brumm, T. 2008. *Erstellung eines Systems thematischer Rollen mit Hilfe einer multiplen Fallstudie*. Trabalho de conclusão de curso. http://www.ipd.kit.edu/Tichy/uploads/arbeiten/135/StudienarbeitBrumm.pdf

Burchardt, Ajoscha, Katrin Erk, Anette Frank, Andrea Kowalski e Sebastian Pado. 2006. SALTO - A Versatile Multi-Level Annotation Tool. In: *Proceedings of LREC 2006*.

Cançado, Márcia. 2009. Argumentos: Complementos e Adjuntos. In: *Revista Alfa*, São Paulo, 53 (1): 35-59.

Cançado, Márcia. 2010. Verbal Alternations in Brazilian Portuguese: a Lexical Semantic Approach. In: *Studies in Hispanic and Lusophone Linguistics*, 3 (1), p. 77-111.

Cançado, Márcia, Luisa Godoy e Luana Amaral. 2012. The construction of a catalog of Brazilian Portuguese verbs. In: *Proceedings of the Workshop on Recent Developments and Applications of Lexical-Semantic Resources (LexSem 2012), in conjunction with KONVENS 2012*. Viena, Itália, pp. 438-445.

Cohen, J. 1960. A coefficient of agreement for nominal scales. In: *Educational and Psychological Measurement*, 20, p. 37-46.

Cohen, J. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. In: *Psychological Bulletin*, 70, p. 213-220.

Davies, Mark e Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. In: *Biometrics*, 38(4), p. 1047–1051.

Dias-Da-Silva, Bento C. 2005. A construção da base da wordnet.br: conquistas e desafios. In: *Proceedings of the Third Workshop in Information and Human Language Technology (TIL*

2005), *in conjunction with XXV Congresso da Sociedade Brasileira de Computação.* São Leopoldo, RS, Brasil, pp. 2238–2247.

Dias-Da-Silva, Bento C., Ariani Di Felippo e Maria das Graças Volpe Nunes. 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 (LREC 2008),* Marrakech, Morocco, pp. 1535-1541.

Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. In: *Language*, Vol. 67, No. 3. (Sep., 1991), pp. 547-619.

Duran, Magali Sanches e Sandra Maria Aluísio. 2011. Propbank-Br: a Brazilian Portuguese *corpus* annotated with semantic role labels. In: *Proceedings of the 8th Symposium in Information and Human Language Technology*, October 24-26, Cuiabá/MT, Brazil.

Duran, Magali Sanches e Sandra Maria Aluísio. 2012. Propbank-Br: a Brazilian treebank annotated with semantic role labels. In: *Proceedings of the LREC 2012*, May 21-27, Istanbul, Turquia.

Fellbaum, C. (1998) *WordNet*: An electronic lexical database. MIT Press. Cambridge, Massachusetts.

Fillmore, Charles J. 1967. The case for case. In: Bach, Emmon e Robert Harms (Eds.). *Proceedings of the Texas Symposium on Language Universals*, April 13-15, 1967.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. In: *Psychological Bulletin*, v. 76, n. 5, p. 378–382.

Fossati, Marco, Claudio Giuliano e Sara Tonelli. 2013. Outsourcing FrameNet to the Crowd. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, p. 742–747, Sofia, Bulgaria.

Franchi, Carlos e Márcia Cançado. 2003. Teoria generalizada dos papéis temáticos. *Revista Estudos da Linguagem*, v. 11, n. 2.

Gelhausen, T. 2010. *Modellextraktion aus natürlichen Sprachen*: Eine Methode zur systematischen Erstellung von Domänenmodellen. Karlsruhe: KIT Scientific Publishing. Tese de doutorado, Karlsruher Institut für Technologie.

Gildea, Daniel e Martin Jurafsky. 2002. Automatic Semantic Role Labeling. In: *Computer Linguistics*, 28(3), p. 245-288. Cambridge: MIT Press.

Gruber, J.S. 1965. *Studies in Lexical Relations*. MIT. Tese de doutorado. Orientador: Edward S. Klima.

Hong, Jisup e Collin F Baker. 2011. How good is the crowd at "real" wsd? In: *Proceedings of the Fifth Law Workshop (LAW V)*, p. 30-37, Portland, Oregon.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw e Ralph Weischedel. (2006) OntoNotes: The 90% solution. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, New York, p. 57–60.

Ienco, Dino, Serena Villata e Cristina Bosco. 2008. Automatic extraction of subcategorization frames for Italian. In: *Proceedings of the LREC 2008*.
http://www.di.unito.it/~ienco/ienco_LREC08.pdf

Jackendoff, R.S. (1990) *Semantic Structures*. Current Studies in Linguistic Series, v. 18. Cambridge: MIT Press.

Kasper, Simon. 2008. *A comparison of 'thematic role' theories*. Philipps-Universität Marburg. Dissertação de mestrado.

Kipper-Schuler, K. 2005. *VerbNet*: a broad-coverage, comprehensive verb lexicon. University of Pennsylvania. Tese de doutorado orientada por Martha S. Palmer.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Levin, Beth e Malka Rappaport-Hovav. 2005. *Argument Realization*. Cambridge, Nova Iorque, Melbourne, Madri, Cape Town, Singapure, São Paulo: Cambridge University Press.

Lima, Bruno de A. F. de. 2007. Valência dos verbos de vitória e derrota em português. Dissertação de Mestrado. Belo Horizonte: UFMG.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In: *Proceedings of the 17th International Conference on Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, EUA, pp. 768-774.

Loper, Edward, Szu-ting Yi e Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In: *Proceedings of the 7th International Workshop on Computational Linguistics*.

Manning, Christopher D. 1993. Automatic aquisition of a large subcategorization dictionary from corpora. In: *ACL '93 Proceedings of the 31st annual meeting on Association for Computational Linguistics*, p. 235-242.

Maziero, Erick G., Tiago A. S. Pardo, Ariani Di Felippo e Bento C. Dias-Da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletronico para o Portugues do Brasil. In: *VI TIL*, pp. 390–392.

Messiant, Cédric. (2008) A subcategorization acquisition system for French verbs. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Techonologies*, Columbus, Ohio, 55-60.

Messiant, Cédric, Anna Korhonen e Thierry Poibeau. 2008. LexSchem: A Large Subcategorization Lexicon for French Verbs. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Marrocos. http://www.lrec-conf.org/proceedings/lrec2008/pdf/142_paper.pdf

Muniz, M. C. M. 2003. *Léxicos Computacionais: Desafios na Construção de um Léxico de Português Brasileiro*. Monografia de Qualificação. Instituto de Ciências Matemáticas de São Carlos, USP. 50p.

Muniz, M. C. M. 2004. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB*. Dissertação de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. 72p.

Palmer, Martha, Daniel Gildea e Paul Kingsbury. 2005. "The Proposition Bank: A *Corpus* Annotated with Semantic Roles", *Computational Linguistics Journal*, 31:1.

Perini, Mário Alberto. 2008. *Estudos de Gramática Descritiva*: as valências verbais. São Paulo: Parábola Editorial.

Preiss, Judita, Ted Briscoe e Anna Korhonen. 2007. A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from *Corpora*. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Praga, República Tcheca, 2007. Disponível em: http://www.cl.cam.ac.uk/~alk23/acl-07.pdf.

Ramisch, Carlos, Aline Villavicencio e Christian Boitet. 2010a. mwetoolkit: a Framework for Multiword Expression Identification. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, Maio de 2010.

Ramisch, Carlos, Aline Villavicencio e Christian Boitet. 2010b. Web-based and combined language models: a case study on noun compound identification. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, Agosto de 2010.

Salomão, Margarida. 2009. FrameNet Brasil: um trabalho em progresso. *Calidoscópio* 7(3), 171-182.

Scarton, Carolina. 2013. *VerbNet.Br*: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. NILC/USP. Dissertação de mestrado orientada por Sandra Maria Aluísio.

Schulte im Walde, Sabine. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In: *Proceedings of the 3rd Conference on Language Resources and Evaluation*, v. IV, Las Palmas de Gran Canaria, Espanha, p. 1351–1357. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.8846&rep=rep1&type=pdf

Scott, M. 2004. *Wordsmith Tools version 4*. Oxford: Oxford University Press.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. In: *Public Opinion Quarterly*, 19(3), p. 321–325.

Zanette, Adriano. 2010. *Aquisição de* Subcategorization Frames *para Verbos da Língua Portuguesa*. Projeto de Diplomação. UFRGS. Orientadora: Aline Villavicencio.

Zanette, Adriano, Carolina Scarton e Leonardo Zilio. 2012. Automatic extraction of subcategorization frames from *corpora*: an approach to Portuguese. In: *Proceedings of PROPOR 2012* - Demonstration Session. Coimbra, Portugal.

Zapiran, B., E. Agirre e L. Màrquez. 2008. Robustness and Generalization of Role Sets: PropBank vs. VerbNet. In: *Proceedings of the ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, June, 2008.

Zilio, Leonardo. 2009. *Colocações especializadas e* Komposita*: um estudo contrastivo alemão-português na área de Cardiologia*. Dissertação de Mestrado. Orientadora: Maria José Bocorny Finatto. Disponível em: http://www.lume.ufrgs.br/bitstream/handle/10183/16877/000706196.pdf?sequence=1

Zilio, Leonardo, Adriano Zanette e Carolina Scarton. 2012. Extração automática de estruturas de subcategorização a partir de *corpora* em português, in: *Anais do ELC 2012*, XI Encontro de Linguística de *Corpus*, São Carlos - SP.

Testuen sinplifikazio automatikoa: arloaren egungo egoera

Automatic Text Simplification: State of Art

Itziar Gonzalez-Dios
Ixa Taldea. Euskal Herriko Unibertsitatea (UPV/EHU)
itziar.gonzalezd@ehu.es

María Jesús Aranzabe
Ixa Taldea. Euskal Herriko Unibertsitatea (UPV/EHU)
maxux.aranzabe@ehu.es

Arantza Díaz de Ilarraza
Ixa Taldea. Euskal Herriko Unibertsitatea (UPV/EHU)
a.diazdeilarraza@ehu.es

Laburpena

Lan honen helburua testuen sinplifikazio automatikoaren arloaren egungo egoera aurkeztea da. Horretarako, ikerketa-lerro honetan egindako sistemak eta prototipoak jaso ditugu hizkuntzaren, helburu duen taldearen eta egiten duen sinplifikazio motaren (sintaktikoa, lexikala edo biak) arabera sailkatuz. Sistemak ebaluatzeko metodoak eta garatu diren baliabideak eta tresnak ere aurkeztuko ditugu.

Gako-hitzak

Testuen sinplifikazio automatikoa, esaldien sinplifikazioa, arloaren egungo egoera

Abstract

The aim of this paper is to give an overview of the state-of-art in automatic text simplification. To that end, we present the systems and prototypes according to the language the are built for, their target audience and the type of simplification (syntactic, lexical or both) they perform. Moreover, we expound the different evaluation methods that have been carried out with these systems and the resources and tools developed so far.

Keywords

Text simplification, sentence simplification, state-of-art survey

1 Sarrera

Testuen sinplifikazioak testu bat testu sinpleagoa lortzea du helburu jatorrizko testuaren esanahia mantenduz; egitura eta hitz konplexuak ordezkatuz sortzen den testua errazagoa izan behar da irakurle jakin batzuentzat. Ezaguna da testu errazek edo sinpleek abantaila ugari eskaintzen dizkietela, bai pertsonei, bai Hizkuntzaren Prozesamenduko tresnei (Chandrasekar, Doran, & Srinivas, 1996).

Testuen sinplifikazioa irakaskuntzan eta batez ere atzerriko hizkuntzen didaktikan landu izan da. Arlo horretan testu sinplifikatuak erabiltzearen helburua ulermena areagotzea eta karga kognitiboa leuntzea da (Crossley, Allen, & McNamara, 2012).

Irakaskuntzan, hain zuzen ere, Allen-ek (2009) eta Crossley-k, Allen-ek, & McNamara-k (2012) sinplifikatzeko bi aukera daudela azaltzen dute:

- Egituraren araberako sinplifikatzea: maila jakin bakoitzari dagozkion hitz zerrendak eta egitura sintaktikoen zerrendak erabilita gauzatzen da, eta kasu batzuetan, testuen konplexutasun maila neurtzen duten (*Readability*) formuletan oinarritzen da.
- Intuitiboki sinplifikatzea: maila jakin bakoitzari dagozkion hitz zerrendak eta egitura sintaktikoen zerrendak kontuan izan ditzaketen arren, oro har, intuizioari jarraiki sinplifikatzen da.

Young-ek (1999) hainbat metodo aipatzen ditu sinplifikatzeko atzerriko hizkuntzaren irakaskuntzan:

- Linguistikoki sinplifikatzea: testua berridaztea esaldiak laburragoak egiteko, esaera idiomatikoak ezabatzea edo parafraseatzea, hitz espezializatuak eta maiztasun gutxikoak ekiditea, eta sintaxi konplexua errebisatzea perpaus bakunak sortzeko.

- Materia sinplifikatzea: testua laburtu, paragrafoak edo atalak kenduz.
- Glosen bitartez sinplifikatzea: itzulpenak edo definizioak gehitzea.
- Prozesamendu kognitiboetan oinarritutako aldaketak eta elaborazioak eginez sinplifikatzea.

Simensen-ek (1987) atzerriko hizkuntzetako liburuak egokitzeko argitaletxeek dituzten gidalerroak aztertu zituen, eta gidalerro horietan oinarrituta, egokitzapenerako hiru printzipio (kontrolaren printzipioak) aurkeztu zituen: informazioaren, hizkuntzaren eta diskurtsoaren kontrola. Hizkuntzaren kontrolaren atalean, Crossley-k, Allen-ek, & McNamara-k (2012) ere azaldutako bi hurbilpenak aurkezten ditu: zerrendetan (egituraren araberako sinplifikazioa) eta intuizioan (intuitiboki sinplifikatzea) oinarritutakoak.

Atzerriko hizkuntzen didaktikaren arloan, testu sinplifikatuak edo originalak erabiltzeak dituen abantailak ere aztertu izan dira, emaitzak kontrajarriak diren arren. Esaterako, Youngen (1999) azterketaren arabera, testu sinplifikatuak ez dira baliagarriak ikasleak irakurketa globala egiten ari badira; are gehiago, kalterako izan daitezke. Oh-ren (2001) arabera, ordea, irakurmenaren ulermen globala erraztu egiten dute testu sinplifikatuek. Crossley-k, Allen-ek, & McNamara-k (2012), berriz, uste dute testu sinplifikatuek erreferentziakidetasun bidezko kohesio altuagoa eta lokailu eta hitz ezagun gehiago eskaintzen dizkietela ingelesa ikasten ari diren ikasleei. Dena dela, arlo honetako autore gehienek gaian sakondu behar dela adierazten dute (Young, 1999; Allen, 2009; Crossley, Allen, & McNamara, 2012).

Hizkuntzaren Prozesamenduari (HP) dagokionez, ikerketa-lerro hau azken urteetan garrantzitsua bihurtu da, ingeleserako ez ezik beste hainbat hizkuntzatarako ere proposatu direlako sistemak eta metodo eta teknika berri ugari argitaratzen ari direlako. Testuak eskuz sinplifikatzeak edo egokitzeak lan handia eta garestia eskatzen du; beraz, HPko tresnak erabiliz testuak sinplifikatzean lana erraztu eta azkartzen da. HPan hartu den sinplifikatzeko ildoa linguistikoki edo estrukturalki sinplifikatzearena izan da, zehazki bi sinplifikazio mota landu dira: sintaktikoa eta lexikoa.

Testuen Sinplifikazio (TS) automatikoa Siddharthan-ek (2002) berridazketa prozesu bat bezala definitzen du; ataza horren helburua konplexutasun lexikala eta sintaktikoa gutxitzea da. Aluísio-k & Gasperin-ek (2010), berriz, HPko ikerketa-lerro bezala definitzen dute PorSimples[1] proiektuan. Bertan Brasilgo portugesezko testuen ulermena areagotzeko, fenomeno lexikalak eta sintaktikoak sinplifikatzen dituzte; lehendabizi, pertsona gutxik ulertzen dituzten hitzak ezagunagoekin ordezkatzen dituzte eta bigarrenik, esaldiak banatu eta esaldien egitura aldatzen dute.

Era berean, TSak HPko beste ataza eta ikerketa-lerro batzuekin elkarrekintza zuzena du, bai atazaren antzekotasunagatik, bai antzeko metodoak erabiltzen dituztelako. Horien artean hurbilen laburpen automatikoa (*summarisation*) dago; bien arteko desberdintasun nagusia da laburpen automatikoak testua trinkotzea duela helburu eta TSak testuen konplexutasun linguistikoa gutxitzea, alegia, ez du testua kondentsatzea helburu. Bi ikerketa-lerro horien arteko muga ezartzea, gainera, kasu batzuetan zaila izaten da erabiltzen diren teknika eta metodo batzuk berdinak direlako. TSaren antzeko beste atazak dira batetik, parafrasiak ikastea eta sortzea (*paraphrase acquisition and generation*), testuak sinplifikatzeko maiz parafrasiak erabili eta berridazketak egiten direlako, eta bestetik, konplexutasuna ebaluatzea (*readability assesement*), TSren aurreprozesu bezala erabili ohi dena.

Lan honetan, HPan egin den testuen sinplifikazioan kontzentratuko gara ikerketa-lerro honen artearen egoera ezagutzera emateko. Sarrera honen ondoren, beraz, 2. atalean HPan zein hizkuntzatarako eta zein helburu taldetarako egin diren lanak aurkeztuko ditugu sortu diren baliabideekin batera. Sinplifikazio motak eta metodoak 3. atalean erakutsiko ditugu. Sistemak ebaluatzeko erabili diren metodoak 4. atalean aurkeztuko ditugu egin diren sistema eta prototipoen bitartez. Amaitzeko, ondorioak aurkeztuko ditugu 5. atalean.

2 Sinplifikazio automatikoa HPan

HPko testuen sinplifikazioan hasierako lanak ingeleserako egin ziren. Lehen lana Chandrasekaren, Doran-en, & Srinivas-ena (1996) izan zen eta TSrako motibazioak azaldu ziren. Hasierako beste lanen artean, PSET (*Practical Simplification of English Text*) proiektuan[2] (Carroll et al., 1998) egindakoa aurki daiteke. Proiektu horretan hizkuntzarekin arazoak zituztenei eta, batez ere, afasia zuten pertsonei zuzendutako sinplifikazioa egin zuten (Carroll et al., 1999). Siddharthan-ek

[1] `http://caravelas.icmc.usp.br/wiki/index.php/Principal` (2011ko irailean atzituta)

[2] `http://www.informatics.sussex.ac.uk/research/groups/nlp/projects/pset.php` (2013ko maiatzean atzituta)

(2002), berriz, testuen sinplifikazio automatikorako oinarrizko arkitektura finkatu zuen.

Ikerketa-lerro hau oso garrantzitsua bilakatu da 2009tik aurrera eta beste hizkuntzetara zabaltzeaz gain, metodo eta teknika berri ugari argitaratu dira, batez ere metodo estatistikoetan eta ikasketa automatikoan oinarrituz.

Esan bezala, TSrako sistema gehienak ingeleserako proposatu eta egin dira; horien artean ditugu Siddharthan-ena (2006) eta Zhu-ren, Bernhard-en, & Gurevych-ena (2010). Azken urteotan beste hizkuntzetarako ere egin dira: japoniera (Inui et al., 2003), Brasilgo portugesa (Candido et al., 2009; Gasperin et al., 2009; Aluísio & Gasperin, 2010), suediera (Rybing, Smith, & Silvervarg, 2010; Keskisärkkä, 2012), arabiera (Al-Subaihin & Al-Khalifa, 2011), gaztelania (Saggion et al., 2011; Saggion, Bott, & Rello, 2013), frantsesa (Seretan, 2012; Brouwers et al., 2012), euskara (Aranzabe, Díaz de Ilarraza, & Gonzalez-Dios, 2012), italiera (Barlacchi & Tonelli, 2013), daniera (Klerke & Søgaard, 2013), bulgariera (Lozanova et al., 2013) eta koreera (Chung et al., 2013).

Jarraian testuen sinplifikazioa zein helburu talderentzat den baliagarria (2.1 azpiatala) eta horientzat egin diren lanak aipatuko ditugu. Ondoren, TSrako sortu diren baliabideak zerrendatuko ditugu atal honen amaieran (2.2 azpiatala).

2.1 Helburu taldeak

Chandrasekar-en, Doran-en, & Srinivas-en (1996) lanean, TSa gizakientzat eta HPko tresnentzat erabilgarria eta onuragarria dela esaten da. Urteak pasa ahala, lan horretan proposatutako ideiak materializatu dira. Jarraian ikusiko ditugu gizakiei eta tresnei zuzenduta egin diren TS lanak:

- Gizakiak. Testu sinpleek informazioa eskuragarriago bihurtzen dute eta horrela testuak ulertzea errazagoa da. Ondoren, gizakientzat egin diren lanak azpitaldeen arabera zerrendatuko ditugu:
  - Urritasunak dituztenentzat (Carroll et al., 1999): afasikoak (Carroll et al., 1998; Max, 2005; Devlin & Unthank, 2006), jaiotzetiko entzumen arazoak dituztenentzat (Inui et al., 2003; Lozanova et al., 2013; Chung et al., 2013), irakurtzeko arazoak dituztenentzat (Bautista, Hervás, & Gervás, 2012), dislexikoak (Rello et al., 2013), adimen-urritasunak dituzten irakurle pobreentzat (Fajardo et al., 2013)
  - Atzerriko hizkuntzen ikasleentzat (Petersen, 2007; Burstein, 2009)
  - Gutxi alfabetatuentzat (Candido et al., 2009)
  - Haurretzat (De Belder & Moens, 2010; Brouwers et al., 2012; Barlacchi & Tonelli, 2013)
  - Orokorrean eta adimen urritasunak dituztenentzat teknologia munduan murgiltzeko (Saggion et al., 2011; Bott & Saggion, 2012)
- Oinarrizko tresnak edo HPko aplikazio aurreratuak. Esaldi luzeak eta konplexuak dituzten testuak zailagoak izaten dira automatikoki prozesatzeko; esaldi laburragoak eta sinpleak erabiltzen diren kasuetan, aldiz, tresnen eta aplikazio aurreratuen errendimendua hobea da. Hori dela eta, testuen sinplifikazioa aurreprozesu bezala erabil daiteke performantzia igotzeko. Hobekuntza hori bai oinarrizko tresnetan, bai aplikazio aurreratuetan gertatzen da. TSa edo esaldien sinplifikazioa erabili duten lanak dira:
  - Dependentzia-gramatikan oinarritutako analizatzaile sintaktiko edo *parser*ak (Chandrasekar, Doran, & Srinivas, 1996)
  - Laburpen automatikoak egiteko (Lal & Rüger, 2002; Siddharthan, Nenkova, & McKeown, 2004; Blake et al., 2007; Vanderwende et al., 2007; Silveira Botelho & Branco, 2012)
  - Informazioa bilatzeko aplikazioak (Beigman Klebanov, Knight, & Marcu, 2004)
  - Azpitituluak egiteko (Daelemans, Höthker, & Sang, 2004)
  - Itzulpen automatikoa (Doi & Sumita, 2004; Poornima et al., 2011)
  - Rol semantikoak etiketatzeko (Vickrey & Koller, 2008)
  - Arlo berezituetako analizatzailea, adibidez biomedikuntzako testuetan (Jonnalagadda et al., 2009)
  - Informazio erauzketa (Jonnalagadda & Gonzalez, 2010b; Evans, 2011)
  - Gertaeren erauzketa (Buyko et al., 2011)
  - Ahozko hizkuntza ulertzen duten sistemak (Tur et al., 2011)
  - Galdera-erantzun sistemak (Bernhard et al., 2012)

– Erlazioen erauzketa (Minard, Ligozat, & Grau, 2012)
– Corpus paraleloetan hitzak lerratzeko (Srivastava & Sanyal, 2012)

2.2 Baliabideak

TSren ikerketa-lerroan HPko beste atazetan bezala corpusak oso baliabide garrantzitsuak dira. Bi motakoak dira bereziki erabilienak: corpus paraleloak eta corpus ez-paraleloak.

Corpus paraleloetan jatorrizko testua eta testu sinplea lerratuta daude, esaldiz esaldi. Hau da, jatorrizko testuko esaldi bakoitzak bere baliokide sinplea du. Mota horretako corpusak Brasilgo portugeserako (Caseli et al., 2009), gaztelaniarako (Bott & Saggion, 2011) eta danierarako (Klerke & Søgaard, 2012) sortu dira.

Corpus ez-paraleloek, berriz, testu sinpleak eta jatorrizkoak gordetzen dituzte lerratu gabe. Modu horretan, bai Dell'Orletta-k, Montemagnik, & Venturi-k (2011) italierarako, bai Hanckek, Vajjala-k, & Meurers-ek (2012) alemanerako alde batetik, haurrei zuzendutako egunkarietako eta aldizkarietako testuak testu sinple bezala jaso dituzte eta bestetik, prentsa arrunteko testuak. Alemanerako ere, corpus ez-paraleloa eraiki dute Klaper-ek, Ebling-ek, & Volk-ek (2013) webeko testuak erabiliz.

Wikipedia entziklopedia ere corpus moduan erabili da. Wikipediaren ingelesezko jatorrizko bertsioaz[3] gain, *Simple English*[4] edo ingeles errazean idatzitako bertsioa dago eskuragarri (Yatskar et al., 2010; Woodsend & Lapata, 2011b; Coster & Kauchak, 2011a; Shardlow, 2013b). Frantseserako, Brouwers et al.- ek (2012) frantsesezko wikipedia erabiltzen dute jatorrizko testuak lortzeko eta Vikidia[5], 8-13 urte tartekoei zuzendutako entziklopedia, testu sinpleak lortzeko.

Medikuntza arloan eta zehatzago esanda erradiologiako txostenetan oinarrituz, Kvistab-ek & Velupillaia-k (2013) suedierako corpusa osatu eta aztertu dute.

Corpusak izan gabe, sinplifikazioa egiten duten tresnak eta sistemak garatzeko eta ebaluatzeko datu-multzoak garatu dira. Zhu-k, Bernhardek, & Gurevych-ek (2010) *Wikipedia*tik eta *Simple Wikipedia*tik hartutako 65.133 artikuluz osatu duten datu-multzoa eraiki dute. Artikuluak parekatzeko *language link* edo hizkuntzen arteko esteka jarraituz, Wikimediako *dump files* erabili dute.

Halaber, Wikipedia eta Simple Wikipedian oinarrituta, 137 k jatorrizko/sinplifikatutako esaldi lerratu dituzte Coster-ek & Kauchak-ek (2011b). 2007ko SemEval-eko lexikoaren ordezkapena (*Lexical substitution*) atazarako sortu zen datumultzoa oinarri hartuta anotazio prozesua eta anotatzaileen arteko adostasun emaitzak azaldu dituzte De Belder-ek & Moens-ek (2012).

Bestelako baliabideen artean Petersen-ek & Ostendorf-ek (2007) corpus bat osatu dute eta ikasketa automatikoaren bitartez eta banatu (*splitting*) behar diren esaldien azterketa egin dute, atzerriko hizkuntzen ikasketan laguntzeko. Štajner-ek & Saggion-ek (2013) ere sinplifikatu behar diren esaldien aukeraketa egiteko algoritmoa aurkezten dute. Algoritmo horrek esaldiak dauden bezala mantendu, banatu edo ezabatu behar diren adierazten du testu-genero eta helburu taldearen arabera.

3 Sinplifikazio motak eta metodoak

Esan bezala, HPan egiten diren testuen sinplifikazioak bi mota nagusikoak dira: sinplifikazio sintaktikoa eta sinplifikazio lexikala. Bi horiek TSko azpiatazatzat hartu izan ohi dira.

Sinplifikazio sintaktikoak testu baten konplexutasun gramatikala murriztea du helburu. Horretarako, egitura sintaktiko konplexuak sinpleagoez ordezkatzen dira (Siddharthan, 2006). Azken urte hauetan, lexiko mailako sinplifikazioak ere bere tokia hartu du eta bere helburua hitzen ulergarritasuna areagotzea da, hitz konplexuak edo maiztasun gutxikoak baliokide diren hitz ezagunagoekin, sinonimoekin edo sintagmekin ordezkatuz (Specia, Jauhar, & Mihalcea, 2012). Badaude sinplifikazio sintaktikoa eta lexikala uztartzen dituzten sistemak ere.

Testuak automatikoki sinplifikatzeko metodoei eta teknikei dagokienez, HPko beste atazetan bezala, hiru multzo nagusi bereizi behar dira: eskuzko erregeletan oinarritutakoak, estatistikan eta ikasketa automatikoan (datuetan) oinarritutakoak eta aurreko biak batzen dituzten sistema hibridoak. Azken urteotan estatistikan oinarritutako metodoek lekua irabazi diete eskuz idatzitako erregeletan oinarritutako sistemei, erregeletan oinarritutako sistemak eraikitzeak denbora eskatzen baitu.

Hiru multzo horietaz gain, badira TSa itzulpen prozesu bat bezala ulertzen dutenak, hau da, jatorrizko testuak testu sinplifikatu bihurtzea itzulpen bat balitz bezala kontsideratzen da, bai estatistikoki, bai eskuzko erregeletan oinarritu-

[3] `http://en.wikipedia.org/wiki/Main_Page` (2013ko irailean atzituta)
[4] `http://simple.wikipedia.org/wiki/Main_Page` (2013ko irailean atzituta)
[5] `http://fr.vikidia.org/wiki/Accueil` (2013ko irailean atzituta)

ta. Horrela, jatorrizko testuaren hizkuntza *A iturri hizkuntza*ren pareko izango litzateke eta, testu sinplearen hizkuntza *B helburu hizkuntza*rena.

Sistemen arkitekturei dagokienez, sistema gehienek modulu hauek dituzte: i) analizatzailea: bertan analisia egiten da, sinplifikazioaren aurreprozesua dena; izan ere, testua analizatu gabe ezin da testua sinplifikatu. Analisi hori gauzatzeko, analizatzaile sintaktikoak osagai-ereduan edo dependentzia-ereduan oinarritzen dira. ii) sinplifikatzailea edo transformatzailea: testuak sinplifikatzeaz arduratzen den modulua, bere gain metodo eta teknikak hartzen ditu eta batzuetan iii) testuen kohesioa bermatzen duen modulua.

