Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 7, Número 2- Dezembro 2015

ISSN: 1647-0818

Volume 7, Número 2 – Dezembro 2015

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Com este número fechamos o sétimo ano da Linguamática. Um ano que introduziu algumas alterações relevantes, como a indexação da revista pela Scopus, a modernização do design de capa, e do conteúdo, com hiperligações que permitem uma melhor navegação nos artigos, quer no PDF que inclui todos os artigos, quer em cada um dos artigos, facilitando o acesso a notas de rodapé, figuras ou citações com um simples clique.

Estas alterações, no entanto, tornaram o processo de edição mais demorado e custoso. Embora continuemos a aceitar contribuições em formato Microsoft Word, os artigos foram todos convertidos para LaTeX, de modo a garantir homogeneidade entre os formatos, e a possibilidade de interligar todos os PDF. Não será, ainda, no oitavo ano da Linguamática que iremos restringir as contribuições a documentos escritos em LaTeX. Mas pedimos que, os autores que tiverem as competências mínimas para o fazer, tentem usar o LaTeX para produzir os seus artigos. No sentido de facilitar este processo existe na plataforma ShareLaTeX um modelo que poderá ser utilizado, on-line, sem necessidade da instalação do processador LaTeX nas respetivas máquinas. Para o ano, outras novidades deverão surgir.

Finalmente, e como não podia deixar de ser, somos obrigados aos autores, que continuam a acreditar na Linguamática como um veículo de informação, contribuindo com trabalhos interessantes, quer aos revisores que se dedicam a analisar cuidadosamente as contribuições, sugerindo correções e melhorias. A todos, o nosso muito obrigado.

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

Antoni Oliver González,
Universitat Oberta de Catalunya,

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,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Joaquim Llisterri,
Universitat Autònoma de Barcelona

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

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade de Santiago de Compostela

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

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

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

Artigos de Investigação

Descoberta de Synsets Difusos com base na Redundância em vários Dicionários

Discovering Fuzzy Synsets from the Redundancy across several Dictionaries

Fábio Santos
CISUC, Departamento de Engenharia Informática
Universidade de Coimbra, Portugal
fasantos@student.dei.uc.pt

Hugo Gonçalo Oliveira
CISUC, Departamento de Engenharia Informática
Universidade de Coimbra, Portugal
hroliv@dei.uc.pt

Resumo

Numa wordnet, conceitos são representados através de grupos de palavras, vulgarmente chamados de synsets, e cada pertença de uma palavra a um synset representa um diferente sentido dessa mesma palavra. Mas como os sentidos são entidades complexas, sem fronteiras bem definidas, para lidar com eles de forma menos artificial, sugerimos que synsets sejam tratados como conjuntos difusos, em que cada palavra tem um grau de pertença, associado à confiança que existe na utilização de cada palavra para transmitir o conceito que emerge do synset. Propomos então uma abordagem automática para descobrir um conjunto de synsets difusos a partir de uma rede de sinónimos, idealmente redundante, por ser extraída a partir de várias fontes, e o mais abrangentes possível. Um dos princípios é que, em quantos mais recursos duas palavras forem consideradas sinónimos, maior confiança haverá na equivalência de pelo menos um dos seus sentidos. A abordagem proposta foi aplicada a uma rede extraída a partir de três dicionários do português e resultou num novo conjunto de synsets para esta língua, em que as palavras têm pertenças difusas, ou seja, *fuzzy synsets*. Para além de apresentar a abordagem e a ilustrar com alguns resultados obtidos, baseamo-nos em três avaliações – comparação com um tesauro criado manualmente para o português; comparação com uma abordagem anterior com o mesmo objetivo; e avaliação manual – para confirmar que os resultados são positivos, e poderão no futuro ser expandidos através da exploração de outras fontes de sinónimos.

Palavras chave

wordnets, synsets, fuzzy clustering, rede léxico-semântica, sinónimos, confiança, dicionários

Abstract

In a wordnet, concepts are typically represented as groups of words, commonly known as synsets, and each membership of a word to a synset denotes a different sense of that word. However, since word senses are complex entities, without well-defined boundaries, we suggest to handle them less artificially, by representing them as fuzzy objects, where each word has its membership degree, which can be related to the confidence on using the word to denote the concept conveyed by the synset. We thus propose an approach to discover synsets from a synonymy network, ideally redundant and extracted from several broad-coverage sources. The more synonymy relations there are between two words, the higher the confidence on the semantic equivalence of at least one of their senses. The proposed approach was applied to a network extracted from three Portuguese dictionaries and resulted in a large set of fuzzy synsets. Besides describing this approach and illustrating its results, we rely on three evaluations – comparison against a handcrafted Portuguese thesaurus; comparison against the results of a previous approach with a similar goal; and manual evaluation – to believe that our outcomes are positive and that, in the future, they might my expanded by exploring additional synonymy sources.

Keywords

wordnets, synsets, fuzzy clustering, lexical-semantic network, synonyms, confidence, dictionaries

1 Introdução

Wordnets são bases de conhecimento léxico-semântico, inspiradas na Wordnet de Princeton (Fellbaum, 1998), a primeira, que definiu este modelo de recurso lexical. Uma wordnet agrupa as palavras de uma língua em conjuntos de sinónimos, normalmente chamados de *synsets*, que representam as possíveis lexicalizações de um conceito nessa língua. A ambiguidade lexical, ou seja, a possibilidade de usar a mesma palavra para transmitir diferentes significados, pode ser representada no modelo da wordnet através da presença da mesma palavra em diferentes synsets, relativos a cada um dos seus sentidos. Ao

This work is licensed under a
Creative Commons Attribution 3.0 License

mesmo tempo, um synset pode incluir um conjunto de palavras que partilhem o mesmo significado. No entanto, na realidade os sentidos não são objectos discretos, mas sim estruturas complexas, sem fronteiras bem definidas (Kilgarriff, 1996), ou seja, ainda que claramente útil ao processamento da língua, esta representação acaba por ser artificial.

Existem actualmente inúmeras wordnets para as mais variadas línguas (ver, por exemplo, Bond & Paik (2012)), e até línguas para as quais há mais de uma wordnet disponível. Para a língua portuguesa, existem pelo menos seis wordnets (ver Gonçalo Oliveira et al. (2015)), construídas por equipas independentes, com licenças diferentes, e seguindo abordagens distintas, cada uma com as suas vantagens e desvantagens. Por exemplo, relativamente a wordnets livres para esta língua, a OpenWN-PT (de Paiva et al., 2012) e a PULO (Simões & Guinovart, 2014) têm ainda uma cobertura limitada ao nível de lemas, sentidos e tipos de relação. No entanto, estão as duas alinhadas com a WordNet de Princeton e, indirectamente com outras wordnets. Isto não só trás benefícios ao nível do processamento multilingue, como permite complementar o conhecimento de cada um destes recursos com informação noutras wordnets (nomeadamente relações, definições ou exemplos). Por outro lado, a Onto.PT (Gonçalo Oliveira & Gomes, 2014) é maior que as anteriores, o que se deve essencialmente à exploração de vários recursos, criados de origem para o português, através da sua abordagem automática de construção, a ECO. Além disso, a Onto.PT abrange um leque de tipos de relação mais alargado que a maior parte das wordnets. Uma limitação, relacionada com a sua construção automática, é que ela não se encontra alinhada com nenhuma outra wordnet. Outra, será o facto da Onto.PT ser potencialmente menos fiável que as demais wordnets, nomeadamente daquelas cuja criação é completamente manual ou que, apesar de tirarem partido de abordagens semi-automáticas, têm uma integração de conteúdos mais controlada.

Para balancear a segunda limitação referida, pretendemos criar uma wordnet com uma cobertura comparável à da Onto.PT, mas onde sejam associadas uma ou várias medidas que transmitam a confiança em cada uma das decisões tomadas na sua criação, incluindo a associação de palavras em synsets ou a ligação de synsets através de uma relação semântica, ambas realizadas em passos da abordagem ECO. O resultado será uma wordnet de grande cobertura que, ao mesmo tempo, será suficientemente flexível para permitir ao utilizador escolher a porção que deseja utilizar, através da aplicação de um ponto de corte na confiança – a escolha por uma porção maior da wordnet englobará tendencialmente conteúdos com confianças mais baixas, enquanto que porções menores terão, em teoria, uma maior fiabilidade. As medidas de confiança podem ainda ser relevantes para a desambiguação do sentido das palavras (Navigli, 2009).

Apresentamos aqui o primeiro passo para a construção do novo recurso, nomeadamente a descoberta de grupos de sinónimos em que a pertença de cada palavra tem um valor associado, que deverá indicar a confiança relativamente à palavra transmitir o mesmo significado que as outras palavras no synset. Para calcular o valor da pertença, propomos tirar partido da redundância presente em redes de palavras relacionadas, obtidas a partir de diferentes fontes, nomeadamente dicionários e wordnets livres. No caso deste artigo, explorou-se para este fim a versão actual do CARTÃO (Gonçalo Oliveira et al., 2011), uma rede léxico-semântica extraída automaticamente a partir de três dicionários da língua portuguesa. No CARTÃO, as palavras estão relacionadas através de um conjunto de relações semânticas, ainda que os seus diferentes sentidos não sejam tratados. Sendo um synset um grupo de sinónimos, esta análise foca-se nas relações de sinonímia, ainda que não se descarte completamente a utilização de outros tipos. Assim, como numa rede de sinonímia as palavras estão ligadas através da relação de sinonímia, a identificação de aglomerados de palavras nestas redes (*clusters*) pode ser aproximada precisamente à descoberta de synsets.

Neste artigo, depois de descrevermos algum trabalho relacionado (secção 2), o que inclui a revisão de alguns algoritmos de *clustering* e de abordagens para descoberta de grupos de palavras relacionadas, propomos uma abordagem para a descoberta dos synsets difusos a partir de redes léxico-semânticas (secção 3). Apresentam-se depois os resultados da aplicação desta abordagem à rede CARTÃO, que resulta num conjunto de synsets difusos para o português (secção 4). Seguem-se alguns números relativos à avaliação dos resultados obtidos, automaticamente, contra os conteúdos de um tesauro referência, criado manualmente, e também através da classificação manual de pares de sinonímia. Os resultados obtidos são colocados lado-a-lado com aqueles obtidos através de uma abordagem anterior (Gonçalo Oliveira & Gomes, 2011) que tinha o mesmo objectivo e que também tinha sido aplicada ao CARTÃO (secção 5). Por fim, antes

de concluir, apresentam-se os resultados de uma nova experiência em que, para além de relações de sinonímia, as relações de hiperonímia também foram consideradas no cálculo das pertenças, o que levou a uma evolução positiva destes valores (secção 6).

2 Trabalho Relacionado

O principal objectivo do trabalho apresentado neste artigo é a identificação de agrupamentos (*clusters*) de sinónimos numa rede de palavras. Dadas as características da relação de sinonímia, estes *clusters* poderão depois ser aproximados a synsets. Para tal, pretende-se definir um algoritmo de *clustering* que, para calcular semelhanças, considere a estrutura da rede e, eventualmente, outras propriedades das palavras envolvidas (por exemplo, relações), que deverão ser tidas em conta no cálculo do valor das pertenças, a nossa confiança. A primeira parte desta secção descreve alguns dos algoritmos que ponderamos utilizar para atingir este objectivo. Na segunda parte, são apresentados alguns trabalhos em que abordagens de *clustering* foram aplicadas precisamente à descoberta de grupos de palavras sinónimas ou relacionadas, utilizadas para descrever conceitos.

2.1 Clustering em grafos

A tarefa de *clustering* tem como objectivo identificar, de forma automática e não supervisionada, agrupamentos de instâncias semelhantes, de acordo com um conjunto de dados a seu respeito e com uma métrica de semelhança sobre esses dados. Entre os vários algoritmos para esta tarefa (Xu & Wunsch, 2005), de acordo com o tipo de partição realizada, há três grandes grupos:

- *Clustering* hierárquico (*hierarchical clustering*): o resultado é uma partição hierárquica onde grupos de instâncias se organizam num estrutura em árvore, cuja raíz será um cluster com todas as instâncias e em que cada instância é uma folha;
- *Clustering* rígido (*hard clustering*): o resultado é uma partição rígida, em que cada instância está contida em um e um só *cluster*;
- *Clustering* difuso (*fuzzy clustering*): o resultado é uma partição em que a mesma instância pode pertencer a mais do que um *cluster*, com diferentes graus de pertença.

A nossa abordagem tem como requisito que o algoritmo actue sobre um grafo (de palavras), e que realize uma partição difusa, cujas pertenças sejam baseadas na confiança que há na associação das instâncias (palavras) aos *clusters* (synsets). De forma a escolher a abordagem a seguir na descoberta de synsets difusos, foi analisado um conjunto de algoritmos de *clustering*, que se apresentam de seguida.

O algoritmo Fuzzy C-Means (FCM) (Bezdek, 1981) é uma abordagem clássica para a descoberta de *clusters* difusos. É a variante difusa do algoritmo K-means (Hartigan & Wong, 1979) onde, dados k pontos aleatórios (centróides), classifica cada instância com a classe do centróide mais próximo. Este cálculo pode ser repetido por várias iterações, até haver convergência. No caso específico do FCM, cada instância pode pertencer a todos os clusters identificados, sendo o grau de pertença calculado com base na sua distância para os respectivos centróides.

Um tipo específico de algoritmos de *clustering* inclui aqueles que representam o domínio do problema como um grafo (ver Schaeffer (2007)), que será a forma óbvia de ver as redes de sinonímia. Ao contrário do FCM, em que é necessário indicar o número de *clusters* pretendidos e a sua posição inicial, nos algoritmos de clustering sobre grafos, o número de *clusters* vai depender essencialmente da estrutura do grafo. Olhando especificamente para aqueles que foram aplicados a problemas no âmbito do processamento de linguagem natural (PLN), destacamos o Markov Clustering (MCL) e o Chinese Whispers (CW), ambos baseados em passeios aleatórios pelo grafo (vulgo, *random walks*).

O MCL (van Dongen, 2000) parte da ideia que os caminhos aleatórios tendem a concentrar-se dentro de um mesmo subgrafo denso e não a saltar entre diferentes subgrafos através de ligações esparsas. O CW (Biemann, 2006) é uma variante do MCL que simplifica as operações do algoritmo anterior, sendo por isso mais eficiente. Inicialmente, para um grafo não direccionado com ou sem pesos, é atribuída uma classe distinta a cada nó. Depois, a cada iteração, os nós podem assumir a classe do vizinho que lhe transmitir maior força, o que se repete até haver estabilidade.

Outro algoritmo também aplicado a problemas de PLN é o Clustering by Committee (CBC, Lin & Pantel (2002)). Este algoritmo começa por encontrar conjuntos de instâncias designados por comités (*committees*), dispersos no espaço. Cada comité é constituído por instâncias que pertencem necessariamente a uma classe, que o comité acaba por definir. As restantes instâncias são de-

pois associadas aos comités mais próximos. Sempre que é realizada uma associação, são removidas todas as características comuns entre os comités e as instâncias que lhe foram associadas, o que permite que nas iterações seguintes essas instâncias possam ser associadas a outros comités.

Nesta análise, verificámos que nenhum dos algoritmos analisados ia ao encontro dos nossos objectivos. Por exemplo, apesar de ter uma utilização bastante generalizada, o FCM requer que o número de *clusters* seja um parâmetro dado inicialmente, quando o que pretendemos é que esta decisão seja tomada de forma automática pelo algoritmo. Para além disso, a posição inicial dos centróides é aleatória, o que torna o algoritmo não determinístico.

Dado que o nosso domínio são redes lexicais, faria sentido optar por um algoritmo que actue sobre grafos e que tire partido da sua estrutura. No entanto, nenhum dos dois algoritmos analisados dentro desta categoria descobre clusters difusos, e nem sequer permite que uma instância pertença a mais do que um cluster. Mesmo quando há nós instáveis, uma decisão acaba por ser tomada relativamente à sua pertença a um cluster que, devido aos caminhos aleatórios, pode não ser sempre o mesmo em diferentes iterações. Ou seja, nem o MCL nem o CW são determinísticos, ainda que o problema seja minimizado em grafos pesados e de maior dimensões (Biemann, 2006). Dada a sua relevância para este trabalho, acrescentamos ainda que, sob o ponto de vista da complexidade temporal, o CW é apresentado como uma variante mais eficiente do MCL (Biemann, 2006), precisamente por ser mais agressivo e considerar apenas o vizinho que transmite mais força e não os restantes. Isto reflecte-se numa complexidade temporal de $\mathcal{O}(s.|E|)$ para o CW, enquanto que para o MCL é $\mathcal{O}(s.|V|^2)$, em que s é o número de iterações, $|E|$ é o número de arcos e $|V|$ o número de vértices do grafo.

Sobre o CBC, que será determinístico, não foi desenhado para operar sobre grafos, ainda que uma adaptação seja possível. No entanto, acaba por sofrer de outros problemas semelhantes. Além disso, apesar de permitir a associação da mesma instância a vários *clusters*, após a sua associação a um comité, são removidas da instância todas as características em comum com esse comité, sendo a verdadeira semelhança com outros comités corrompida.

Apesar de nenhum destes algoritmos satisfazer os nossos requisitos, a abordagem que propomos na secção 3 acaba por combinar características dos algoritmos aqui revistos.

2.2 Descoberta de grupos de palavras

A tarefa de desambiguação do sentido das palavras (em inglês, *word sense disambiguation*) (Navigli, 2009) tem como objectivo associar a ocorrência de uma palavra, em contexto, ao seu sentido mais adequado, dentro de um repositório de sentidos (por exemplo, um dicionário). Para o inglês, é comum utilizar-se a WordNet (Fellbaum, 1998) ou, de forma a cobrir mais conhecimento sobre o mundo, uma extensão deste, como a BabelNet (Navigli & Ponzetto, 2012).

Uma tarefa próxima, é a indução dos sentidos das palavras (em inglês, *word sense induction*) (Nasiruddin, 2013). Aí, não existe um repositório e os sentidos são descobertos de forma normalmente não supervisionada, através da análise de semelhanças entre palavras, tendo em conta os contextos em que ocorrem e as relações em que estão envolvidas.

O nosso trabalho está ligado à indução do sentido das palavras, porque queremos identificar precisamente os sentidos possíveis de cada palavra e os sinónimos de cada um, de forma automática, recorrendo simplesmente a uma rede léxico-semântica, onde as palavras são identificadas apenas pela sua ortografia e classe gramatical e não existe divisão entre sentidos. Há alguma relação com o trabalho de Lin & Pantel (2002), em que o algoritmo CBC foi usado para descobrir conceitos, representados através de palavras que co-ocorrem frequentemente em texto e partilham um conjunto de relações sintácticas. Por isso, estes agrupamentos vão para além de grupos de sinónimos. Considere-se por exemplo o conceito com melhor qualidade descoberto por Lin & Pantel (2002), "arma de fogo", que inclui as seguintes palavras: *handgun, revolver, shotgun, pistol, rifle, machine gun, sawed-off shotgun, submachine gun, gun, automatic pistol, ...*. Para além do CBC, o algoritmo MCL foi também utilizado para detectar ambiguidades (Dorow et al., 2005). Mais precisamente, ao extrair uma rede de co-ocorrências a partir de um texto, as palavras ambíguas correspondem a vértices mais instáveis, ou seja, que ligam dois subgrafos densos.

Enquanto que os trabalhos anteriores exploram texto corrido, há outros que, tal como nós, usam redes de sinonímia extraídas precisamente de dicionários para identificar grupos de palavras sinónimas. Por exemplo, Gfeller et al. (2005) propõem uma forma de solucionar uma limitação do algoritmo MCL: não permitir que uma palavra seja incluída em mais do que um cluster. Para tal, o MCL é executado várias vezes, com ruído estocástico aleatório, de forma a identificar em que diferentes *clusters* os vértices mais instáveis

da rede aparecem. Estes vértices corresponderão, mais uma vez, a palavras ambíguas ou que poderão necessitar de ser desambiguadas. Um procedimento inspirado no anterior foi também aplicado ao português, na descoberta automática de synsets (Gonçalo Oliveira & Gomes, 2010). Contudo, este procedimento, que poderá aumenta o não-determinismo do MCL, resultou numa maioria de synsets demasiado grandes para que tivesse utilidade efectiva.

A ideia de utilizar conjuntos difusos para representar conceitos também não é nova. A nosso ver, ela vai ao encontro da ideia de que os sentidos das palavras não têm fronteiras muito bem definidas (Kilgarriff, 1996). Neste âmbito, Velldal (2005) apresenta uma abordagem para descobrir, a partir de texto corrido, conjuntos de palavras que podem ajudar a descrever conceitos, e em que as suas semelhanças contextuais são usadas como graus de pertença. O resultado é que, dada uma palavra (por exemplo, *cavalo*), é possível observar diferentes conjuntos difusos, cada um correspondente a um dos seus possíveis sentidos (por exemplo, meio de transporte – *carro (0.97), autocarro (0.80), barco (0.72), ...* – ou animal – *pássaro (0.86), cão (0.83), gato (0.80), ...*). Há também quem tenha atribuído graus de pertença de palavras a synsets, com base em vários julgamentos humanos (Borin & Forsberg, 2010).

No que diz respeito à criação de uma wordnet com medidas de confiança associadas, que é o nosso objectivo a longo prazo, existe para a língua inglesa trabalho na extensão da WordNet para outros domínios através de associações difusas (Araúz et al., 2012). Isto inclui não só um grau de pertença das palavras a synsets, mas também um valor difuso para o estabelecimento de relações entre synsets.

Num trabalho anterior, aplicamos uma abordagem simplista na descoberta de synsets difusos para o português, também a partir de redes de sinonímia extraídas de dicionários (Gonçalo Oliveira & Gomes, 2011). O algoritmo aplicado assume que cada palavra é um *cluster* potencial, que pode atrair nós semelhantes. Para obter as pertenças, é calculado o cosseno entre cada palavra e cada uma das outras, representadas pelo seu vector na matriz de adjacências da rede, que tem 1 nas adjacências e 0 nas restantes palavras. Esta foi a primeira abordagem para a descoberta automática de synsets difusos para o português que, contudo, originou mais uma vez, em média, synsets demasiado grandes, cuja utilização seria impraticável, pelo menos sem a aplicação de um ponto de corte, que se tornou obrigatório, e cujas pertenças nem sempre faziam muito sentido.

Após analisar melhor a abordagem, identificamos uma das causas do último problema, que seria a utilização dos vectores de adjacência completos, ao calcular o cosseno. Estando perante uma matriz esparsa, a maior parte das entradas é 0, ou seja, a pertença de palavras com muitas ligações (muito ambíguas ou com muitos sinónimos) é penalizada perante as outras, por as primeiras terem menos entradas nulas. Para além disso, apesar da abordagem anterior permitir a exploração de várias fontes de sinoníma, ela acabava por não explorar suficientemente a redundância de informação para reforçar a decisões tomadas. Um dos objectivos da abordagem proposta neste artigo é também melhorar o trabalho anterior. Assim, para além da realização de avaliações automática e manual, sempre que possível, foi feita uma comparação com os resultados obtidos anteriormente.

3 Abordagem Proposta

Como nenhum dos algoritmos revistos vai ao encontro dos nossos requisitos, propomos uma abordagem que combina características de mais do que um algoritmo. Para descobrir um conjunto de synsets difusos a partir de uma rede léxico-semântica, a abordagem proposta tem dois passos principais:

1. Identificação de um conjunto de centróides, onde as palavras já têm uma ligação forte e partilham semelhanças;

2. Cálculo dos graus de pertença, com base na proximidade de cada palavra aos centróides.

No nosso caso, os centróides são nada mais nada menos que *clusters* base, identificados a partir da estrutura do grafo e onde não há sobreposição. De certa forma, podem ser vistos como uma estrutura inicial, tal como os comités no CBC, que será numa segunda fase aumentada. Para a sua identificação, contudo, deve ser utilizado um algoritmo eficiente que tire partido da estrutura do grafo, tal como o CW.

No segundo passo, os graus de pertença de cada palavra são calculados com base na semelhança entre as características (palavras relacionadas) da palavra que são relevantes para o *centróide* e as palavras do próprio centróide, o que de certa forma se assemelha ao cálculo das pertenças no FCM. No entanto, não será necessário realizar novas iterações, precisamente porque cada *centróide* já incluirá palavras com um elevado grau de proximidade.

Formalizando, a abordagem proposta é aplicada a uma rede de sinonímia $G = (P, R)$, onde P é o conjunto de palavras e R é o conjunto de pares de sinonímia. A rede G pode ser representada através de uma matriz de adjacências $A(|P| \times |P|)$, onde $A_{ij} = \omega_{ij}$, um peso que reflecte o número de vezes que um par de sinónimos, $R(P_i, P_j)$, ocorre nas fontes utilizadas (dicionários, por exemplo). O peso máximo, m, é portanto uma constante, igual ao número de fontes utilizadas.

No primeiro passo, o algoritmo de *clustering* aplicado resulta num conjunto de clusters centróide C. No segundo, o valor de pertença da palavra P_i ao centróide C_k, $\mu(P_i, C_k)$, é calculado através da equação 1, onde T é o conjunto de palavras relevantes para o cálculo, ou seja, todas as palavras do centróide C_k e a palavra P_i, que pode ou não estar no centróide (ver equações 2). A multiplicação do denominador por m serve apenas para normalizar o valor da pertença no intervalo $[0-1]$. No final, o número de synsets difusos é igual ao número de clusters base.

$$\mu(P_i, C_k) = \frac{\sum_{j=0}^{|C_k|} A_{i[C_kj]}}{m \times |T|} \quad (1)$$

$$\begin{aligned} &T = \{C_k \cup P_i\}, \text{ou seja} \\ &|T| = \{|C_k|, P_i \in C_k\} \vee \{|C_k| + 1, P_i \notin C_k\} \end{aligned} \quad (2)$$

A abordagem é ilustrada com auxílio do grafo na figura 1, centrado na palavra *banana*. Em português europeu, esta palavra tanto pode ser o nome de uma fruta, como pode ter o sentido figurado de uma pessoa sem iniciativa. Suponha-se que o grafo é extraído a partir de três dicionários ($m = 3$) e que o algoritmo CW identifica os dois *clusters* centróide representados na tabela 1.

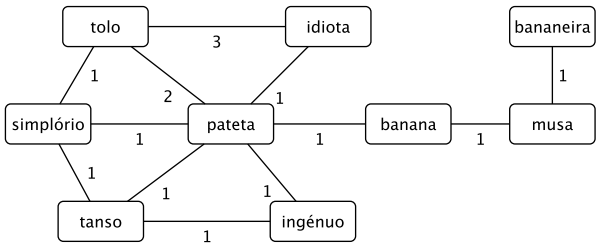


Figura 1: Rede de sinonímia com palavras e pesos das ligações.

Para calcular o valor de pertença de *banana* ao centróide C_A, devem ser consideradas as ligações às palavras *musa* e *bananeira*, ou seja, apenas 1. Este número é dividido por $3 \times |T|$, em que T incluir as palavras relevantes, $T = \{musa, bananeira, banana\}$. Portanto,

C_A	*musa*, *bananeira*
C_B	*banana*, *pateta*, *idiota*, *tolo*, *simplório*, *tanso*, *ingénuo*

Tabela 1: Centróides descobertos a partir da rede da figura 1, com o algoritmo Chinese Whispers.