Hurrengo azpiataletan, sinplifikazio mota bakoitza egiteko erabili diren tekniken eta metodoen laburpena egingo dugu. Azaldutako lan guztiak 1. eta 2. tauletan sailkaturik ikusiko ditugu, 1. taulan sinplifikazio sintaktikoa edo lexikala egiten dituztenak eta 2. taulan bi sinplifikazioak edo bestelako sinplifikazioak egiten dituztenak. Taula horietako lehenengo zutabean, hizkuntzak eta sistemak aurkeztuko dira lanei erreferentzia eginez, eta hurrengo zutabeetan egin duten sinplifikazio mota adieraziko da. Sistemak hizkuntzaren, kronologiaren eta alfabetoaren arabera zerrendatu dira.

3.1 Sinplifikazio sintaktikoa

Sinplifikazio sintaktikoa izan zen hasiera batean testuen sinplifikazio automatikoan garrantzi gehien eman zitzaion sinplifikazio mota edo azpiataza. Hasierako lan gehienak eskuzko erregeletan oinarritutakoak ziren, baina azken urteetan Wikipedia bezalako baliabideei esker, metodo estatistikoak ugaritu egin dira.

3.1.1 Eskuzko erregeletan oinarritutakoak

Sinplifikazio sintaktikoa aurkeztu zen lehendabiziko lanean (Chandrasekar, Doran, & Srinivas, 1996), bi metodo aurkeztu ziren: lehenengoa *chunk*etan oinarrituta eta bigarrena dependentzietan. Hurrengo lanean, erregela horiek ikasketa automatikoaren bitartez erauztea proposatu zuten (Chandrasekar & Srinivas, 1997).

Lehendabiziko lan horretan TSak bi helburu talde zituela ere azaldu zen: gizakiak eta tresnak. Banaketa horri jarraiki azalduko ditugu egin diren lanak.

Gizakiekin hasiz, afasikoei egunkariko testuak irakurterrazagoak egiteko sistema proposatu dute aipatutako PSET proiektuan. Sistema anaforaren tratamenduan oinarritzen da eta 3 modulu ditu: i) anaforaren ebazpena ii) sinplifikazioa eta iii) lehendabiziko moduluak detektatutako izenordainei dagozkien izen sintagmen ordezkapena (Canning & Tait, 1999).

Max-ek (2005; 2006) sinplifikazio motorra testu editore batean integratzen du gaixotasun kognitiboak dituzten pertsonei testuak egokitzen dizkieten autoreei laguntzeko.

Brasilen alfabetatze arazo larriak dituztenez, Brasilgo portugesean linguistikoki zailak diren fenomenoak aztertu ondoren, sinplifikazio proposamenak eman zituzten Aluísio et al.-ek (2008). Sinplifikazioa gauzatzeko erregelak eskuliburu batean deskribatu zituzten (Specia, Aluísio, & Pardo, 2008). Candido et al.-ek (2009) Siddharthan-en (2002) sintaxia sinplifikatzeko hiru mailako arkitekturan oinarrituz zazpi eragiketa azaltzen dituzte. Eragiketa horiek esaldiak banatzea, diskurtso markatzailea aldatzea, ahots pasiboa aktibo bihurtzea, perpausen hurrenkera aldatzea, hurrenkera kanonikoa errespetatzea, adizlagunak galdegai/ez-galdegai bihurtzea eta ez sinplifikatzea dira. Sistema horrek sinplifikatu aurretik testuak konplexuak diren ala ez ebaluatzen du (Gasperin et al., 2009).

Suedierarako garatu duten CogFLUX sistemak (Rybing, Smith, & Silvervarg, 2010) irakurketa errazeko eta testu normaletako corpus azterketa batean oinarritutako 25 transformazio erregela erabiltzen ditu. Erregela horiek bi motatakoak dira: esaldietan sintagmak ezabatu edo ordezkatzen dituztenak eta informazio sintaktikoa gehitzen dutenak.

Haurrek ipuinetako gertaerak hobeto uler ditzaten, ERNESTA *(Enhanced Readability through a Novel Event-based Simplification Tool)* (Barlacchi & Tonelli, 2013) izeneko sistema garatu dute italierarako. Anafora ebatzi ondoren, eta gertaerak kontuan hartuta, esaldiak sinplifikatzen ditu informazio psikolinguistikoan oinarrituz.

Entzumen arazoak dituztenak helburu izanik, bulgarierarako Lozanova et al.-ek (2013) multzo ezberdinetan (anaforaren ebazpena, perpaus mugak, subjektuen berreskurapena. etab.) banatzen diren 23 erregeletako sistema aurkezten dute. Aplikatzen dituzten eragiketak dira esaldiak banatzea, esaldi konplexuen sinplifikazioa, anaforaren ebazpena, subjektuen berreskurapena, perpausen hurrenkera zehaztea eta sintagma osagarrien txertatzea.

Gorrei koreerazko albisteen ulermena errazteko, Chung et al.-ek (2013) esaldi laburrago eta sinpleagoak sortzeaz gain errepresentazio grafikoak erabiltzen dituzte. Sinplifikatzeko esaldiak banatzen dituzte eta argumentuak tokiz aldatzen dituzte dagokien aditzetik gertuago egon daite-

zen.

HPko tresnak helburu dituzten sistemei dagokienez, tresnek informazioa eraginkortasun handiz prozesa dezaten, Beigman Klebanov-ek, Knight-ek, & Marcu-k (2004) *Easy Access Sentences* kontzeptua proposatu zuten. Sortzen diren esaldi horiek jatorrizko testuaren informazioa mantendu behar dute, eta aditz jokatu bakarraz eta entitate batez osatuak izan behar dira. Gainera, esaldi horiek gramatikalak izan behar dira. Horrelako esaldiak erabilita tresnek errazago aurkitu eta prozesatuko dute informazioa.

Azpitituluak egitean esaldiak sinplifikatzeko Daelemans-ek, Höthker-ek, & Sang-ek (2004) bi hurbilpen aurkezten dituzte, bata aurrerago azalduko dugun ikasketa automatikoan oinarritutakoa eta bestea erregeletan oinarritutakoa. Azken honetan nederlandera eta ingelesa sinplifikatzeko erregelak konpilatu dituzte. Bi faseetan egiten dute sintagmen ezabatzea: lehendabizi, erredundanteak diren esaldiak aukeratzen dituzte eta bigarrenik, trinkotasun maila bateko esaldiak ezabatzen dituzte. Ezabatzeko hautagaiak adberbioak, adjektiboak, izen propioak, sintagma preposizionalak, egitura parentetikoak, erlatibozko perpausak, zenbakiak eta denbora-adierazpenak dira.

Biomedikuntzako artikuluen laburpenetako analizatzaile sintaktikoaren emaitzak hobetzeko eta testu horietako informazioa erauzteko algoritmoa garatu dute Jonnalagadda et al.-ek (2009). Algoritmo horrek *Link Grammar* analizatzaile sintaktikoaren bitartez hitz pareen arteko erlazio gramatikal jakinak eta puntuazioa erabiltzen ditu; horrela, esaldiak perpausetan banatzen dituzte. BioSimplify sisteman (Jonnalagadda & Gonzalez, 2010a) izen-sintagmen ordezkatzeak ere gehitzen dituzte.

Biomedikuntzarekin jarraituz, iSimp sistemak (Peng et al., 2012) alor horretako testu zientifikoen laburpenak sinplifikatzen ditu testu meatzaritza egiteko helburuarekin. Patroiak erabiliz koordinazioa, erlatibozko perpausak eta aposizioak tratatzen ditu.

Itzulpen automatikoari laguntzeko Poornima et al.-ek (2011) esaldiak sinplifikatzeko bi metodo proposatzen dituzte: perpaus koordinatuak eta mendeko perpausak banatzea eta erlatibozko izenordaina dutenak sinplifikatzea.

Ahozko hizkuntza ulertzen duten sistemen errendimendua hobetzeko Tur et al.-ek (2011) esaldiak sinplifikatzen dituzte. Beraien helburua sailkatzaile bat denez, eta ez gizakiak, sortzen dituzten esaldiek ez dute zertan gramatikalak izan.

Laburpen automatikoak egiteko, Bawakid-ek & Oussalah-ek (2011) Tregex patroiak erabiliz esaldiak banatzeko erregelak aplikatzen dituzte. Algoritmo baten bitartez esaldiak trinkotzen (*sentence compression*) dituzte laburpena egin aurretik.

Laburpen automatikoak egiten dituzten sistemekin jarraituz, sistema batean integratuta dagoen bi mailako (analisia eta transformazioa) sinplifikatzailea aurkezten dute Silveira Botelho-k & Branco-k (2012) portugeserako. Erlatibozko perpausak, aposizioak eta egitura parentetikoak lantzen dituzte horiek testutik ezabatuz.

Corpus paraleloen hitzak lerratzeko, gakohitzetan oinarritutako zerrendak erabiliz banatzen dituzte esaldiak Srivastava-k & Sanyal-ek (2012). Zerrenda horiek ingeleserako eta Hindirako osatu dituzte.

Helburu taldea irekita uzten duten lanak ere badira. Esaterako, sintaxia sinplifikatzearekin batera kohesioari garrantzia emanez hiru mailako exekuzio-hodia duen arkitektura proposatzen du sintaxia sinplifikatzeko Siddharthan-ek (2006). Sintaxiko hiru maila horiek analisia, transformazioa eta birsorkuntza dira. Azkeneko maila horretan bermatzen da testu berriaren kohesioa.

Sistema irekita uzten dute ere Aranzabek, Díaz de Ilarrazak, & Gonzalez-Diosek (2012) corpus azterketan oinarrituz proposatzen dituzten erregeletarako. Dependentzia-zuhaitzetan oinarrituz esaldi konplexuetatik esaldi sinpleen zuhaitzen transformazioak egiten dituzte (Aranzabe, Díaz de Ilarraza, & Gonzalez-Dios, 2013).

3.1.2 Datuetan oinarritutakoak

Chandrasekar-en, Doran-en, & Srinivas-en (1996) lanari jarraituz, erregelak ikasketa automatikoaren bitartez ikastea proposatzen dute Chandrasekar & Srinivas-ek (1997). Une bakoitzean esaldi bana prozesatzen duen bi mailako arkitektura aurkezten dute: analisia eta transformazioak. Analisiak osagaien eta dependentzien informazioa erabiltzen du, eta transformazioak egiteko, erregelak automatikoki erauztea proposatzen dute domeinuetara errazago egokitzeko.

Azpitituluak egiteko Daelemans-en, Höthker-en, & Sang-en (2004) bigarren hurbilpena (3.1.1 azpiatalean aurkeztu dugu lehenengoa) ikasketa automatikoan oinarritzen da. Esaldiak sinplifikatzeko prozesua hitzen transformazio ataza bezala ulertzen dute; prozesu horretan testuko hitzak kopiatu, ezabatuak edo ordezkatu egiten dira.

Medero-k & Ostendorf-ek (2011) testuen sinplifikazioan egindako aldaketa sintaktikoak identifikatu eta deskribatzen dituen sistema bat aurkezten dute ingelesa irakurterrazagoa izateko.

Estatistika erabiliz ere, Bach et al.-ek (2011) *log-linear* ereduan oinarritutako sistema eraiki dute. Metodo horrek ezaugarri sorta baten gainean lan egiten duen *margin-based discriminative learning* algoritmo bat erabiltzen du. *Stack decoding* algoritmo batekin sinplifikazio hipotesiak sortu eta bilatzen dituzte.

Biomedikuntzako testuetan gertaeren erauzketa helburu izanik, Minard-ek, Ligozat-ek, & Grau-k (2012) ataza horretarako beharrezkoa den informazioarekin geratzeko sinplifikatzen dituzte esaldiak. Horretarako, corpus txiki baten etiketazioan oinarrituta CRF (*Conditional Random Fields*) sailkatzailea erabiltzen dute etiketatzeko, SVMen (*Super Vector Machine*) sarrera izango direnak. Esaldiak etiketatu ahala, corpusa handitzen joaten dira.

Danierarako, corpus handiak erabili gabe, Klerke-k & Søgaard-ek (2013) azpiesaldien aukeraketa eginez esaldi sinpleak lortzen dituzte. Azpiesaldi hautagaiak lortzeko, dependentzia analizatzaile sintaktiko batean oinarritzen diren heuristikoak, esaldiaren osagaiak mantentzeko erabiltzen dituztenak, ezabatzeko ausazko prozedura batekin konbinatzen dituzte. Hautagaien artean aukeratzeko funtzio-galera bat erabiltzen dute.

3.1.3 Hibridoak

Rol semantikoak etiketatzeko, esaldia banatzen dute Vickrey-k & Koller-ek (2008). Horretarako, eskuzko erregela sintaktikoak aplikatzen dituzte eta ondoren ikasketa automatikoaren bidez zein erregela aplikatu erabakitzen dute.

Koordinazioaren fenomenoan zentratuz Evans-ek (2011) esaldi koordinatuen anotazio sakon batean oinarrituta 4 sailkatzaile probatzen ditu. Esaldiak berridazteko corpus azterketan oinarritutako eskuzko erregelak erabiltzen ditu.

Frantseseko egitura sintaktikoak sinplifikatzeko arauak erauztean erabiltzen den eskuzko metodoa osatzeko, Seretan-ek (2012) estatistikoki nabarmentzen diren egitura linguistikoak proposatzen ditu etiketatzaileei laguntzeko.

Frantseserako ere modulu batek corpus azterketan lortutako erregelak programazio linealaren bitartez aukeratzen ditu (Brouwers et al., 2012).

3.2 Sinplifikazio lexikala

Azkeneko urte hauetan lexiko mailako sinplifikazioak ere bere tokia hartu du. Bere helburua hitzen ulergarritasuna areagotzea da, hitz konplexuak edo maiztasun gutxikoak hitz ezagunagoekin, sinonimoekin edo sintagmekin ordezkatuz. Hain garrantzitsua bihurtu da azken urte hauetan sinplifikazio lexikala non eta SemEval lehiaketan ataza bat antolatu zuten (Specia, Jauhar, & Mihalcea, 2012). Ataza horretan 5 sistemek hartu zuten parte eta teknika ezberdinak erabili zituzten (Amoia & Romanelli, 2012; Jauhar & Specia, 2012; Johannsen et al., 2012; Ligozat et al., 2012; Sinha, 2012).

Bi metodo nagusi erabil daitezke lexikoa sinplifikatzeko: hiztegietan eta datu-base lexikaletan oinarrituz ala estatistika erabiliz. Lan gehienek bien konbinazioak erabiltzen dituzte. Gehien erabili diren baliabideak, berriz, *WordNet* (Fellbaum, 2010) datu-base lexikala eta Wikipedia izan dira. Bi horien erabileraren arabera sailkatuko ditugu aurkeztuko ditugu lanak.

WordNet erabiliz, sinplifikazio lexikala aztertzen duten hainbat lan aurki ditzakegu, adibidez De Belder-en, Deschacht-en, & Moens-en (2010) metodoak bi motako hitz alternatiboen multzoak sortzea proposatzen du ordezkatu nahi den hitzari zuzenduak. Lehendabiziko hitz multzoa sinonimoak dituen hiztegi batetik edo WordNetetik lortzen da, eta bigarrena *Latent Words* hizkuntza-eredua erabiliz. Amaierako hitzaren aukeraketa probabilitate bidez kalkulatzen dute hiru baliabide hauetan oinarrituz: psikolinguistikako neurriak dituen datu-basea, testu errazen corpuseko unigramen probabilitatea eta silaba kopurua.

Bott et al.-ek (2012) ere bi mailatan oinarritzen den ordezkapen-eragiketak inplementatu dituzte gaztelaniako lexikoa sinplifikatzeko. Kasu honetan *OpenThesaurus* baliabide lexikalean oinarritzen dira eta ordezkapenerako hautagai hoberena aukeratzen dute hitzei sinpletasun neurri bat eman ondoren. WordNet eta *OpenThesaurus*en arteko konbinazioarekin ere esperimentuak egin dituzte (Saggion, Bott, & Rello, 2013).

WordNeten oinarrituz, Thomas-ek & Anderson-ek (2012) 6 algoritmo probatzen dituzte sinplifikazio lexikal egokiena lortzeko. Algoritmo horiek *Personalized Page Rank* eta informazio maximizazioaren printzipioak erabiltzen dituzte.

Nunes et al.-ek (2013) 4 pausotan banatutako metodoa aurkezten dute: kategoriak etiketatzea, sinonimoak identifikatzea, testuinguruaren maiztasunaren araberako ordezkapena eta esaldia zuzentzea. Erabiltzen dituzten baliabide lexikalak WordNet eta sinonimoen datu-base bat dira. Ordezkapenak egitean hitzen maiztasunak bilatzeko, haurren literaturako liburuekin osatutako hiztegi batean eta web bilatzaileetan oinarritzen dira.

Hiztegiak ere erabiliz, suedierarako Keskisärkkä-k (2012) sinonimoen ordezkapenaren bi-

tartez sinplifikatu du lexikoa. Ordezko sinonimoak aukeratzeko hiru estrategia erabili ditu: hitzen maiztasuna, hitzen luzera eta sinonimia.

Ingelesera itzuliz, medikuntzako lexikoa sinplifikatzeko, Leroy et al.-ek (2013) ordezkapen hitzak proposatzen dituen algoritmo bat testu editore batean integratzeko helburua dute, ondoren aditu batek balidazio edo gainbegiratze urratsean hitzik egokiena aukera dezan gramatikaltasuna bermatzearekin batera. Algoritmo horrek bi pausotan egiten du lan: lehenik, termino zailak identifikatzen ditu *Google Web Corpus*ean[6] n-grama kontaketak eginez eta agerpen gutxi dituztenak hitz zailagoak direla onartuz. Ondoren, termino horien hitz alternatibo errazak proposatzen ditu: sinonimoak eta hiperonimoak WordNetetik, definizioak eta mota semantikoak *Unified Medical Language System*etik[7] (UMLS), eta definizioak ingeles *Wiktionary*tik[8] eta *Simple Wiktionary*tik[9] erauziz, soilik kategoria gramatikal bera duten hitzak proposatzen dituelarik.

Aipatu beharreko beste baliabide bat Wikipedia da. Simple Wikipediako edizioen historiala erabiliz Yatskar et al.-ek (2010) bi hurbilpen proposatzen dituzte: i) edizio eragiketa guztiekin modelo probabilistiko bat sortzen dute eta ii) sinplifikazioak ez diren errebisioak iragazteko metadata erabiltzen dute.

Wikipediarekin jarraituz, Ligozat et al.-ek (2013) lexikoa sinplifikatzean kontuan hartzeko hiru irizpide aurkezten dituzte eta bakoitzari dagokion eredua aurkezten dute, Simple Wikipediako terminoen frekuentziak erabiliz, n-grametan oinarrituz eta kookurrentzien informazioa hartuz.

Ildo berean baina WordNet erabiliz, hitzen testuingurua kontuan hartzen duen bi mailako sistema proposatzen dute Biran-ek, Brody-k, & Elhadad-ek (2011). Lehenengo mailak erregelak erauzten ditu eta bigarrenak sinplifikazioa egiten du. Erregelak erauztean, lehenik, sinplifikatzeko hautagaiak diren edukizko hitz guztientzat (*stop words*, zenbakiak eta puntuazioa baztertuz) bektore bana eraikitzen dute eta, ondoren, hitz bakoitza ordezkatuko duen hautagaiak lortzeko WordNet erabiltzen dute. Sinplifikatzean jatorrizko esaldiaren testuinguruak bi modutan eragiten du: hitz-esaldien antzekotasuna eta testuinguruaren antzekotasuna. Egileen arabera sistema hori zazpi hitz baino gehiagoko esaldientzat da egokia.

Bayes teoreman oinarrituta, Shardlow-ek (2012) hitz bat testuinguru sinple batean agertzen den probabilitatea kalkulatzen du hitzen frekuentziaren gaineko kontaketak Wikipedia eta Simple Wikipediatik hartuz. Hitzak ordezkatzeko erabiltzen duten baliabide lexikala WordNet da. Shardlow-ek (2013a), berriz, sinplifikatzeko hautagaiak diren hitz konplexuak identifikatzeko metodoak azaltzen ditu.

Azpiatal honekin bukatzeko, Kauchak-ek (2013) hizkuntza eredu bat egokitzean sinplifikatu gabeko datuak (datu normalak) erabiltzearen eragina aztertu du eta ondorioztatu du datu normal gehigarriek ingeles errazaren hizkuntza-ereduen performantzia hobetzen dutela.

3.3 Bi sinplifikazioak

Atal honetan sinplifikazio sintaktikoa eta lexikala batera aztertzen dituzten lanak azalduko ditugu. Sistema horiek bi motatako sinplifikazioak egiten dituzte, hau da, testuaren konplexutasun lexikala eta sintaktikoa murrizten dute. Arkitektura orokorraren transformazio moduluan, beraz, sistema horiek bi sinplifikatzaile dituzte, sintaxia tratatzen duena eta lexikoa tratatzen duena.

3.3.1 Eskuzko erregeletan oinarritutakoak

Gizakia helburu duten lanekin hasiz, afasia dutenei albisteak egokitzeko Carroll et al.-en (1998) lanean bi mailatako arkitektura proposatzen da: analizatzailea eta sinplifikatzailea. Analizatzaile moduluak hiru azpimodulu ditu: lexiko etiketatzailea, analizatzaile morfologikoa eta analizatzaile sintaktikoa. Sinplifikatzaileak, berriz, bi atal dauzka: sintaxi sinplifikatzailea eta lexiko sinplifikatzailea. Sintaxi sinplifikatzaileak esaldi pasiboen aktiborako bihurketa, txertatutako perpausen erauzketa eta esaldien banaketa tratatzen ditu. Lexiko sinplifikatzaileak, aldiz, WordNeteko sinonimoak hartu eta Oxfordeko Psikolinguistikako datu-basean galdetuz, sinonimo bakoitzaren frekuentziak lortzen ditu. Ondoren, sinplifikazio mailaren arabera, sinonimo bat edo beste aukeratzen da. Sistema hori afasikoek dituzten fenomenoak kontuan izanda eraiki da.

Japonierarako eta jaiotzetiko entzumen arazoak dituzten pertsonei zuzenduta, 28.000 erregela baino gehiago inplementatu dituzte Inui et al.-ek (2003). Erregela horiek bai parafrasi lexikalak (sinonimoen ordezkapena), bai parafrasi sintaktikoak (banatutako egitura bat kendu, esaldiak banatu, e.a.) egiten dituzte.

[6] `http://catalog.ldc.upenn.edu/LDC2009T25` (2013ko urrian atzituta)

[7] `http://www.nlm.nih.gov/research/umls/` (2013ko irailean atzituta)

[8] `http://en.wiktionary.org/wiki/Wiktionary:Main_Page` (2013ko irailean atzituta)

[9] `http://simple.wiktionary.org/wiki/Main_Page` (2013ko irailean atzituta)

Medikuntzako literatura sinplifikatzeko asmoz, SIMTEXT sistemak (Damay et al., 2006; Ong et al., 2007), lexikoan sinonimoak ordezkatuz eta sintaxian perpausak banatzeko erregelak eta transformazio erregelak erabiliz osasun informazioa eskuragarriago egiten dute.

Biomedikuntzaren domeinuan, baina itzulpen automatikorako testuinguru baten barnean, medikuntza arloko testuak ingelesetik txinerara itzultzen dituen sisteman, eskuzko erregelen bitartez sinplifikatzen dute sintaxia Chen et al.-ek (2012) eta lexikoa terminoak ordezkatuz.

Analfabetismoari aurre egiteko, Brasilgo portugeserako 3.1.1. atalean aurkeztutako sistemari sinplifikazio lexikala gehituta sortu dute SIMPLIFICA sistema (Scarton et al., 2010), testu sinplifikatzailea integratuta duen testu-editorea. Lexikoa sinplifikatzeko, hitz konplexuak eta sinpleak dituzten hiztegietan oinarritzen dira.

Alfabetatze baxuaren arazoari aurka egiteko, Al-Baseet-ek (Al-Subaihin & Al-Khalifa, 2011), arabierarako sistemak, 4 moduluetako arkitektura proposatzen du: konplexutasuna ebaluatzea, lexikoa sinplifikatzea, sintaxia sinplifikatzea eta arabiar hizkuntzaren tipologia dela eta diakritizazioa. Lexiko mailan, sinonimoak bilatzeko WordNet proposatzen dute eta sintaxi mailan, elipsia, subjektuen, objektuen eta aditzen banaketa eta esaldi pasiboak bezalako fenomeno konplexuak lantzea proposatzen dute.

Larrialdien kudeaketaren domeinuetako testuak ulerterrazagoak egiteko, Temnikova-k, Orasan-ek, & Mitkov-ek (2012) hizkuntza kontrolatu bat proposatzen dute eta instrukzioak dituzten testuak sinplifikatzeko 5 motatako erregelak ematen dituzte: orokorrak, formatuaren gainekoak, sintaktikoak, lexikalak eta puntuazioaren gainekoak. Sinplifikatutako testuek egitura jakin batzuk jarraitu behar dituzte larrialdi kasuetan eraginkorrak izan daitezen: izenburua, azpitituluak, baldintzak, egin behar diren ekintzak (instrukzioak), oharrak (azalpenak) eta zerrendatzeak. Izenburuak eta instrukzioak ezinbestean azaldu behar badute ere, beste elementuak aukerakoak dira.

Helburu taldea irekia duen hiru mailako exekuzio-hodia duen arkitektura proposatzen du sintaxirako eta bikoa lexikorako Siddharthan-ek (2002). Sintaxiko hiru maila horiek analisia, transformazioa eta birsorkuntza dira eta lexikokoak parafrasiak eta sinonimoen ordezkapenak.

REGENT sistema (Siddharthan, 2011) dependentzia motatuetan oinarrituz, koordinazioa, mendeko perpausak, erlatibozko perpausak, eta aposizioak sinplifikatzeko eta ahots pasiboa aktibo bihurtzen dituen 63 erregelaz osatzen da. Esaldiak sortzeko bi aukera dauzka: i) transformatutako dependentzia grafoak hitzen hurrenkerarekin eta jatorrizko esaldiaren morfologiarekin lerratzea, *gen-light* eta ii) Stanfordeko dependentziak DSyntS errepresentazioak bihurtu ondoren esaldiak RealPro azalerako errealizatzailearekin sortzea, *gen-heavy*.

3.3.2 Datuetan oinarritutakoak

Tresnei begira, itzulpenaren kalitatea hobetzeko Doi-k & Sumita-k (2004) esaldiak banatzen dituzte. Bi pausotan egiten dute: lehendabizi hautagaiak lortzeko n-grametan oinarritutako hizkuntza-ereduak (NLM, *N-gram Language Model*) erabiltzen dituzte, ondoren hautagaien artean aukeratzeko NLMa eta esaldien antzekotasuna erabiltzen dituzte.

Estatistikan oinarritutako itzulpen automatikoko (SMT, *statistical machine translation*) teknikak oinarri hartuta zuhaitzen transformazioak egiten dituen eredua aurkezten dute Zhu-k, Bernhard-ek, & Gurevych-ek (2010). Eredu horrek 4 eragiketa egiten ditu: perpausak banatzea (*splitting*), hitzak ezabatzea edo erortzen uztea (*dropping*), ordenatzea (*reordering*) eta sintagma/hitz ordezkapena (*substitution*). Eredua iteratiboki entrenatzeko *expectation maximization* (EM) algoritmoa erabiltzen dute eta entrenatzeprozesu hori bizkortzeko hizkuntza bakarreko hitzen mapaketan oinarritutako metodoa aplikatzen dute. Azkenik, dekodifikatzaile bat erabiltzen dute esaldi sinplifikatuak sortzeko estrategia irenskorrak (*greedy*) erabiliz eta hizkuntza-ereduak integratuz.

Coster-ek & Kauchak-ek (2011a) itzulpen automatikoko hurbilpenak erabiltzen dituzte, baina sintagmetan oinarritutako aldaera gehitzen dute, itzulpen automatikoan aldaera horrek hobekuntzak lortu ez dituen arren, TSan sintaxian oinarritutakoak baino emaitza hobeak lortu ditu.

Wubben-ek, van den Bosch-ek, & Krahmerek (2012) ere esaldiak sinplifikatzeko sintagmetan oinarritutako itzulpen automatikoa erabiltzen dute, antzekotasun-ezan (*dissimilarity*) oinarritutako *re-ranking* heuristiko batekin areagotuz eta hizkuntza bakarreko corpus paralelo batean entrenatuz.

Portugueserako Specia-k (2010) ere SMT teknikak erabiltzen ditu. Metodo horrekin emaitza onak lortzen dira batez ere sinplifikazio lexikalean eta berridazketa sinpleetan.