$\mu(banana, C_A) = \frac{1}{9}$. Para o cálculo da pertença da palavra *banana* ao centróide C_B, as características relevantes são o número de ligações com todas as palavras do centróide C_B, apenas 1, para a palavra *pateta*. Considera-se ainda que cada palavra tem o número máximo de "ligações" a si própria, por isso, neste caso, como $banana \in C_B$, soma-se 3 ao número de ligações relevantes. Ou seja, o numerador será 4, e assim, $\mu(banana, C_B) = \frac{4}{21}$.

C'_A	*bananeira* (0.666), *musa* (0.666), *banana* (0.111)
C'_B	*pateta* (0.476), *tolo* (0.428), *idiota* (0.333), *simplório* (0.285), *tanso* (0.285), *ingénuo* (0.285), *banana* (0.190), *musa* (0.041)

Tabela 2: Synsets difusos com pertenças calculadas com base nos clusters discretos da tabela 1

4 Descoberta de synsets difusos para o português

Esta secção apresenta os resultados da aplicação da abordagem proposta à rede léxico-semântica CARTÃO, que se começa por descrever, seguida de uma visão numérica dos resultados e, por fim, de exemplos ilustrativos, com alguns dos synsets difusos obtidos.

4.1 Rede Léxico-Semântica

A rede léxico-semântica utilizada para a descoberta de synsets difusos foi o CARTÃO (Gonçalo Oliveira et al., 2011), disponível gratuitamente, e extraída de forma automática a partir de três dicionários da língua portuguesa, com base em padrões textuais nas suas definições. Para ajudar a caracterizar esta rede, a tabela 3 apresenta algumas das suas propriedades numéricas, mais propriamente para os sub-grafos de sinonímia entre substantivos (N), verbos (V), adjetivos (Adj) e advérbios (Adv).

Para cada sub-grafo, é indicado o número de vértices (palavras, $|P|$) e arestas distintas (relações de sinonímia, $|R|$), o grau médio dos vértices ($\overline{deg}(P)$) – equação 3 calcula o grau de um vértice – o coeficiente médio de *clustering* ($\overline{CC}$) – equação 4 calcula este coeficiente para

um vértice, sendo que $viz(P_i)$ representa o conjunto dos vizinhos do vértice P_i – o número de componentes conectadas[1] ($|Comp|$), e o número de palavras da componente maior ($|P|_{mc}$).

$$deg(P_i) = |R(P_i, P_j)| : P_i, P_j \in P \quad (3)$$

$$CC(P_i) = \frac{2.|R(P_j, P_k)|}{|viz(P_i)|.(|viz(P_i)| - 1)} : P_j, P_k \in viz(P_i) \quad (4)$$

Tal como outros investigadores (por exemplo, Gfeller et al. (2005)), também nos apercebemos que estes sub-grafos, extraídos de dicionários, são constituídos por uma grande componente, e várias pequenas. Além disso, os coeficientes de *clustering* são comparáveis aos de outras redes de pequeno mundo (em inglês, *small-world networks*), em que a distância média entre dois vértices é curta. Comparando os três sub-grafos, o sub-grafo dos verbos possui um grau médio mais elevado, o que significa que os verbos terão mais sinónimos e/ou serão mais ambíguos e/ou vagos. Também se observa que o sub-grafo de advérbios é significativamente mais pequeno que os demais, por isso acabou por não ser utilizado nas experiências apresentadas nas próximas secções.

4.2 Propriedades dos synsets descobertos

Ao correr o algoritmo proposto no CARTÃO, obtemos um conjunto com quase 15 mil synsets difusos C', a que chamamos CLIP 2.0, com as propriedades apresentadas na tabela 4 para cada categoria gramatical, nomeadamente: número de palavras (#pals), média de sentidos por palavra ($\overline{sents}$), palavra com mais sentidos (max(#sents)), total de synsets (#synsets), média de palavras por synset ($\overline{|synset|}$), synsets com apenas duas palavras ($|synset| = 2$), synsets com mais de 25 palavras ($|synset| > 25$), e tamanho do maior synset (max($|synset|$)). Na mesma tabela, incluem-se as mesmas propriedades para a nossa abordagem anterior, onde foi utilizada uma versão anterior do CARTÃO (originalmente Padawik (Gonçalo Oliveira & Gomes, 2011), depois rebaptizado como CLIP (Gonçalo Oliveira, 2013)), e onde, durante a geração original, foi aplicado um ponto de corte sobre as pertenças (θ) de 0,01. Ainda na mesma tabela, apresentam-se as propriedades dos synsets no tesauro TeP 2.0 (Maziero et al., 2008), criado manualmente para o português do Brasil.

Quando comparado com o CLIP 1.0, parece haver menos ruído, mesmo sem a aplicação de nenhum ponto de corte no CLIP 2.0. Isto porque existem menos synsets, em média mais pequenos para os nomes e adjetivos, e de tamanho comparável para os verbos. As palavras são também menos ambíguas. No TeP, o número médio de palavras por synset é mais baixo, tal como o número médio de sentidos por palavra, o que já era esperado, não só pelo TeP ter sido criado manualmente, mas também devido à nossa abordagem difusa, e pelo maior grau de cobertura do nosso tesauro. Recordamos, no entanto, que pode ser aplicado um ponto de corte às pertenças dos synsets difusos, de modo que estes fiquem pequenos e, tendencialmente, mais confiáveis. Por outro lado, no TeP o número de synsets de verbos e adjetivos é mais do dobro, e ligeiramente mais baixo para os substantivos. No entanto, os nossos synsets cobrem quase o dobro das palavras do TeP (cerca de 70 mil contra 40 mil), mais propriamente um número próximo de verbos, ligeiramente superior de adjectivos, e mais do dobro de substantivos. O número inferior de synsets de verbos e adjetivos pode, por um lado, indicar que o CLIP 2.0 não cobre tantos sentidos quanto o TeP mas, por outro, que o CLIP 2.0 agrupará significados mais próximos, que muitas vezes nem fará sentido separar. Esta capacidade está relacionada com a chamada "ambiguação" do sentido das palavras (Dolan, 1994).

4.3 Alguns resultados

A tabela 5 ilustra os resultados obtidos através de uma selecção manual de palavras polissémicas da língua portuguesa e alguns dos synsets difusos que as incluem, organizados de acordo com o conceito que transmitem (frequentemente clarificado pela palavra com maior pertença) e onde as palavras são apresentadas por ordem decrescente do grau de pertença. Numa observação global, tanto a constituição dos synsets como os graus de pertença parecem fazer sentido.

5 Avaliação

Nesta secção, os resultados obtidos são avaliados, primeiro através da sua confirmação no TeP, aqui usado como recurso dourado por ter sido criado manualmente e, segundo, manualmente. Os resultados de cada avaliação são comparados com os mesmos resultados obtidos para o CLIP 1.0.

[1] Uma componente de um grafo é um subgrafo no qual todos os pares de vértices estão ligados através de pelo menos um caminho, sem que estejam ligados a mais nenhum vértice do grafo.

POS	$\|P\|$	$\|R\|$	$\overline{deg}$(G)	$\overline{CC}$	$\|Comp\|$	$\|P_{mc}\|$
N	43,724	65,127	2.98	0.21	5,812	28,734
V	10,380	26,266	5.06	0.25	362	9,549
Adj	31,014	17,368	3.57	0.23	2,049	12,343
Adv	1,271	1,296	2.04	0.18	160	819

Tabela 3: Propriedades dos sub-grafos de cada categoria gramatical na rede CARTÃO.

	Cat	**Palavras**			**Synsets**				
		#pals	$\overline{sents}$	max(#sents)	#synsets	$\overline{\|synset\|}$	$\|synset\| = 2$	$\|synset\| > 25$	max($\|synset\|$)
	N	43.721	1,92	42	9.881	8,49	4.147	632	554
CLIP 2.0	**V**	10.380	3,15	54	1.438	22,76	289	370	500
	Adj	17.368	2,28	44	3.571	11,.07	1.530	367	322
	N	39.354	7,78	46	20.102	15,23	3,885	3,756	109
CLIP 1.0	**V**	11.502	14,31	42	7.775	21,17	307	2,411	89
($\theta = 0,01$)	**Adj**	15.260	10,36	43	8.896	17,77	1,326	2,157	109
	N	17.158	1,71	21	8.254	3,56	3.079	0	21
TeP 2.0	**V**	10.827	2,08	41	3.978	5,67	939	48	53
	Adj	14.586	1,46	19	6.066	3,50	3.033	19	43

Tabela 4: Propriedades numéricas dos synsets.

Além disso, procuramos validar os graus de pertença através da observação do seu valor para pares de sinonímia confirmados/correctos e não confirmados/incorrectos.

5.1 Comparação com um tesauro criado manualmente

Como o TeP é um tesauro criado manualmente para o português, temos alguma confiança nos seus conteúdos. Para além disso, foi desenvolvido de forma completamente independente do CARTÃO. Daí ter sido o TeP a nossa primeira opção para verificar a qualidade dos synsets descobertos.

Para facilitar a comparação, transformou-se cada conjunto de synsets descobertos num conjunto de pares de sinonímia, que seriam depois confirmados no TeP. Considera-se que um par de sinonímia, $R(w_a, w_b)$, é um conjunto de duas palavras que pertencem ao mesmo synset C_x, ou seja, $R(w_a, w_b) \rightarrow \exists C_x : w_a \in C_x \wedge w_b \in C_x$. Então, para cada par presente nos tesauros descobertos, verificou-se se existia pelo menos um synset no TeP que contivesse as duas palavras. A tabela 6 apresenta a proporção de pares confirmados para os synsets de cada categoria gramatical, não só para os resultados da abordagem actual, mas também para o CLIP 1.0.

Como, considerando todos os pares, a proporção de pares confirmados é muito baixa, na mesma tabela apresenta-se a evolução desse número para diferentes pontos de corte aplicados às pertenças (θ) – ao aplicar um ponto de corte, descartam-se de cada synset todas as palavras cuja pertença é inferior ao valor do corte. Nomeadamente, para os pontos de corte 0,105, 0,225 e 0,510, é apresentado: o número de pares do tesauro (Total); o número de pares com ambas as palavras no TeP (PalavrasNoTeP) e respectiva proporção relativamente ao número total; e o número de pares confirmados no TeP (ParNoTeP) e respectiva proporção relativamente ao número de pares com palavras cobertas. Ainda a este respeito, a figura 2 mostra, para o CLIP 1.0 e 2.0, a evolução das proporções PalavrasNoTeP (Palavras) e ParNoTeP (Pares). Para referência, o TeP inclui 51.533 pares de substantivos, 89.456 de verbos, e 51.645 de adjectivos.

É possível verificar que, tal como esperado numa medida de confiança, a proporção de pares confirmados aumenta para pontos de corte mais elevados, quer para o CLIP 1.0 como para o CLIP 2.0. No entanto, para cortes superiores a 0,6, mais de 90% dos pares do CLIP 2.0 são confirmados, enquanto que o CLIP 1.0 nunca chega a 80% de pares confirmados. Curioso também é a proporção de pares com ambas as palavras no TeP, que desce de forma mais consistente no CLIP 1.0 do que no 2.0. Aliás, a partir de um certo ponto, o CLIP 2.0 modifica mesmo a sua tendência e o número de pares desse tipo deixa de descer. Ou seja, se tomarmos o TeP como referência absoluta, estes números levam-nos a crer que a abordagem aqui proposta não só resulta em synsets mais coerentes, mas a uma medida de confiança mais fiel.

No entanto, apesar de mais confiável que um recurso criado de forma automática, o TeP está longe de ser uma referência absoluta. Aliás, TeP e CARTÃO são recursos, até certo ponto, complementares, não só relativamente a lemas, mas também a pares de sinonímia (veja-se a comparação realizada em Gonçalo Oliveira et al. (2011)). Para além de ter sido criado de forma manual, o TeP foca-se no português do Bra-