Woodsend-ek & Lapata-k (2011a) jatorrizko eta helburu testuak kontuan izanik, berridazketa konplexuak egiten dituzten erregelak ikasten dituzte. Hurbilpen hori *quasi-synchronous gram-*

Hizkuntzak eta sistemak	**Sinplifikazio sintaktikoa**			**Sinpl.**
	Esk. erregelak	Datuak	Hibridoak	**lexikala**
Ingelesa				
(Chandrasekar, Doran, & Srinivas, 1996)	✔	-	-	-
(Chandrasekar & Srinivas, 1997)	-	✔	-	-
(Beigman Klebanov, Knight, & Marcu, 2004)	✔	-	-	-
(Daelemans, Höthker, & Sang, 2004)	✔	✔	-	-
(Doi & Sumita, 2004)	-	✔	-	-
(Max, 2005; Max, 2006)	✔	-	-	-
(Siddharthan, 2006)	✔	-	-	-
(Vickrey & Koller, 2008)	-	-	✔	-
(Jonnalagadda et al., 2009)	-	✔	-	-
(De Belder, Deschacht, & Moens, 2010)	-	-	-	✔
BioSimplify (Jonnalagadda & Gonzalez, 2010a)	✔	-	-	-
(Yatskar et al., 2010)	-	-	-	✔
(Bawakid & Oussalah, 2011)	✔	-	-	-
(Biran, Brody, & Elhadad, 2011)	-	-	-	✔
(Bach et al., 2011)	-	✔	-	-
(Evans, 2011)	-	-	✔	-
(Medero & Ostendorf, 2011)	-	✔	-	-
(Poornima et al., 2011)	✔	-	-	-
(Tur et al., 2011)	✔	-	-	-
(Amoia & Romanelli, 2012)	-	-	-	✔
(Jauhar & Specia, 2012)	-	-	-	✔
(Johannsen et al., 2012)	-	-	-	✔
(Ligozat et al., 2012)	-	-	-	✔
(Minard, Ligozat, & Grau, 2012)	-	✔	-	-
iSimp (Peng et al., 2012)	✔	-	-	-
(Shardlow, 2012)	-	-	-	✔
(Silveira Botelho & Branco, 2012)	✔	-	-	-
(Sinha, 2012)	-	-	-	✔
(Specia, Jauhar, & Mihalcea, 2012)	-	-	-	✔
(Srivastava & Sanyal, 2012)	✔	-	-	-
(Thomas & Anderson, 2012)	-	-	-	✔
(Nunes et al., 2013)	-	-	-	✔
(Kauchak, 2013)	-	-	-	✔
(Ligozat et al., 2013)	-	-	-	✔
(Shardlow, 2013a)	-	-	-	✔
Portugesa (Br eta Pt)				
PorSimples proiektua (Candido et al., 2009)	✔	-	-	-
(Silveira Botelho & Branco, 2012)	✔	-	-	-
Suediera				
CogFLUX (Rybing, Smith, & Silvervarg, 2010)	✔	-	-	-
(Keskisärkkä, 2012)	-	-	-	✔
Gaztelania				
(Bott et al., 2012; Saggion, Bott, & Rello, 2013)	-	-	-	✔
Frantsesa				
(Brouwers et al., 2012)	-	-	✔	-
(Seretan, 2012)	-	-	✔	-
Euskara				
(Aranzabe, Díaz de Ilarraza, & Gonzalez-Dios, 2012)	✔	-	-	-
Italiera				
ERNESTA (Barlacchi & Tonelli, 2013)	✔	-	-	-
Bulgariera				
(Lozanova et al., 2013)	✔	-	-	-
Koreera				
(Chung et al., 2013)	✔	-	-	-

1 taula: Sinplifikazio sintaktikoa edo lexikala egiten dituzten sistemak eta prototipoak hizkuntzaren, sinplifikazio motaren eta teknikaren arabera sailkatuta

*mar*ean (QG) oinarrituta dago. Programa lineal osoa bezala formulatuta dago eta QGa erabiltzen du berridazketa posible guztien espazioa harrapatzeko. Egileen arabera, eredu hori kontzeptualki sinplea eta konputazionalki efizientea da.

Feblowitz-ek & Kauchak-ek (2013) zuhaitzen transformazioa egiten dute lerratutako corpus baten analisiari probabilistikoki *synchronous tree substitution grammar* (STSG)-ekin ordezkapenak eginez. Hauek dira jarraitzen dituzten pausoak:

Hizkuntzak eta sistemak	Bi sinplifikazioak			Bestelakoak
	Esk. erregelak	Datuak	Hibridoak	
Ingelesa				
PSET, Systar (Carroll et al., 1998; Canning & Tait, 1999)	✔	-	-	-
(Siddharthan, 2002)	✔	-	-	-
SIMTEXT (Damay et al., 2006; Ong et al., 2007)	✔	-	-	-
(De Belder & Moens, 2010)	-	-	✔	-
(Kandula, Curtis, & Zeng-Treitler, 2010)	-	-	-	✔
(Siddharthan, 2010)	-	-	-	✔
(Zhu, Bernhard, & Gurevych, 2010)	-	✔	-	-
(Coster & Kauchak, 2011a)	-	✔	-	-
REGENT (Siddharthan, 2011)	✔	-	-	-
(Woodsend & Lapata, 2011a)	-	✔	-	-
(Chen et al., 2012)	✔	-	-	-
(Temnikova, Orasan, & Mitkov, 2012)	✔	-	-	-
(Wubben, van den Bosch, & Krahmer, 2012)	-	✔	-	-
(Feblowitz & Kauchak, 2013)	-	✔	-	-
(Paetzold & Specia, 2013)	✔	-	-	-
Japoniera				
(Inui et al., 2003)	✔	-	-	-
PorSimples, SIMPLIFICA (Gasperin et al., 2009)	✔	-	-	-
(Specia, 2010)	-	✔	-	-
Arabiera				
(Al-Subaihin & Al-Khalifa, 2011)	✔	-	-	-
Gaztelania				
Simplex proiektua (Bott, Saggion, & Figueroa, 2012)	-	-	✔	-
(Bautista et al., 2012)	-	-	-	✔
(Fajardo et al., 2013)	-	-	-	✔
Daniera				
(Klerke & Søgaard, 2013)	-	✔	-	-

2 taula: Bi sinplifikazioak eta bestelako sinplifikazioak egiten dituzten sistemak eta prototipoak hizkuntzaren, sinplifikazio motaren eta teknikaren arabera sailkatuta

i) gramatika ikasi zuhaitzen osagaiak lerratuz, ii) gramatika osatu informazio lexikala gehituz, iii) egoera finituko transduktore bat aplikatu, entrenatutako *log-linear* eredu batekin *n-best* sinplifikazio hoberenen zerrenda osatzeko eta iv) puntuazio altuena duena aukeratu.

Paetzold-ek & Specia-k (2013) Wikipediarekin ere zuhaitzen transformazioa egiten dituzten erregelak ikasten dituzte, bai sintaxia, bai lexikoa sinplifikatuko dituzten erregela bolumen handiak lortzeko. Hurbilpen horrek 3 osagai nagusi ditu: entrenatzeko modulua, sinplifikazio modulua eta *ranking* modulua. Erregelak ikasteko *Tree Transducer Toolkit* (T3) erabiltzen dute.

3.3.3 Hibridoak

Haurrei zuzendutako sistema garatzeko, lexikoa sinplifikatzeko WordNetetik lortutako hitz alternatiboak hizkuntza-eredu batekin konbinatzen dituzte De Belder-ek & Moens-ek (2010). Sintaxian aposizioak, erlatibozko perpausak, *Prefix subordination* (mendeko perpausak eta txertatzeko elementua gehituz) eta *infix coordination and subordination* (mendeko perpausak eta koordinatuak, soilik esaldiak banatuz) egiturak erregelen bidez sinplifikatzen dituzte eta ondoren programazio lineal osoaren bitartez testuan oro har izan dezaketen eragina kalkulatzen dute sinplifikazio hoberena aukeratuz.

Gizakiei testuak irisgarriagoak egiteko, gaztelaniako Bott-en, Saggion-en, & Figueroa-ren (2012) sistemak hiru pauso ditu: lehenik gramatika batek sinplifikatu behar diren egiturak bilatu eta etiketatzen ditu; bigarrenik, iragazki estatistiko batek berresten du ea benetan etiketatutako esaldiak sinplifikatu behar diren ala ez eta azkenik, manipulazio sintaktikoak egiten dituzte: ezabatzeak, txertatzeak eta nodo sintaktikoak kopiatzea. Erregelen bidez sinplifikatzen dituzten egiturak erlatibozko perpausak, gerundiozko eta partizipiozko egiturak, perpaus koordinatuak eta objektuen koordinazioa dira. Sistema horretan sinplifikazio lexikala eta materia sinplifikatzailea integratzeko asmoa dutela adierazten dute eta Drndarević et al.-en (2013) sintaxia eta lexikoa sinplifikatzen dituzten moduluak ebaluatzen dituzte.

3.4 Bestelako sinplifikazioak

Bestalde, badira ere esaldiak konektoreen erreformulazioen bitartez sinplifikatzen dituzten lanak (Siddharthan, 2010). Lan horretan, ingelesezko

because of lokailua duen esaldi batetik bi esaldi sortzeko erregeletan oinarritutako hiru hurbilpen ezberdinen konparazioa egiten da. Beste mota bateko sinplifikazioa zenbakidun adierazpenena da, esaterako gaztelaniazko *1,9 millones de hogares* kateari zenbaki bidezko adierazpenak eta lexikoa sinplifikatu ondoren *2 millones de casas* lortzea (Bautista et al., 2012). Zenbakidun adierazpenak sinplifikatzeko 5 eragiketa aurkezten dituzte: parentesi arteko zenbakiak ezabatzea, hizkiz daudenak zifraz ematea, kantitate handiak hitzen bitartez adieraztea, biribiltzea eta hamarrekoak ezabatuz biribiltzea.

Amaitzeko sintaktikoarekin batera semantikoa (Kandula, Curtis, & Zeng-Treitler, 2010) egiten duten lanak aurki ditzakegu. Semantikoa egiten duten lanak gutxiago dira HPan eta azalpenak gehitzea izango litzateke horien ataza, adibidez ingelesezko *Humerus* hitzari azalpena parentesi artean gehitzean *Humerus (a part of arm)* lortzea.

4 Ebaluaziorako metodoak

Testuen sinplifikazioa egiten duten sistemen ebaluazioa nola egin komunitatean irekita dagoen galdera bat da. Ataza horrentzat zehazki metrika edo metodoren bat proposatzen ez den bitartean, atal honetan aurkeztuko ditugu orain arte erabili diren metodoak. Metodo erabilienak itzulpen automatikoan erabiltzen diren neurriak, irakurketaren konplexutasun neurriak (*readability measures*) eta erabiltzaileei edo anotatzaileei galdeketak egitea dira. Normalean autoreek metodo bat baino gehiago erabiltzen dituzte sistemak ebaluatzeko.

4.1 Moduluz moduluko ebaluazioa

Hasierako lanetan moduluz modulu ebaluatzen ziren sistemak, analisia egiten zuen moduluari garrantzia emanez. Chandrasekar-ek, Doranek, & Srinivas-ek (1996) egin zuten ebaluazioan beren analisirako bi hurbilpenak (*chunk*etan eta dependentzietan oinarritutakoak) alderatu zituzten.

Siddharthan-ek (2002) moduluz moduluko ebaluazioa egiten du, baina etorkizuneko lanetan bi metodo proposatzen ditu sistemaren errendimendua oro har ebaluatzeko. Bi metodo horiek dira intrinsekoa (*intrinsically*), erabiltzaileen ebaluazioa erabiltzea, eta estrintsekoa (*extrinsically*), bere errendimendua beste sistema batean analizatzaile sintaktikoa, itzultzaile automatikoa) duen eragina neurtzea Jonnalagadda et al.-ek (2009) sinplifikatutako esaldiak eta jatorrizko esaldiak analizatzean egiten duten bezala. Siddharthan-ek (2006) erabiltzaileekin egindako ebaluazioan sinplifikatutako gramatikaltasuna eta antzekotasun semantikoa (ea jatorrizko esaldia eta sinplifikatutako esaldia baliokideak diren) neurtzen ditu.

Lan berriagoen artean, Aranzabek, Díaz de Ilarrazak, & Gonzalez-Diosek (2013) analisia eta esaldiak banatzen dituen modulua ebaluatzen dute sortu duten urre-patroi baten kontra konparatuz.

4.2 Erabiltzaileen bidezkoa

Ebaluatzeko beste metodo bat sistema hori helburu duen erabiltzaileekin ebaluatzea da. Carroll et al.-ek (1998) beraien sistema ebaluatzeko irakurketa esperimentuak egin dituzte ikusmen arazoak ez dituzten afasikoekin. Horrez gain, esaldien ulergarritasuna eta sistemaren baliagarritasuna aztertzeko subjektuak elkarrizketatu dituzte.

Begi mugimenduaren neurtzailea edo *Eyetracker*a erabiliz, hau da, begia eta buruaren arteko mugimendua eta begirada non kokatuta dagoen neurtzen duen tresna, gaztelanian sinplifikazio lexikalaren eragina aztertu da. Alde batetik, Rello et al.-ek (2013) lexikoa sinplifikatzeko bi estrategia probatu dituzte: hitzak sinonimo errazagoekin ordezkatzea eta sinonimo errazagoak eskaintzea hitz konplexuarekin batera. Bestetik, Rello-k, Baeza-Yates-ek, & Saggionek (2013) sinplifikazio lexikalaren eragina gaztelaniazko aditzen parafrasien bitartez (*confiar* eta *tener confianza* bezalako aditz eta kolokazio pareak kontrastatuz) neurtu dute. Rello et al.-ek (2013) adierazpen numerikoa letraz edo hitzen bitartez ematea aztertu zuten. Hiru esperimentu horiek dislexia diagnostikatuta duten pertsonekin egin zituzten, baina datuak kontrastatu ahal izateko kontrol talde bat ere osatu zuten.

Beste sistema batzuk ere erabiltzaileekin ebaluatuak izan dira, baina erabiltzaile horiek ez dira beti sistema garatzean izan zuten helburu taldekoak. De Belder-ek & Moens-ek (2010) ebaluazioa Wikipedia (jatorrizkoa eta sinplea) entziklopediako eta *Literacyworks*[10] web orriko testuekin eta *Amazon's Mechanical Turk crowdsourcing*erako[11] web zerbitzua erabiliz egin zuten. Sinplifikazio lexikoa ebaluatzeko, ordezkapena zuzena zen ala ez galdetzen zuten eta sinplifikazio sintaxikoa ebaluatzeko, esaldiak zuzenak

[10] `http://literacynet.org/cnnsf/` (2013ko urrian atzituta)

[11] `https://www.mturk.com/mturk/welcome` (2013ko irailean atzituta)

ziren ala ez galdetzen zuten.

Sistemak helburu talde jakinik ez duen kasuetan, Yatskar et al.-ek (2010) lexikoa sinplifikatzen duten sistema ezberdinak ebaluatzeko ingeleseko jatorrizko hiztunak eta jatorrizko hiztunak ez direnen anotazioak erabili dituzte. Hala, sistema horietako bakoitzak sortzen dituen ehun hitz-pare eta ausazko beste ehun hitz-pare hartu dituzte eta etiketatzaileei eskatu zaie bikote horietako bakoitzean adierazteko zein den sinpleagoa, zein konplexuagoa, antzekoak diren, erlaziorik gabekoak diren, zalantza sorrarazten dien edo erabakitzeko zaila gertatu zaien.

Mechanical Turk web zerbitzuaren bitartez Leroy et al.-ek (2013) medikuntza arlokoen testuen sinplifikazio lexikala ebaluatzeko bi parametro neurtu dituzte: nabaritutako zailtasuna eta benetako zailtasuna. Lehenengoa neurtzeko, 1-5 bitarteko Likert eskala erabiltzen dute, 1 oso erraza izanik eta 5 oso zaila. Bigarrena neurtzeko, hiru neurri erabiltzen dituzte: ulermena neurtzeko testuarekin batera agertzen diren 5 aukera anitzeko galdera, ikasketarako testurik gabeko beste 7 aukera anitzeko galdera eta informazioaren oroimena neurtzeko, 2 estaldura-galdera libre.

4.3 Ebaluazio automatikoa

Corpusaren kontrako neurketak egitea ebaluatzeko beste metodo bat da. Ebaluazio mota horrek eskatzen duen baliabidea da eskuzko corpus sinplifikatu bat izatea urre-patroi bezala erabiltzeko. Modu honetan sinplifikazio eragiketak eta erregelak ondo aplikatzen diren aztertu ohi da.

Candido et al.-ek (2009) eskuz sinplifikatutako corpus baten kontra eragiketa guztiak bananbanan ebaluatzen dituzte doitasuna, estaldura eta F neurria erabiliz. Horretaz gain, esaldiak zuzen sinplifikatuak izan diren ebaluatzeko eskuzko ebaluazioaren beharra aurreikusten dute esaldiak benetan sinplifikatu diren jakiteko, Aluísio et al.-en (2008) aipatu bezala. Sistema ebaluatzeko, 143 esaldiz osatutako erreferentzia corpusa eraiki dute eta bertan interesgarriak diren egitura sintaktikoak jaso dituzte Gasperin-ek, Maziero-k, & Aluisio-k (2010). Corpuseko esaldiak eskuz sinplifikatuak izan dira sinplifikazio erregelak jarraituz. Perpaus adberbialen sinplifikazio-erregelak ebaluatzeko erreferentzia corpuseko esaldi sinplifikatuen eta jatorrizko esaldien konparazioa bi etiketatzaileek egin zuten eta horiei hiru etiketa jartzea eskatu zieten: 0) esaldi sinplifikatuaren esanahia aldatzen da 1) esaldi sinplifikatuaren esanahia ez da aldatzen baina ez da irakurtzeko errazagoa 2) esaldi sinplifikatuaren esanahia ez da aldatzen eta irakurtzeko errazagoa da. Sinplifikatutako esaldiak ebaluatzeko, berriz, *Levenshtein* (edizio) distantzia erabiliz konparatzen dituzte erreferentzia corpuseko esaldi sinplifikatuekin. Erregelen aplikazio-hurrenkera ere ebaluatzen dute *Levenshtein* distantzia erabiliz.

Bott-ek, Saggion-ek, & Mille-k (2012) erregelak aplikatu diren kasu guztiak kontuan hartuz, erregela testuinguru egokian aplikatu den eta emaitza ona izan duen ebaluatzen dute. Ondoren, doitasuna, F neurria eta erregela aplikatu den maiztasuna kalkulatzen dituzte etiketatu dituzten 262 esaldietan.

Sinplifikazio lexikala ebaluatzeko Cohen-en *kappa* indizean oinarritutako anotatzaileen arteko adostasun neurria proposatzen dute Specia-k, Jauhar-ek, & Mihalcea-k (2012) bai anotatzailen arteko adostasuna kontrastatzeko, bai sistemen arteko konparazioa egiteko urre-patroiaren kontra.

4.4 Itzulpen automatikoko neurriak erabiliz

Testuen sinplifikazio automatikoa itzulpen prozesu bat bezala uler daitekeenez, itzulpen automatikoko sistemak bezala ere ebaluatua izan da arlo horretan erabiltzen diren metrikak aplikatuz.

Daelemans-ek, Höthker-ek, & Sang-ek (2004) ebaluazioak duen zailtasunari erreparatuta eta jakinda oso garestia dela eskuz ebaluatzea, itzulpen automatikoan erabiltzen den BLEU metrika proposatzen dute. Ildo horretatik jarraituz, Zhuk, Bernhard-ek, & Gurevych-ek (2010) ere MTko BLEU and NIST neurriak erabiltzen dituzte beraien sistemaz gain sortu dituzten beste 4 oinlerro sistema ebaluatzeko.

Specia-k (2010) ere neurri horiek erabiltzeaz gain kate-parekatzea eta eskuzko ebaluazioa egiten ditu. Horrela segmentuak egokiak eta naturalak diren eta espero zen sinplifikazioa egiten duten egiaztatzen du.

Coster-ek & Kauchak-ek (2011a) BLUEz gain testuen trinkotasuna neurtzeko erabiltzen diren beste bi neurri erabiltzen ditu: *Simple String Accuracy* (SSA) neurria eta hitzen gainean kalkulatutako F neurria.

Bach et al.-ek (2011) itzulpen eta laburpen automatikoetan erabiltzen diren AveF10, ROUGE-2 eta ROUGE-4 metrikekin batera Flesch-Kincaid konplexutasun neurria erabiltzen dute.

Barlacchi-k & Tonelli-k (2013) MTko neurria den TER erabiltzen dute, eta TER-Plus tresna.

4.5 Irakurketaren konplexutasun neurriak erabiliz

Aurreprozesu bezala erabiltzen diren konplexutasuna ebaluatzen duten sistemak ere erabili izan dira sinplifikatutako testu horien konplexutasun maila baxuagoa den aztertzeko. Adibidez, Siddharthan-ek (2006) Flesch konplexutasun neurriak erabiltzen ditu egunkarietako testuen sinplifikazioak ebaluatzeko.

Lehen aipatu dugun galdeketan metodoarekin batera konplexutasun formulak ere erabiltzen dituzte Drndarević et al.-ek (2013). Konplexutasun neurriak ausaz aukeratutako 100 testuri aplikatzen dizkiete; testu horiek hiru mailatan sinplifikatuak izan dira: lexikala, sintaktikoa eta biak. Galdeketetan, berriz, 25 etiketatzaileri hiru galdera erantzuteko eskatzen diete. Galdera horiek jatorrizko esaldien gramatikaltasunari, sinplifikatutako esaldien gramatikaltasunari eta jatorrizko esaldiaren eta sinplifikatutako esaldiaren esanahien arteko ezberdintasunei buruzkoa dira. Multzo bakoitzarentzat erdiko joera jakiteko batezbestekoa eta mediana kalkulatzen dituzte eta aldakortasunaren indikatzaile bezala frekuentzien distribuzioa.

Temnikova-k, Orasan-ek, & Mitkov-ek (2012) bi motako ebaluazioa egiten dute intrintsekoa eta estrintsekoa. Intrintsekoa konplexutasuna neurtzen duten neurrien bitartez egiten dute. Estrintsekoak, berriz, hiru modutan egiten du: irakurmenezko ulermenean duen eragina, eskuzko itzulpenean eta itzulpen automatikoan duen eragina eta amaierako erabiltzaileen onargarritasuna aztertzen ditu. Hirurak erabiltzaileekin egin zituzten; lehenengoa eta hirugarrena galdeketa bidez, eta bigarrena postedizio esfortzua automatikoki neurtuz (denbora, ikuspuntu teknikoa eta ikuspuntu kognitiboa).

Ebaluaziorako neurriak eta metodoak bateratzeko asmoarekin, eta goian aipatutako esperimentuetan oinarrituta, Temnikova-k & Manevak (2013) C-neurria *(Comprehension Score, C-score)* proposatzen dute testuen ulermena ebaluatzeko. C-neurria testuz testu kalkulatzen da eta hiru formula ditu: sinplea, osoa eta testu tamainakoa. Formula bakoitzak aldagai batzuk hartzen ditu kontuan, eta testuen tamaina ezberdinen arabera aplika daiteke bata edo bestea.

5 Ondorioak

Lan honetan testuen sinplifikazio automatikoaren arloaren egungo egoera aurkeztu dugu HPko ikerketa-lerro honen ikuspegi orokorra emateko. Hizkuntza eta helburu talde ezberdinetarako egin diren lanak aurkeztearekin batera sistemen deskribapenen laburpenak egin ditugu eta sistema horiek ebaluatzeko erabili diren metodoak azaldu ditugu. Orokorrean sistemek jarraitzen duten arkitektura deskribatu dugu. Sistema horiek kronologikoki egiten duten sinplifikazio motaren arabera (sintaktikoa, lexikala edo biak) eta metodo nagusiaren arabera (erregeletan, datuetan oinarrituta edo biak) sailkatu ditugu, 1. eta 2. tauletan laburburbildu dugun moduan. Horrez gain, ikerketa honetatik sortu diren tresnak eta baliabideak ere ezagutzera eman ditugu.

Deskribapen hori kontuan hartuta, testuen sinplifikazioaren bilakaera ikusi ahal izan dugu hasierako lanetatik azkeneko argitalpenetaraino. Nabarmentzekoa da azken urte hauetan ikerketalerro honek izan duen emankortasuna, bai hizkuntza gehiagotara zabaltzen ari delako, bai teknika eta metodo ugari esperimentatzen ari direlako. Ohartu gara ere, hizkuntza asko tipologiaren aldetik oso ezberdinak diren arren, sinplifikatzeko pausoak eta eragiketak oso antzekoak direla; izan ere, hizkuntza batentzat baliagarria dena beste batentzat ere baliagarria gertatu delako.

Amaitzeko, testu sinplifikatuen alorrean gero eta jende gehiago lan egiten ari dela ikus dezakegu, mota horretako testuek eskaintzen dituzten abantailak handiak baitira bai pertsonentzat, bai HPko tresnentzat.

Eskerrak

Itziar Gonzalez-Diosen lana Eusko Jaurlaritzak doktoreak ez diren ikertzaileak prestatzeko Doktoratu Aurreko Programako laguntza bati esker izan da. Ikerketa hau Eusko Jaurlaritzak IXA taldea, A motako ikertalde finkatua (IT344-10) eta MICCINek Hibrido Sint (TIN2010-20218) proiektuei emandako finantziazioagatik gauzatu da.

Erreferentziak

Al-Subaihin, Afnan A. & Hend S. Al-Khalifa. 2011. Al-Baseet: A proposed Simplification Authoring Tool for the Arabic Language. In *International Conference on Communications and Information Technology (ICCIT)*, pages 121–125, March.

Allen, David. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599.

Aluísio, Sandra M. & Caroline Gasperin. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplifica-

tion of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics.

Aluísio, Sandra M., Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, & Renata P.M. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 240–248, New York, NY, USA. ACM.

Amoia, Marilisa & Massimo Romanelli. 2012. SB: mmSystem-Using Decompositional Semantics for Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 482–486. Association for Computational Linguistics.

Aranzabe, María Jesús, Arantza Díaz de Ilarraza, & Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In Luz Rello & Horacio Saggion, editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8.

Aranzabe, María Jesús, Arantza Díaz de Ilarraza, & Itziar Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68.

Bach, Nguyen, Qin Gao, Stephan Vogel, & Alex Waibel. 2011. TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 474–482.

Barlacchi, Gianni & Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 476–487.

Bautista, Susana, Biljana Drndarevic, Raquel Hervás, Horacio Saggion, & Pablo Gervás. 2012. Análisis de la Simplificación de Expresiones Numéricas en Español mediante un Estudio Empírico. *Linguamática*, 4(2):27–41.

Bautista, Susana, Raquel Hervás, & Pablo Gervás. 2012. Simplificación de textos centrada en la adaptación de expresiones numéricas. In *I Congreso Internacional Universidad y Discapacidad*, Madrid, 11/2012.

Bawakid, Abdullah & Mourad Oussalah. 2011. Sentences Simplification for Automatic summarization. In *Cybernetic Intelligent Systems (CIS), 2011 IEEE 10th International Conference on*, pages 59–64. IEEE.

Beigman Klebanov, Beata, Kevin Knight, & Daniel Marcu. 2004. Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 735–747.

Bernhard, Delphine, Louis De Viron, Véronique Moriceau, & Xavier Tannier. 2012. Question Generation for French: Collating Parsers and Paraphrasing Questions. *Dialogue and Discourse*, 3(2):43–74.

Biran, Or, Samuel Brody, & Noemie Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA, June. Association for Computational Linguistics.

Blake, Catherine, Julia Kampov, Andreas K Orphanides, David West, & Cory Lown. 2007. UNC-CH at DUC 2007: Query expansion, lexical simplification and sentence selection strategies for Multi-Document summarization. In *Proceedings of the Document Understanding Conference 2007*.

Bott, Stefan, Luz Rello, Biljana Drndarevic, & Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING*, pages 357–373.

Bott, Stefan & Horacio Saggion. 2011. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bott, Stefan & Horacio Saggion. 2012. Automatic simplification of spanish text for e-accessibility. In *Computers Helping People with Special Needs*. Springer, pages 527–534.

Bott, Stefan, Horacio Saggion, & David Figueroa. 2012. A Hybrid System for Spanish Text Simplification. In *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 75–84, Montreal, Canada.

Bott, Stefan, Horacio Saggion, & Simon Mille. 2012. Text Simplification Tools for Spanish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Brouwers, Laetitia, Delphine Bernhard, Anne-Laure Ligozat, & Thomas François. 2012. Simplification syntaxique de phrases pour le français. In *Actes de la Conférence Conjointe JEP-TALN-RECITAL, Montpellier, France*, pages 211–224.

Burstein, Jill. 2009. Opportunities for Natural Language Processing Research in Education. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 5449 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 6–27.

Buyko, Ekaterina, Erik Faessler, Joachim Wermter, & Udo Hahn. 2011. Syntactic Simplification and Semantic Enrichment-Trimming Dependency Graphs for Event Extraction. *Computational Intelligence*, 27(4):610–644.

Candido, Jr., Arnaldo, Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, & Sandra M. Aluisio. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Canning, Yvonne & John Tait. 1999. Syntactic Simplification of Newspaper Text for Aphasic Readers. In *ACM SIGIR'99 Workshop on Customised Information Delivery*, pages 6–11. Citeseer.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, & John Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, & John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL*, volumen 99, pages 269–270. Citeseer.

Caseli, Helena M., Tiago F. Pereira, Lucia Specia, Thiago. A. S. Pardo, Caroline. Gasperin, & Sandra Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *the Proceedings of CICLing*, pages 59–70.

Chandrasekar, Raman, Christine Doran, & Bangalore Srinivas. 1996. Motivations and Methods for Text Simplification. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 1041–1044, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chandrasekar, Raman & Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Chen, Han-Bin, Hen-Hsen Huang, Hsin-Hsi Chen, & Ching-Ting Tan. 2012. A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. In *COLING*, pages 545–560.

Chung, Jin-Woo, Hye-Jin Min, Joonyeob Kim, & Jong C Park. 2013. Enhancing Readability of Web Documents by Text Augmentation for Deaf People. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 30:1–30:10, New York, NY, USA. ACM.

Coster, William & David Kauchak. 2011a. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coster, William & David Kauchak. 2011b. Simple English Wikipedia: A New Text Simplification Task. In *ACL (Short Papers)'11*, pages 665–669.

Crossley, Scott A, David Allen, & Danielle S McNamara. 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1):89–108.

Daelemans, Walter, Anja Höthker, & Erick Tjong Kim Sang. 2004. Automatic Sentence Simplification for Subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Damay, Jerwin Jan S., Gerard Jaime D. Lojico, Kimberly Amanda L. Lu, Dex B. Tarantan, & Ethel C. Ong. 2006. SIMTEXT. Text Simplification of Medical Literature. In *3rd National Natural Language Processing Symposium - Building Language Tools and Resources*.

De Belder, Jan, Koen Deschacht, & Marie-Francine Moens. 2010. Lexical Simplification. In *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

De Belder, Jan & Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

De Belder, Jan & Marie-Francine Moens. 2012. A Dataset for the Evaluation of Lexical Simplification. *Computational Linguistics and Intelligent Text Processing*, pages 426–437.

Dell'Orletta, Felice, Simonetta Montemagni, & Giulia Venturi. 2011. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pages 73–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Devlin, Siobhan & Gary Unthank. 2006. Helping Aphasic People Process Online Information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06, pages 225–226, New York, NY, USA. ACM.

Doi, Takao & Eiichiro Sumita. 2004. Splitting Input Sentence for Machine Translation Using Language Model with Sentence Similarity. In *Proc. of the 20th international conference on Computational Linguistics*.

Drndarević, Biljana, Sanja Štajner, Stefan Bott, Susana Bautista, & Horacio Saggion. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 488–500.

Evans, Richard J. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and linguistic computing*, 26(4):371–388.

Fajardo, Inmaculada, Gema Tavares, Vicenta Ávila, & Antonio Ferrer. 2013. Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities*, 34(4):1267–1279.

Feblowitz, Dan & David Kauchak. 2013. Sentence Simplification as Tree Transduction. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10, Sofia, Bulgaria, August. Association for Computational Linguistics.

Fellbaum, Christiane. 2010. WordNet. In Roberto Poli, Michael Healy, & Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*. Springer Netherlands, pages 231–243.

Gasperin, Caroline, Erick Maziero, & Sandra M Aluisio. 2010. Challenging Choices for Text Simplification. In *Computational Processing of the Portuguese Language*. Springer, pages 40–50.

Gasperin, Caroline, Erick Maziero, Lucia Specia, Thiago A.S. Pardo, & Sandra M. Aluisio. 2009. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *the Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401.

Hancke, Julia, Sowmya Vajjala, & Detmar Meurers. 2012. Readability Classification for German using lexical, syntactic, and morphological features. In *COLING 2012: Technical Papers*, page 1063–1080.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, & Tomoya Iwakura. 2003. Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics.

Jauhar, Sujay Kumar & Lucia Specia. 2012. UOW-SHEF: SimpLex–Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 477–481. Association for Computational Linguistics.

Johannsen, Anders, Héctor Martínez, Sigrid Klerke, & Anders Søgaard. 2012. EMNLP@

CPH: Is frequency all there is to simplicity? In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 408–412. Association for Computational Linguistics.

Jonnalagadda, Siddhartha & Graciela Gonzalez. 2010a. BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. In *AMIA Annual Symposium Proceedings*, volumen 2010, page 351. American Medical Informatics Association.

Jonnalagadda, Siddhartha & Graciela Gonzalez. 2010b. Sentence Simplification Aids Protein-Protein Interaction Extraction. *Arxiv preprint arXiv:1001.4273*.

Jonnalagadda, Siddhartha, Luis Tari, Joerg Hakenberg, Chitta Baral, & Graciela Gonzalez. 2009. Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 177–180. Association for Computational Linguistics.

Kandula, Sasikiran, Dorothy Curtis, & Qing Zeng-Treitler. 2010. A Semantic and Syntactic Text Simplification Tool for Health Content. In *AMIA Annual Symposium Proceedings*, volumen 2010, page 366. American Medical Informatics Association.

Kauchak, David. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria, August. Association for Computational Linguistics.

Keskisärkkä, Robin. 2012. Automatic Text Simplification via Synonym Replacement. Master's thesis, Linköping.

Klaper, David, Sarah Ebling, & Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Klerke, Sigrid & Anders Søgaard. 2012. DSim, a Danish Parallel Corpus for Text Simplification. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Klerke, Sigrid & Anders Søgaard. 2013. Simple, readable sub-sentences. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 142–149, Sofia, Bulgaria, August. Association for Computational Linguistics.

Kvistab, Maria & Sumithra Velupillaia. 2013. Professional Language in Swedish Radiology Reports–Characterization for Patient-Adapted Text Simplification. In *Scandinavian Conference on Health Informatics 2013*, page 55.

Lal, Partha & Stefan Rüger. 2002. Extract-based Summarization with Simplification. In *Proceedings of the Workshop on Text Summarization at DUC 2002 In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization*.

Leroy, Gondy, James E Endicott, David Kauchak, Obay Mouradi, & Melissa Just. 2013. User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of medical Internet research*, 15(7).

Ligozat, Anne-Laure, Anne Garcia-Fernandez, Cyril Grouin, & Delphine Bernhard. 2012. ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 487–492. Association for Computational Linguistics.

Ligozat, Anne-Laure, Cyril Grouin, Anne Garcia-Fernandez, & Delphine Bernhard. 2013. Approches à base de fréquences pour la simplification lexicale. In *Actes TALN-RÉCITAL 2013*, pages 493–506. ATALA.

Lozanova, Slavina, Ivelina Stoyanova, Svetlozara Leseva, Svetla Koeva, & Boian Savtchev. 2013. Text Modification for Bulgarian Sign Language Users. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 39–48, Sofia, Bulgaria, August. Association for Computational Linguistics.

Max, Aurélien. 2005. Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*.

Max, Aurélien. 2006. Writing for Language-Impaired Readers. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 3878 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 567–570.

Medero, Julie & Mari Ostendorf. 2011. Identifying Targets for Syntactic Simplification. In *Proceedings of the SLaTE 2011 workshop*, pages 69–72.

Minard, Anne-Lyse, Anne-Laure Ligozat, & Brigitte Grau. 2012. Simplification de phrases pour l'extraction de relations. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 1–14, Grenoble, France, June. ATALA/AFCP.

Nunes, Bernardo Pereira, Ricardo Kawase, Patrick Siehndel, Marco A. Casanova, & Stefan Dietze. 2013. As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on*, pages 128–132.

Oh, Sun Young. 2001. Two Types of Input Modification and EFL Reading omprehension: Simplification Versus Elaboration. *TESOL Quarterly*, 35(1):69–96.

Ong, Ethel, Jerwin Damay, Gerard Lojico, Kimberly Lu, & Dex Tarantan. 2007. Simplifying Text in Medical Literature. *J. Research in Science Computing and Eng*, 4(1):37–47.

Paetzold, Gustavo H & Lucia Specia. 2013. Text Simplification as Tree Transduction. In Sociedade Brasileira de Computaçao, editor, *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 116–125.

Peng, Yifan, Catalina O Tudor, Manabu Torii, Cathy H Wu, & K Vijay-Shanker. 2012. iSimp: A Sentence Simplification System for Biomedical Text. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE.

Petersen, Sarah E. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Ph.D. thesis, Citeseer.

Petersen, Sarah E & Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *In Proceedings of Workshop on Speech and Language Technology for Education. SLaTE*, pages 69–72. Citeseer.

Poornima, C., V. Dhanalakshmi, K.M. Anand, & KP Soman. 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42.

Rello, Luz, Ricardo Baeza-Yates, Stefan Bott, & Horacio Saggion. 2013. Simplify or Help? Text Simplification Strategies for People with Dyslexia. *Proc. W4A*, 13.

Rello, Luz, Ricardo Baeza-Yates, & Horacio Saggion. 2013. The Impact of Lexical Simplification by Verbal Paraphrases for People with and without Dyslexia. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 501–512.

Rello, Luz, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, & Horacio Saggion. 2013. One Half or 50%? An Eye-Tracking Study of Number Representation Readability. In *Proc. INTERACT*, volumen 13, pages 1–17.

Rybing, Jonas, Christian Smith, & Annika Silvervarg. 2010. Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*, pages 17–18.

Saggion, Horacio, Stefan Bott, & Luz Rello. 2013. Comparing Resources for Spanish Lexical Simplification. In *SLSP 2013: 1st International Conference on Statistical Language and Speech Processing*, pages 1–12. Springer.

Saggion, Horacio, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, & Lorena Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

Scarton, Carolina, Matheus de Oliveira, Arnaldo Candido Jr, Caroline Gasperin, & Sandra Maria Aluísio. 2010. SIMPLIFICA: a

tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 41–44. Association for Computational Linguistics.

Seretan, Violeta. 2012. Acquisition of Syntactic Simplification Rules for French. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shardlow, Matthew. 2012. Bayesian Lexical Simplification. Txosten teknikoa, Short Taster Research Project. The University of Manchester.

Shardlow, Matthew. 2013a. A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shardlow, Matthew. 2013b. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria, August. Association for Computational Linguistics.

Siddharthan, Advaith. 2002. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC'02)*, pages 64–71, Washington, DC, USA. IEEE Computer Society.

Siddharthan, Advaith. 2006. Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109.

Siddharthan, Advaith. 2010. Complex Lexico-Syntactic Reformulation of Sentences using Typed Dependency Representations . In *Proceedings of the 6th International Natural Language Generation Conference*, pages 125–133. Association for Computational Linguistics.

Siddharthan, Advaith. 2011. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11. Association for Computational Linguistics.

Siddharthan, Advaith, Ani Nenkova, & Kathleen McKeown. 2004. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 896. Association for Computational Linguistics.

Silveira Botelho, Sara & António Branco. 2012. Enhancing Multi-document Summaries with Sentence Simplification. In *ICAI 2012: International Conference on Artificial Intelligence*.

Simensen, Aud Marit. 1987. Adapted Readers: How are they Adapted. *Reading in a Foreign Language*, 4(1):41–57.

Sinha, Ravi. 2012. UNT-SimpRank: Systems for Lexical Simplification Ranking. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 493–496. Association for Computational Linguistics.

Specia, Lucia. 2010. Translating from Complex to Simplified Sentences. *Computational Processing of the Portuguese Language*, pages 30–39.

Specia, Lucia, Sandra M. Aluísio, & Thiago A.S. Pardo. 2008. Manual de Simplificação Sintática para o Português. Txosten teknikoa NILC-TR-08-06, São Carlos-SP.

Specia, Lucia, Sujay Kumar Jauhar, & Rada Mihalcea. 2012. Semeval-2012 Task 1: English Lexical Simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

Srivastava, Jyoti & Sudip Sanyal. 2012. Segmenting Long Sentence Pairs to Improve Word Alignment in English-Hindi Parallel Corpora. In *Advances in Natural Language Processing*. Springer, pages 97–107.

Štajner, Sanja & Horacio Saggion. 2013. Adapting Text Simplification Decisions to Different Text Genres and Target Users. *Procesamiento del Lenguaje Natural*, 51:135–142.

Temnikova, Irina & Galina Maneva. 2013. The C-Score – Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29, Sofia, Bulgaria, August. Association for Computational Linguistics.

Temnikova, Irina, Constantin Orasan, & Ruslan Mitkov. 2012. CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3007–3014, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Thomas, S. Rebecca & Sven Anderson. 2012. WordNet-Based Lexical Simplification of a Document. In *Empirical Methods in Natural Language Processing*, pages 80–88.

Tur, Gokhan, Dilek Hakkani-Tur, Larry Heck, & S. Parthasarathy. 2011. Sentence Simplification for Spoken Language Understanding. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5628–5631. IEEE.

Vanderwende, Lucy, Hisami Suzuki, Chris Brockett, & Ani Nenkova. 2007. Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion. *Information Processing & Management*, 43(6):1606–1618.

Vickrey, David & Daphne Koller. 2008. Sentence Simplification for Semantic Role Labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008: HLT)*, pages 344–352.

Woodsend, Kristian & Mirella Lapata. 2011a. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

Woodsend, Kristian & Mirella Lapata. 2011b. WikiSimple: Automatic Simplification of Wikipedia Articles. In *Proceedings of the TwentyFifth AAAI Conference on Artificial Intelligence*, pages 927–932.

Wubben, Sander, Antal van den Bosch, & Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1015–1024. Association for Computational Linguistics.

Yatskar, Mark, Bo Pang, Cristian Danescu-Niculescu-Mizil, & Lillian Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Young, Dolly N. 1999. Linguistic Simplification of SL Reading Material: Effective Instructional Practice? *The Modern Language Journal*, 83(3):350–366.

Zhu, Zhemin, Delphine Bernhard, & Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, pages 1353–1361. Association for Computational Linguistics.

Hacia un tratamiento computacional del Aktionsart

Towards a Computational Treatment of Aktionsart

Juan Aparicio
Universitat de Barcelona
juanapariciomera@yahoo.es

Irene Castellón
Universitat de Barcelona
icastellon@ub.edu

Marta Coll-Florit
Universitat Oberta de Catalunya
mcollfl@uoc.edu

Resumen

En el área del Procesamiento del Lenguaje Natural (PLN), a la hora de crear aplicaciones inteligentes, el tratamiento semántico es fundamental. Sin embargo, la investigación que actualmente se está llevando a cabo en PLN está todavía lejos de conseguir niveles profundos de compresión del lenguaje. El objetivo principal de nuestra investigación es la representación del Aktionsart (la manera como se construye el evento expresado por un verbo en su desarrollo temporal). Una de las dificultades básicas que presenta el tratamiento semántico del lenguaje es el establecimiento de clases, debido principalmente a la naturaleza gradual del significado y la alta incidencia del contexto en la interpretación de las diferentes unidades. En este artículo nos centraremos en la presentación de las clases aspectuales léxicas de nuestra propuesta. El total de clases definidas se clasifica en dos grupos, las clases simples: estados, procesos y puntos, cuya combinación da lugar a las clases complejas: culminaciones, realizaciones y graduales. Esta presentación se llevará a cabo tanto desde el punto de vista teórico, como de su implementación computacional.

Palabras clave

Lingüística computacional, Aktionsart, clases, implementación.

Abstract

In the area of Natural Language Processing (NLP), when creating intelligent applications, semantic processing is essential. However, research currently being conducted in NLP is still far from achieving deep levels of understanding of language. The main goal of our research is the representation of Aktionsart (how the event expressed by a verb is construed as unfolding over time). One of the basic difficulties presented by the semantic processing of language is establishing classes, mainly due to the gradual nature of meaning, and the high incidence of context in the interpretation of the different units. In this work we focus on the presentation of the lexical aspectual classes of our proposal. The total number of defined classes is classified into two groups, simple classes: states, processes and points, the combination of which gives rise to the complex classes: culminations, accomplishments and graduals. This presentation will take place both from the theoretical point of view, and its computational implementation.

Keywords

Computational linguistic, Aktionsart, classes, implementation.

1 Introducción

El tratamiento semántico en el área del Procesamiento del Lenguaje Natural (PLN) es fundamental para la creación de cualquier aplicación inteligente, como por ejemplo los sistemas de pregunta-respuesta, la extracción de información o la implicación textual. No obstante, la investigación actual en PLN está lejos de llegar a una comprensión profunda del lenguaje, ya que se basa mayoritariamente en métodos estadísticos superficiales y se dispone de pocos recursos anotados a nivel semántico.

Una de las dificultades básicas que presenta el tratamiento semántico del lenguaje es el problema del establecimiento de clases, principalmente por la naturaleza gradual del significado y la alta incidencia del contexto en la interpretación de las unidades. Precisamente una de las tareas básicas de investigación en PLN que avanza con mayor dificultad es la resolución automática de la ambigüedad semántica del léxico (*Word Sense Disambiguation*, WSD) (Agirre y Edmonds, 2007), en la que los resultados descienden dramáticamente en la desambiguación de unidades verbales. Así, parece evidente la necesidad de una caracterización profunda de las unidades léxicas y de las relaciones que se establecen entre ellas para obtener una representación del significado y, posteriormente, aplicar procesos de razonamiento.

El objetivo general de nuestra investigación es el tratamiento formal del Aktionsart, una de las características semánticas menos tratadas en la representación computacional. El Aktionsart, también denominado aspecto léxico o modo de

acción, se refiere a la manera en que el evento expresado por un verbo se desarrolla y se distribuye en el tiempo: si es estático o dinámico, si es durativo o puntual, si es homogéneo o implica una culminación, entre otras distinciones. A partir de la combinación de estas oposiciones nocionales básicas se han propuesto tipologías de clasificación verbal que se consideran útiles para predecir el comportamiento sintáctico de los predicados. A modo de ejemplo, los verbos que expresan eventos durativos generalmente no admiten modificadores temporales puntuales (p.ej. **La policía persiguió al ladrón en un instante*), mientras que este contexto es perfectamente plausible con un verbo que exprese un evento puntual o de escasa duración (p.ej. *La policía atrapó al ladrón en un instante*).

Aunque existen numerosos trabajos lingüísticos sobre este tipo de información (Vendler, 1957; Verkuyl, 1989; Pustejovsky, 1991; Smith, 1991; Levin y Rappaport Hovav, 1995; De Miguel, 1999, 2004; Croft, 2008; Coll-Florit, 2011, 2012; entre otros), lejos de ser un ámbito de estudio con unos principios teóricos y metodológicos consensuados, la bibliografía sobre Aktionsart se caracteriza por la multiplicidad de propuestas que difieren en cuanto al número y organización de las clases aspectuales. Además, existe un amplio debate a la hora de explicar por qué un mismo verbo puede admitir más de una interpretación aspectual. Por ejemplo, comparemos las siguientes oraciones (1-2):

1. María se bebió un vaso de agua en cinco minutos /*durante cinco minutos.
2. María bebió agua durante cinco minutos / *en cinco minutos.

Observamos que en este caso obtenemos una interpretación télica o atélica dependiendo de las propiedades del objeto directo. Existen diferentes razones por las que un verbo puede moverse de una clase aspectual a otra: tiempo verbal, sujeto plural, perífrasis verbales, etc. (De Miguel, 1999; Rothstein, 2004). Así, una propuesta de clasificación aspectual debe considerar operaciones de cambio o coerciones. En nuestra propuesta, defenderemos que la clase aspectual del verbo determina de qué operaciones de cambio podría ser input y en qué contextos.

Desde el punto de vista metodológico, uno de los problemas que presentan estos estudios para el tratamiento computacional es que la mayoría suelen tratar pocas clases eventivas con escasos ejemplos. Son pocos los trabajos de amplia cobertura que, además de un sistema completo de representación, proporcionen una clasificación de un conjunto de predicados extenso. La información representada en un sistema computacional que trate el lenguaje real no puede ser parcial y los predicados tratados no pueden ser pocos si lo que se pretende es comprobar su viabilidad tanto descriptiva como predictiva.

El objetivo final de nuestra investigación es establecer una propuesta representacional de amplia cobertura, que defina las posibles operaciones de cambio aspectual y que esté expresada formalmente para su aplicación computacional. En este artículo nos centraremos en la presentación de las clases aspectuales léxicas de nuestra propuesta, tanto desde el punto de vista teórico (§2), como de su implementación computacional (§3).

2 Descripción de las clases léxicas

Uno de los autores de referencia en el estudio de la estructura eventiva es Vendler (1957). Vendler propone cuatro clases aspectuales de los predicados verbales: estados, actividades, realizaciones y logros. En concreto, el autor establece una distinción genérica entre clases que implican progresión temporal, esto es, sucesión de diferentes fases temporales (actividades y realizaciones), y clases que están formadas por una sola fase temporal (estados y logros).

Nuestra propuesta de clasificación eventiva se sitúa en el marco de los modelos de descomposición semántica (Dowty, 1979; Tenny, 1994; Moens y Steedman, 1988; Grimshaw, 1990; Pustejovsky, 1991, 1995; Engelberg, 1999; Levin y Rappaport Hovav, 1995, 2005; Rappaport Hovav y Levin, 1998, 2000; De Miguel, 2004; entre otros). Según esta aproximación, los eventos denotados por los predicados verbales no constituyen entidades atómicas, sino que están dotados de una estructura subléxica o subeventiva, por eso, se tratan de manera separada los eventos simples y los eventos complejos, en función del número de subeventos implicados. Más concretamente, se considera que un evento simple consiste en un único subevento, mientras que un evento complejo está compuesto por más de un subevento que, independientemente, está bien formado.

Uno de los trabajos más relevantes que parte de la descomposición semántica es el de Dowty (1979[1]. Según este autor, las diferentes propiedades aspectuales de un evento se pueden explicar a partir de una clase homogénea de predicados estativos,

[1] La propuesta de Dowty (1979) fue ampliada posteriormente por Levin y Rappaport Hovav (1995) y Rappaport Hovav y Levin (1998).

más tres conectores aspectuales: DO, BECOME y CAUSE. Las estructuras lógicas propuestas por Dowty para cada una de las clases eventivas de Vendler se configuran tal como se muestra en la Tabla 1.

Estado	**predicado' (x)**
Logro	**BECOME predicado' (x)**
Actividad	**DO (x, [predicado' (X)])**
Realización	**Ø CAUSE ψ (en que Ø es normalmente una actividad y ψ un logro**

Tabla 1: Estructuras lógicas de Dowty (1979).

Otro de los modelos más representativos de esta aproximación es el de Pustejovsky (1991, 1995), quien asume que los eventos están dotados de una estructura interna que se puede descomponer en diferentes etapas o subeventos. En concreto, Pustejovsky propone tres clases eventivas: dos clases simples, Estados (E) y Procesos (P), y una clase compleja, las Transiciones (T). La estructura de las diferentes clases propuestas por Pustejovsky se puede representar de la siguiente manera (Fig. 1)

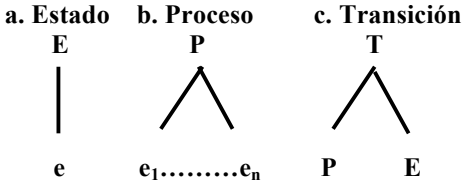


Fig. 1: Modelo de Pustejovsky.

Basándonos en los modelos de descomposición semántica, nuestra propuesta de clasificación eventiva también establece la distinción entre eventos simples y eventos complejos. En particular, proponemos tres clases de eventos simples. Así, además del Estado (E) y el Proceso (Pr) del modelo de Pustejovky, incluimos una tercera clase simple, el Punto (Pu), que se refiere a un evento que ocurre de forma instantánea, sin implicar una consecuencia o estado resultante. No obstante, se considera dinámico porque ocurre (implica un cambio cualitativo). La descripción general de las tres clases simples de nuestra clasificación es la siguiente:

- *Estado (E):* situación homogénea que no ocurre, sólo se limita a mantenerse durante el periodo temporal en el cual se da. Ejemplos: *equivaler*, *caber*, *pertenecer.*
- *Proceso (Pr):* evento dinámico que implica sucesión de diferentes fases temporales (con progresión), pero no tiene una culminación temporal inherente. Ejemplos: *caminar*, *buscar*, *perseguir*.
- *Punto (Pu):* evento dinámico y puntual (ocurre en breves instantes), que no implica un cambio o consecuencia. Ejemplos: *toser*, *pestañear*, *saltar*[2].

Estas tres clases simples focalizan los diferentes estadios básicos de un núcleo eventivo (Moens i Steedman, 1988): el proceso previo, el punto de culminación del evento y el estado resultante, tal como se representa gráficamente en la Figura 2.

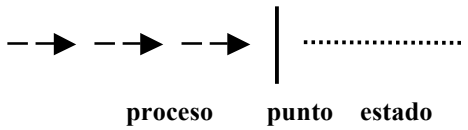


Fig. 2. Estadios básicos de un núcleo eventivo.

Asimismo, contemplamos tres clases de eventos complejos que resultan de la combinación de dos o más clases simples:

- *Culminación* (C): evento complejo compuesto por un punto (Pu) y una consecuencia, generalmente un estado (E). Ejemplos: *superar*, *marearse*, *cerrar*[3].
- *Realización* (R): evento complejo formado por un proceso (Pr) y una culminación (C). Ejemplos: *construir*, *aprender*, *instalar*.
- *Gradual* (G): evento complejo formado por una iteración de culminaciones (C), con un cambio gradual. Ejemplos: *enfriar*, *secar*, *engordar.*

La Tabla 3 presenta la formalización de la estructura interna de estas tres clases complejas.

A su vez, entendemos que estos grupos eventivos son clases genéricas que, en algunos casos, se estructuran internamente en diferentes subclases.

[2] Esta clase eventiva equivale a los llamados verbos semelfactivos (Smith, 1991). No obstante, en nuestro trabajo hemos optado por usar la terminología de Moens y Steedman (1988).

[3] Esta clase corresponde en gran medida a los logros tradicionales. Sin embargo, hemos optado por utilizar la terminología de Moens y Steedman (1988).

Culminación	**C = Pu + E**
Realización	**R = Pr + C [Pu + E]**
Gradual	**G = C[Pu + E]$_1$ ...C[Pu + E]$_n$**

Tabla 3. Formalización de las clases complejas.

A continuación presentamos detalladamente la caracterización y el modelo de representación de cada una de estas clases. El sistema de representación que utilizaremos está basado en Croft (2008). Según este autor, se hace necesario un sistema de representación bidimensional que sea capaz de definir las propiedades aspectuales en términos de las propiedades geométricas de la representación. En este modelo, los eventos se representan en dos dimensiones: el tiempo (T) y el cambio cualitativo (C). Los eventos puntuales son puntos en T, mientras que los eventos durativos se extienden en T. Los eventos estáticos son puntos en C, mientras que los eventos dinámicos se extienden en C (representando cambios de un estado cualitativo a otro).

Con este conjunto de distinciones básicas, las cuales se pueden representar geométricamente, junto con los conceptos de PERFIL y CONTORNO, podemos representar cognitivamente el conjunto de clases aspectuales que describimos a continuación. Un verbo en un contexto gramatical particular denota o PERFILA (Langacker, 1987) una (o más) de las distintas fases que componen el CONTORNO ASPECTUAL de un evento. Así, la representación está compuesta de dos partes: la primera se corresponde con la estructura del contorno como un todo, mientras que la segunda con la parte del contorno que se perfila. En este modelo de representación, tanto el PERFIL como el CONTORNO son parte del significado de la forma lingüística.

2.1 Estados

Los verbos estativos tradicionalmente se han descrito como verbos que expresan eventos homogéneos, estables y durativos, que no implican sucesión de fases temporales ni culminaciones intrínsecas. Se limitan a mantenerse durante el periodo temporal en el cual se dan (p.ej. *ser catalán* o *pertenecer a una asociación*).

En la Figura 3 podemos ver la representación bidimensional (C/T) de la estructura interna de un estado. En esta representación, el estado *la puerta está cerrada* se extiende en el tiempo (T). Las líneas discontinuas representan el estado previo *no estar cerrada* y el cambio hacia el estado actual. En este modelo cognitivo de representación, sólo las líneas continuas representan aquello que se perfila, en este caso un estado. El conjunto total de líneas (continuas y discontinuas) representa el contorno eventivo como un todo que también forma parte del significado lingüístico.

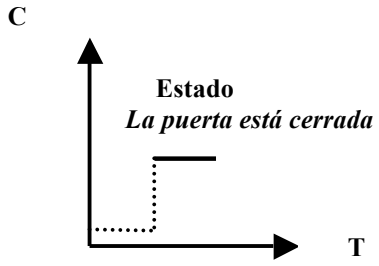


Fig. 3. Representación de un estado.

2.2 Procesos

Con respecto a la categoría de los procesos (o actividades), tradicionalmente se han descrito como verbos que describen situaciones constituidas exclusivamente por un proceso o desarrollo que se extiende en el tiempo, sin un límite temporal inherente. P.ej. *correr*, *caminar*.

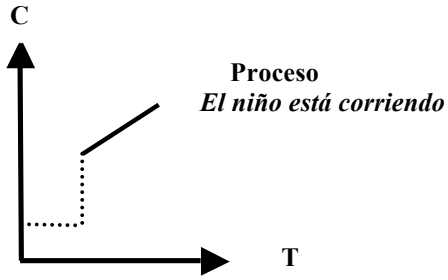


Fig. 4: Representación de un proceso

En la Figura 4, podemos observar como el proceso *el niño está corriendo* perfila el progreso en el tiempo (es durativo) y el progreso en el cambio (implica cambios cualitativos). El resto del contorno (líneas discontinuas) representa el estado previo *no está corriendo* y el cambio hacia el inicio del proceso *correr*.

En el caso de los procesos, se han distinguido subtipos de la categoría. Hay, Kennedy y Levin (1999: 132) distinguen entre actividades directas (del tipo *cool* <enfriar> o *age* <envejecer>), que expresan eventos no delimitados pero con un cambio directo en una escala, y las actividades indirectas (del tipo *dance* <bailar>), que expresan procesos que avanzan con cambios cíclicos. Esta

distinción también es asumida por Croft (2008). De todas maneras, son muchos los autores que consideran que los verbos del tipo 'enfriarse' y 'envejecer' forman una clase diferenciada de los procesos: 'gradient verbs' (Talmy 1985: 77), 'gradual completion verbs' (Bertinetto y Squartini, 1995). En este trabajo los llamamos graduales, una clase compleja, tal como expondremos detalladamente más adelante.

2.3 Puntos

Una de las clases eventivas no contempladas por Vendler (1957) es la de los *puntos* o *predicados semelfactivos* (Smith, 1991), que se definen como eventos dinámicos, instantáneos y sin consecuencia. Este tipo aspectual fue identificado por Carlson (1981: 39), que los llamó 'momentaneous'; Talmy (1985: 77) los describió como la clase 'full-cycle'; Moens y Steedman (1988: 95) se refieren a ellos como 'puntos', terminología que hemos adoptado en nuestro trabajo; Jackendoff (1991: 40) los llamó 'point events'; y finalmente Croft (1998, 2008) se refiere a este tipo aspectual como 'cyclic achievements'.

Smith (1991: 55) considera que los puntos son eventos instantáneos y, en consecuencia, no pueden aparecer con el modificador adverbial 'en X tiempo'. Sin embargo, Rothstein (2004, 2008a) considera que los puntos sí que pueden aparecer con este modificador (p.ej. *John jumped in three seconds*.[4] <John saltó en tres segundos>). De todas maneras, creemos, de acuerdo con Smith (1991), que aunque los puntos pueden consumir un cierto tiempo en su realización (conocimiento del mundo), son conceptualizados como instantáneos.

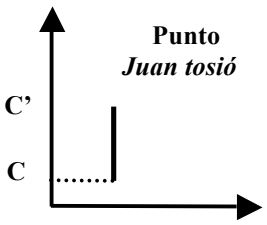


Fig. 5. Representación de un punto

[4] Ejemplo extraído de Rothstein (2008a).

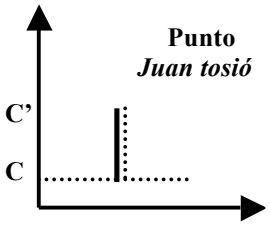


Figura 6. Representación de un punto

En la Figura 5 podemos ver como el verbo *toser* perfila un evento puntual que no tiene consecuencias, por lo tanto, no nos lleva hacia un estado resultado diferente. Si miramos ahora la Figura 6, donde se representa el contorno completo, se puede observar como después de *toser*, Juan vuelve a su estado normal de *no toser*. Esto es, el cambio de C hacia C' implica reversión hacia C después de llegar a ese punto. Las líneas discontinuas indican la reversión hacia el estado inicial de *no toser*.

2.4 Culminaciones

La clase de las culminaciones se corresponde en gran medida con los logros tradicionales, que se han definido como eventos dinámicos, télicos y puntuales (Vendler, 1957). En la bibliografía sobre aspecto, cuando la noción de puntualidad se aplica a los logros crea problemas, ya que algunos predicados que pertenecen a esta clase son compatibles con la expresión 'en X tiempo' (p.ej. *alcanzar la cima de una montaña*), lo que implica una duración que se refiere al proceso que tiene lugar antes de conseguirse el punto de culminación.

Otro problema que se ha planteado en la bibliografía a la hora de definir los logros es la aceptación de su propia existencia. Algunos autores consideran que la puntualidad es un rasgo que tiene que ver con la pragmática y no es pertinente lingüísticamente[5]. Ante todos estos problemas, se han propuesto dos soluciones: o bien reducir las clases de Vendler (1957), o bien aumentarlas.

[5] Es importante mencionar que estudios recientes en psicolingüística han demostrado que los hablantes atribuimos diferentes grados de duración a los eventos expresados por verbos. Además, estos diferentes grados de duración se corresponden con el tiempo de procesamiento mental de las unidades lingüísticas: los verbos que expresan eventos durativos tardan más tiempo en procesarse que los verbos que expresan eventos puntuales (Coll-Florit y Gennari, 2011).