Palavra	Conceito	Synsets difusos
pasta	mistura	*mistura*(0.333), *amálgama*(0.127), *mescla*(0.111), *matalotagem*(0.079), *anguzada*(0.079), *comistão*(0.079), *misto*(0.079), *landoque*(0.079), *salsada*(0.0758), *confusão*(0.0758), *enovelamento*(0.063), *cacharolete*(0.063), *macedónia*(0.063), *mexedura*(0.063), *caldeação*(0.063), *mixagem*(0.063), *pasta*(0.063), *angu*(0.063), *amalgamação*(0.063), *comistura*(0.063), *impurezas*(0.063), *mistão*(0.063), *estricote*(0.063), *fusão*(0.045), *temperamento*(0.03), *pot-pourri*(0.015), *imisção*(0.015), *cocktail*(0.015), *ensalsada*(0.015), *envolta*(0.015), *agrupamento*(0.015), *baralha*(0.015), *marinhagem*(0.015), *salgalhada*(0.015), *misturada*(0.015), *miscelânea*(0.015), *têmpera*(0.015), *imperfeição*(0.015), *conjunto*(0.015), *combinação*(0.015), *logro*(0.015), ...
	dinheiro	*dinheiro*(0.28), *bufunfa*(0.069), *caroço*(0.053), *tutu*(0.042), *pataco*(0.037), *bagalhoça*(0.037), *guines*(0.037), *cobre*(0.032), *pecúnia*(0.032), *gaita*(0.032), *cacique*(0.032), *pílula*(0.026), *morubixaba*(0.026), *pila*(0.026), *cacau*(0.026), *arame*(0.026), *calombo*(0.026), *patacaria*(0.026), *gimbo*(0.026), *maco*(0.026), *bubão*(0.026), *chelpa*(0.026), *roço*(0.026), *levação*(0.026), *íngua*(0.026), *vénus*(0.021), *verdinha*(0.021), *mondrongo*(0.021), *pírula*(0.021), *dindim*(0.021), *trocado*(0.021), *curaca*(0.021), *pataca*(0.021), *massaroca*(0.021), *bagalho*(0.021), *carcanhol*(0.021), *pilim*(0.021), *encórdio*(0.021), *teca*(0.021), *coronel*(0.021), *matambira*(0.021), *mussuruco*(0.021), *cinco-réis*(0.021), *metal*(0.021), *cunques*(0.021), *janda-cruz*(0.021), *boro*(0.021), *cum-quibus*(0.021), *bilhestres*(0.021), *calique*(0.021), *parrolo*(0.021), *zerzulho*(0.021), *caronha*(0.021), *nhurro*(0.021), *baguines*(0.021), *pecuniária*(0.021), *pecunia*(0.021), *marcaureles*(0.021), *china*(0.021), *fanfa*(0.021), *dieiro*(0.021), *influente*(0.021), *guino*(0.021), *grana*(0.02), *tostão*(0.01), *riqueza*(0.01), ...
planta	vegetal	*vegetal*(0.667), *plantas*(0.667), *planta*(0.111)
	plano	*plano*(0.379), *projecto*(0.23), *tenção*(0.207), *desígnio*(0.207), *traçado*(0.161), *propósito*(0.161), *intenção*(0.149), *pressuposto*(0.138), *intento*(0.138), *prospecto*(0.126), *desenho*(0.126), *planta*(0.126), *programa*(0.115), *traça*(0.0.115), *mente*(0.092), *risco*(0.089), *resolução*(0.089), *prospeto*(0.08), *arquitectura*(0.08), *ideia*(0.078), *pressuposição*(0.069), *traçamento*(0.069), *prepósito*(0.069), *presuposto*(0.069), *intuito*(0.067), *vista*(0.067), *alçado*(0.057), *planificação*(0.057), *design*(0.057), *pranta*(0.057), *esboço*(0.055), *planejamento*(0.045), *fundição*(0.046), *gizamento*(0.046), *caruru*(0.046), *aspecto*(0.044), *medida*(0.044), *fim*(0.044), *vontade*(0.044), *desejo*(0.044), ...
sede	centro	*centro*(0.6), *núcleo*(0.4), *sensório*(0.333), *foco*(0.333), *club*(0.267), *sede*(0.222), *âmago*(0.222), *meio*(0.167), *coração*(0.167), *metrópole*(0.111), *escol*(0.056), *pólo*(0.056), *clube*(0.056), *umbigo*(0.056), *cérebro*(0.056), *fundo*(0.056), *gema*(0.056), *cadeira*(0.056), *casco*(0.056), *aglomeração*(0.056), *grupo*(0.056), *empório*(0.056), *essência*(0.056), *casino*(0.056), ...
	secura	*sede*(0.429), *secura*(0.333), *sequidão*(0.286), *seda*(0.238), *cerdas*(0.19), *sieda*(0.19), *seeda*(0.19), *aridez*(0.083), *centro*(0.083), *cerda*(0.042), *foco*(0.042), *impassibilidade*(0.042), *mortalha*(0.042), *cadeira*(0.042), *núcleo*(0.042), *diocese*(0.042), *ambição*(0.042), *impaciência*(0.042), *banco*(0.042), *apetite*(0.042), *avidez*(0.042), *ânsia*(0.042), *insensibilidade*(0.042), *capital*(0.042), *polidipsia*(0.042), *luxo*(0.042), *frieza*(0.042), *seta*(0.042), *magreza*(0.042)
	impaciência	*impaciência*(0.533), *frenesi*(0.467), *rabujice*(0.267), *despaciência*(0.267), *farnesia*(0.267), *inquietação*(0.222), *sofreguidão*(0.167), *pressa*(0.167), *desespero*(0.111), *nervosismo*(0.111), *ansiedade*(0.111), *exaltação*(0.111), *cócegas*(0.111), *freima*(0.111), *freimaço*(0.056), *formigueiro*(0.056), *precipitação*(0.056), *agastamento*(0.056), *impertinência*(0.056), *sofreguice*(0.056), *sede*(0.056), *inguinação*(0.056), *ira*(0.056), *furor*(0.056), *excitação*(0.056), *prurido*(0.056), *fúria*(0.056),...
verde	cor verde	*verde*(0.274), *virente*(0.137), *verdejante*(0.137), *relvoso*(0.118), *gramíneo*(0.098), *esmeraldino*(0.098), *prásino*(0.098), *desassazonado*(0.098), *viridente*(0.098), *ervoso*(0.098), *verdoso*(0.098), *ecológico*(0.078), *dessazonado*(0.078), *graminoso*(0.078), *viridante*(0.078), *herboso*(0.078), *porráceo*(0.078), *viçoso*(0.055), *inoportuno*(0.037), *fresco*(0.037), *esverdeado*(0.037), ...
	amador	*inexperiente*(0.917), *noviço*(0.067), *novato*(0.067), *inexperto*(0.417), *novel*(0.267), *ingénuo*(0.267), *inocente*(0.267), *principiante*(0.133), *novo*(0.133), *viçoso*(0.133), *matumbo*(0.067), *incompetente*(0.067), *amador*(0.067), *verde*(0.067), *moço*(0.067), *bisonho*(0.067), *ingênuo*(0.067), ...
limpar	tornar limpo	*limpar*(0.262), *purificar*(0.126), *enxugar*(0.098), *expurgar*(0.066), *mundificar*(0.06), *desinfectar*(0.06), *purgar*(0.055), *secar*(0.055), *depurar*(0.049), *mirrar*(0.049), *lavar*(0.049), *descontaminar*(0.044), *despoluir*(0.038), *desinçar*(0.038), *virginizar*(0.038), *esburgar*(0.038), *dessecar*(0.038), *assear*(0.038), *luir*(0.038), *varrer*(0.038), *esmirrar*(0.033), *desensopar*(0.033), *desenxovalhar*(0.033), *absterger*(0.033), *tamisar*(0.027), *virginalizar*(0.027), *desparasitar*(0.027), *vassourar*(0.027), *desenxamear*(0.027), *emundar*(0.027), *desecar*(0.027), *desempestar*(0.027), *desenodoar*(0.027), *desenfarruscar*(0.027), *perlavar*(0.027), *detergir*(0.027), *achicar*(0.027), ...
	podar	*desramar*(0.778), *escamondar*(0.556), *mondar*(0.556), *limpar*(0.25), *petelar*(0.083), *desgalhar*(0.083), *derramar*(0.083), *alveitarar*(0.083), *carpir*(0.083), *capinar*(0.083), *corrigir*(0.083)
	peneirar	*joeirar*(0.533), *escrivar*(0.333), *utar*(0.267), *acrivar*(0.267), *outar*(0.267), *peneirar*(0.111), *limpar*(0.111), *tamisar*(0.056), *crivar*(0.056), *cirandar*(0.056), *brocar*(0.056)
	roubar	*ripar*(0.533), *bifar*(0.467), *ripançar*(0.4), *surrupiar*(0.267), *palmar*(0.267), *surripiar*(0.222), *furtar*(0.111), *limpar*(0.111), *pifar*(0.056), *raspar*(0.056), *arrancar*(0.056), *puxar*(0.056)
estimar	apreciar	*apreciar*(0.444), *valorar*(0.333), *estimar*(0.333), *avaliar*(0.333), *cotar*(0.222), *valorizar*(0.222), *admirar*(0.222), *ponderar*(0.19), *considerar*(0.143), *amar*(0.095), *discernir*(0.095), *julgar*(0.095), *equacionar*(0.048), *ustir*(0.048), *trutinar*(0.048), *estranhar*(0.048), *qualificar*(0.048), *apreçar*(0.048), *gostar*(0.048), *desfrutar*(0.048), *adular*(0.048), *conhecer*(0.048), *recensear*(0.048), *aquilatar*(0.048), *numerar*(0.048), *desejar*(0.048), *sentir*(0.048), *reputar*(0.048), ...
	avaliar	*avaliar*(0.625), *aquilatar*(0.375), *quilatar*(0.292), *apreçar*(0.292), *equacionar*(0.208), *almotaçar*(0.208), *conceituar*(0.208), *aderar*(0.208), *julgar*(0.185), *estimar*(0.148), *apreciar*(0.148), *pesar*(0.111), *conhecer*(0.111), *louvar*(0.111), *calcular*(0.111), *ajuizar*(0.074), *quantiar*(0.074), *aferir*(0.074), *computar*(0.074), *aperfeiçoar*(0.074), *ponderar*(0.074), *reputar*(0.074), *cotar*(0.037), *valorar*(0.037), *arbitrar*(0.037), *mensurar*(0.037), *qualificar*(0.037), *contrastar*(0.037), *orçar*(0.037), *montar*(0.037), *taxar*(0.037), *apurar*(0.037), *discernir*(0.037), *examinar*(0.037), *tomar*(0.037)

Tabela 5: Synsets difusos de palavras polissémicas no CLIP 2.0.

sil, e acaba por não cobrir várias palavras da língua portuguesa, nem alguns sentidos das palavras que inclui, principalmente aqueles menos comuns. Esta será mesmo a principal razão para a proporção de pares confirmados ser muito baixa quando não é aplicado qualquer ponto de corte.

5.2 Avaliação manual

Devido às limitações já referidas do TeP, decidimos efectuar uma avaliação adicional, desta vez manual, seguindo as mesmas regras que na avaliação feita ao CLIP 1.0, detalhada em Gonçalo Oliveira & Gomes (2011) e Gonçalo Oliveira (2013). Mais precisamente, esta avaliação passou pelas seguintes fases:

1. Remoção (automática) dos synsets de todas as palavras que não ocorrem nos corpos acessíveis a partir do serviço AC/DC (Santos & Bick, 2000);
2. Selecção (automática) apenas dos synsets onde todas as palavras têm uma frequência superior a 100, nos mesmos corpos;
3. Escolha (automática), de n pares de palavras, sendo que cada par tem duas palavras provenientes do mesmo synset;
4. Classificação manual de cada par como sinónimos (correto) ou não (incorreto).

Os dois primeiros passos foram feitos para tornar a avaliação mais rápida e focada em palavras conhecidas, por serem frequentes. No terceiro passo, optamos por gerar três conjuntos aleatórios: 150 pares de nomes, 150 pares de verbos e 150 pares de adjetivos. No quarto passo, cada par foi classificado por dois avaliadores humanos, de forma independente, a quem foi sugerido a consulta de dicionários na rede, em caso de dúvida. A tabela 7 apresenta os resultados obtidos por avaliador e a sua concordância κ, assim como os resultados da avaliação manual do CLIP 1.0, mas apenas para os nomes, tal como apresentada em Gonçalo Oliveira (2013). Apresentam-se ainda as médias das medidas de pertença dos pares classificados como correctos por ambos os avaliadores ($\overline{\mu_c}$), pares onde não houve concordância entre avaliadores ($\overline{\mu_d}$), e pares classificados como incorrectos por ambos ($\overline{\mu_i}$). Não nos foi possível recuperar os dados de avaliação manual do CLIP 1.0, o que não nos permite fazer a análise dos graus de pertença para a abordagem anterior.

Embora exista margem para melhorias, a proporção de pares correctos foi, uma vez mais, superior ao mesmo valor no CLIP 1.0. Nota-se que os verbos são a categoria com mais pares incorretos, provavelmente por ser também o sub-grafo com maior grau médio, ou seja, maior número de ligações por vértice (ver tabela 3), o que dará origem a mais confusão.

A média das pertenças de palavras em pares classificados como corretos ($\overline{\mu_c}$), incorretos ($\overline{\mu_{!c}}$) e discordantes ($\overline{\mu_d}$) têm um comportamento consistente para todas as categorias. Ou seja, o seu valor é mais elevado para os pares classificados como correctos por ambos os avaliadores, seguidos pelos pares em que não houve concordância e pelos pares classificados como incorrectos por ambos.

A título de exemplo, apresentam-se na tabela 8 alguns pares de palavras presentes no mesmo synset (Pal_1 e Pal_2), a pertença de cada uma ao synset (μ_1 e μ_2), e a classificação do par (sinónimos possíveis ou não?) por cada um dos avaliadores (Class_A e Class_B).

6 Utilização de relações de hiperonímia

Os resultados apresentados anteriormente constituíram a primeira experiência na aplicação da abordagem proposta a relações de sinonímia extraídas a partir de três dicionários da língua portuguesa. No entanto, relações de outros tipos podem também transmitir informação relevante no cálculo da pertença (confiança) das palavras a synsets. Nesta secção apresenta-se uma das primeiras experiências onde, para além da utilização das relações de sinonímia, da mesma forma que o anteriormente relatado, as relações de hiperonímia são também consideradas no cálculo da pertença. Este tipo de relação foi escolhido não só por ter muitas instâncias no CARTÃO (cerca de 115 mil, 95 mil distintas), mas principalmente por indicar uma generalização/especificação. Ou seja, hipónimos partilham um conjunto de características com os seus hiperónimos, e por isso podem considerar-se semanticamente próximos. Há mesmo várias medidas para o cálculo de similaridade semântica com base nestas relações, na WordNet (por exemplo, Resnik (1995) ou Leacock & Chodorow (1998)).