Algunos autores como Verkuyl (1989, 1993), Mourelatos (1978), Pustejovsky (1991, 1995) o Marín (2000) consideran que la distinción entre realizaciones y logros no es pertinente lingüísticamente, por lo que reconocen una gran clase de verbos télicos que incluye tanto realizaciones como logros: los eventos. Otros autores como Bertinetto (1986), Smith (1991), Croft (1998, 2008) o Rothstein (2004, 2008b), no aceptan la vía reduccionista de unificar logros y realizaciones, sino que toman el camino contrario e identifican más de una clase de logros: *logros progresivos* o *runup achievements's* y *logros puntuales*. Los logros progresivos admiten un estadio preparatorio que se puede medir a través de un adverbial temporal como 'en X tiempo' (*morir*, *alcanzar la meta*, *desmayarse*, *caer dormido*, etc.), mientras que los logros puntuales (*caer*, *explotar*, etc.) son consistentemente menos durativos. Otra autora que no acepta la aproximación reduccionista es De Miguel (2004) que considera dos grandes grupos de logros: logros simples y logros complejos (seguidos de un estado o un proceso).

Queremos hacer notar que la aproximación reduccionista, al no tener en cuenta la diferencia entre realizaciones y logros, pierde la evidencia empírica que aporta la compatibilidad adverbial, en particular, la interpretación que aporta el adverbial temporal 'durante X tiempo'. La interpretación de este adverbial es muy diferente si se combina con un logro o con una realización. Con una realización (3), 'durante X tiempo' delimita una parte de la situación transformándola en un proceso. Por otro lado, los logros o bien no admiten esta construcción (4) o, si la admiten, focalizan la consecuencia de un logro, esto es, un estado (5) o un proceso (6).

3. Juan escribió una carta durante dos horas
4. *La policía atrapó al ladrón durante dos horas
5. Cerraron las instalaciones durante dos horas
6. El agua hirvió durante dos horas

Así, en nuestro trabajo adoptamos la aproximación no reduccionista y establecemos la distinción entre logros y realizaciones. Es más, hemos optado por utilizar el término 'culminación' de Moens y Steedman (1988) para referirnos a los logros. Esta decisión está motivada por dos razones. Por un lado, consideramos más pertinente el término 'culminación', en lugar de 'logro', ya que se utiliza como primitivo para formar clases más complejas. Por otro lado, algunos autores consideran la existencia de logros simples (De Miguel, 2004), mientras que en nuestro sistema de clasificación todas las culminaciones se consideran complejas, ya que implican un punto y una consecuencia[6]. Así, verbos como *llegar* o *superar*, indican el instante preciso en que una entidad pasa a estar en una nueva situación. De hecho, en algunos casos el contexto morfosintáctico permite focalizar únicamente el estado durativo resultante, tal como se observa en los siguientes ejemplos (7-8):

7. El agua llega hasta la ventana (estado actual del agua)
8. La temperatura supera los treinta grados (estado actual de la temperatura)

En definitiva, en nuestro sistema de clasificación eventiva identificamos una única clase genérica de culminaciones, concebidas como eventos complejos formados por un punto y una consecuencia (que generalmente es un estado (p.ej. *abrir*), aunque en casos residuales también puede ser un proceso (p.ej. *hervir*)).

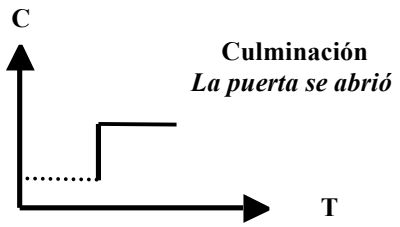


Fig. 7. Representación de una culminación

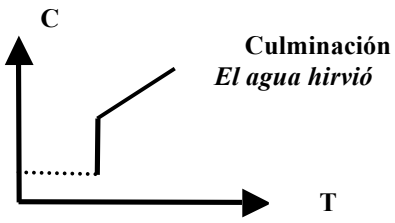


Fig. 8. Representación de una culminación

Tal como podemos ver en las Figuras (7-8), una culminación perfila la transición puntual hacia un estado o hacia un proceso. Ambos son ejemplos de un cambio desde un estado inicial a un estado resultado (en el caso de *abrir*) o hacia un proceso resultado (en el caso de *hervir*).

[6] Autores como Binnick (1991), Smith (1991) y Rothstein (2004) también consideran que las culminaciones siempre implican un resultado (afectación de una entidad).

2.5 Realizaciones

La clase de las realizaciones se caracteriza por expresar eventos complejos formados por un proceso y una culminación, esto es, un punto seguido de una consecuencia o estado resultante (p.ej. *construir un puente* o *beber un vaso de vino*). En palabras de Smith (1991: 26): "Accomplishments have successive stages in which the process advances to its natural final endpoint. They result in a new state. When a process with a natural final endpoint reaches its outcome, the event is completed and cannot continue".

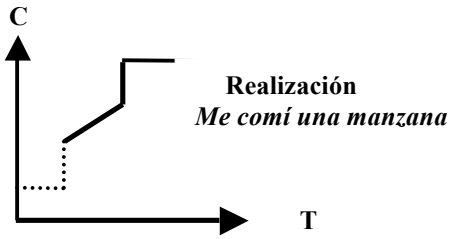


Fig 9.Representación de una realización

En la Figura 9 podemos ver como una realización perfila un proceso seguido de una culminación (líneas continuas).

Es importante apuntar que las realizaciones generalmente son transitivas. Es precisamente a partir de esta observación que en la bibliografía sobre aspecto se debate sobre el nivel lingüístico de las realizaciones: ¿son clases léxicas y/o oracionales? De acuerdo con Rothstein (2004), entendemos que 'construir una casa' es una realización léxica ya que expresa la culminación a través de un argumento. En cambio, 'correr hasta la estación', estaría formado por un proceso léxico, 'correr', que se reinterpreta como realización a nivel oracional a partir de un sintagma preposicional no argumental. Esta propuesta también es próxima a la de Tenny (1994), autora que considera que los procesos del tipo 'correr' lexicalizan un *path* (ruta) que puede tener un *terminus* (límite temporal), mientras que las realizaciones requieren un argumento interno que realiza la función de *measure* (medida), que englobaría la ruta y el límite temporal.

Así, en la línea apuntada por Rothstein (2008b), establecemos la distinción entre realizaciones léxicas y realizaciones no-léxicas. Algunos ejemplos prototípicos de realizaciones léxicas son *construir, leer, comer* o *beber*, que tienen un objeto directo medible, verbos denominales como *ensillar*, y verbos con un objeto transversal como *cruzar* o *atravesar*. Finalmente, entendemos que las realizaciones no-léxicas son culminaciones que adoptan un carácter procedural a partir de la perífrasis progresiva (*la puerta se está cerrando*) o bien procesos delimitados (*correr hasta la estación*).

2.6 Graduales

Hay dos clases de verbos que aparentemente no encajan en la clasificación cuatripartita de Vendler (1957), por un lado, tenemos los puntos (*golpear*, *saltar*, *disparar*), verbos que denotan eventos instantáneos que no tienen consecuencia y, por otro lado, tenemos los llamados eventos graduales (*ensanchar*, *envejecer*, *engordar*), verbos que parece pueden pertenecer a varias clases aspectuales. Los eventos graduales han sido tratados por autores como Dowty (1979), Abusch (1985, 1986), Bertinetto y Squartini (1995), Hay (1998), Hay, Kennedy, Levin (1999) y Rothstein (2008a).

Dowty (1979) ya vio que los graduales son eventos que pueden denotar cambios instantáneos, tal como se puede ver en (9). Además de la lectura de culminación, los graduales también pueden denotar eventos con extensión temporal en los cuales la interpretación es ambigua entre un proceso y una realización. (10) puede ser interpretado como que la sopa estaba cada vez más fría (proceso) o que la sopa finalmente pasó a estar fría (realización) (Abusch 1985, 1986).

9. La sopa se enfrió en un instante
10. La sopa se enfrió

Además, *enfriarse* cuando aparece con extensiones temporales, puede aparecer tanto con expresiones télicas (11) como atélicas (12).

11. La sopa se enfrío en media hora
12. La sopa se enfrío durante horas

Para analizar la semántica de esta clase aspectual es importante señalar que un gran subconjunto de estos verbos derivan de adjetivos. Abusch (1986: 4) considera que el significado de este tipo de verbos es ambiguo, por lo que establece dos reglas semánticas para derivar el significado del verbo del adjetivo. Así, el verbo *enfriarse* tiene un significado ambiguo entre *estar cada vez más frío* o *pasar a estar finalmente frío.*

Otros autores como Bertinetto y Squartini (1995) consideran que un verbo de este tipo es un híbrido entre un proceso y una realización. Si hacemos referencia a la ambigüedad que vio Abusch en el significado de estos verbos, Bertinetto y Squartini dicen que la única diferencia en significado depende del grado del cambio obtenido (el límite

final, o un estadio intermedio). Las dos alternativas dependen de consideraciones de tipo pragmático.

Llegados a este punto se hace necesario explicar la semántica léxica de los graduales para ver como estas interpretaciones se pueden derivar unas de otras. Nuestra propuesta se basa en la de Rothstein (2008a). Creemos que los graduales son eventos complejos formados por una iteración de culminaciones, con un cambio gradual. Así, los graduales denotan eventos que cambian, ahora bien, el cambio no está caracterizado como un cambio de α a χ, donde $\alpha \rightarrow \neg \chi$, sino como un cambio de valor en una escala. A este cambio lo llamamos cambio gradual. Así, según Rothstein (2008a: 188) el verbo *enfriarse* "denotes the set of events in which the temperature of x at the minimal final interval of e is lower than the temperature of x at the minimal initial interval of e."

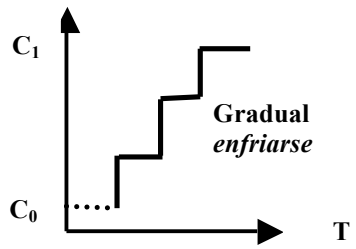


Fig. 10. Representación de *enfriarse* como un gradual

Si miramos la Figura 10 podemos ver como un gradual está formado por una iteración de culminaciones. En el eje del cambio (C), el cambio gradual supone una gradación de la propiedad que caracteriza el estado final. C[0,1] es un rango numérico que describe el grado en que se da un cambio de estado en un punto dado del tiempo (T). Así, C_0 y C_1 denotan el límite más bajo y el más alto, respectivamente, de un cambio gradual en un intervalo temporal dado. Con este análisis sólo asignamos un valor a enfriarse, el de 'estar cada vez más frío'.

Veamos ahora como a partir de esta estructura se pueden derivar, a través de procesos de coerción, el resto de interpretaciones que tradicionalmente se asigna a los graduales. *Enfriarse* también puede denotar un conjunto de cambios, donde el cambio es de una situación en la que a x se le asigna un valor d en la escala del frío, a una situación en la que a x se le asigna un valor más bajo que d. Este conjunto de cambios es conceptualizado como instantáneo (inherentemente no tiene extensión), así el ejemplo (13) es perfectamente compatible con esta interpretación.

13. Cuando eché el hielo, el líquido se enfrió al instante (aunque no mucho).[7]

En este caso, *enfriarse* tiene todas las propiedades de una *culminación*. Sin embargo, en otras ocasiones *enfriarse* puede aparecer con el progresivo o con modificadores atélicos. En estos casos, como vemos en los ejemplos (14-15), se implica la paradoja del imperfectivo. (14) implica que la sopa se ha enfriado (algo) y (15) implica que la sopa se enfrió durante un intervalo de tres horas y durante todos los subintervalos que componen esas tres horas.

14. La sopa se estaba enfriando
15. La sopa se enfrió durante tres horas

Con esta interpretación el verbo *enfriarse* denota un conjunto de eventos iterativos con extensión temporal: el conjunto de los eventos sucesivos de *enfriarse*. En estos casos, *enfriarse* denota un *proceso.*

Finalmente, *enfriarse* puede interpretarse como una *realización*: 'pasar a estar finalmente frío'. La telicidad de un evento puede estar determinada por modificadores temporales delimitadores (16), o por modificadores de grado (17).

16. La sopa se enfrió en media hora
17. La sopa se enfrió cinco grados

En resumen, los graduales son eventos complejos, formados por una iteración de culminaciones, con cambios graduales situados en una escala. Ahora bien, cuando los predicados graduales aparecen sin valores de delimitación definidos explícitamente y sin extensiones temporales son *culminaciones*, cuando aparecen con modificadores temporales atélicos son *procesos* y finalmente cuando aparecen con modificadores temporales delimitadores o con modificadores de grado son siempre *realizaciones*. Las diferentes interpretaciones aspectuales de los verbos graduales (culminación/ proceso / realización) se dan bajo los efectos de la coerción.

3 Representación computacional

Una vez presentada la caracterización de las clases eventivas, nos centraremos en su representación computacional. La implementación de las clases se ha realizado en Prolog, concretamente en SWI-Prolog (Wielemaker et al

[7]Ejemplo adaptado de Rothstein (2008a).

2012), una implementación de Prolog en código abierto.

El objetivo final de la implementación es la representación de los procesos de coerción que ejercen las perífrasis aspectuales de fase sobre los predicados verbales, para ello nuestro sistema se compone de diversos módulos: una gramática, un léxico (Freeling, Padró 2011), la especificación de las clases y las reglas de combinación aspectual. Presentamos aquí únicamente la primera parte de la implementación, la especificación de las clases eventivas a nivel léxico.

Como hemos explicado en el apartado anterior, para la representación de los eventos utilizamos dos dimensiones: el tiempo (T) y el cambio cualitativo (C) (Croft 2008). Estas dos dimensiones las representamos mediante conjuntos de pares que representan las coordenadas de los gráficos presentados en el apartado anterior, (t y c), con valores iniciales y finales que indican los cambios sufridos en el evento en cuanto al tiempo y al cambio cualitativo. Así, en el caso de un evento puntual en nuestra representación utilizamos la siguiente notación: t(0,0). En los casos en que el evento sea durativo lo expresaremos mediante el par t(0,1). Lo mismo se aplica para el cambio cualitativo, utilizamos c(0,0) para los casos en que no se da el cambio (por ejemplo: los estados) y c(0,1) para los casos en los que sí se da el cambio.

En la Tabla 4 se puede consultar la correspondencia entre el sistema de representación presentado en el apartado 2 y su implementación.

Representación gráfica	**Representación computacional**	
	Cambio cualitativo	**Progresión en el tiempo**
—	**c(n,n)**	**t(n, n+1)**
|	**c(n,n+1)**	**t(n,n)**
/	**c(n,n+1)**	**t(n,n+1)**

Tabla 4. Correspondencias entre la representación gráfica y la representación computacional.

En nuestro sistema, todo evento se representa mediante un PERFIL y un CONTORNO, y estos dos aspectos configuran la CLASE. Todos estos elementos son los atributos de cualquier 'evento':

- En primer lugar, la clase. Se trata de una etiqueta que indica qué interpretación recibe el evento en el nivel léxico. Sus posibles valores son: estado, proceso, punto, culminación, realización o gradual.

- En segundo lugar, un perfil que especifica la interpretación del evento a nivel léxico y que el proceso de coerción puede modificar o cambiar.

- En tercer lugar, el contorno que representa la estructura completa del evento, por lo que siempre incluye el perfil. Para su definición hemos utilizado siempre clases simples, ya que en las agrupaciones sintácticas esta forma de representación nos permite perfilar diferentes interpretaciones según el contexto.

Veamos ahora como se implementan cada una de las clases definidas.

La clase simple de los estados se define por su nombre, su perfil y su contorno. En cuanto al perfil, su valor es un estado que denota un evento sin cambio y durativo: c(1,1), t(0,1). En lo referente al contorno, éste se compone de tres clases simples: el estado anterior al evento -estado(c(0,0), t(-1,0))-, un punto -punto(c(0,1),t(0,0))- que daría paso a la última fase, y el estado que consideramos perfilado en el nivel léxico: estado(c(1,1),tiempo(0,1)). La combinación de esta información construye la clase 'estado' que se puede observar en la Figura 11.

```
evento (estado,
        perfil([estado(c(1,1),t(0,1))]),
        contorno([estado(c(0,0),t(-1,0)),
                  punto(c(0,1),t(0,0)),
                  estado(c(1,1),t(0,1))])).
```

Fig. 11. Representación de la clase simple estado.

De forma similar, la clase simple de los procesos se define por su nombre, su perfil -la propia estructura interna del proceso que incluye cambio y progresión en el tiempo: proceso(c(1,2),t(0,1))- y por último su contorno, básicamente el estado anterior y el punto que inicia el proceso. Podemos observar el resultado final de la representación de la clase proceso en la Figura 12.

```
evento(proceso,
        perfil([proceso(c(1,2),t(0,1))]),
        contorno([estado(c(0,0),t(-1,0)),
                  punto(c(0,1),t(0,0)),
                  proceso(c(1,2),t(0,1))])).
```

Fig. 12: Representación de la clase simple proceso.

Para finalizar la descripción de las clases simples, nos centraremos en los puntos. En este caso, el perfil de los puntos se representa mediante un cambio c(0,1) y la no progresión del tiempo t(0,0), mientras que el contorno añade al perfil el estado previo al evento, que aparece con rasgos negativos por ser anterior al evento, y la reversión hacia el estado inicial (representado por un punto y un estado). Podemos ver la estructura final de la clase punto en la Figura 13.

```
evento(punto,
       perfil([punto(c(0,1),t(0,0))]),
       contorno([estado(c(0,0),t(-1,0)),
                 punto(c(0,1),t(0,0)),
                 punto(c(1,0),t(0,0)),
                 estado(c(0,0),t(0,1))])).
```

Fig. 13. Representación de la clase simple punto.

Una vez formalizadas las clases simples, nos adentramos ahora en las clases complejas. Estas clases combinan en su perfil alguna de las clases simples. En primer lugar, tenemos la culminación (Figura 14). El perfil de una culminación se compone de un punto y un estado resultante, mientras que el contorno añade a éstos el estado inicial correspondiente al momento previo al evento, representado mediante el rasgo temporal con valores negativos, por ser anterior a la estructura perfilada.

```
evento(culminación,
       perfil([culminación([punto(c(0,1),t(0,0)),
                            estado(c(1,1),t(0,1))])]),
       contorno([estado(c(0,0),t(-1,0)),
                 punto(c(0,1),t(0,0)),
                 estado(c(1,1),t(0,1))])).
```

Fig. 14. Representación de la clase compleja culminación.

Otra de las clases complejas es la realización, representada mediante un perfil compuesto por un proceso y una culminación. De la misma forma que en las culminaciones, en el contorno se incluye el estado anterior y el punto que da lugar al proceso (Figura 15).

```
evento(realizacion,
       perfil([realización ([proceso(c(1,2),t(0,1)),
                             culminación( punto(c(2,3),t(1,1)),
                                          estado(c(3,3),t(1,2)))])]),
       contorno([estado(c(0,0),t(-1,0)),
                 punto(c(0,1),t(0,0)),
                 proceso(c(1,2),t(0,1)),
                 punto(c(2,3),t(1,1)),
                 estado(c(3,3),t(1,2))])).
```

Fig. 15. Representación de la clase realización.

Por último, la clase de los graduales estructura su perfil mediante una iteración de culminaciones, tal como se puede ver en la Figura 16, donde los valores de c y t se actualizan en cada una de las iteraciones[8]. Por otro lado, el contorno añade al perfil el estado previo a la realización del evento que aparece con valores negativos.

```
evento(gradual,
       perfil([gradual([culminación([
                                     punto(c(0,1),t(0,0),
                                     estado(c(1,1), (0,1))]),
                                    culminacion([
                                     punto(c(1,2),t(1,1),
                                     estado(c(2,2), (1,2))]),
                                    culminacion([
                                     punto(c(2,n),t(2,2)),
                                     estado(c(n,n), t(2,n))])])]),
       contorno([estado(c(0,0),t(-1,0)),
                 punto(c(0,1),t(0,0)),
                 estado(c(1,1), t(0,1)),
                 punto(c(1,2),t(1,1)),
                 estado(c(2,2), t(1,2)),
                 punto(c(2,n),t(2,2)),
                 estado(c(n,n),t(2,n))])).
```

Fig. 16. Representación de la clase gradual.

Cada una de estas clases se asocia a los predicados verbales correspondientes, de forma que nuestro léxico se enriquece con estas especificaciones.

[8] Hemos representado la iteración mediante tres culminaciones, asignando el valor *n* a la última iteración.

4 Conclusiones y trabajo futuro

En este artículo hemos presentado el establecimiento de 6 clases aspectuales: 3 simples (estados, procesos y puntos) y 3 complejas (culminaciones, realizaciones y graduales).

Estas clases se han definido siguiendo a Croft (2008), concretamente utilizando un sistema de representación bidimensional. Estas dos dimensiones son el tiempo (T) y el cambio cualitativo (C) que permiten representar tanto la progresión en el tiempo como la realización de un cambio cualitativo. Junto con estas dos dimensiones utilizamos los conceptos de perfil y contorno para representar cognitivamente las diferentes clases aspectuales. Por último hemos formalizado computacionalmente este análisis mediante la implementación en SWI-Prolog de estas dos dimensiones.

Actualmente, estamos trabajando en el fenómeno de la coerción, de forma que estas clases combinadas con una perífrasis aspectual produzcan una nueva interpretación a nivel sintáctico. Para ello hemos desarrollado una gramática lógica, que asociada al léxico de Freeling para el español, nos proporciona un análisis sintáctico para poder realizar el proceso de coerción.

Agradecimientos

Este trabajo se ha realizado gracias al proyecto: *Adquisición de escenarios de conocimiento a través de la lectura de textos: Lingüística y cognición (SKATER)* del Ministerio de Economía y Competitividad (TIN2012-38584-C06-06).

Referencias

Abusch, Dorit. 1985. *On Verbs and Times*, Tesis doctoral, Amherst, University of Massachussets.

Abusch, Dorit. 1986. Verbs of change, causation and time, Technical Report CSLI-86-50, *Center for the Study of Language and Information*, Standford University.

Agirre, Eneko, Philip Edmons (Eds.) 2007. *Word sense disambiguation algorithms and applications*. *Text, Speech and Language Technology*, vol., 33. Berlin, Heidelberg, New York: Springer-Verlag.

Alturo, Núria. 2001. Les activitats no són accions (situacions i tipus de text en anglès i en català), *Caplletra*, 30, 111-134.

Bach, Emmon. 1981. The Algebra of Events, *Linguistics and Philosophy* 9, 5-16.

Bennet, Winfield. S., Tanya Herlick, Katherine Hoyt, Jospeh Liro y Ana Santisteban. 1990. Toward a Computational Model of Aspect and Verb Semantics, *Machine Translation*, 4, 217-250.

Bertinetto, Pier Marco. 1986. *Tempo, Aspetto e Azione nel verbo italiano, Il sistema dell'indicativo*, Firenze, Accademia della CRusca.

Bertinetto. Pier Marco y Mario Squartini. 1995, An Attempt at Defining the Class of 'Gradual Conmpletion' Verbs. En *Temporal Reference Aspect and Actionality, 1: Semantic and Syntactic Perspectives*, eds. Pier Marco Bertinetto, Valentina Biachi, James Higginbotham y Mario Squartini, Torino, Rosenberg and Sellier, pp. 11-26.

Binnich, Robert I. 1991. *Time and the Verb. A guide to Tense and Aspect*, Oxford, Oxford University Press.

Carlson, Laurie. 1981. Aspect and quantification. En *Syntax and Semantics. Tense and Aspect*, Tedeschi, P.J y Zaenen, A. (eds.), pp. 31-64.

Coll-Florit, Marta. 2011. Aproximación empírica a los modos de acción del verbo: un estudio basado en corpus, *Revista Signos: Estudios de Lingüística*, 77: 233-250.

Coll-Florit, Marta. 2012. Sobre la naturaleza gradual de los modos de acción del verbo: prototipos y polisemia en el cálculo aspectual, *ELUA. Estudios de Lingüística*, 26: 145-162.

Coll-Florit, Marta y Silvia Gennari. 2011. Time in language: Event duration in language comprehension, *Cognitive Psychology*, 62, 41-79.

Croft, Willian. 1998. Event structure in argument linking. En *The projection of arguments: lexical and compositional factors*, Butt, M. y Geuder, W. (eds.), Standford, Centre for the Study of Language and Information, pp. 1-43.

Croft, William. 2008. Aspectual and causal structure in event representations. En V. Gathercole (Ed.), *Routes to language development. Studies in honor of Melissa Bowerman* (pp. 139-166). Mahwah, NJ: Erlbaum.

De Miguel, Elena. 1999. El aspecto léxico. En Bosque, I. y V. Demonte (Eds.): *Gramática descriptiva de la lengua española* (pp. 2977-3060). Madrid: Espasa Calpe.

De Miguel, Elena. 2004. Qué significan aspectualmente algunos verbos y qué pueden llegar a significar, *Estudios de Lingüística*, 18: 167-206.

Dowty, David. 1979. *Word meaning and Montague gramma*r. Dordrecht: Reidel.

Engelberg, S. 1999. The magic of the moment: What It Means to Be a Punctual Verb. En *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistic Society*, Chang, S., Liav, L. y Ruppenhofer, J. (eds), Berkeley, Berkeley Linguistic Society, pp. 109-121.

Grimshaw, Jane. 1990. *Argument structure.* Cambridge, The MIT Press.

Hay, L. 1998. The Non-Uniformity of Degree Achievements, ponencia presentada en el *72 Annual Meeting of the LSA*, New York.

Hay, Jenifer, Christopher Kennedy y Beth Levin. 1999. Scalar Structure Underlies Telicity en *Degree Achievements*, *SALT9*, 127-144.

Jackendoff, Ray. 1991. Parts and Boundaries, *Cognition*, 41, 9-45.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar, vol. I: theoretical prerequisites*. Standford: Standford University Press.

Levin, Beth, y Malka Rappaport Hovav. 1995. *Unaccusativity: At the lexical syntax-semantics interface.* Cambridge, MA: MIT Press.

Levin, Beth, y Malka Rappaport Hovav. 2005. *Argument Realization*, Cambridge, Cambridge University Press.

Levin, Beth, y Malka Rappaport Hovav. 2010. Lexicalized Scales and Verbs of Scalar Change, ponencia presentada en *46th Annual Meeting of the Chicago Linguistic Society*, Chicago, University of Chicago.

Marín, R. 2000. *El Componente Aspectual de la Predicación.* Tesis doctoral. Barcelona: Universitat Autònoma de Barcelona

Moens, Marc y Mark Steedman. 1988. Temporal ontology and Temporal reference, *Computational Linguistics*, 14, 15-28.

Mourelatos, A. 1978. Events, Processes and States. *Linguistics and Philosophy*, 2, 415-34.

Padró, Lluís. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, vol. 3, n. 2, pg. 13--20.

Pustejovsky, James (1991). The syntax of event structure, *Cognition*, 41: 47-81.

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

Rappaport Hovav, Malka, y Beth Levin. 1998. Building Verb Meaning. En *The Projection of Arguments: Lexical and Composicional Factors*, Butt, M. y Geuder, W. (eds.), Standford, Center for the Study of Language and Information Publications, pp. 96-134.

Rappaport Hovav, Malka, y Beth Levin. 2000. Classifying Single Argument Verbs. En *Lexical Specification and Insertion*, Coopmans, P., Everaert, M. y Grimshaw, J. (eds.), Amsterdam, John Benjamins, pp. 269-304.

Rothstein, Susan. 2004. *Structuring Events: A Study in the Semantics of Lexical Aspect*. Oxford: Blackwell.

Rothstein, Susan. 2008a. Two puzzles for a theory of lexical aspect: the case of semelfactives and degree adverbials. En *Event Structures in Linguistic Form and Interpretation*, Dölling, J., Heyde-Zybatow, T. y Shaefer, M. (eds.), Berlin, Mouton De Gruyter, pp. 175-198.

Rothstein, Susan. 2008b.Telicity and atomicity, *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*, Rothstein, S. (ed.), Amsterdam, John Benjamins, pp. 43-78.

Smith, Carlota. 1991. *The parameter of aspect.* Dordrecht: Kluwer.

Talmy, Leonard. 1985. Lexicalization patterns; semantic structure in lexical forms. *Language typology and syntactic description, vol, 3: grammatical categories and the lexicon*, ed. Timothy Shopen, 57-149. Cambridge: Cambridge University Press.

Tenny, Carol Lee. 1994. *Aspectual Roles and The Syntax-Semantics Interface*, Dordrecht: Kluwer.

Vendler, Zeno. 1957. Verbs and Times, *The Philosophical Review*, LXVI: 143-160.

Verkuyl, Henk J. 1989. Aspectual Classes and Aspectual Composition, *Linguistics and Philosophy*, 12: 9-64.

Verkuyl, Henk J. 1993. *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure.* Cambridge, Cambridge University Press.

Wielemaker, Jan, Tom Schrijvers, Markus Triska, Torbjörn Lager. 2012. SWI-Prolog. Cambridge Journals. CUP.

Novas Perspetivas

La subjetivización del *de que* en el español de Colombia

The subjectification of *de que* in Colombian Spanish

Matías Guzmán Naranjo
Westfälische Wilhelms-Universität Münster
m_guzm01@uni-muenster.de

Resumen

En este trabajo analizo el fenómeno conocido como dequeísmo en español, y en particular las diferencias semánticas entre las oraciones canónicas con *que* y las oraciones dequeístas. Reviso estudios de corpus previos sobre dequeísmo, e intento confirmar las predicciones de estos con un corpus independiente de español oral colombiano. Finalmente, con una regresión logística pruebo nuevos posibles parámetros que podrían influenciar la elección de una u otra construcción. El resultado es que las conclusiones y predicciones de la mayoría de análisis previos no son replicables con el corpus usado en este estudio, y que el dequeísmo parece estar influenciado por el contexto discursivo (situaciones reales o no reales), y la subjetividad expresada en la oración.

Palabras clave

Lingüística de corpus, dequeísmo, subjetivización

Abstract

This paper deals with the phenomenon known as *dequeísmo* in Spanish, in particular with the semantic differences between canonical sentences with *que* and dequeísta sentences. I analyze previous corpus studies of dequeísmo, test their predictions with an independent corpus of spoken Colombian Spanish, and finally carry out a logistic regression to test new possible parameters that might influence speaker's choice. The result is that most previous accounts of dequeísmo are not consistent with the corpus used for this study, and that dequeísmo seems to be influenced by the discourse context (real or non real situations), and speaker's subjectivity.

Keywords

Corpus linguistics, dequeísmo, subjectification

1 Introducción

El fenómeno conocido como dequeísmo en español consiste en la alternación entre las formas *que* y *de que* para introducir oraciones subordinadas, oraciones relativas, y en algunos casos para introducir oraciones principales como respuesta a una pregunta. El dequeísmo, a pesar de ser considerado como un error gramatical por la mayoría de prescriptivistas (Avila, 2003; Silva Villar and Gutiérrez Rexach, 2012), está extremadamente extendido por el mundo hispanohablante y puede encontrarse en la mayoría de dialectos de América y España, como lo evidencian la multitud de estudios realizados sobre el tema en diferentes regiones (Battini, 1949; Rabanales, 1974; Quilis Sanz, 1986; Prieto, 1995; Serrano, 1998; Schwenter, 1999; Del Moral, 2008, y referencias mencionadas en estos).

En estudios recientes algunos autores (Schwenter, 1999; Cornillie y Delbecque, 2008; Del Moral, 2008) han propuesto que el dequeísmo debe analizarse como un producto de un proceso de gramaticalización y subjetivización. Según estos estudios, el uso de *de que* no es producto de factores pragmáticos como cuando los hablantes intentan sonar importantes, o cuando intentan evitar errores (ver Bentivoglio, 1975; Suárez, 2009, para algunos ejemplos de esta perspectiva pragmática), sino que es un elemento con una función gramatical específica. De acuerdo con esta posición, *de que* no es una forma sinónima con *que*, y expresa diferencias semánticas o semántico-pragmáticas. No obstante, no ha sido posible hasta el momento establecer la naturaleza exacta de esta diferencia, y hay opiniones encontradas al respecto. García (1986) y Martínez-Sequeira (2000) ven el dequeísmo como un mecanismo que usa el hablante para crear distancia icónica entre manifestación gramatical del sujeto y lo dicho, y así distanciarse él mismo de la afirmación. Schwenter (1999) analiza el dequeísmo como un marcador evidencial que expresa información de segunda mano. Del Moral (2008) por su parte, toma una posición diferente y propone que *de que* no es un marcador evidencial ni una forma de distanciamiento, sino una forma de expresar información más subjetiva para el hablante, o que el hablante intenta mostrar que el sujeto de la oración tiene una posición subjetiva

frente a la información o la acción que se le atribuye. La hipótesis de que *de que* es un elemento gramaticalizado (de ahora en adelante H1) es supremamente interesante porque si es acertada, es posible explicar con ella por qué los hablantes son capaces de alternar entre las dos formas en un mismo texto y en una misma oración, y por qué *de que*, aun siendo más larga, es usada junto con *que*.

A pesar de que esta explicación para el desarrollo del dequeísmo funciona en la teoría, solo ha habido unos cuantos estudios empíricos (García, 1986; Martínez-Sequeira, 2000; Del Moral, 2008) para determinar si las oraciones con *de que* sí llevan alguno de las cargas semánticas mencionadas en el párrafo anterior, y entre estos estudios, Suárez (2009) no encontró resultados positivos que apoyaran H1. En el presente trabajo intentaré evaluar algunas de las diferentes propuestas de análisis del dequeísmo con un experimento en forma de encuesta y un análisis cuantitativo de un corpus oral para de español colombiano.

La sección 1. explica de forma más detallada los aspectos sintácticos del dequeísmo, y los tipos de construcciones en los que se encuentra. En la sección 2. reviso algunos de los estudios de corpus más relevantes realizados hasta la fecha sobre el dequeísmo. La sección 3. explica la metodología para ambos experimentos; y en la sección 4. presento los resultados obtenidos. En la sección 5. evalúo los resultados de ambos experimentos y las implicaciones de estos. Finalmente, la sección 6. presenta algunas consideraciones finales e implicaciones del presente estudio, así como algunas recomendaciones para futuros experimentos sobre el dequeísmo.

2 Aspectos sintácticos del dequeísmo

La primera fuente que hace uso de la palabra dequeísmo en Google Books es Estudios filológicos y lingüísticos publicado en 1974 (Rosenblat, 1974); y una de las primeras fuentes que menciona el fenómeno (sin darle el nombre de dequeísmo) es Banttini (1949). Otros trabajos pioneros sobre el dequeísmo fueron llevados a cabo por Bentivoglio (1975; 1980), quien realizó estudios sobre este fenómeno en Venezuela y Chile[1]. No obstante, las oraciones dequeístas son bastante antiguas en la historia del español, en parte evidenciado por su distribución geográfica, y en parte por documentos históricos que demuestran su uso desde el siglo XIV (para algunos ejemplos ver Del Moral, 2008). A pesar de esto, el dequeísmo suele ser asociado en varios países con dialectos de las clases sociales más bajas (Rabanales, 1974; Prieto, 1995; Serrano, 1998).

El dequeísmo no es un cambio en la forma en la que los verbos codifican los argumentos, de objeto directo a objeto preposicional, *de que* es una conjunción que sustituye a *que*. Los siguientes ejemplos (y todos los demás ejemplos presentados en este artículo) fueron tomados del corpus PRESEEA Medellín (Rátiva, 2007)[2]. El corpus PRESEEA es un corpus del español oral de Medellín, Colombia; cuenta con alrededor de 800000 palabras, y fue recogido con entrevistas semidirigidas. Los ejemplos en (1) muestran que los hablantes en la segunda ocurrencia de *ser* usan como una cópula en la forma canónica sin una oración subordinada, pero introduce la subordinada anterior con *de que*:

(1) la juventud no es **de que** tengo la piel lisa, la juventud es lo **que** uno respire.

Este primer ejemplo muestra que los hablantes no han cambiado el paradigma verbal, y que no necesariamente hacen uso de *de que*. Esto indica que *de que* no es analizado como una conjunción más una preposición, sino como un solo elemento.

Un ejemplo similar en (2) muestra aún mejor el hecho de que el dequeísmo es una elección entre dos construcciones diferentes, y que *de que* no ha remplazado realmente a *que*:

(2) de todas maneras ellos no compartían, **de que** después de cuatro años, y mi forma de ser mía y todo eso, no compartían que de pronto las cosas ya no estuvieran funcionado como, como debían funcionar.

Como podemos ver en (2), el hablante puede usar *que* y *de que* en la misma construcción con el mismo verbo (*compartir (de) que*...), en el mismo contexto y en el mismo hilo discursivo. Esta característica del dequeísmo es extremadamente difícil de explicar, y constituye uno de los mayores desafíos al elaborar una teoría sobre este fenómeno. En este caso no parece haber ninguna diferencia

1 Bentivoglio tomó un acercamiento un poco diferente al dequeísmo, ella propone que el uso de oraciones dequeístas se da porque los hablantes lo ven como un uso de las clases altas, no por razones funcionales. Si bien es posible que este factor juegue un papel en el uso del *de que*, este artículo no explora esta posibilidad.

2 Este corpus es de acceso libre y todos los datos y experimentos presentados en este artículo pueden ser replicados: http://comunicaciones.udea.edu.co/corpuslinguistico/

semántica entre ambas oraciones, la actitud del hablante no parece haber cambiado de la primera a la segunda oración, y aun así el hablante decide alternar entre ambas. Casos como estos hacen que una teoría puramente semántica del dequeísmo sea poco probable.

De la misma manera, un hablante dequeísta puede elegir no alternar en una serie de oraciones y mantenerse consistente con una de las dos formas:

(3) a. [el] problema del narcotráfico, el problema de las de los paras, el problema de la guerrilla, los enfrentamientos las pandillas, todas esas cosas, todo eso ha hecho **de que** realmente se acentúe más la división que hay en la ciudad [...] y eso ha hecho **de que** realmente pues se pierda como ese calor que uno tenía o sentía antes de una ciudad muy acogedora.
b. Para que la gente no diga pues **de que** es el beneficio **de que** es la familia.

Ambos ejemplos son interesantes. En (3a) el hablante usa en ambos casos *de que* para el mismo verbo pero en oraciones separadas, y en (3b) el hablante usa dos veces *de que* como cabezas de dos subordinadas de la misma oración principal.

En (4) vemos otro ejemplo de alternación entre *que* y *de que* por el mismo hablante, en este caso con verbos diferentes:

(4) yo haría algo pero por los niños, por los de escasos recursos, por ellos sí haría, porque considero **de que**, espacio para mucha gente hay, me parece **que** falta mucho espacio para los niños, de escasos recursos.

Otro contexto sintáctico en el que *de que* puede substituir a *que* es cuando actúa como un pronombre relativo. Este caso ha sido poco discutido en la literatura, y parece ser que solo Del Moral (2008) lo menciona. No obstante, es especialmente relevante cuando se considera que *de que* puede ser un caso de gramaticalización. En (5) a-d se presentan ejemplos del uso pronominal de *de que*:

(5) a. El metro [...] es **algo de** que va directo, para en la estación
b. me imaginaba de que sí yo seguía soltero iba a ser **una persona de que** iba a estar muy sola en mi casa.
c. Yo no cambio el Doce de Octubre por Bello, por Envigado, El Poblado, no, o sea, es **un barrio de que** ha venido de menos a más.
d. por decir yo tengo **un amigo de que** tiene un niño.

Del Moral (2008) encontró construcciones similares en su corpus. Estos casos sugieren que *de que* está fuertemente asociado a *que* en el sistema de los hablantes dequeístas, y estos han extendido su uso a otras construcciones en las que *que* se usaría. Si este análisis es correcto, estos ejemplos representan fuerte evidencia en favor de H1.

Algunas construcciones con *que* aún no pueden sustituirse por *de que* (al menos en el corpus no se encontraron ejemplos de estos tipos):

Oraciones comparativas:

(6) a. más grande que él.
b. * más grande de que él.

como un pronombre relativo después de un artículo:

(7) a. el que, la que, lo que.
b. * el/la/lo de que.

como un pronombre interrogativo:

(8) a. ¿qué es eso?.
b. ¿de qué es eso?

En este caso, aunque el ejemplo (8b) es gramatical, el *de que* no es interpretado como un solo elemento, sino como dos, y con el significado de la preposición *de* aún presente.

Con estos ejemplos podemos ver que el dequeísmo es un fenómeno supremamente complejo, y que no es claro a simple vista por qué los hablantes a veces escogen *que* y a veces *de que*. Si en realidad hay una diferencia semántica entre ambas no es posible encontrarla mirando tan solo algunos ejemplos, es necesario estudiar muchos casos en conjunto.

3 Aspectos semántico-pragmáticos sobre el dequeísmo

En esta sección presentaré algunos estudios que han tratado de identificar las propiedades semántico-pragmáticas del dequeísmo, y los dividiré en tres hipótesis que intentan explicar el fenómeno.

García (1986) estudió el dequeísmo en Santiago, Caracas y Buenos Aires. Ella propuso que *de que* crea distancia entre el hablante y la oración

(Hipótesis de Distancia e Incertidumbre). Esta hipótesis también significa que cuando un hablante usa *de que*, éste se encuentra menos seguro de la afirmación. Para verificar su hipótesis, García realizó un estudio de corpus que encontró que *de que* se usa con mayor frecuencia con verbos en tercera persona y en formas impersonales que en primera persona, y que se usa más con conjugaciones no presentes. Estos datos son consistentes con su teoría porque indican que *de que* se usa cuando el hablante no está demasiado involucrado con la afirmación (cuando la oración no es en primera persona), o cuando ésta se encuentra más distante y es más incierta (en pasado o futuro).

De forma similar, Schwenter (1999) propuso que *de que* es un marcador evidencial (Hipótesis de Evidencialidad). Los datos del estudio de Schwenter también mostraron preferencia por la tercera persona con las oraciones dequeístas, lo cuál es interpretado como confirmación de la Hipótesis de evidencialidad.

Kanwit (2012), en un estudio aún no publicado, intentó corroborar las afirmaciones de Schwenter con el mismo corpus usado por aquel. Además de considerar las variables Tiempo y Persona, Kanwit también consideró el contexto discursivo. Para evaluar el contexto Kanwit propone tres niveles: cargado emocionalmente (temas posiblemente ofensivos), importante (temas de importancia para el hablante), y neutral. En su estudio encontró que *de que* se usa con más frecuencia en contextos cargados emocionalmente que en los demás contextos. Kanwit interpreta estos resultados como evidencia de que *de que* funciona como un marcador evidencial al hacer que la afirmación sea más débil, pero también podrían entenderse desde la Hipótesis de Distancia e Incertidumbre.

Del Moral (2008) realizó un estudio con un corpus histórico en el que analizó cómo *de que* ha evolucionado en los últimos siete siglos. Del Moral propone que el dequeísmo es un producto de subjetivización (Hipótesis de Subjetivización), y para verificar esta hipótesis comparó el número de oraciones dequeístas en primera persona y no primera persona en cada siglo. Del Moral encontró que la proporción de oraciones dequeístas en primera persona no solo es mayor que en no primera persona, sino que la proporción de éstas ha ido incrementando constantemente desde el siglo XIII. Del Moral también encontró que los verbos sicológicos eran más frecuentes con oraciones dequeístas, y concluye que esto es evidencia a favor de su teoría. La propuesta de Del Moral se enmarca en la teoría de gramaticalización (Bybee, 2003; Bybee, Perkins, Pagliuca, 1994; Company C., 2004; Haspelmath, 2002, 2004 and 2006). Dentro de esta teoría, uno o varios elementos léxicos incrementan su frecuencia, sufren erosión fonética y semántica, y empiezan a actuar como elementos gramaticales. En este caso *de* ha perdido su componente semántico y forma una sola palabra fonológica en *de que*, y actúa como una conjunción independiente.

A pesar de ser una propuesta con una sólida base teórica, el estudio de Del Moral tiene algunos problemas. El más evidente es que el corpus usado, Corpus del Español (Davis, 2002), no es consistente en el tipo de textos que recoge entre siglos. Los documentos más antiguos no son comparables ni en tipo ni en extensión a los más modernos, por lo que cualquier generalización resulta algo problemática. En segundo lugar, Del Moral no explica qué contó exactamente, ni qué algoritmo (o búsqueda) usó para extraer las oraciones dequeístas. Tampoco explica cómo distinguió entre el *de que* canónico y el dequeísta.

Es importante dejar en claro que la Hipótesis de Subjetivización es totalmente opuesta a la Hipótesis de Distancia e Incertidumbre. En la primera *de que* indica que el hablante está mucho más involucrado con la afirmación que en las oraciones con *que*, mientras que la segunda se marca lo contrario, que el hablante se quiere distanciar de lo dicho.

Finalmente, un estudio por Suárez (2009) afirma que no hay evidencia positiva para ninguna de estas tres hipótesis (Suárez las llama Hipótesis Funcionalistas, p. 164), es decir, H1. Suárez no encontró evidencia de que haya diferencias significativas en el uso de presente-no presente, ni entre primera persona-no primera persona en las oraciones dequeístas. Los resultados de su estudio se presentan a continuación:

Primera persona			No primera persona		
N	T	%	N	T	%
20	934	2,10%	13	875	1,50%
Presente			No presente		
N	T	%	N	T	%
30	1584	1.9%	3	221	1,40%

Tabla I: Datos para Persona y Tiempo en oraciones dequeístas (Suárez, 2009, p. 164)

Si se comparan de esta forma, Suárez tendría razón y las diferencias entre las proporciones no serían significativas ($p<0.001$). No obstante, no es claro cuáles son los totales reales en este caso.

Suárez no explica si contó el número total de oraciones, o el número total de oraciones con *de que* producidas por todos los hablantes, o solo las producidas por los hablantes dequeístas. La forma apropiada habría sido solo contar las oraciones de los hablantes dequeístas, pero sin los datos del estudio no es posible verificar cuál es el caso.

En resumen hay tres hipótesis sobre la función y significado de *de que*: La Hipótesis de Distancia e Incertidumbre, la Hipótesis de Evidencialidad y la Hipótesis de Subjetivización. Estudios diferentes con corpus diferentes han producido resultados que apoyan o una u otra hipótesis, y en general son inconsistentes los unos con los otros. Además, tanto la Hipótesis de Distancia e Incertidumbre como la Hipótesis de Evidencialidad hacen predicciones iguales sobre la distribución de los datos, lo que dificulta mucho poder distinguirlas empíricamente. En lo que sigue propongo una forma tanto diferente para evaluar la validez de estas tres hipótesis.

4 Metodología

En esta sección presento los métodos usados para verificar las tres hipótesis descritas en las sección anterior, y doy una descripción simple de las técnicas estadísticas empleadas.

4.1 Encuesta

Para probar la Hipótesis de Evidencialidad se realizó una encuesta con hablantes de español de Colombia. En la encuesta a los participantes se les presentaron 20 oraciones en las que una supuesta persona estaba diciendo algo sobre alguien más, y ellos debían marcar de 1 a 5 (1 muy inseguro, y 5 muy seguro) qué tan seguro estaba la supuesta persona de información expresada en la oración. La mitad de las oraciones eran distractores, de las 10 oraciones restantes, 5 eran oraciones canónicas con *que* y 5 eran oraciones con *de que*; todos los verbos en todas las oraciones eran diferentes. Para evitar cualquier efecto producido por la elección particular de las oraciones, la muestra de hablantes se dividió en dos y a la mitad se le asignó la encuesta A, y a la otra mitad la encuesta B. Las oraciones dequeístas en A aparecían como canónicas en B, y las oraciones canónicas en A aparecían como dequeístas en B; los distractores no se alteraron entre A y B. Además se recogieron los datos Edad, Sexo, y Estrato (o Clase Social). La mayoría de los participantes que respondió la encuesta resultó estar entre los rangos de edad 20-35 años, por lo que esta variable no se consideró en el análisis estadístico.

Después de que los informantes concluyeron la encuesta, se les preguntó de forma informal y sin que tuvieran que anotar la respuesta, si ellos consideraban que las oraciones dequeístas expresaban duda sobre lo dicho. Se escogió hacer esta parte de forma informal porque en las pruebas los hablantes comunicaron muy poco cuando se les pidió que escribieran su respuesta, pero intentaron dar explicaciones más detalladas cuando se les pidió que simplemente dieran su opinión de forma oral. Esta decisión hace que los resultados de la pregunta complementaria no puedan ser analizados estadísticamente, y solo presentaré las impresiones generales de los hablantes.

Para analizar los datos comparé si en general, las oraciones dequeístas habían sido marcadas como más inciertas (tendiendo a 1) que las oraciones canónicas. También comprobé si había diferencia entre las clases sociales, y entre hombres y mujeres. La estadística escogida fue el test Chi cuadrado para proporciones.

4.2 Estudio de corpus

Para probar la Hipótesis de Distancia e Incertidumbre y la Hipótesis de Subjetivización se condujo un estudio de corpus con el corpus PRESEEA. Para procesar el corpus se utilizó la libraría FreeLing con su API de Java (Padró y Stanilovsky, 2012).

Se extrajeron todas las oraciones dequeístas (56 en total), y se etiquetaron manualmente para las variables Sexo (del hablante), Tiempo (presente-no presente), Modo (indicativo, no indicativo), Número (plural, singular), Polaridad (si la oración era positiva o negativa), Pronominal (uso de pronombre personal o pro-drop), Número de Palabras (número de palabras que aparecen después de *de que* en la oración), Contexto (condicional o real), Tipo de Verbo (sicológico, percepción, decir, u otros marcados como 'do'), Persona (primera persona, no primera persona). Para la variable Contexto, las oraciones que expresaban hechos reales o de los que el hablante parecía estar seguro se marcaron como 'real', mientras que las oraciones que hacía relación a situaciones o eventos posibles, de los que el hablante no estaba seguro se marcaron como 'condicional'. La razón para esta variable es que las afirmaciones sobre eventos condicionales son más subjetivas que las oraciones sobre situaciones reales porque requieren que el hablante tome posición frente a ellas. La motivación para la variable PRONOMINAL es similar, asumo que si se hace uso explícito del pronombre, el hablante está haciendo énfasis en la persona que está haciendo la

afirmación y es por tanto más subjetiva[3]. Adicionalmente, 57 oraciones con uso de que en forma canónica se extrajeron aleatoriamente del corpus de los textos de los hablantes dequeístas (los hablantes no dequeístas se ignoraron para esta parte del estudio), y se etiquetaron bajo los mismo criterios. Esto permite tener un punto de comparación sólido, algo que la mayoría de estudios mencionados en la sección anterior no presentan.

Primero que todo, se evalúan las propuestas de que ciertas categorías gramaticales como tiempo, modo, o persona ocurren con mayor o menor frecuencia en las oraciones dequeístas. Para esto se emplea el test Chi-cuadrado para proporciones.

Para evaluar la Hipótesis de Subjetivización se construyó un modelo con el método de regresión logística para variables binarias (ver Sheather (2009) para la explicación matemática de esta técnica). En una regresión logística la variable dependiente es categorial y binaria, a un valor se le asigna 1 y al otro valor se le asigna 0. El modelo intenta predecir a partir de las variables independientes la probabilidad de que la variable dependiente sea 1 o 0. Con este método es posible evaluar el efecto total de cada variable, y asignar un valor de qué tan bien una variable puede predecir la variable dependiente. Los resultados se resumen en una tabla de coeficientes. Finalmente, para ver qué tan bien logra predecir el Modelo datos nuevos se realiza una validación bootstrap que selecciona aleatoriamente un subconjunto de los datos, crea el modelo con el resto de los datos, y luego evalúa el nuevo modelo contra los datos no considerados; este proceso se repite 200 veces. Los resultados se presentan en una tabla de coeficientes y en una gráfica.

El propósito de este análisis es determinar qué predictores (variables independientes) muestran una fuerte correlación con la variable dependiente, en este caso si la oración es dequeísta o no. La ventaja de este método sobre otros es que permite analizar varias variables al mismo tiempo y las posible interacciones entre estas.

5 Resultados

5.1 Resultados de la encuesta

Un total de 118 hablantes respondió la encuesta, 60 la encuesta A y 58 la encuesta B. La mayoría de los hablantes, en especial los que se encontraban entre las clases sociales 3, 4 y 5, afirmaron que todas las oraciones eran más o menos iguales, pero que las oraciones dequeístas eran un poco raras. Para la mayoría, la elección de un nivel de certeza dependió del verbo usado y no de si la oración era o no dequeísta. Cuando se les preguntó después de que habían contestado la encuesta, si las oraciones con *de que* expresaban duda, todos contestaron negativamente. Como se mencionó en la sección anterior, esta respuesta no se anotó, por lo que no es de relevancia estadística, pero sí hace dudar que la Hipótesis de Evidencialidad sea acertada.

Una primera mirada a los resultados parece indicar que la encuesta no apoya la hipótesis propuesta, y que no hay diferencia en seguridad entre las oraciones canónicas y las dequeístas. Si los datos se analizan como un todo, tenemos dos distribuciones idénticas para *que* y *de que*:

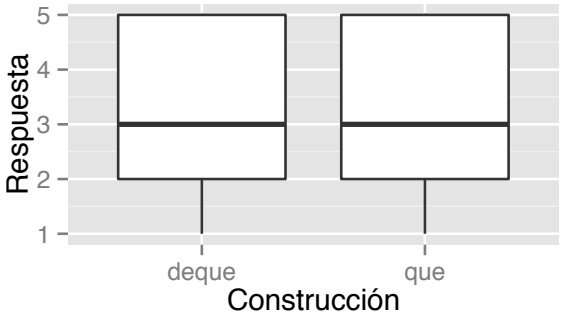


Figura 1: Diagrama de cajas de la distribución de respuestas para ambos cuestionarios.

Como podemos observar en la Figura 1., no parece haber diferencia en la distribución de respuestas para las oraciones canónicas y dequeístas. Ambas distribuciones tienen una media[4] de 3, no hay valores extremos, y ambas distribuciones tienden hacia el límite superior, lo cual quiere decir que la mayoría de las respuestas estuvieron más cerca de 5 que de 1. Un histograma representando las distribuciones por respuesta se presenta en la Figura 2:

3 Hay que notar que autores como Company C. (2004) ven esta variable al revés, y afirman que las oraciones más subjetivas no tienen un sujeto agentivo. No habiendo evidencia empírica de que este es el caso, rechazo la propuesta de Company C.

4 Como los datos son ordinales, el promedio no puede usarse como una medida de tendencia central. La diferencia entre 1 y 2 no es necesariamente la misma que la diferencia entre 3 y 4.

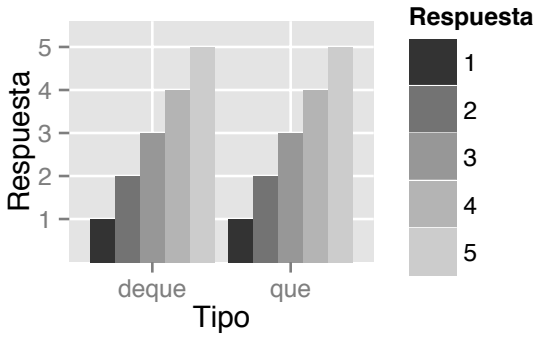


Figura 2: Histograma de la distribución de respuestas para ambos cuestionarios.

De la Figura 2. podemos concluir que no parece haber diferencias entre ambas construcciones. Si los datos se dividen según los diferentes factores considerados en el estudio las proporciones cambian ligeramente. En la Figura 3. podemos ver la frecuencia de que cada respuesta por construcción según el estrato de los hablantes.

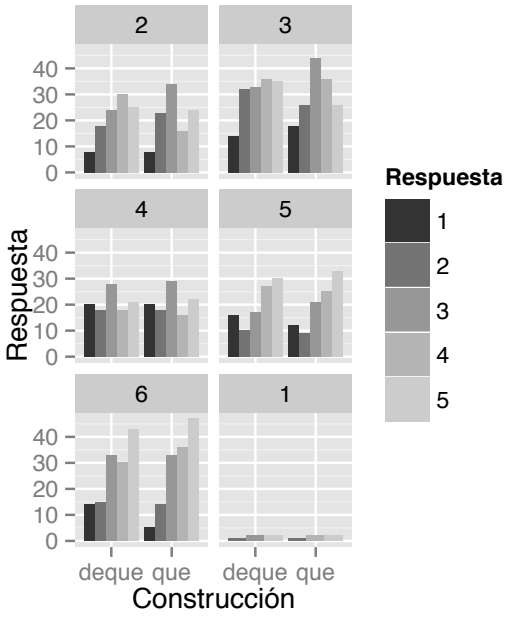


Figura 3: Histograma de la distribución de respuestas para ambos cuestionarios dividido por estrato.

La Figura 3. muestra que para las clases sociales (estratos) 4, 5 y 6 las distribuciones son casi idénticas, pero en los estratos 2 y 3 sí se observa algo de variación entre las oraciones con *que* y *de que*. Estos son los dos estratos en los que se podría esperar mayor variación por lo comentado anteriormente, que el dequeísmo es asociado principalmente con las clases sociales más bajas. No obstante, si se comparan las distribuciones para cada construcción por estrato con el test Chi-cuadrado de proporciones, obtenemos que no hay diferencias estadísticamente relevantes entre éstas: para el estrato 2: X=20, df=16, p=0.2202; para el estrato 3: X=15, df=12,p=0.2414.

Una gráfica similar en la Figura 4. muestra las proporciones para hombres y mujeres.

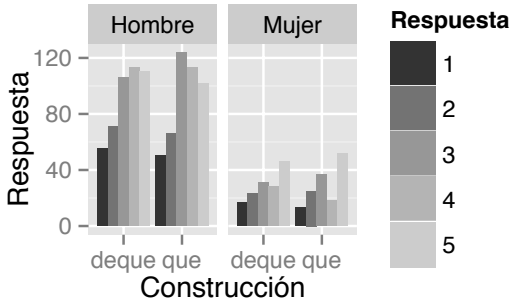


Figura 4: Histograma de la distribución de respuestas para ambos cuestionarios dividido por sexo.

Ambas distribuciones son bastante similares, ambas tienen una media de 3, y el test Chi-cuadrado de proporciones no revela diferencias estadísticamente relevantes (Para mujeres: X=20, df=16, p=0.2202; para hombres: X=20, df=16, p=0.2202).

Las distribuciones para cada clase social en ambos cuestionarios se presentan en la Figura 5.

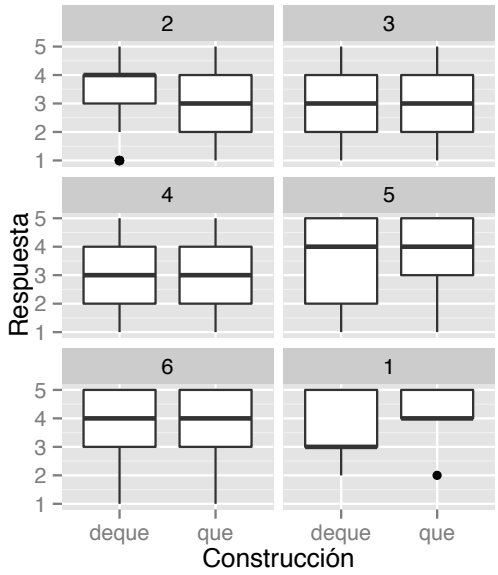


Figura 5: Diagramas de cajas de la distribución de respuestas para ambos cuestionarios dividido por estrato.

De las Figura 5. podemos concluir que la variación entre ambas construcciones es mínima y no parece estar presente en ninguna clase social.

5.2 Resultados del estudio de corpus

Para el análisis de corpus las proporciones de las 10 variables etiquetadas se presentan en la Figura 6.

En las Figura 6. podemos ver que la mayoría de las variables muestran proporciones casi idénticas para ambas construcciones. Las únicas variables que parecen ser diferentes son Contexto (condicional o real), Tipo de Verbo (percepción, sicológico, decir, otros) y Pronominal (uso de pronombre personal o pro-drop). Para intentar verificar las propuestas presentadas en la sección 3. se realizó un test Chi-cuadrado de proporciones para cada variable presentada como significativa en las Hipótesis de Evidencialidad e Hipótesis de Distancia e Incertidumbre. Para Modo los resultados son que la diferencia en proporciones no es estadísticamente significativa (X=1.472, df=1, p=0.225); para Tiempo los resultados son que la diferencia en proporciones no es estadísticamente significativa (X=3.0473, df=1, p=0.08105); para Número los resultados son que la diferencia en proporciones no es estadísticamente significativa (X=0.0012, df=1, p=0.9729). Ninguna de las predicciones de las Hipótesis de Evidencialidad e Hipótesis de Distancia e Incertidumbre se pudo confirmar con el corpus analizado.