Há, no entanto, que distinguir casos em que, nos dicionários, a relação é apresentada como sendo de sinonímia (equivalência) de casos em que é apresentada como hiperonímia (digamos, semelhança). Como num synset todas as palavras devem partilhar um significado, a primeira deve ter mais peso. Além disso, a nosso ver, quando uma palavra não está numa relação de sinonímia com nenhuma das palavras de um synset, ela simplesmente não deve pertencer a esse

Cat	Corte (θ)	Pares CLIP 1.0 ($\theta = 0,01$) Total	PalavrasNoTeP	ParNoTeP	Pares CLIP 2.0 Total	PalavrasNoTeP	ParNoTeP
N	0.000	664.559	293.970 (44.2%)	25.893 (08.8%)	2.317.478	1.081.018 (46.6%)	30.407 (02.8%)
	0.105	126.287	27.251 (21.6%)	10.466 (38.4%)	279.882	126.422 (45.2%)	16.588 (13.1%)
	0.225	74.639	11.726 (15.7%)	6.061 (51.7%)	62.607	23.141 (37.0%)	7.331 (31.7%)
	0.510	51.698	5.296 (10.2%)	2.988 (56.4%)	7.012	1.362 (19.4%)	1.127 (82.7%)
V	0.000	399.614	241.886 (60.5%)	33.818 (14.0%)	1.385.293	1.008.012 (72.8%)	49.476 (04.9%)
	0.105	44.688	16.856 (37.7%)	7.839 (46.5%)	28.378	18.101 (63.8%)	8.654 (47.8%)
	0.225	21.019	6.614 (31.5%)	3.871 (58.5%)	5.443	3.353 (61.6%)	2.519 (75.1%)
	0.510	11.528	2.819 (24.5%)	1.587 (56.3%)	794	423 (53.3%)	375 (88.7%)
Adj	0.000	346.076	212.104 (61.3%)	222.96 (10.5%)	1.149.294	685.983 (59.7%)	27.902 (04.1%)
	0.105	52.005	21.211 (40.8%)	8.446 (39.8%)	33.296	17.044 (51.2%)	7.885 (46.3%)
	0.225	26.283	9.203 (35.0%)	4.927 (53.5%)	9.722	4.420 (45.5%)	3.128 (70.8%)
	0.510	16.222	4.643 (28.6%)	2.621 (56.5%)	2.319	822 (35.4%)	754 (91.7%)

Tabela 6: Confirmação de pares de sinonímia no TeP.

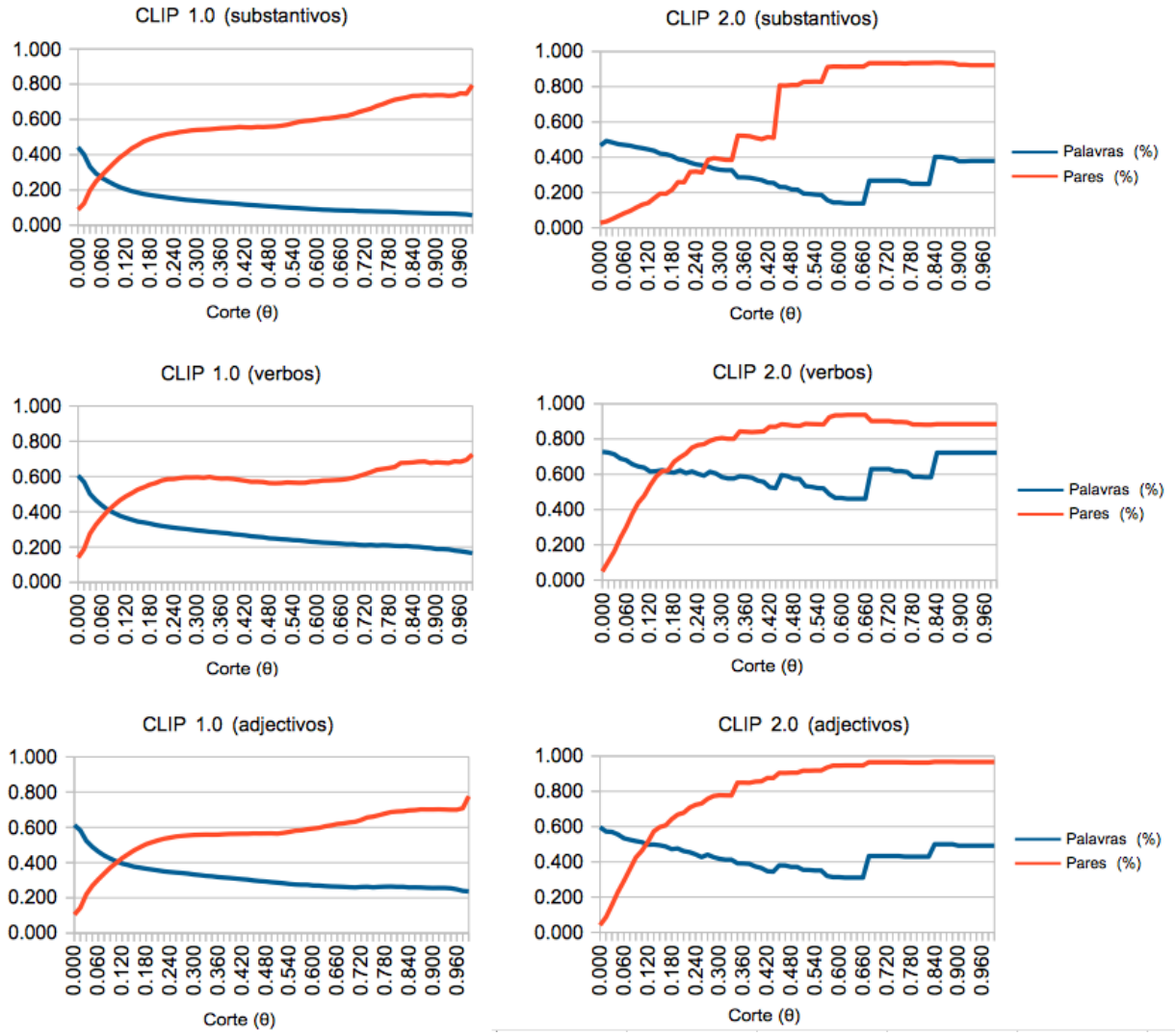


Figura 2: Proporção entre pares do CLIP 1.0 e CLIP 2.0 confirmados no TeP e pares com ambas as palavras cobertas pelo TeP.

Cat	CLIP 1.0 ($\theta = 0,01$) Correctos	κ	CLIP 2.0 Correctos	κ	$\overline{\mu_c}$	$\overline{\mu_d}$	$\overline{\mu_i}$
N	75.0%	0.43	84.7-88.0%	0.75	0.30	0.26	0.22
V	N/A	N/A	68.7-68.7%	0.65	0.25	0.18	0.17
Adj	N/A	N/A	74.7-77.3%	0.74	0.25	0.19	0.15

Tabela 7: Resultados da avaliação manual e média dos graus de pertença para cada classe de pares.

Class$_A$	**Class$_B$**	**Cat**	**Pal$_1$**	μ_1	**Pal$_2$**	μ_2
×	×	N	*sede*	0,429	*diocese*	0,042
✓	×	N	*necessidade*	0,055	*mando*	0,111
✓	✓	N	*vaqueiro*	0,555	*ganadeiro*	0,444
×	×	ADJ	*natal*	0,416	*patriótico*	0,066
✓	×	ADJ	*húmido*	0,055	*hídrico*	0,055
✓	✓	ADJ	*patriótico*	0,666	*nacionalista*	0,555
×	×	V	*vigiar*	0,083	*civilizar*	0,166
✓	×	V	*dormir*	0,133	*roncar*	0,583
✓	✓	V	*remodelar*	0,833	*reformular*	0,111

Tabela 8: Exemplos de pares de sinónimos e sua classificação manual pelos avaliadores.

synset, mesmo que seja um hipónimo ou hiperónimo de alguma. A relação de hiperonímia propriamente dita, estabelecida entre dois synsets (generalização e especialização) será integrada numa fase posterior deste trabalho.

Com base nas considerações anteriores, as relações de hiperonímia vão apenas aumentar a pertença em casos em que uma palavra está em relações de sinonímia com algumas das palavras do synset base, mas também em relações de hiperonímia com outras dessas palavras. Estes casos serão, acreditamos, situações em que, na própria linguagem, a utilização do hipónimo e do hiperónimo se confundem e acabam por ser usadas para referir o mesmo. Ao mesmo tempo, num dicionário que já inclua uma relação de sinonímia entre duas palavras, não serão consideradas relações de hiperonímia entre as mesmas palavras. Assim, o peso vindo de cada fonte nunca pode ser superior a 1. Devemos acrescentar que, como nos dicionários utilizados as relações de hiperonímia se estabelecem apenas entre substantivos, esta experiência foi aplicada somente a palavras desta categoria gramatical.

Para confirmar rapidamente que a consideração dos hiperónimos desta forma alterava as pertenças da forma desejada, aproveitamos os dados da avaliação manual anterior. A tabela 9 apresenta os valores médios das pertenças de pares de palavras do mesmo synset, primeiro, sem a consideração das relações de hiperonímia e, segundo, quando as relações de hiperonímia são consideradas com um peso que é metade dos das relações de sinonímia. Ou seja, para calcular a pertença, antes de aplicar a equação 1, a matriz de adjacências da rede, A, é alterada, de forma a que, sempre que haja uma relação de hiperonímia entre duas palavras $H(P_i, P_j)$, se também existir pelo menos uma relação de sinonímia, $R(P_i, P_j)$, a ligação entre as palavras é reforçada, $A_{ij} += 0.5$.

De forma a observar a evolução nas pertenças médias, a tabela 9 apresenta também a diferença entre o valor destas antes ($peso = 0,0$) e depois de considerar as relações de hiperonímia

Peso hiperonímia	$\overline{\mu_c}$	$\overline{\mu_d}$	$\overline{\mu_i}$
0,0	0,29960	0,26132	0,21957
0,5	0,30368	0,26234	0,22096
Diferença	0,00408	0,00102	0,00139
Ganho	0,01362	0,00390	0,00633

Tabela 9: Diferenças e ganhos nas pertenças médias de pares de sinonímia correctos, discordantes e incorrectos.

($peso = 0,5$), e mostra ainda o ganho em cada média (equação 5).

$$Ganho = \frac{Valor_{nova} - Valor_{anterior}}{Valor_{anterior}} \quad (5)$$

Como apenas os casos em que existiam relações de hiperonímia era valorizados, e não havia mais nenhuma alteração, os valores das pertenças ou se mantinham, ou aumentavam. Ou seja, o ganho seria zero ou positivo. Na tabela verifica-se que, apesar do ganho ser sempre positivo, é ligeiramente superior nos casos em que ambos os anotadores concordaram que as duas palavras do par eram sinónimos, ou seja, tornou o valor das pertenças um pouco mais fiel a uma medida de confiança.

Com base nos valores obtidos, decidimos começar a utilizar também as relações de hiperonímia no cálculo das pertenças aos synsets difusos. A título de exemplo, a tabela 10 apresenta três synsets difusos antes e depois de serem consideradas as relações de hiperonímia.

7 Conclusões e trabalho futuro

Com vista à descoberta de conceitos, descritos por conjuntos de palavras com pertenças variáveis, apresentamos uma nova abordagem para a descoberta de synsets difusos através de redes léxico-semânticas. Esta abordagem tira partido da redundância em redes extraídas a partir de várias fontes, neste caso dicionários, por isso o valor da pertença pode, de certa forma, quantificar a confiança na utilização da palavra para se referir ao conceito que emerge do synset.

Antes	**Depois**
ramada(0.67), *ramagem*(0.52), *rama*(0.52), *enramada*(0.29), *ramosidade*(0.24), *arramada*(0.19), *fronde*(0.19), *parreira*(0.13), *latada*(0.083), *França*(0.042), *ramaria*(0.042), ***folhagem*(0.042)**	*ramada*(0.67), *ramagem*(0.52), *rama*(0.52), *enramada*(0.29), *ramosidade*(0.24), *arramada*(0.19), *fronde*(0.19), *parreira*(0.13), *latada*(0.083), ***folhagem*(0.063)**, *França*(0.042), *ramaria*(0.042)
panfleto(0.83), *libelo*(0.83), *querela*(0.11), ***folheto*(0.11)**	*panfleto*(0.83), *libelo*(0.83), ***folheto*(0.17)**, *querela*(0.11)
apelido(0.46), ***nome*(0.46)**, ***alcunha*(0.40)**, *cognome*(0.31), ***epíteto*(0.23)**, ***sobrenome*(0.23)**, *designação*(0.17), *denominação*(0.17), *qualificação*(0.15), ...	***nome*(0.48)**, *apelido*(0.46), ***alcunha*(0.41)**, *cognome*(0.31), ***sobrenome*(0.25)**, ***epíteto*(0.24)**, *designação*(0.17), *denominação*(0.17), *qualificação*(0.15), ...

Tabela 10: Exemplos de synsets difusos com pertenças das palavras antes e depois de considerar as relações de hiperonímia.

A abordagem proposta diferencia-se de uma abordagem anterior para o mesmo fim por ser realizada em dois passos e por considerar apenas as adjacências relevantes para o cálculo das pertenças de cada palavra a um synset. Isto diminuiu o ruído e tornou o valor das pertenças mais facilmente interpretável, o que se confirma não só pela avaliação manual de ambas as abordagens, mas também pela comparação do valor das pertenças de diferentes pares de palavras. Como esperado numa medida de confiança, pares de palavras que devem estar no mesmo synset (sinónimos) têm em média uma pertença superior a pares que, de acordo com anotadores humanos, não são sinónimos.

Ainda assim, apesar dos resultados positivos, os valores da avaliação mostram que há ainda muita margem de melhoria. Por exemplo, enquanto cerca de 88% dos pares de substantivos pertencentes ao mesmo synset são efectivamente sinónimos, para os verbos, este número desce para 68%. Nos próximos passos a realizar neste âmbito, pretendemos realizar novas experiências para averiguar a melhor forma de considerar outros tipos de relação. Por exemplo, uma ideia a seguir é que palavras sinónimas devem estar relacionadas da mesma forma com as mesmas palavras (por exemplo, tanto *carro*, como *automóvel* devem ser hipónimos de *veículo* e ter como partes *roda* ou *motor*). Por outro lado, pretendemos aplicar esta abordagem a outras fontes de sinonímia, que permitirão não só ampliar o recurso, mas também reforçar a medida de confiança. Entre os recursos candidatos encontram-se outras wordnets livres, como o próprio TeP (Maziero et al., 2008), a OpenWN-PT (de Paiva et al., 2012) ou a PULO (Simões & Guinovart, 2014).

O recurso resultante deste trabalho será uma wordnet para a língua portuguesa, criada de forma automática, e em que haverá valores de confiança associados a algumas das decisões tomadas, incluindo não só a inclusão de palavras em synsets, como também o estabelecimento de relações entre synsets, que será uma das próximas fases do trabalho. Acreditamos que este recurso, a ser disponibilizado em breve, possa ser de grande utilidade para aqueles que procuram uma wordnet para o português em que o balanço entre cobertura e confiança possa ser personalizado de acordo com as necessidades da aplicação.

Apesar de ser possível realizar um exercício de alinhamento da versão actual do recurso a outra wordnet, uma prática cada vez mais comum, isso não é uma das nossas preocupações actuais, como não foi para o Onto.PT. Isto porque, a cada versão, não só os conteúdos, mas a própria estrutura do recurso podem ser substancialmente alterados. Por exemplo, para além da exploração de diferentes recursos, os vários passos da abordagem ECO podem ser implementados de forma diferente e levar a diferenças ao nível das fronteiras dos synsets e da granularidade dos sentidos de cada palavra. Ou seja, para cada nova versão, seria necessário realizar um novo alinhamento, quer devido à aplicação de diferentes implementações de cada passo da abordagem ECO, ou simplesmente à utilização de diferentes recursos. Para minimizar este trabalho, seria necessário definir um núcleo fixo de synsets que se manteriam estáveis de versão para versão, ou então esperar que o recurso atinja uma fase menos experimental.

Agradecimentos

Este trabalho foi parcialmente realizado no âmbito do projecto ConCreTe – *Concept Creation Technology*.

The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

Referências

Araúz, P. León, J. Gómez-Romero & F. Bobillo. 2012. A fuzzy ontology extension of wordnet and eurowordnet for specialized knowledge. Em *Proceedings of Terminology and Knowledge Engineering Conference* TKE 2012, Madrid, Spain.

Bezdek, James C. 1981. *Pattern recognition with fuzzy objective function algorithms*. Norwell, MA, USA: Kluwer Academic Publishers.

Biemann, Chris. 2006. Chinese Whispers: An efficient graph clustering algorithm and its application to natural language processing problems. Em *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing* TextGraphs-1, 73–80. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. Em *Proceedings of the 6th Global WordNet Conference* GWC 2012, 64–71.

Borin, Lars & Markus Forsberg. 2010. From the people's synonym dictionary to fuzzy synsets - first steps. Em *Proceedings of LREC 2010 workshop on Semantic relations. Theory and Applications*, 18–25. La Valleta, Malta.

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

van Dongen, Stijn Marinus. 2000. *Graph clustering by flow simulation*: University of Utrecht. Tese de Doutoramento.

Dorow, Beate, Dominic Widdows, Katarina Ling, Jean-Pierre Eckmann, Danilo Sergi & Elisha Moses. 2005. Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. Em *Proceedings of MEANING-2005, 2nd Workshop organized by the MEANING Project*, Trento.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database (language, speech, and communication)*. The MIT Press.

Gfeller, David, Jean-Cédric Chappelier & Paulo De Los Rios. 2005. Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. Em *Proceedings of International Symposium on Applied Stochastic Models and Data Analysis* ASMDA 2005, 106–113. Brest, France.

Gonçalo Oliveira, Hugo. 2013. *Onto.pt: Towards the automatic construction of a lexical ontology for portuguese*: University of Coimbra. Tese de Doutoramento. http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_PhdThesis2012.pdf.

Gonçalo Oliveira, Hugo, Leticia Antón Pérez, Hernani Costa & Paulo Gomes. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* 3(2). 23–38.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2010. Automatic creation of a conceptual base for Portuguese using clustering techniques. Em *Proceedings of 19th European Conference on Artificial Intelligence (ECAI 2010)*, 1135–1136. Lisbon, Portugal: IOS Press.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2011. Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. Em *Proceedings of 22nd International Joint Conference on Artificial Intelligence* IJCAI 2011, 1801–1806. Barcelona, Spain: IJCAI/AAAI.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2014. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation* 48(2). 373–393.

Gonçalo Oliveira, Hugo, Valeria de Paiva, Cláudia Freitas, Alexandre Rademaker, Livy Real & Alberto Simões. 2015. As wordnets do português. Em Alberto Simões, Anabela Barreiro, Diana Santos, Rui Sousa-Silva & Stella E. O. Tagnin (eds.), *Linguística, Informática e Tradução: Mundos que se Cruzam*, vol. 7(1) (OSLa: Oslo Studies in Language 1), 397–424. University of Oslo.

Hartigan, J. A. & M. A. Wong. 1979. A K-means clustering algorithm. *Applied Statistics* 28. 100–108.

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

Leacock, Claudia & Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. Em Christiane Fellfaum (ed.), *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, 265–283. Cambridge, Massachusetts: The MIT Press.

Lin, Dekang & Patrick Pantel. 2002. Concept discovery from text. Em *Proceedings of 19th International Conference on Computational Linguistics* COLING 2002, 577–583.

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

Nasiruddin, Mohammad. 2013. A state of the art of word sense induction: A way towards word sense disambiguation for under resourced languages. *TALN/RECITAL 2013* .

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2). 1–69.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250.

de Paiva, Valeria, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. Em *Proceedings of 24th International Conference on Computational Linguistics* COLING (Demo Paper), 353–359.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. Em *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* IJCAI'95, 448–453. San Francisco, CA, USA.

Santos, Diana & Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em *Proceedings of 2nd International Conference on Language Resources and Evaluation* LREC 2000, 205–210.

Schaeffer, Satu Elisa. 2007. Graph clustering. *Computer Science Review* 1(1). 27–64.

Simões, Alberto & Xavier Gómez Guinovart. 2014. Bootstrapping a portuguese wordnet from galician, spanish and english wordnets. *Advances in Speech and Language Technologies for Iberian Languages* 8854. 239–248.

Velldal, Erik. 2005. A fuzzy clustering approach to word sense discrimination. Em *Proceedings of 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Denmark.

Xu, Rui & D. Wunsch, II. 2005. Survey of clustering algorithms. *Transactions on Neural Networks* 16(3). 645–678. doi:10.1109/TNN.2005.845141.

Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus

Recognition of Terms in Spanish by Applying a Contrastive Approach

Olga Acosta
Departamento de Ciencias del Lenguaje
Pontificia Universidad Católica de Chile
oacostal@uc.cl

César Aguilar
Departamento de Ciencias del Lenguaje
Pontificia Universidad Católica de Chile
caguilara@uc.cl

Tomás Infante
Magíster en Procesamiento y Gestión de la Información
Pontificia Universidad Católica de Chile
tomasinfante@gmail.com

Resumen

En este artículo presentamos una metodología para la identificación y extracción de términos a partir de fuentes textuales en español correspondientes a dominios de conocimiento especializados mediante un enfoque de contraste entre corpus. El enfoque de contraste entre corpus hace uso de medidas para asignar relevancia a palabras que ocurren tanto en el corpus de dominio como en corpus de lengua general o de otro dominio diferente al de interés. Dado lo anterior, en este trabajo realizamos una exploración de cuatro medidas usadas para asignar relevancia a palabras con el objetivo de incorporar la de mejor desempeño a nuestra metodología. Los resultados obtenidos muestran un desempeño mejor de las medidas diferencia de rangos y razón de frecuencias relativas comparado con la razón *log-likelihood* y la medida usada en Termostat.

Palabras clave

Término, *unithood*, *termhood*, extracción terminológica, lenguaje especial

Abstract

In this article we present a methodology for identifying and extracting terms from text sources in Spanish corresponding specialized-domain corpus by means of a contrastive approach. The contrastive approach requires a measure for assigning relevance to words occurring both in domain corpus and reference corpus. Therefore, in this work we explored four measures used for assigning relevance to words with the goal of incorporating the best measure in our methodology. Our results show a better performance of rank difference and relative frequency ratio measures compared with *log-likelihood* ratio and the measure used by Termostat.

Keywords

Term, *unithood*, *termhood*, term extraction, special language

1 Introducción

Desde el punto de vista del aprendizaje de ontologías a partir de textos, el reconocimiento automático de términos (en inglés, ATR) se considera un prerrequisito para tareas más complejas como son, por ejemplo, la extracción de conceptos y taxonomías (Buitelaar et al., 2005). A grandes rasgos, un término es una representación lingüística de conceptos de dominio específico (Kageura & Umino, 1996; Pazienza, 1998; Vivaldi et al., 2001), y una terminología construida de forma coherente puede, por tanto, ser útil como plataforma básica para construir ontologías y además usada para otras aplicaciones importantes (diccionarios, traducción automática, indexación, tesauros, etc.). En este sentido Pazienza et al. (2005) señalan también el valor que tiene la extracción automática de términos como punto de partida para desarrollar sistemas inteligentes y con ello mitigar el cuello de botella en la adquisición de conocimiento. Los enfoques usados para la extracción automática de términos son los siguientes: i) lingüístico, ii) estadístico, iii) aprendizaje automático, y iv) métodos híbridos (Ananiadou & Mcnaught, 2005; Lossio-Ventura et al., 2014; Kockaert & Steurs, 2015). Por un lado, se han definido técnicas lingüísticas para filtrar candidatos no relevantes, por ejemplo, vía la configuración de patrones morfosintácticos, mientras que la parte estadística y/o pro-

This work is licensed under a
Creative Commons Attribution 3.0 License

babilística conlleva la aplicación de medidas estadísticas para asignar relevancia a términos candidatos. Por otro lado, los enfoques de aprendizaje automático (en inglés, ML) usan datos de entrenamiento para aprender rasgos útiles para la extracción de términos. Finalmente, los métodos actuales consisten en híbridos que incorporan algunos o potencialmente todos los enfoques anteriores para identificar y reconocer términos (Vivaldi & Rodríguez, 2007). Los enfoques actuales están basados preponderantemente en métodos probabilísticos y lingüísticos debido a que el principal reto en aprendizaje automático es seleccionar un conjunto de rasgos discriminantes que caractericen los términos (Lossio-Ventura et al., 2014), lo que representa una tarea compleja.

La enorme cantidad de información digital disponible y áreas de conocimiento que evolucionan rápidamente, como es el caso de la biomedicina (Kageura & Umino, 1996; Poesio, 2005; Ananiadou & Mcnaught, 2005), influyen directamente en el interés por mejorar los métodos actuales para la extracción automática de términos e implementarlos en sistemas computacionales con la meta de agilizar el trabajo de identificación y extracción del vocabulario de un dominio. Los rasgos o propiedades que caracterizan a los términos son *unithood* y *termhood*, tal como lo proponen Kageura & Umino (1996). Estos rasgos se han explorado en la literatura sobre la extracción automática de términos (Vivaldi et al., 2001; Ananiadou & Mcnaught, 2005; Kit & Liu, 2008; Barrón-Cedeño et al., 2009; Gelbukh et al., 2010; Spasić et al., 2013; Kockaert & Steurs, 2015). Asimismo, se han considerado otras cuestiones relevantes, como es el caso de la ambigüedad y variación de los términos (Daille et al., 1996; Spasić et al., 2013), que dependiendo de la aplicación para la que se realiza la extracción de terminología adquirirán mayor o menor importancia (Kockaert & Steurs, 2015). Con relación a las variantes de términos, tanto Ananiadou (1994) como Vivaldi & Rodríguez (2007) señalan que a los términos se les atribuyen rasgos de no ambigüedad y mono-referencialidad para designar conceptos en un dominio, sin embargo, esto dista mucho de la realidad debido a que los problemas de polisemia, homonimia y sinonimia ocurren con más frecuencia de lo esperado.

En este artículo presentamos una metodología para reconocer y extraer términos candidatos, tanto unipalabra como multipalabra. Nuestra propuesta consiste en un enfoque de comparación entre corpus para identificar las palabras relevantes del dominio analizado, así como asignarles una ponderación que refleje su relevancia. Para lograr lo anterior, comparamos cuatro medidas diferentes para calcular el *termhood* de cada palabra común al corpus de dominio y de referencia: razón de frecuencia relativa (Manning & Schütze, 1999), razón log-likelihood (Gelbukh et al., 2010), diferencia de rangos (Kit & Liu, 2008) y la medida usada en *TermoStat* (Drouin, 2003). Bajo este enfoque contrastivo de corpus, asumimos que una palabra estrechamente relacionada con el dominio debe tener una probabilidad de ocurrencia más alta en dicho dominio que en un corpus de referencia. Así, si este proceso de asignación de relevancia es eficaz, palabras de dominio tendrán ponderaciones mayores que palabras no relacionadas con el dominio. En una fase posterior, la relevancia de la palabra se puede usar para extraer términos candidatos multipalabra, de modo que las palabras con ponderaciones altas contribuirán a incrementar la relevancia de sintagmas nominales cuando están presentes (*termhood* multipalabra). En el caso de la propiedad de *unithood*, consideramos que los patrones morfosintácticos constituyen una buena evidencia de *unithood* (Vivaldi & Rodríguez, 2007; Kockaert & Steurs, 2015). Además, como parte de la metodología, proponemos también la implementación de heurísticas lingüísticas para construir automáticamente una lista de adjetivos no relevantes del dominio analizado. Esto último es relevante ya que adjetivos (principalmente adjetivos relacionales) tienen una interpretación composicional, por lo que medidas tradicionales (por ejemplo, información mutua) fallan en la tarea de mostrar *unithood* de términos candidatos multipalabra.