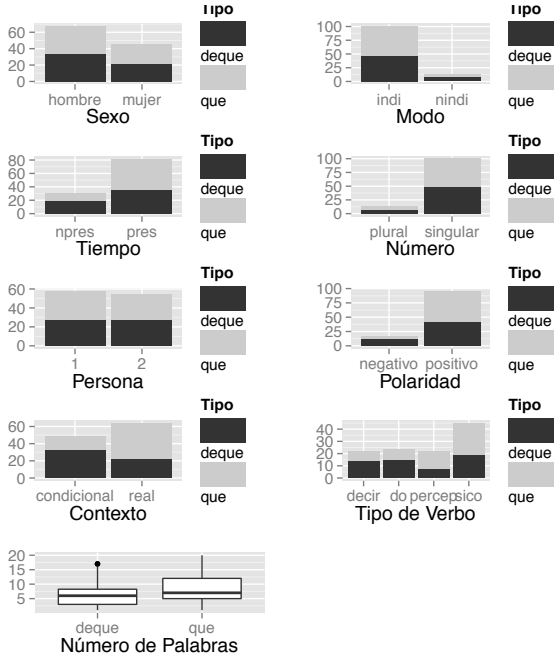


Figura 6: Distribución de las categorías gramaticales analizadas para el corpus PRESEEA.

A continuación se presentan los resultados del análisis logístico sobre los datos recolectados, así como la validación bootstrap del modelo. Las implicaciones de estos resultados serán discutidas en la siguiente sección.

La Tabla II presenta los índices del modelo adaptado a los datos. Para el modelo solo las variables significantes ($p<0.05$) se mantuvieron, estas fueron: Contexto * Pronominal, y en cierta medida Tipo de Verbo.

La validación bootstrap para el modelo se presenta en la Tabla III. La validación retuvo 3 factores en 178 experimentos, 2 en 18, y solamente 1 en 4 experimentos.

En las Tablas II y III podemos ver que las variables Contexto, Ponominal y Tipo de Verbo todas fueron clasificadas como predictores significativos de la variable dependiente por el modelo de regresión logística. Ninguna de las otras variables tuvo relevancia como predictor de la variable dependiente.

	Modelo			
	Coeficient	S.E.	Wald Z	Pr(Z)
Intercept	-2.3921	0.7982	-3.00	0.0027*
Contexto=real	3.04113	0.7342	4.14	0.0001***
Pronominal=sí	-0.1923	0.6746	-0.29	0.7756
TipoVerbo=percepción	322.876	0.7813	1.64	0.1018
TipoVerbo=sicológico	706.777	0.7586	3.14	0.0017**
TipoVerbo=decir	0.5152	0.7896	0.65	0.6233
Contexto * Pronominal	-2.4685	0.9736	-2.54	0.0112*

Probabilidad del modelo e índices de discriminación	
C	0.813
Pseudo R2	0.384
L.R.	38.40
D.f.	6
Pr (chi2)	0.0001***
Num. obs.	0.113
Dyx	0.626
Brier	0.175

*** p < 0.001, ** p < 0.01, * p < 0.05, ·p < 0.1

Tabla II: Coeficientes de la regresión logística.

Factores	origen	training	test	optimism	corrected
Dxy	0.6259	0.6625	0.5890	0.0735	0.5524
R2	0.3842	0.4341	0.3382	0.0959	0.2883
Intercept	0.0000	0.0000	-0.0012	0.0012	-0.0012
Slope	1.0000	1.0000	0.7803	0.2197	0.7803
Brier	0.1746	0.1672	0.1896	-0.0224	0.1970

Tabla III: Coeficientes de la validación bootstrap para la regresión logística.

La estadística D de Somers (Dyx en este caso) mide el efecto de las variables independientes sobre

la variable dependiente. Dyx puede tomar valores de -1 a 1, -1 indica un perfecta correlación negativa, mientras que 1 indica una perfecta correlación positiva. Un valor para Dyx de 0.626 es índice de un efecto moderado.

El índice de discriminación general C (Overal Discrimination Index) es una concordancia entre la probabilidad predicha y la respuesta real. Un valor C de 0.5 marca una predicción totalmente aleatoria, y un valor de 1 una capacidad de predicción perfecta. Un C de 0.813 indica una buena capacidad predictiva (0.8 indica una capacidad predictiva muy alta).

El Brier Score es un índice de qué tan exacta es una predicción probabilística. Puede tomar valores de 0 (el mejor) hasta 1 (el peor). En este caso 0.175 es indicativo de que el modelo tiene una capacidad predictiva alta (para explicaciones más detalladas de estos índices ver: Newson, 2006; Baayen, 2008; Gries, 2009).

En la Tabla III vemos que el optimismo para Dyx y Brier es relativamente alto (por encima de 0.02), y que los valores corregidos son un poco peores que en el modelo original. Esto mide qué tan optimista es el modelo, entre menos optimismo, mejor puede predecir el modelo datos nuevos. Estos valores parecieran indicar que el modelo no es muy capaz de predecir nuevos datos; pero, por otro lado la validación retuvo las tres variables independientes la gran mayoría de las veces, y consideró Contexto como significativa en 199 experimentos. En general podemos tener confianza en el modelo. Además, el modelo logró predecir correctamente la construcción en un 75% de las veces. Esto significa que 25% o más de la variación está dada por estas tres variables (es posible predecir el 50% de forma aleatoria). Un 25% de la variación se debe a otros factores no capturados en el modelo.

Las Figuras 7. y 8. a continuación muestra los efectos parciales de cada variable independiente sobre la variable dependiente. En la Figura 8. podemos ver que la interacción entre Contexto y Pronominal es significativa, y que el Contexto tiene más impacto sobre la probabilidad de la construcción si la oración no es pronominal.

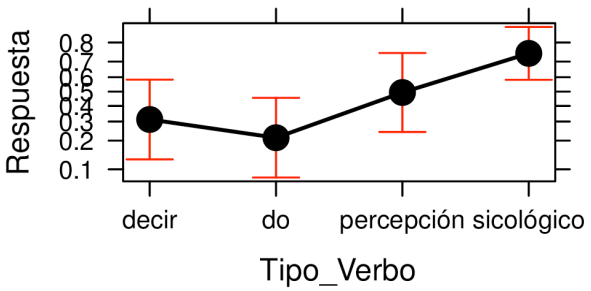


Figura 7: Efectos parciales de la variable Tipo de Verbo.

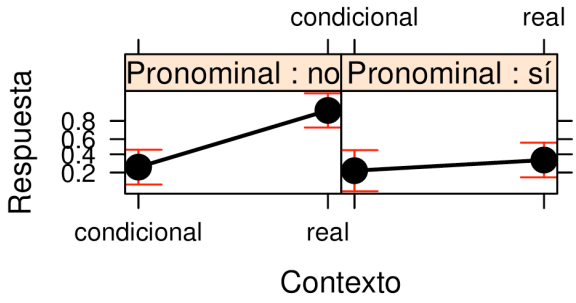


*Figura 8: Efectos parciales de la interacción de variables Contexto * Pronominal.*

6 Discusión

La encuesta realizada, tal como se describe en 5., no reveló ningún efecto estadísticamente positivo. La construcción *de que* no parece influenciar la forma en la que los hablantes ven la fuente de la información, o la seguridad que tiene el hablante sobre lo dicho. Por esta razón parece improbable que el dequeísmo sea una forma de evidencialidad.

Los resultados del estudio de corpus son más reveladores. Aunque no se pudieron confirmar las predicciones iniciales propuestas por los autores de las Hipótesis de Distancia e Incertidumbre e Hipótesis de Subjetivización, el modelo demostró que hay una fuerte correlación entre la alternación entre *que* y *de que* y el contexto de la oración así como el uso de pronombres personales. Ambos factores apoyan la Hipótesis de Subjetivización. La semántica verbal parece jugar un papel en la elección de la construcción pero no es claro cómo puede ser interpretado, o si favorece alguna de las hipótesis, especialmente porque Del Moral (2008) predice que los verbos sicológicos deberían aparecer con mayor frecuencia con oraciones dequeístas.

Situaciones irreales, preguntas hipotéticas y opiniones personales requieren un alto nivel de subjetividad por parte del hablante, y el sujeto de la oración debe estar más involucrado con la afirmación que si se reportan situaciones concretas y reales. No obstante, también se puede manifestar gran nivel de subjetividad por un evento real si el hablante así lo desea, es por esta razón que no existen pares de oraciones en las que *de que* no sea posible en lugar de *que* y viceversa. Esto también explica por qué los hablantes, en medio de una oración, pueden cambiar de *que* a *de que* o

viceversa con el propósito de aumentar o disminuir el nivel de subjetividad que quieren expresar.

Finalmente, el hecho de que el modelo no pudiera predecir más del 75% de las oraciones indica que aparte de los factores indicados en estos, hay otros motivos que juegan un papel importante en la alternación.

7 Consideraciones finales

Después de revisar los estudios más significativos sobre el dequeísmo, los resultados presentados en este artículo apoyan la hipótesis presentada por Del Moral (2008) de que el dequeísmo es un caso de gramaticalización de subjetividad, aunque desafían algunas de sus predicciones y su interpretación del fenómeno. Como Gutiérrez Rexach (2012), no se encontraron correlaciones sugeridas en estudios previos entre Tiempo, Modo o Persona y el uso de *de que*. Una razón para dudar de estos estudios es la falta de claridad metodológica y el hecho de que no se reportan datos totales. No obstante, hay que reconocer que el corpus usado para este trabajo es relativamente pequeño, y un corpus más grande produciría resultados más confiables.

Referencias

Avila, Fernando. 2003. Donde Va La Coma Where Is the Bed Going. Editorial Norma.

Baayen, R Harald. 2008. Analyzing linguistic data: A practical introduction to statistics using R. Cambridge University Press.

Battini, Berta Elena Vidal de. 1949. El habla rural de San Luis. Vol. 1. Universidad de Buenos Aires. Facultad de Filosofía y Letras.

Bentivoglio, Paola. 1975. Queísmo y dequeísmo en el habla culta de Caracas. En: Colloquium on Hispanic linguistics 1975. Georgetown University: 1–18.

Bentivoglio, Paola. 1980. El dequeismo en Venezuela: ¿un caso de ultracorreccion? En: Homenaje a Ambrosio Rabanales. En: Boletin de Filologia Santiago de Chile 31: 705–719.

Boye, Kasper and Peter Harder. 2012. A usage-based theory of grammatical status and grammaticalization. *Language* 88(1): 1–44.

Bybee, Joan L. 2003. Mechanisms of Change in Grammaticalization: The Role of Frequency. Oxford: Oxford University Press.

Bybee, Joan, Revere Perkins, and William Pagliuca. 1994. The evolution of grammar: Tense, aspect, and modality in the languages of the world. University of Chicago Press.

Company C., Concepción. 2004. Gramaticalización por subjetivización como prescindibilidad de la sintaxis. *Nueva Revista de Filología Hispánica* 52(2): 1–28.

Cornillie, Bert and Nicole Delbecque. 2008. Speaker commitment: Back to the speaker. Evidence from Spanish alternations. *Belgian Journal of Linguistics* 22(1): 37–62. DOI: 10.1075/bjl.22.03cor. http://benjamins.com/#catalog/journals/bjl.22.03cor/details (consultado el 02/22/2013).

Davis, Mark. (2002). Corpus del Español: 100 million words, 1200s-1900s. http://www.corpusdelespanol.org (consultado el 02/22/2013).

Del Moral, Gabriel. 2008. Spanish dequeísmo a case study of subjectification. *Nueva Revista de Lenguas Extranjeras.* 10: 183–214. http://bdigital.uncu.edu.ar/2643 (consultado el 02/22/2013).

García, Erica C. 1986. El fenómeno (de) queísmo desde una perspectiva dinámica del uso comunicativo de la lengua. *Actas del II Congreso Internacional sobre el español de América:* 46–65.

Geurts, Bart. 2000. Explaining grammaticalization (the standard way). *Linguistics* 38(4): 781–788.

Gries, Stefan Th. 2009. Statistics for linguistics with R: a practical introduction. De Gruyter Mouton.

Haspelmath, Martin. 2002. On Directionality in Language Change, with particular reference to unidirectionality in grammaticalization. En: International Conference: New Reflections on Grammaticalization.

Haspelmath, Martin. 2004. On directionality in language change with particular reference to grammaticalization. *Typological Studies in Language* 59: 17–44.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Cambridge Journal of Linguistics* 42(1): 25–70.

Hopper, Paul J. and Elizabeth Closs Traugott. 2003. Grammaticalization. Cambridge: Cambridge University Press.

Kanwit, Matthew. 2012. Discourse Topic and (De)queísmo: A Variationist Study of the Spanish of Caracas. https://www.indiana.edu/~iulcwp/pdfs/12-Kanwit.pdf (consultado el 02/22/2013).

Lee, Binna. 2006. A Corpus-based Approach on the Development of Conjunction while. En: Lehmann, Christian. 2002. Thoughts on grammaticalization. Universität Erfurt, Philosophische Fakultät, Seminar für Sprachwissenschaft.

Martínez-Sequeira, Ana Teresa. 2000. El dequeísmo en el español de Costa Rica: Un análisis semántico-pragmático. PhD thesis. University of Southern California.

Newmeyer, Frederick J. 2000. Deconstructing grammaticalization. *Language Sciences* 23(2): 187–229.

Newson, Roger. 2006. Confidence intervals for rank statistics: Somers' D and extensions. *Stata Journal* 6(3): 309.

Norde, Muriel. 2002. The final stages of grammaticalization: affixhood and beyond. *Typological Studies in Language* 49: 45–66.

Padró, Lluís and Evgeny Stanilovsky. May 2012. FreeLing 3.0: Towards Wider Multilingual- ity. *Proceedings of the Language Resources and Evaluation Conference* (LREC 2012). ELRA. Istanbul, Turkey.

Prieto, Luis. 1995. Análisis sociolingüístico del dequeísmo en el habla de Santiago de Chile. *Boletín de Filología*: 379–452.

Quilis Sanz, María José. 1986. El dequeísmo en el habla de Madrid y en la telerradio difusión española. *Boletín de la Academia Puertorriqueña de la Lengua Española*: 14–139.

Rabanales, Ambrosio. 1974. Queísmo y dequeísmo en el español de Chile. *Estudios filológicos y lingüísticos. Homenaje a Ángel Rosenblat*: 413–444.

Rátiva, González. 2007. PRESEEA-Medellín-Co. Informe sobre el estado de la investigación. En: El español hablado en las comunidades hispánicas. Informe PRESEEA.

Real Academia Española and Asociación de Academias de la Lengua Española. 2005a. Diccionario panhispánico de dudas. Santillana. `http://lema.rae.es/dpd/srv/search?id=vTr05If13D6tGOqCWV` (consultado el 02/22/2013).

Real Academia Española and Asociación de Academias de la Lengua Española. (2005b). Diccionario panhispánico de dudas. Santillana. `http://lema.rae.es/dpd/srv/search?id=0WI0lLaCjD655ud6n5` (consultado el 02/22/2013).

Rosenblat, Angel. 1974. Estudios filológicos y lingüísticos: homenaje a Angel Rosenblat en sus 70 años. Caracas: Instituto Pedagógico.

Schwenter, Scott A. 1999. Evidentiality in Spanish morphosyntax: a reanalysis of (de) queísmo. *Estudios de variación sintáctica*. Iberoamericana: 65–87.

Serrano, María José. 1998. Estudio sociolingüístico de una variante sintáctica: el fenómeno dequeísmo en el español canario. *Hispania*: 392–405.

Sheather, Simon. 2009. A modern Approach to Regression with R. Vol. 13. Springer.

Silva Villar, Luis and Javier Gutiérrez Rexach. 2012. Predication, Complementation and the Grammar of Dequeísmo Structures. In: Current Formal Aspects of Spanish Syntax and Semantics. Ed. by Melvin González-Rivera and Sandro Sessarego. Cambridge: Cambridge Scholars Publishing: 8–41.

Suárez, Manuel Almeida. 2009. Funcionalismo y antifuncionalismo en la teoría lingüística: a propósito del dequeísmo. In: XI Jornadas de Lingüística: Homenaje Al Profesor José Luis Guijarro Morales: Cádiz, 22 y 23 de Abril de 2008: 149.

Traugott, Elizabeth Closs. 2003. From subjectification to intersubjectification. En: Motives for language change 124: 139.

Traugott, Elizabeth Closs and Richard B. Dasher. 2001. Regularity in semantic change. Cambridge: Cambridge University Press.

Hacia un modelo computacional unificado del lenguaje natural

Towards an unified computational model of natural language

Benjamín Ramírez González
Estudiante de doctorado UCM
benjaminramirezg@gmail.com

Resumen

¿Qué tipo de formalismo debe utilizarse para representar el lenguaje natural? Es necesario un formalismo capaz de describir adecuadamente todas las secuencias de las lenguas naturales. Pero, además, en la medida de lo posible, debe ser un formalismo sencillo, de un coste computacional reducido. Esta pregunta ha generado mucha controversia entre las principales escuelas generativas: la Gramática Transformacional y las Gramáticas de Unificación. En este artículo se defiende que, pese a las diferencias existentes, en última instancia, tales escuelas formalizan el lenguaje humano mediante un mismo tipo de formalismo bien definido: lo que Noam Chomsky llamó lenguaje independiente del contexto. Bajo el prisma de este artículo, la Lingüística actual está en condiciones de ofrecer un modelo computacional unificado del lenguaje natural.

Palabras clave

Jerarquía de Chomsky, lenguajes regulares, lenguajes dependientes del contexto, lenguajes independientes del contexto, coste computacional, complejidad computacional, lenguaje natural, formalismo, gramática generativa, Gramática Generativa Transformacional, Gramáticas de Unificación, HPSG, estructuras de rasgos, estructura compartida.

Abstract

What formalism should be used in order to formalize natural language? That formalism must be able to describe all sequences of natural languages in a right way. Moreover, as long as possible, that formalism must be simple, with a reduced computational cost. This question has triggered a great controversy among the main branches of generative Linguistics: Transformational Grammar and Unification Grammars. The claim of this paper is that, despite discrepancies, these linguistic models formalize natural language by means of the same formal language. This is a well-defined formal language: the context-sensitive language of Noam Chomsky's hierarchy. So, from the point of view of this paper, nowadays, Linguistics can offer an unified computational model of natural language.

Keywords

Chomsky's Hierarchy, regular languages, context-sensitive languages, context-free languages, computational cost, computational complexity, natural language, formalism, generative grammar, Transformational Grammar, Unification Grammars, HPSG, feature structures, feature sharing.

1 La Jerarquía de Chomsky

Toda ciencia utiliza un lenguaje formal para representar los fenómenos que estudia. En Lingüística y en Procesamiento del Lenguaje Natural, la decisión de qué tipo de lenguaje formal utilizar para modelar las secuencias de las lenguas humanas es una cuestión de suma importancia. La Lingüística es una ciencia empírica que se pregunta por la naturaleza del lenguaje humano como fenómeno natural que está representado de forma concreta (mediante una formalización determinada) en el cerebro[1]. Por su parte, en Procesamiento del Lenguaje Natural es crucial el tipo de formalización mediante el cual se codifican las secuencias de una lengua: un formalismo sencillo puede conducir a la elaboración de herramientas eficientes, mientras que un formalismo más complejo puede comprometer tal eficiencia.

Un lenguaje formal Λ es un conjunto (quizá infinito) de secuencias σ. Para formar estas secuencias se utiliza un vocabulario Υ: un conjunto finito de símbolos atómicos υ. La gramática Γ de Λ es la definición de qué concatenaciones de elementos υ son posibles en las secuencias de Λ, es decir Γ define Λ.

Es comúnmente aceptado en Lingüística y en Computación que existen cuatro clases fundamentales de lenguajes formales, con distinto nivel de complejidad. Son los cuatro tipos que estableció Noam Chomsky en la llamada Jerarquía de Chomsky[2]: lenguajes regulares, lenguajes inde-

[1] Véase (Chomsky, 2003), p. 106.

[2] (Chomsky, 1956). Véanse también (Chomsky, 1957),

pendientes del contexto, lenguajes dependientes del contexto y lenguajes recursivamente enumerables. Estos tipos de lenguajes están ordenados en orden creciente de complejidad y cada uno de ellos engloba a todos los lenguajes de complejidad menor. Chomsky estableció para cada uno de estos cuatro tipos de lenguaje un tipo de gramática correspondiente: el tipo de gramática necesaria para generar (definir) tal lenguaje. Así pues, los lenguajes regulares se pueden generar mediante gramáticas regulares, los lenguajes independientes del contexto pueden generarse mediante gramáticas independientes del contexto, etc.

El objetivo de Chomsky al definir su jerarquía era decidir qué tipo de lenguaje formal es el más adecuado para representar las secuencias de las lenguas naturales. Parece adecuado elegir el tipo más sencillo (el de menor coste computacional) de entre aquellos capaces de representar adecuadamente tales secuancias.

De forma paralela al trabajo de Chomsky, en Teoría de la Computación se definieron una serie de máquinas abstractas capaces de definir distintos tipos de lenguaje[3]. Se ha demostrado que estos tipos de lenguaje son, precisamente, los que propuso Chosmky en su jerarquía. Es decir, que estas máquinas abstractas y las gramáticas de la jerarquía de Chomsky son descripciones equivalentes de los mismos tipos de lenguaje. Los lenguajes regulares se describieron en términos de autómatas finitos, los lenguajes independientes del contexto se describieron en términos de autómatas a pila y los dependientes del contexto como autómatas acotados linealmente. Además, los lenguajes no recursivos se pueden describir mediante máquinas de Turing.

Estas máquinas abstractas son algoritmos capaces de decidir en un número finito de pasos, y con el uso de una determinada cantidad de memoria, si una secuencia forma o no parte de un lenguaje. Es por esa naturaleza algorítmica por lo que, en este artículo, se ha decidido representar los lenguajes formales por medio de este tipo de máquinas abstractas. Se espera que, con esta representación resulte fácil entender cómo el coste computacional de usar un tipo de lenguaje (un tipo de máquina) es sensiblemente mayor del de usar otro lenguaje más sencillo.

A continuación se verá en qué consisten los tres primeros tipos de lenguaje de la jerarquía, y cuál es la naturaleza y coste computacional de las gramáticas necesarias para definirlos[4].

(Sánchez León, 2006) y (Serrano, 1975).

[3]Se sigue aquí a (Hopcroft, Motwani, y Ullman, 2001) y (Aranda et al., 2006).

[4]El cuarto tipo de la jerarquía, el más complejo, se obvia en adelante, pues queda fuera del interés de este artículo. El lector interesado en tal tipo de lenguajes puede consultar (Hopcroft, Motwani, y Ullman, 2001).

1.1 Lenguajes regulares

Los lenguajes regulares son aquellos cuyas secuencias σ son una mera concatenación lineal de elementos υ ($\sigma = \upsilon_1, \upsilon_2 ... \upsilon_n$). Podría decirse que, en los lenguajes regulares, cada σ es un objeto unidimensional, en el cual cada constituyente υ_i solo se relaciona con el resto de constituyentes de la secuencia en términos de su posición lineal relativa. En concreto, los lenguajes regulares solo admiten secuencias σ de tipo υ^n, consistentes en la concatenación de n elementos υ. Esta caracterización de las secuencias contrastará en los apartados sucesivos con la propia de las secuencias de otros lenguajes más complicados. Las gramáticas necesarias para generar lenguajes regulares tienen un coste computacional muy bajo, pues no requieren memoria alguna, salvo la necesaria para determinar el momento exacto del proceso de generación en el que se encuentran.

Un lenguaje es regular si existe un autómata de estados finitos capaz de generarlo. Un autómata de estados finitos consta de un conjunto Q de estados, uno de los cuales es el estado de inicio i. Se define también un subconjunto Φ de Q con los estados finales del autómata. Además, el autómata cuenta con un conjunto Σ de símbolos de entrada, el vocabulario a partir del cual se forman las secuencias del lenguaje que se define. Por último, el autómata cuenta con una función δ. Esta función toma un estado de Q y un símbolo de Σ y devuelve un estado de Q.

Por ejemplo, imagínese un autómata de estados finitos Γ_r que define el lenguaje Λ_r, donde Λ_r solo consta de una secuencia: la oración del español *Juan visitó Madrid*. La definición de Γ podría ser la siguiente: $Q = \{q^0, q^1, q^2, q^3\}$, $i = q^0$, $\Phi = \{q^3\}$ y $\Sigma = \{Juan, Madrid, visitó\}$. Además, δ sería la función de transición representada en la figura 1 en forma de grafo.

Figura 1: Función de transición δ en Γ_r

El funcionamiento de este autómata Γ_r en labores de procesamiento podría ser el siguiente. A Γ_r se le pasa como *input* una cadena de símbolos de Σ, por ejemplo, la cadena *Juan visitó Madrid*. La labor de Γ_r es determinar si tal cadena es una

secuencia del lenguaje Λ_r. Γ_r se encuentra en su estado de inicio q^0 y toma el primer símbolo de la cadena de entrada: *Juan*. De acuerdo con δ (figura 1), desde q^0, dada la aparición de *Juan* en la cadena de entrada, Γ_r pasa al estado q^1. A continuación, ya desde q^1, se analiza el siguiente símbolo de la cadena de entrada: *visitó*. La función δ establece que, desde q^1, Γ_r pasa a q^2 al recibir *visitó*. Del mismo modo, Γ_r pasará de q^2 a q^3 al recibir el símbolo siguiente de la cadena de entrada: *Madrid*. Se considera que el proceso termina cuando la cadena de entrada se ha consumido, o cuando δ no define transición alguna desde el estado en que se encuentra para el símbolo de la cadena de entrada que se le ha pasado. Se considera que la secuencia analizada es parte del lenguaje que define Γ_r si, cuando el proceso termina, Γ_r está en un estado final (q^3 en este caso) y la cadena de entrada se ha consumido.

Este tipo de gramática tiene un coste computacional bajo, pues solo hace una lectura secuencial de la cadena de entrada. Los pasos del proceso de análisis son tantos como símbolos haya en la cadena de entrada. Pero, ¿es este tipo de análisis adecuado para las secuencias de las lenguas naturales? No lo es. Las gramáticas regulares (los autómatas de estados finitos) no son capaces de dar cuenta del hecho de que las secuencias de las lenguas naturales obedecen a una estructura sintagmática. Las estructuras sintagmáticas son objetos bidimensionales, ordenados en una dimensión lineal y en otra de dependencia estructural. De acuerdo con ello, la oración *Juan visitó Madrid* no es solo la concatenación de palabras de la figura 1, sino que tales palabras obedecen a una estructura que pudiera representarse mediante el sistema de corchetes del ejemplo (1).

(1) $[_O\ [_{SN}\text{Juan }]\ [_{SV}\text{visitó }[_{SN}\text{Madrid }]]]$

Las gramáticas regulares (los autómatas de estados finitos) no son capaces de dar cuenta de este tipo de estructuras. Obsérvese que los corchetes del ejemplo (1) no son una mera concatenación de signos [y], sino que tales signos obedecen a ciertas reglas gramaticales. En concreto, por cada signo [de la secuencia debe haber un signo] correspondiente. Una gramática regular no tiene la memoria necesaria para generar este tipo de secuencias: en su lectura secuencial de la cadena de entrada, no tiene forma de recordar en un momento dado del proceso generativo cuántos signos [han aparecido en la secuencia para generar solo el mismo número de elementos].

Por tanto, cabe concluir que las lenguas naturales no pueden ser formalizadas adecuadamente mediante lenguajes regulares.

1.2 Lenguajes independientes del contexto

El segundo tipo de lenguajes formales que estableció Chomsky en su jerarquía es el de los llamados lenguajes independientes del contexto. Son lenguajes cuyas secuencias se pueden entender como objetos bidimensionales. Si las secuencias de los lenguajes regulares eran de tipo υ^n, las de los lenguajes independientes del contexto son de tipo ${\upsilon_a}^n{\upsilon_b}^n$. Es decir, en las secuencias de los lenguajes independientes del contexto los elementos υ no solo están concatenados en una dimensión lineal, sino que pueden estar agrupados en subsecuencias entre las que se pueden establecer determinadas relaciones. Por ejemplo, en una secuencia de tipo ${\upsilon_a}^n{\upsilon_b}^n$, puede establecerse que el número n de elementos υ de tipo a debe ser idéntico al número de elementos b que aparece a continuación. En definitiva, estas secuencias son concatenaciones lineales de elementos υ (primera dimensión), pero además, entre estos elementos υ se pueden establecer relaciones estructurales locales (varios elementos υ adyacentes en la dimensión lineal pueden entenderse como constituyentes de una subsecuencia ϕ de σ).

Este tipo de lenguajes sí pueden caracterizar adecuadamente la naturaleza sintagmática de las secuencias de las lenguas naturales. Pero las gramáticas necesarias para generar lenguajes independientes del contexto son más complejas que las gramáticas regulares. Una gramática independiente del contexto podría formalizarse como un autómata de estados finitos aumentado con una pila.

Un autómata a pila consta del mismo conjunto de estados Q que un autómata de estados finitos, con su estado de inicio i, con un subconjunto Φ de estados finales, el vocabulario Σ que determina las palabras posibles de las secuencias del lenguaje que se define, y la conocida función de transición δ. Pero además, el autómata a pila cuenta con un vocabulario de pila Υ. La función δ de un autómata a pila toma tres argumentos: un estado de Q, un símbolo de Σ y un símbolo de pila Υ. El resultado de la función es un par formado por un estado de Q y un nuevo símbolo de pila de Υ. En definitiva, una pila es una memoria que se va alterando según el autómata hace transiciones. Es un modo sencillo de recordar en un momento dado del análisis cuántos elementos de un determinado tipo se han visto ya en la cadena de entrada.

Por ejemplo, imagínese un autómata a pila Γ_{ic} que define un lenguaje Λ_{ic}, donde Λ_{ic} es un con-

junto de secuencias de corchetes [ ]. En una secuencia σ de Λ_{ic}, por cada $[_i$ que se abre, debe aparecer después un $]_i$ que lo cierre. Tal autómata Γ_{ic} podría obedecer a lo siguiente: $Q = \{q^0, q^1\}$, $i = q^0$, $\Phi = \{q^1\}$, $\Sigma = \{[,]\}$ y $\Upsilon = \{p\}$. Además, δ sería la función de transición representada en la figura 2 en forma de grafo.

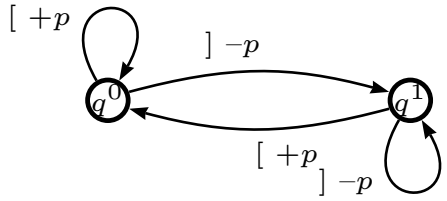


Figura 2: Función de transición δ en Γ_{ic}

Esta función δ establece que, cada vez que se encuentra un símbolo [en la cadena de entrada, se añade un símbolo p a la pila (véase $+p$), y cada vez que se encuentra], se elimina un símbolo p de la pila. Las pilas son estructuras de almacenamiento de datos en las cuales el último elemento que entra es el primero en salir (de ahí su nombre). En este caso, la pila es una memoria que recuerda en cada momento cuántos elementos [están abiertos, a fin de saber cuántos elementos] correspondientes se deben añadir. De este modo Γ_{ic} establece que las secuencias deben tener el mismo número de símbolos de apertura que de cierre. Imagínese que los símbolos [,] y p tienen un índice, de modo que cada símbolo $[_i$ debe cerrarse por el correspondiente $]_i$. Además, por cada $[_i$ se añade a la pila un p_i, y la lectura de p_i en la pila es la única que admite la aparición del correspondiente corchete de cierre $]_i$. Esto permitiría establecer que los pares de corchetes de las secuencias deben anidarse y no cruzarse. Un autómata de este tipo permite analizar las estructuras sintagmáticas de las secuencias de las lenguas naturales, como la representada en el ejemplo (1), que se repite aquí en (2).

(2) $[_O\ [_{SN}$Juan] $[_{SV}$visitó $[_{SN}$Madrid]]]

Obviamente, la máquina utilizada para hacer un análisis adecuado de la estructura sintagmática de (2) sería mucho más compleja que Γ_{ic}. Pero la cuestión crucial es que tal autómata no puede ser un mero autómata de estados finitos, sino que deberá contar, al menos, con una memoria equivalente a la pila de Γ_{ic}. Como parece obvio, un autómata a pila tiene, en principio, un coste computacional añadido sobre el propio de un autómata de estados finitos (un coste de memoria: la pila). No obstante, es importante señalar que los autómatas a pila, como ocurriera con los de estados finitos, solo hacen una lectura secuencial de la cadena de entrada. Es cierto que cada paso que efectúan es más costoso que el correspondiente paso en un autómata de estados finitos, pero el número de pasos no es mayor: un autómata a pila ejecuta tantos pasos de análisis como elementos tenga la cadena de entrada.

1.3 Lenguajes dependientes del contexto

Se ha visto que un autómata a pila sí es capaz de modelar estructuras sintagmáticas. Pero hay ciertos fenómenos sintácticos de las lenguas naturales que este tipo de lenguajes no pueden modelar: el movimiento y el ligamiento. Véanse los siguientes ejemplos del español. En (3) el pronombre *lo* está ligado por el antecedente *Juan*: *lo* y *Juan* se refieren, por ello, al mismo individuo. De un modo similar, en (4), el pronombre interrogativo *Qué* y la huella *h* están vinculados: gracias a ello, *Qué* se linealiza en la margen izquierda de la oración, pero conserva su valor de objeto del verbo *descubrirán*.

(3) [Juan$_i$ [cree [que [lo$_i$ descubrirán]]]]

(4) [Qué$_i$ [crees [que [descubrirán [h$_i$]]]]]

Estas relaciones i de (3) y (4) —ligamiento y movimiento— van más allá de la complejidad propia de una estructura sintagmática: son relaciones que se establecen, dentro de una estructura sintagmática, entre elementos alejados tanto en la dimensión lineal como en la jerárquica[5]. Se suele decir de estos fenómenos que son no locales.

Por tanto, las secuencias de (3) y (4), con sus identidades i, no son meras estructuras sintagmáticas, no son meros objetos bidimensionales. Si se quisiera representar en términos espaciales que el índice i de *Juan* y el de *lo* es el mismo objeto, sería necesaria una tercera dimensión[6].

[5] La dimensión que forma la jerarquía de constituyentes sintagmáticos.

[6] Imagínese una secuencia unidimensional σ formada por la concatenación de n elementos de tipo υ ($\sigma = \upsilon_1, ..., \upsilon_n$). En tal secuencia unidimensional, el elemento υ_i solo puede ocupar una posición: la posición i. Pero imagínese que σ se concibe como una de las dos coordenadas (dimensiones) que permiten definir el plano ϕ. Junto a $\sigma = \upsilon_1, ..., \upsilon_n$ existe entonces $\sigma' = \upsilon'_1, ..., \upsilon'_n$, y cualquier punto del plano ϕ se define como $\phi(\upsilon_i, \upsilon'_j)$. Entonces, en este nuevo objeto bidimensional ϕ, al elemento υ_i de σ (que en σ, objeto unidimensional, solo ocupaba la posición i) le corresponden n posiciones distintas: las posiciones $\phi(\upsilon_i, \upsilon'_1), ..., \phi(\upsilon_i, \upsilon'_n)$. Del mismo modo, el punto $\phi(\upsilon_i, \upsilon'_j)$ solo cuenta con una posición en el plano ϕ. Pero si ϕ forma parte de la definición del cubo Υ (un objeto tridimensional), de nuevo, en Υ, a $\phi(\upsilon_i, \upsilon'_j)$ le corresponden n posiciones: $\Upsilon(\upsilon_i, \upsilon'_j, \upsilon''_1), ..., \phi(\upsilon_i, \upsilon'_j, \upsilon''_n)$. Recuérdese que las estructuras sintagmáticas de (3) y (4) son objetos bidimensionales. En tales planos bidimensionales no es posible que un punto —el índice i— ocupe dos posiciones distintas. Por tanto, si el índice i ha de concebirse como un

Precisamente, si los lenguajes regulares definían secuencias unidimensionales y los independientes del contexto secuencias bidimensionales, los lenguajes independientes del contexto permiten formalizar las secuencias como objetos tridimensionales (multidimensionales, en realidad). Es decir, estos lenguajes pueden estar formados por secuencias del tipo $v_a{}^n v_b{}^n v_c{}^n$, donde es necesario coordinar tres dimensiones: el número de elementos de tipo v_a, de tipo v_b y de tipo v_c debe ser el mismo, debe ser n.

Para generar estos lenguajes, la Teoría de la Computación define un nuevo tipo de máquina abstracta: el autómata acotado linealmente. Estas máquinas cuentan, como las anteriores, con un conjunto Q de estados (con su estado de inicio i y sus estados finales Φ), un vocabulario Σ y una función de transición δ. No tienen una pila, pues, como se verá, su memoria supera con creces a la de la pila. Lo característico de los autómatas acotados linealmente es el modo en el que leen la cadena de entrada. Los autómatas anteriores hacían una mera lectura secuencial de esta. Los autómatas acotados linealmente, en cambio, pueden también retroceder en la lectura, volver a avanzar, etc. Incluso son capaces de modificar los símbolos de la cadena de entrada. Dicho en otros términos, la función δ de estos autómatas toma un estado de Q y un símbolo de Σ y devuelve tres elementos: un estado de Q, un símbolo de Σ y una instrucción $\leftarrow$ o $\rightarrow$. El estado de Q es el estado al que se desplaza el autómata como consecuencia de la transición. El símbolo de Σ es el símbolo con el que la transición rescribe la posición de la cadena de entrada que se acaba de leer. Y la instrucción $\leftarrow$ o $\rightarrow$ determina si el siguiente elemento de la cadena de entrada que se va a leer es el que se encuentra a la derecha del actual o el que se encuentra a su izquierda.

Sería muy complicado crear y explicar un autómata acotado linealmente capaz de generar las estructuras sintagmáticas de (3) y (4). Baste, entonces, con explicar su lógica. En los análisis de (3) y (4), *lo* y *h* deben compartir índice con *Juan* y *Qué*, respectivamente. De un modo u otro, la aparición de *lo* y *h* está legitimada por la existencia de un antecedente. Esta relación no puede regularse haciendo una mera lectura secuencial de la cadena de entrada. Es necesario un análisis que, cuando observa la aparición de *lo* o *h*, compruebe que estos tienen un antecedente legítimo. Para ello es necesaria una memoria más compleja que pila de los autómatas a pila. La memoria de los autómatas acotados linealmente es, precisamente, su capacidad de recordar la cadena de entrada y de volver sobre ella para hacer las comprobaciones oportunas. Por tanto, un hipotético autómata acotado linealmente Γ_{dc} que definiese el lenguaje Λ_{dc} formado por las dos estructuras de (3) y (4) podría leer secuencialmente la cadena de entrada y, al encontrar *lo* o *h*, volver sobre sus pasos para comprobar si existen los antecedentes *Juan* o *Qué*.

Este tipo de lenguaje formal, el lenguaje dependiente del contexto, sí es capaz de dar cuenta de esta clase de fenómenos sintácticos. No obstante, la complejidad de estos lenguajes es sensiblemente mayor a la propia de los lenguajes regulares y también a la de los lenguajes independientes del contexto. La observación crucial que permite hacer tal afirmación es el hecho de que las máquinas utilizadas para reconocer estos lenguajes (los autómatas acotados linealmente) permiten lecturas no secuenciales de la cadena de entrada. Con las máquinas anteriores, el tiempo de análisis de una secuencia de entrada de longitud n estaba siempre definido por una función lineal: la máquina empleaba siempre n pasos. En cambio, con un análisis no secuencial de las cadenas de entrada, el tiempo puede ser polinómico. Una máquina puede leer los n elementos de la secuencia de entrada m veces. Por tanto, los pasos del análisis son n^m.

2 La Gramática Transformacional

Se han visto en 1 ciertos fenómenos (estructura sintagmática, ligamiento y movimiento) que suponen un reto para la elaboración de formalizaciones adecuadas de las secuencias de las lenguas naturales. Se ha sugerido que el lenguaje formal necesario para modelar adecuadamente esos fenómenos debe ser lo que Chomsky llamó un lenguaje independiente del contexto. Este tipo de lenguajes formales tiene una cierta complejidad. Por tanto, su aplicación al lenguaje natural quizá comprometa la eficiencia computacional de las formalizaciones. Esta ha sido una preocupación central de ciertas escuelas lingüísticas generativas (Gramáticas de Unificación, entre otras) que han intentado desarrollar modelos más sencillos, más adecuados en términos de coste de procesamiento.

No obstante, la Gramática Generativa Transformacional que fundara Noam Chomsky en los años 50 del siglo XX[7] sí ha utilizado de forma sistemática formalizaciones de las lenguas naturales

único objeto, este debe definirse en una tercera dimensión. En ese objeto tridimensional resultante, i sí puede actuar como índice de dos elementos diferentes.

[7] (Chomsky, 1957)

que, de un modo u otro, consisten en lenguajes independientes del contexto. Para la Gramática Transformacional, las secuencias de las lenguas naturales son estructuras de constituyentes como las vistas en (3) y (4) —se repiten aquí en (5) y (6)—.

(5) [Juan$_i$ [cree [que [lo$_i$ descubrirán]]]]

(6) [Qué$_i$ [crees [que [descubrirán [h$_i$]]]]]

Es decir, en este modelo gramatical, tales estructuras no son solo la representación de la historia derivativa de las secuencias. Para el modelo transformacional, las estructuras de (5) y (6) son objetos sobre los que las reglas de la gramática operan. A las reglas gramaticales que trabajan sobre este tipo de estructura se les ha llamado transformaciones o reglas transformacionales. Por ejemplo, una transformación podría tomar como *input* una estructura como la de (7) y mover *Qué* desde la posición de base a la periferia izquierda, para dar lugar a (6).

(7) [Crees [que [descubrirán [qué]]]]

Como se ha visto en 1.3, este tipo de operación solo es posible para gramáticas equivalentes a autómatas acotados linealmente. Por tanto, el modelo transformacional entiende las lenguas naturales como lenguajes dependientes del contexto[8].

3 Gramáticas de Unificación: HPSG

Las Gramáticas de Unificación nacieron en los años 80, como solución a los problemas de procesamiento en tiempo real de los modelos gramaticales previos. Las más importantes son GPSG, LFG y HPSG. GPSG (*Generalized Phrase Structure Grammar*) fue desarrollado por Gazdar, Pullum y Sag —(Gazdar et al., 1985)—. Hoy se encuentra en desuso. En cambio, LFG (*Lexical Functional Grammar*) se encuentra en pleno desarrollo —(Kaplan, Ronald, y Bresnan, 1982), (Dalrymple, Lamping, y Saraswat, 1995) y (Bresnan, 2001)—. HPSG (*Head-Driven Phrase Structure Grammar*) es la evolución de GPSG, y será el modelo que se use como referencia a continuación —(Pollard y Sag, 1987), (Pollard y Sag, 1994) y (Sag, Wasow, y Bender, 2002)—.

A diferencia de lo que ocurre en la Gramática Transformacional, en HPSG, las estructuras de constituyentes (como las de (5) o (6)) no se contemplan propiamente como objetos gramaticales. Los sintagmas de tales estructuras se crean a partir de sus constituyentes inmediatos, pero la relación entre el sintagma creado y sus constituyentes no se recuerda: el sintagma no guarda la información de cuáles son los constituyentes que lo formaron. En este sentido, las estructuras del tipo de (5) o (6), en HPSG, no existen. No es posible, por tanto, para una gramática de tipo HPSG, tomar una estructura como la de (7) para efectuar directamente sobre ella el movimiento necesario para generar (6). No es posible, sencillamente, porque, en HPSG, la estructura de constituyentes de (7) no existe como objeto gramatical.

Esta limitación hace de HPSG un modelo gramatical más sencillo que la Gramática Transformacional, en términos computacionales. En principio, cabría pensar que este modelo es equivalente a una gramática independiente del contexto, pues solo permite analizar estructuras sintagmáticas. Entonces, ¿cómo se consige en HPSG modelar aquellos fenómenos cuya complejidad va más allá, como el ligamiento y el movimiento? Se hace mediante el uso generalizado de estructuras de rasgos.

En HPSG, toda realidad gramatical (las unidades léxicas, las reglas gramaticales, etc.) se formaliza como tipo τ consistente en una estructura de rasgos[9]. Una característica crucial de las estructuras de rasgos que usa HPSG es el hecho de que dos o más rasgos de una estructura pueden compartir valor. Es decir, el valor de tales rasgos es el mismo objeto. Si dos rasgos comparten valor, sus valores deben ser idénticos, y si el valor de uno se modifica, el del otro también ha de hacerlo de forma consecuente. Véase en la figura 3 un ejemplo de estructura de rasgos caracterizadora

[8] En realidad, (Peters y Ritchie, 1973) demostraron que una gramática transformacional, tal como se definía en sus primeras versiones —(Chomsky, 1957), (Chomsky, 1965)— generan lenguajes aún más complejos: lenguajes recursivamente enumerables, según la jerarquía de Chomsky. La demostración de estos autores estaba fundamentada en el hecho de que las transformaciones de esas primeras versiones del modelo eran capaces de eliminar elementos de una secuencia. Las formulaciones posteriores de lo que es una regla transformacional —(Chomsky, 1986)— eliminan esta posibilidad. Se puede pensar que la Gramática Transformacional, con esta nueva formulación, solo genera lenguajes dependientes del contexto. (Chomsky, 1965) da por sentado que la complejidad computacional del modelo transformacional es la propia de un lenguaje dependiente del contexto.

[9] Un tipo τ consiste en una estructura de rasgos: un conjunto de 0 o más rasgos ϕ. Por ejemplo, el tipo τ_i se puede definir como una estructura de rasgos con los rasgos ϕ_a y ϕ_b. Además, cada rasgo ϕ se concibe como un par atributo valor, donde el atributo es un identificador del rasgo y el valor debe ser uno de los tipos τ definidos en la gramática. Se entiende que los tipos τ_a y τ_b en τ_0, en tanto que tipos, son estructuras de rasgos que, a su vez, pueden tener también sus rasgos y tipos anidados. Véase (Shieber, 1986).

de un verbo transitivo. Como puede verse, la estructura recoge la categoría gramatical verb, los rasgos de persona, género y número y los dos argumentos propios de una verbo transitivo. Como se ve, el primer argumento del verbo y el verbo mismo comparten el valor de los rasgos de persona y número. Con ello se formaliza el hecho de que sujeto y verbo concuerdan en persona y número.

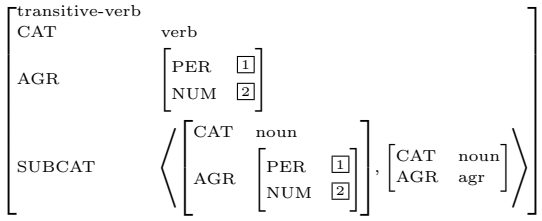


Figura 3: Estructura de rasgos en HPSG

Gracias a la estructura compartida, cada vez que una regla de la gramática crea un nuevo sintagma a partir de sus constituyentes inmediatos, en tal sintagma se puede recoger toda la información de dichos constituyentes que pudiera ser útil. Imagínese que el sintagma γ tiene la información i codificada en su estructura de rasgos. Tal sintagma γ se toma como constituyente de un nuevo sintagma β. Gracias a la estructura compartida, i puede pasar a la estructura de rasgos de β. De nuevo, si β se toma como constituyente inmediato de α, i puede pasar de β a α. A este proceso sucesivo de ascenso de información a lo largo de la estructura sintáctica se le llama habitualmente en la bibliografía de HPSG percolación de rasgos. Es esta percolación de rasgos la que permite establecer relaciones (como las propias del ligamiento y el movimiento) entre elementos alejados lineal y jeráquicamente en las secuencias de las lenguas naturales.

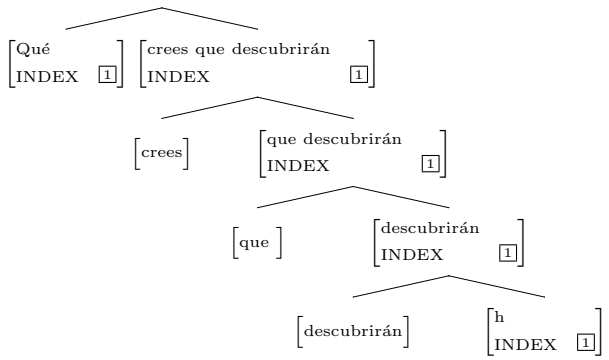


Figura 4: Percolación de rasgos en HPSG

Por ejemplo, en la figura 4 se representa el proceso derivativo que lleva a generar la oración de (6). La relación entre la huella h y *Qué* se establece mediante la percolación del índice 1. La unión del verbo *descubrirán* a una huella desencadena el ascenso del índice de la huella al sintagma resultante. Cada vez que se crea un sintagma a partir de un constituyente con este índice, el índice se recoge en dicho sintagma resultante. De este modo, *Qué* puede identificar su índice con el de su huella de forma local, pues este índice ha ascendido paso a paso hasta el sintagma al que *Qué* se une.

En resumen, la Gramática Transformacional contempla las estructuras de constituyentes como objetos gramaticales reales, sobre los que una regla puede operar: una regla puede mover elementos destro de la estructura, puede identificar elementos alejados lineal y jerárquicamente, etc. Obsérvese la semejanza entre este tipo de procesos y los propios de los autómatas acotados linealmente vistos en 1.3. La cadena de entrada de estos autómatas se recordaba, íntegra, durante toda la derivación, y toda ella era accesible en todo momento. El entender las estructuras sintagmáticas completas como *input* de las reglas gramaticales tiene una complejidad equivalente. Se concluye, entonces, como ya se ha dicho anteriormente, que la Gramática Transformacional formaliza las lenguas humanas como lenguajes dependientes del contexto.

En cambio, en HPSG y el resto de Gramáticas de Unificación, las estructuras de constituyentes no se contemplan como objetos gramaticales. Las reglas de estas gramáticas no pueden tomar una estructura de constituyentes y establecer directamente una relación no local entre distintos elementos de esta. Una gramática de este tipo es equivalente a un autómata a pila: la lectura de la cadena de entrada es secuencial, pues, en cada momento de la derivación, la gramática no tiene acceso a la estructura de constituyentes previamente derivada. Solo se necesita una memoria equivalente a una pila que permita regular la correcta apertura y cierre de constituyentes sintagmáticos.

Para establecer relaciones no locales, en HPSG la información relevante debe ser recogida sucesivamente en cada sintagma tal como se muestra en la figura 4. Este proceso de percolación de rasgos se ha formalizado mediante estructuras de rasgos donde (esto es crucial) dos o más rasgos pueden tomar por valor el mismo objeto (se dice que tales rasgos comparten estructura). En principio, se entiende que este modelo gramatical, dada esta restricción fundamental, es más económico en términos computacionales que la

Gramática Transformacional.

4 Hacia un modelo unificado

En este apartado se defiende que los modelos gramaticales anteriormente expuestos no son propuestas irreconciliables. Se asume que HPSG es un modelo más sencillo desde el punto de vista de coste computacional. Pero se defiende también que tal diferencia no es en realidad una diferencia fundamental, sino, más bien, una diferencia de grado.

Esta reflexión está fundamentada en dos observaciones. La primera consiste en que las estructuras de rasgos con estructura compartida son, en sí, objetos tridimensionales. Obsérvese que una estructura de constituyentes con relaciones no locales —(5) y (6) con su coindización i— es perfectamente equivalente a una estructura de rasgos donde unos rasgos se anidan dentro de otros y donde dos rasgos, quizá en niveles de anidación distintos, comparten valor. Por tanto, tanto las estructuras sintagmáticas con relaciones no locales como las estructuras de rasgos que usa HPSG son objetos tridimensionales, o secuencias propias de un lenguaje dependiente del contexto. Para manejar tales estructuras complejas serán necesarias gramáticas dependientes del contexto, equivalentes a autómatas acotados linealmente: gramáticas capaces de recordar y manipular la estructura completa en todo momento de la derivación.

De hecho, en las gramáticas computacionales basadas en HPSG, las relaciones de dependencia sintagmática se representan en forma de estructuras de rasgos[10], tal como se muestra en la figura 5, donde ARGS es el rasgo cuyo valor es la lista de constituyentes de un signo.

$$\left[\begin{array}{l}\textsc{s}\\ \textsc{args}\left\langle\left[\begin{array}{l}\textsc{np}\\ \textsc{args}\left\langle[\ldots],[\ldots]\right\rangle\end{array}\right],\left[\begin{array}{l}\textsc{vp}\\ \textsc{args}\left\langle[\ldots],[\ldots]\right\rangle\end{array}\right]\right\rangle\end{array}\right]$$

Figura 5: Dependencia sintagmática con estructuras de rasgos

Sería perfectamente posible representar las estructuras sintagmáticas de (5) y (6) del modo esbozado en esta figura 5, y marcar la coindización i de tales estructuras mediante una relación de estructura compartida. Asúmase, entonces, que es posible formalizar tanto una Gramática Transformacional como una Gramática de Unificación (HPSG) mediante este tipo de estructuras de rasgos con estructura compartida. Es fácil observar, entonces, que la única diferencia real entre estas gramáticas consiste en las restricciones que presentan sus reglas en cuanto a qué rasgos anidados tienen acceso. Si una Gramática Transformacional se caracteriza por recordar en cada momento toda la estructura previamente derivada, sus reglas tendrán acceso a todos los rasgos ARGS anidados en la estructura de rasgos. En cambio, las reglas de una gramática de tipo HPSG tienen vedado el acceso a la información de tales rasgos[11].

Obsérvese que, desde este punto de vista, la complejidad computacional de los dos modelos gramaticales no es esencialmente distinta. Ambos, de un modo u otro, tienen que gestionar objetos complejos, tridimensionales, para lo cual tendrán que utilizar gramáticas equivalentes, en ambos casos, a autómatas acotados linealmente, a gramáticas dependientes del contexto. Esta naturaleza dependiente del contexto estará en el corazón mismo del algoritmo de análisis —llámese *parser*— de las Gramáticas Transformacionales. Las gramáticas de tipo HPSG, por su parte, utilizarán *parsers* dotados de algún sistema que compruebe si las estructuras de rasgos de dos constituyentes son compatibles, y que calcule cuál ha de ser la estructura del sintagma resultante dadas las estructuras de sus constituyentes. Estos cálculos, dada la complejidad de las estructuras de rasgos con estructura compartida, los debe afrontar una gramática dependiente del contexto. En definitiva, en ambos casos, la complejidad del proceso alcanza el nivel propio de un lenguaje dependiente del contexto. Esto no implica que la diferencia de coste computacional de ambos modelos no sea significativa, pero sí se deduce de aquí que tal diferencia no es fundamental.

La segunda observación que sugiere la posibilidad de alcanzar un modelo computacional unificado para el lenguaje natural es la siguiente. Se observa en la Gramática Transformacional una preocupación por acotar el dominio de acción de las reglas de la gramática. Por ejemplo, en el modelo de fases de (Chomsky, 2008), la estructura sintagmática se divide en tramos a los que

[10] Véase (Bender, Flickinger, y Oepen, 2002). Cabe preguntarse por qué se representan las relaciones de dependencia sintagmática en gramáticas de tipo HPSG si los constructos formados por sintagma y constituyentes no se contemplan como objetos gramaticales. Se hace por motivos técnicos, como la representación de la ordenación lineal de los constituyentes de un sintagma. Esto no compromete el carácter independiente del contexto de la gramática, pues las reglas siguen sin tener acceso a la información de tales constituyentes anidados.

[11] Valga esta formulación personal de lo que definieron los fundadores de HPSG en (Pollard y Sag, 1987) como *Locality Principle*.

se llama fases. Mientras se está derivando una fase, las reglas de la gramática sí tienen acceso a todos los constituyentes anidados en esta. Pero una vez la fase se ha completado, las reglas no tienen acceso a sus constituyentes. Utilizando la formalización de la figura 5, en el modelo de fases, las reglas solo tienen acceso a la información de los rasgos ARGS anidados en la misma fase en la que la derivación opera en ese momento; pero no tendrán acceso a los rasgos ARGS más anidados, los pertenecientes a las fases anteriores. Esta limitación debe ir acompañada de un proceso de ascenso de información que recuerda a la percolación de rasgos de HPSG vista en 3.

En resumen, las estructuras de rasgos usadas en HPSG son objetos tridimensionales, secuencias propias de un lenguaje dependiente del contexto. Dada esta realidad, tanto la Gramática Transformacional como HPSG deben hacer frente a una complejidad computacional que, en esencia, es del mismo tipo. Llegada la discusión a este punto, surge la siguiente pregunta: dentro de la estructura sintagmática, ¿cuál debe ser el ámbito de acción de las reglas de la gramática? Las distintas escuelas dan distintas soluciones a esta pregunta. Pero, en favor de la sencillez computacional, todas pretender acotar tal ámbito.

5 Conclusión

Se ha discutido en los apartados previos cuál es el formalismo adecuado para dar cuenta de las lenguas naturales: cuál es el formalismo de menor coste computacional de entre los capaces de describir adecuadamente tales lenguas. Se han presentado los distintos tipos de lenguajes formales posibles y se ha discutido qué formalización han utilizado la Gramática Transformacional y las Gramaticas de Unificación (en concreto, HPSG). Se ha llegado a la conclusión de que las diferencias entre las formalizaciones de estas escuelas no son fundamentales. El panorama resultante es el siguiente. Tanto la Gramática Transformacional como las Gramáticas de Unificación formalizan las secuencias de las lenguas naturales mediante lenguajes dependientes del contexto. En este sentido, la Lingüística ha alcanzado un cierto consenso. Por razones de economía computacional, parece conveniente, no obstante, formalizar tales secuencias como objetos en los cuales las relaciones no locales se reduzcan, en la medida de lo posible, a procesos locales cíclicos. Es decir, parece deseable que, en cada paso local de creación de estructura, el sintagma resultante recoja, de un modo u otro, la información de sus constituyentes que, presumiblemente, pudiera ser necesaria en un momento posterior de la derivación. La alternativa sería que la derivación recordara en todo momento toda la estructura creada y pudiese acceder a ella, lo cual resulta de un coste computacional inasumible. HPSG opta por una interpretación radical de esta idea: todos los procesos gramaticales son locales, pues ninguna regla tiene acceso a los constituyentes del sintagma sobre el que opera. El modelo de fases de la Gramática Transformacional opta por una interpretación menos exigente: solo dentro de ciertos tramos estructurales (las fases), las reglas tienen acceso a los constituyentes anidados.

Bibliografía

Aranda, Joaquín, Natividad Duro, José Luis Fernández, José Jiménez, y Fernando Morilla. 2006. *Fundamentos de lógica matemática y computación*. Sanz y Torres.

Bender, Emily M., Dan Flickinger, y Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precission grammars. CSLI, Stanford University.

Bresnan, J. 2001. *Lexical-Functional Syntax*. Oxford, Basil Blackwell.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions PGIT*, 2:113–124.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton and co., N.Y. publishers.

Chomsky, Noam. 1965. *Aspects of the Theory of Sintax*. The MIT Press, Cambridge, Massachusetts.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origins and Use*. Praeger Publishers, N.Y., USA.

Chomsky, Noam. 2003. *Sobre la naturaleza y el lenguaje*. Cambridge University Press.

Chomsky, Noam. 2008. On phases. En *Foundational Issues in Linguistic Theory*. The MIT Press, Cambridge, Massachusetts, páginas 133–166.

Dalrymple, M., J. Lamping, y V. Saraswat. 1995. *Formal Issues in Lexical Functional Grammar*. CSLI Publications.

Gazdar, G., E. Klein, G. Pullum, y I. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Hopcroft, John E., Rajeev Motwani, y Jeffrey D. Ullman. 2001. *Introduction to Automata Theory, Languages and Computation*. Pearson Education.

Kaplan, Ronald, y Joan Bresnan. 1982. Lexical-functional grammar: a formal system for grammatical representation. En Joan Bresnan, editor, *The mental representation of grammatical relations*. Cambridge: The MIT Press, páginas 173–281.

Peters, P. Stanley y R. W. Ritchie. 1973. On the generative power of transformational grammar. *Information Science*, 6:49–83.

Pollard, Carl y Ivan A. Sag. 1987. *Information-Based Syntax and Semantics*. The University of Chicago Press.

Pollard, Carl y Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

Sag, Ivan A., Thomas Wasow, y Emily Bender. 2002. *Syntactic Theory: a Formal Introduction*. CSLI Publications.

Serrano, Sebastián. 1975. *Elementos de lingüística matemática*. Editarial Anagrama.

Shieber, Stuart M. 1986. *An Introduction to Unification Based Approaches to Grammar*. CSLI Publications.

Sánchez León, Fernando. 2006. Gramáticas y lenguajes formales. Departamento de Lingüística Computacional de la Real Academia Española.

Volume 5, Número 2- Dezembro 2013

linguaMATICA

ISSN: 1647-0818

Dossier

Artigos de Investigação

Novas Perspetivas

http://linguamatica.com