En la sección 2 presentamos trabajos relacionados con la extracción automática de términos. En la sección 3, discutimos algunas cuestiones relacionadas con términos y su comportamiento. En la sección 4, describimos nuestra propuesta metodológica para la extracción automática de términos. En la sección 5 presentamos resultados de nuestros experimentos. Finalmente, en la sección 7 bosquejamos nuestras conclusiones.

2 Trabajo relacionado

La extracción automática de términos se ha utilizado para construir recursos lexicográficos como diccionarios, glosarios y vocabularios, así como recursos computacionales que sean útiles en el procesamiento automático de textos. Asimismo, tareas como la recuperación de información, clasificación textual, traducción automática, etc., se han beneficiado de los avances en la extracción automática de términos. Fi-

nalmente, si asumimos que los términos denotan conceptos, una terminología bien construida puede ser un punto de partida importante para el aprendizaje de ontologías (Kageura & Umino, 1996; Buitelaar et al., 2005; Poesio, 2005; Wong, 2008; Kockaert & Steurs, 2015).

Como se mencionó en párrafos anteriores, existen por lo menos tres enfoques diferentes para la extracción automática de terminología, sin embargo, ninguno considerado de forma independiente, ha sido completamente exitoso. Por un lado, los enfoques lingüísticos o basados en reglas intentan filtrar términos candidatos mediante patrones de formación de términos como evidencia de *unithood* (Ananiadou, 1994; Justeson & Katz, 1995; Bourigault et al., 1996; Heid, 1998; Jacquemin, 2001). Debido a que el uso de patrones morfosintácticos no ayuda a discernir entre palabras de dominio y de uso general, el enfoque común es generar una lista de palabras vacías (en inglés, *stopword list*) como una forma de filtrar candidatos no relacionados con el dominio. Una desventaja importante del enfoque lingüístico es la gran cantidad de ruido que produce y el hecho de que no es directamente aplicable a diferentes dominios y lenguajes.

Los enfoques estadísticos, por otro lado, usan un número diferente de medidas y distribuciones estadísticas para calcular *unithood* y *termhood*. En el caso específico de *unithood*, este tipo de enfoques considera dos propiedades que son comunes en términos multipalabra: combinaciones de palabras relativamente estables y cuya ocurrencia es alta. Los enfoques estadísticos, dada esta estabilidad sintagmática, y la variación nula en el orden de las palabras, pueden enfocarse en analizar n-gramas sin considerar la estructura lingüística subyacente. Las medidas usadas para el cálculo de colocaciones representan un buen ejemplo de cálculo de *unithood*. En términos formales, estas medidas cuantifican cuánto se desvía lo observado de lo que se espera como producto del azar, dadas las frecuencias individuales de las palabras. Entre las medidas para cuantificar la divergencia entre lo observado y esperado están el estadístico X^2 usado en Drouin (2003), así como en Matsuo & Ishizuka (2004). Otras medidas para el cálculo de *unithood* incluyen el puntaje-t (*t-score*), razón *log-likelihood* (Dunning, 1993), información mutua (Church & Hanks, 1990) y el coeficiente phi. Por su parte, Wermter & Hahn (2005) consideran el grado de remplazamiento de las palabras constituyentes de un término multipalabra por otras (modificabilidad paradigmática). Existen pocos ejemplos de enfoques donde únicamente se aplique la estadística para el proceso de extracción automática de términos, generalmente las medidas para el cálculo de *unithood* se combinan con una fase lingüística y se calculan para las combinaciones que pasaron ya el filtro lingüístico. Un experimento que sólo considera una fase estadística es el realizado por Pantel & Lin (2001).

Respecto a la medición del rasgo *termhood*, que refiere al grado en que una unidad léxica denota conceptos del dominio, el enfoque inicial fue el uso de la frecuencia del término candidato en el dominio como un indicio de su importancia (Daille et al., 1994). Algunos autores fijan el origen de la extracción automática de términos al campo de la recuperación de información y esta relación estrecha se evidencia al considerar medidas como el TF-IDF para el cálculo de *termhood* (Evans & Lefferts, 1995; Medelyan & Witten, 2006). En este sentido, se aplica la fórmula TF-IDF para ponderar alto aquellos candidatos con un nivel de especificidad mayor. Un tercer método se enfoca en el uso contextual de los candidatos a término que ya pasaron un filtro lingüístico para después analizar su co-ocurrencia con palabras de contexto adicionales (Maynard & Ananiadou, 1999; Frantzi et al., 2000). Un cuarto método se enfoca en los candidatos unipalabra y analiza su estructura morfológica interna (Aubin & Hamon, 2006). Áreas de conocimiento, como la medicina, derivan en gran parte su terminología de raíces griegas y latinas, lo que se puede explotar como un rasgo de *termhood* (Ananiadou, 1994). Finalmente, otro enfoque consiste en contrastar el comportamiento de un candidato dentro del dominio con información de un corpus de referencia o lengua general. En este método se asume que los términos son específicos de dominio, y como consecuencia ocurren con mayor frecuencia en su dominio que en otros dominios o en lengua general. Bajo esta premisa, se compara la frecuencia de un candidato en un corpus de dominio específico con su frecuencia en un corpus de referencia o de otro dominio diferente. Por ejemplo, el método *contrastive weight* de Basili et al. (1997) es una adaptación del TF-IDF porque en lugar de considerar la ocurrencia en diferentes documentos de una colección se usa la dispersión de los candidatos en dominios diferentes. Por su parte Ahmad et al. (1999) usan una medida para hacer referencia al concepto de *weirdness* de una palabra mediante la comparación de las frecuencias normalizadas de la palabra entre un corpus especializado y uno de referencia. Chung (2003) usa una razón de frecuencia normalizada para medir *termhood*. Por su parte, Wong (2008) usa el comportamiento distribucional de una palabra en otro corpus para medir la distribución

intra-dominio y el comportamiento distribucional multi-dominio. Drouin (2003) compara precisión y cobertura para la clasificación de métodos de prueba de hipótesis diferentes, en un intento por determinar el mejor método. Por último, Kit & Liu (2008) y Gelbukh et al. (2010) miden *termhood* de palabras simples mediante la medida de diferencia de rangos y la razón *log-likelihood*, respectivamente.

Finalmente, los sistemas de aprendizaje automático usan datos de entrenamiento para aprender rasgos que sean útiles y relevantes para el reconocimiento y clasificación de términos. Varias técnicas de aprendizaje automático se han usado para identificar y clasificar términos, los que incluyen HMMs (Collier et al., 2000), enfoques Bayesianos, SVMs (Kazama et al., 2002; Yamamoto et al., 2003), y árboles de decisión (Lopez & Romary, 2010).

3 Términos y su comportamiento

Los términos son palabras o unidades léxicas que denotan conceptos en un dominio restringido (Ananiadou, 1994; Daille, 1996; Jacquemin, 1997; Pazienza, 1998; Frantzi et al., 2000; Vivaldi et al., 2001; Buitelaar et al., 2005; Wong, 2008; Spasić et al., 2013). De acuerdo con Kageura & Umino (1996) y Daille et al. (1996), los términos son expresiones principalmente multipalabra de tipo nominal, caracterizadas por propiedades morfológicas, sintácticas y semánticas. Términos como *dominio restringido, lenguaje de especialidad, lenguaje especial, sublenguaje, dominio especializado, dominio especialista*, por otro lado, refieren a un subsistema lingüístico con términos especializados y otros recursos lingüísticos usados para comunicar de manera precisa y sin ambigüedad en un área de conocimiento determinada (Vivanco, 2006; Ananiadou, 1994).

Aunque los términos son representaciones lingüísticas para denotar conceptos en dominios especializados, no es posible distinguirlos completamente de palabras comunes por su forma debido a que los lenguajes de especialidad se derivan del lenguaje general, y por ello siguen las mismas reglas de formación de palabras (L'Homme, 2004). No obstante, Pazienza et al. (2005), señalan que hay un gran interés dentro de la lingüística computacional por establecer una definición más profunda de lo que es un término, con el fin de desarrollar algoritmos para mejorar el desempeño de los enfoques actuales de extracción.

Con la meta de aclarar lo que es un término y aquello que se requiere para su identificación y extracción, se han propuesto dos propiedades: *unithood* y *termhood* (Kageura & Umino, 1996). La propiedad de *unithood* refiere al grado de estabilidad sintagmática de un candidato a término y es relevante solo para el caso términos multipalabra. Por otro lado, *termhood* refiere al grado de relevancia de un candidato al dominio y se enfoca en ambos, unipalabra y multipalabra. Particularmente, en la propiedad de *unithood*, estamos de acuerdo con Kit & Liu (2008) respecto a que es una condición necesaria pero no suficiente, ya que un término verdadero debe tener un *unithood* alto. Empero, pueden existir muchos candidatos que lo tengan, pero no por ello serán términos verdaderos.

Kageura & Umino (1996) mencionan que el origen de la extracción automática de términos se puede encontrar en el campo de la recuperación de información. En la recuperación de información, los términos índice o palabras clave se usan para indexar o recuperar documentos. Estos términos índice tienen algún significado en sí mismos, y regularmente tienen la categoría gramatical de nombres (Baeza & Rivera, 2011). En este escenario, no todos los nombres son relevantes para indexar documentos, es decir, algunos representan mejor los documentos que otros y pueden discriminar más efectivamente entre ellos. En el caso de la extracción automática de términos, no todos los nombres denotan conceptos relevantes en el dominio, por lo que es necesario asignar relevancia más alta a los más significativos que a los menos importantes, lo cual no es una tarea fácil.

3.1 Estructura de términos

De acuerdo con Daille (1996), Kageura & Umino (1996), así como Spasić et al. (2013), para propósitos prácticos, los términos se definen como sintagmas nominales que ocurren con frecuencia en textos de un dominio específico donde tienen un significado especial. En la extracción automática de términos para el español (Vivaldi, 2004; Vivaldi & Rodríguez, 2007), unidades como los nombres, adjetivos y la preposición *de*, son los más comunes que participan en la formación de términos multipalabra. Vivaldi & Rodríguez (2007) presentan datos respecto al uso de estos patrones en lenguajes de especialidad de dos subcorpus en español del corpus técnico del IULA. De 2,145 términos en estos dos subcorpus, 48 % son nombres únicos; 45 % son sintagmas nominales multipalabra, es decir, `<nombre+adjetivo>`, y 7 % tienen el patrón `<nombre+preposición+nombre>`, donde la preposición de tiene un uso mucho mayor que el resto

de las preposiciones. A partir de los datos se puede ver que, por lo menos en español, los términos de una sola palabra constituyen un grupo importante, contrario a la observación hecha por Daille et al. (1996) respecto a que la mayoría de los términos son sintagmas nominales multipalabra.

3.2 Patrones con frase preposicional

Daille et al. (1996) argumenta que los términos son secuencias que muestran diferentes tipos de variaciones, contrario a la concepción tradicional de que los términos son secuencias fijas. Para estos autores, una variante de un término es un enunciado, que es semántica y conceptualmente relacionado a un término original. Las variaciones se dividen en dos clases principales: morfológicas y sintácticas. Por ejemplo, una variante morfológica de *célula epitelial* es *célula asociada con el epitelio*. Por otro lado, una variación sintáctica de la misma expresión es *célula de tumor epitelial*. Adicionalmente, las variaciones sintácticas se pueden dividir en variaciones que preservan significado: célula sanguínea — *célula de la sangre*, y aquellas que incluyen un cambio en significado: *célula sanguínea* — *célula mononuclear sanguínea*. En el caso de variaciones que preservan significado, Daille et al. (1996) argumenta:

> The most constant signs of permutation occur around the preposition of. It is with this preposition, as opposed to the others, that a strict permutation without insertion leads to the best results, i.e., less noise. We can verify that the terms variance analysis or chromosome duplication are completely equivalent to textual sequences under which they have been identified (analysis of variance, duplication of chromosome). From this point of view, the sole permutation demonstrates a synonymous connection between the term in its basic form and its transformed form (p. 222).

En lo que respecta al uso de preposiciones, Marchis (2010) señala que los compuestos nominales (por ejemplo, *kidney diseases*) difícilmente se presentan en lenguas romance (por lo menos en rumano y español), por lo que se pueden combinar los nombres con preposiciones como *de* o *a*, sin embargo, los nombres con adjetivos relacionales se priorizan para evitar el abuso en el uso de frases preposicionales. Por otro lado, como Vivanco (2006) menciona, las preposiciones inherentes a las lenguas romances, que se usan con frecuencia en términos, aunque van en contra de la brevedad, sirven como un factor aclaratorio en español, esto podría, como Marchis (2010) sugiere, ser una causa importante de la alta productividad de frases nominales como `<nombre+adjetivo>`.

De acuerdo con Daille et al. (1996), las preposiciones usadas en las variantes de permutación se dividen en dos categorías: preposiciones *de, con, para*, y *en*, que sólo tienen una función relacional y el resto que son preposiciones locativas, de las cuales de es la preposición semánticamente menos informativa debido a que representa una gran cantidad de relaciones. En el cuadro 1, mostramos datos sobre el uso de preposiciones en frases nominales con estructura:

```
<Nombre + adjetivo? + preposición +
      determinante? + nombre>
```

en tres colecciones de textos diferentes: Corpus de Ingeniería Lingüística (Medina et al., 2004), textos extraídos de MedlinePlus en español y un corpus de referencia general recolectado automáticamente de la Web (extraído de un periódico mexicano). Como se puede observar en el cuadro 1, las preposiciones que tienen una función relacional representan más del 85 % de uso comparado con el resto. La preposición *de* en español también es por mucho la preposición más usada.

Preposición[1]	**Referencia** %	**CLI** %	**MedlinePlus** %
De	68.3	72.8	65.3
En	11.7	10.6	14
Con	3.4	3.3	6.5
Para	2.0	3.2	3.1
Resto	14.6	10.1	11.1

Cuadro 1: Uso de preposiciones en español.

3.3 Adjetivos no relevantes en terminología

Un adjetivo es una categoría gramatical cuya función es modificar nombres (Demonte, 1999; Fábregas, 2007). Existen dos tipos de adjetivos que asignan propiedades a los nombres: adjetivos calificativos o descriptivos y adjetivos relacionales. Los adjetivos calificativos refieren a rasgos constitutivos del nombre modificado. Estos rasgos se exhiben o caracterizan por medio de una propiedad física única: color, forma, predisposición..., (*el libro azul, la señora delgada*). Por otro lado, los adjetivos relacionales asignan un conjunto de propiedades, es decir, todas las características que conjuntamente definen nombres

[1] Las preposiciones consideradas por las Real Academia de la Lengua Española son: *a, ante, bajo, cabe, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, so, sobre* y *tras*.

(*puerto marítimo, paseo campestre*). En terminología, los adjetivos relacionales representan un elemento importante para construir términos especializados, por ejemplo: *hernia inguinal, enfermedad venérea, desorden psicológico*, se consideran términos en medicina. En contraste, *hernia rara, enfermedad seria*, y *desorden crítico* parecen juicios más descriptivos y estrechamente relacionados con un contexto específico.

Identificación sintáctica de adjetivos no relevantes

Con base en lo anterior, si consideramos la estructura interna de adjetivos, se pueden identificar dos tipos: adjetivos permanentes y episódicos (Demonte, 1999). El primer tipo de adjetivo representa situaciones estables, propiedades permanentes que caracterizan individuales. Estos adjetivos se ubican fuera de cualquier restricción espacial o temporal (*psicópata, egocéntrico, apto*). Por otro lado, los adjetivos episódicos refieren a situaciones transitorias o propiedades que implican cambio y con limitaciones de espacio-tiempo. Casi todos los adjetivos descriptivos derivados de participios pertenecen a esta última clase, así como participios adjetivales (*harto, limpio, seco*). El español es uno de los pocos lenguajes que en su sintaxis representa esta diferencia en el significado de adjetivos. En muchos lenguajes esta diferencia solo es reconocible a través de la interpretación. En español, las propiedades individuales se predican con el verbo *ser*, y las episódicas con el verbo *estar*, lo que es esencial para probar a qué clase pertenece un adjetivo. Con la meta de identificar y extraer adjetivos no relevantes, proponemos extraer los adjetivos predicados con el verbo *estar*.

Otra heurística lingüística para identificar adjetivos descriptivos es que solo estos tipos de adjetivos aceptan adverbios de grado, y pueden ser parte de construcciones comparativas, por ejemplo, *muy alto, extremadamente grave*. Finalmente, solo los adjetivos calificativos pueden preceder un nombre porque —en español— los adjetivos relacionales siempre se posponen (*la antigua casa*).

4 Metodología

En este trabajo proponemos una metodología para extraer términos de un corpus de dominio especializado. La entrada debe ser un corpus con etiquetado morfosintáctico. En este caso, el corpus se ha etiquetado con FreeLing (Carreras et al., 2004). Los etiquetadores más usados para español son TreeTagger Schmid (1994) y FreeLing. En este experimento usamos FreeLing porque es más preciso. Lo cuadro 2 muestra las etiquetas más usadas.

Etiqueta	**Significado**
N.* (NC, NP)	Nombre
A.* (AQ, AO)	Adjetivo
V.* (VM, VA, VS)	Verbo
RG	Adverbio
SP	Preposición
D.* (DD, DP,DA,DI, DT,DE)	Determinante
P.* (PP, PD, PX, PI, PT, PR, PE)	Pronombre
C.* (CC, CS)	Conjunción
F.* (FA, FC, FZ, FG, FS)	Puntuación

Cuadro 2: Las etiquetas FreeLing más comunes.

4.1 Estandarización de etiquetas

La mayoría de las etiquetas fueron truncadas a dos caracteres, excepto en el caso del verbo *estar*, cuya etiqueta es VAE.

4.2 Análisis sintáctico superficial

El análisis sintáctico superficial es el proceso de identificar y clasificar segmentos de una oración vía la agrupación de las etiquetas morfosintácticas principales que forman frases no recursivas básicas. Una gramática para el análisis sintáctico superficial es un conjunto de reglas que indican cómo deben agruparse las oraciones. Las reglas de una gramática usan los patrones de etiquetas para describir secuencias de palabras etiquetadas, por ejemplo, `<DA>?<NC><AQ>*`. Los patrones de etiquetas son similares a patrones de expresión regular, donde los símbolos como "`*`" significan cero o más ocurrencias, "`+`" significa una o más ocurrencias y "`?`" representa un elemento opcional.

En este trabajo nos enfocamos en la extracción de términos base a partir de información textual. Lingüísticamente, como sucede en inglés, los patrones de términos más productivos consisten de un nombre y cero o más adjetivos (Vivaldi et al., 2001; Barrón-Cedeño et al., 2009). Estos términos se denominan términos base, de los cuales se derivan términos más complejos (Daille et al., 1996). Mediante el uso de etiquetas FreeLing, estos patrones se pueden representar como una expresión regular:

```
<NC><AQ>*
```

Como Daille et al. (1996) mencionan, si consideramos el patrón:

```
<nombre><preposición De><nombre>
```

En muchos casos, los términos con una frase preposicional con el núcleo *de* son variantes de formas básicas, en este caso, del patrón `<nombre><adjetivo>` (por ejemplo, *enfermedad renal-enfermedad del riñón*). Finalmente, la expresión regular usada para extraer adjetivos no relevantes de acuerdo con las heurísticas mencionadas en la sección 3 son:

```
<RG><AQ>
<VAE><AQ>
<D.*|P.*|F.* |S.*><AQ><NOUN>
```

Donde RG, AQ y VAE, corresponden a las etiquetas para adverbios, adjetivos, y el verbo estar, respectivamente. Las etiquetas `<D.*|P.*|F.*|S.*>` corresponden a determinantes, pronombres, signos de puntuación y preposiciones. La expresión `<D.*|P.*|F.*|S.*>` es una restricción para reducir *ruido*, ya que elementos erróneamente etiquetados por FreeLing como adjetivos se extraen sin esta restricción.

4.3 Reducción de ruido

Con el objetivo de eliminar palabras no relevantes de frases nominales antes de asignar relevancia a términos candidatos multipalabra utilizamos los adjetivos descriptivos obtenidos mediante heurísticas lingüísticas. Sumado a lo anterior, los sintagmas candidatos que tienen como núcleo nombres muy comunes como: *caso, mayoría, vez, superficie, área, tamaño, tipo, subtipo, forma, parte, término, clase* y *subclase*, son eliminados en esta fase.

Adjetivos no relevantes

De acuerdo con Pazienza et al. (2005), es posible utilizar un conjunto de palabras no relevantes al dominio para refinar la terminología que se deriva de un proceso automático. Barrón-Cedeño et al. (2009) considera una lista de palabras vacías con alta frecuencia en un corpus que se espera no formen parte de términos en un dominio específico. Consideramos que una lista construida de esta forma tiene desventajas debido a que además de la selección por frecuencia de ocurrencia se requiere la supervisión humana para determinar si una palabra es relevante o no al dominio.

Dado lo anterior, consideramos que la frecuencia de una palabra no basta como indicador y que se pueden tomar en cuenta heurísticas lingüísticas que operan en un lenguaje específico para automatizar la selección de palabras no relevantes dentro del dominio, sin embargo, una de las desventajas es que esto conduce a dependencia del lenguaje. En el caso del español, Demonte (1999) propone un conjunto de rasgos característicos para distinguir entre adjetivos calificativos y relacionales. Estas heurísticas se mencionaron en la sección 3.

Consideramos cómo se usan los adjetivos en el corpus de dominio en lugar de su frecuencia de uso, por lo que las heurísticas implementadas en Acosta et al. (2013) —un adverbio que precede a un adjetivo, un adjetivo que precede a un nombre, y el verbo *estar* precediendo un adjetivo— se usan para extraer adjetivos no relevantes del dominio como se mencionó en párrafos anteriores con relación a la distinción entre adjetivos permanentes y episódicos.

Finalmente, los adjetivos del corpus de referencia se obtuvieron también con estas tres heurísticas en mente y fueron manualmente revisados para determinar su relevancia a cualquier dominio de conocimiento especializado (por ejemplo, adjetivos como *relevante, importante, necesario, apropiado, correspondiente*, etc., se consideraron en la lista de adjetivos no relevantes). Esta es una lista de tamaño fijo que puede considerarse como lista base donde se pueden agregar los adjetivos extraídos del dominio. En el apéndice A presentamos un subconjunto de los mejores términos candidatos multipalabra, antes y después de reducir *ruido*. Como se puede observar de este subconjunto de candidatos antes de remover *ruido*, existen juicios descriptivos que están estrechamente relacionados con un contexto específico. Estos casos se recuperan con los patrones morfosintácticos implementados, por lo que es necesario aplicar una fase de reducción de ruido. Los adjetivos extraídos del dominio con heurísticas lingüísticas se presentan en el apéndice A.

4.4 Relevancia de palabras

Siguiendo el enfoque propuesto por Enguehard & Pantera (1995), primero evaluamos *termhood* de palabras simples con cuatro medidas propuestas en la literatura sobre la extracción automática de términos (Manning & Schütze, 1999; Drouin, 2003; Kit & Liu, 2008; Gelbukh et al., 2010). Dado el patrón sintáctico usado para términos en este estudio, tomamos en cuenta sólo nombres y adjetivos en ambos corpus porque son el tipo de palabras más usadas para la construcción de términos.

Gelbukh et al. (2010) y Kit & Liu (2008) sólo se enfocan en la extracción de candidatos unipa-

labra, por lo que únicamente ponderan las palabras que ocurren en ambos corpus. En nuestro experimento también consideramos las palabras que ocurren sólo en el corpus de dominio. En este sentido, las palabras con una frecuencia absoluta de al menos 1 se ponderan exclusivamente con la frecuencia de ocurrencia, como en 1, para el caso de las medidas razón de frecuencias relativas y diferencia de rangos. Para el caso de la medida razón *log-likelihood*, la ponderación únicamente se realiza con la frecuencia de la palabra para hacerla más compatible con la escala de ponderaciones de esta medida. En este trabajo, asumimos que el corpus de referencia es lo suficientemente grande para filtrar palabras no relevantes, por tanto las palabras que solo ocurren en el dominio tienen una probabilidad mayor de ser relevantes y su frecuencia refleja su importancia.

$$\text{Peso}\,(w_i) = 1 + \log_2\,(f_{w_i}) \qquad (1)$$

Consideramos que mientras más grande sea el corpus de referencia, mayor *exhaustividad*[2] tendrá de palabras de clase abierta de uso general, así como una probabilidad mayor de que ocurran términos de especialidad por lo menos una vez (el corpus de referencia fue recolectado de un periódico online donde se publican noticias respecto a ciencia y tecnología, así como otros rubros de información), por lo que, al igual que Drouin (2003), consideramos es lo suficientemente heterogéneo para contribuir a lograr una precisión más alta en la asignación de relevancia.

En resumen, como Ananiadou (1994) lo señala, la terminología se interesa en formas de palabra que ocurren con alta y baja frecuencia, o también aquellas en un rango medio, es decir, en todas las unidades que podrían ser términos en una colección de textos particular. Por su parte, Vivaldi et al. (2001) mencionan que incluso cuando se analizan corpus grandes, existe siempre la posibilidad de encontrar un término que sólo ocurra una vez. Por tanto, si esto representa o no un problema depende de la meta de la extracción. Por ejemplo, si solo deseamos caracterizar un documento, podríamos esperar que un término representativo ocurra muchas veces en el documento. Sin embargo, si deseamos hacer un mapa conceptual del texto, serían necesarios todos los términos.

[2] En el contexto de la recuperación de información, Baeza & Rivera (2011) describen la *exhaustividad* de un documento como la cobertura que proporciona para los temas principales del documento. Así, si agregamos nuevo vocabulario a un documento, la *exhaustividad* de la descripción del documento se incrementa.

4.5 Relevancia de términos candidatos

La relevancia de términos candidatos multipalabra se calcula como la suma de los pesos individuales (*termhood* unipalabra) de las palabras que forman el candidato.

Formalmente, si un sintagma nominal candidato s tiene una longitud de n palabras, $w_1 w_2 \ldots w_n$, donde $n > 1$, entonces la relevancia del candidato s es la suma de las ponderaciones individuales w_i de las palabras presentes en el sintagma:

$$\text{Termhood}\,(s) = \sum_{i=1}^{n} \text{termhood}\,(w_i) \qquad (2)$$

5 Resultados

Esta sección presenta los resultados de nuestro experimento con un conjunto de 200,000 tokens de un corpus extraído del sitio Web MedlinePlus en español.

5.1 Fuentes de información textual

Corpus de dominio

La fuente de información textual se constituye de un conjunto de documentos del dominio médico, básicamente enfermedades del cuerpo humano y temas relacionados (cirugías, tratamientos, etc.). Estos documentos se recolectaron del sitio Web MedlinePlus en español. MedlinePlus es un sitio que se enfoca en proporcionar información respecto a enfermedades, tratamientos y condiciones.

El tamaño del corpus es de 1.2 millones de tokens, pero realizamos nuestro experimento con un subconjunto de 200,000 tokens que refieren a enfermedades, tratamientos, etc., exclusivamente de los ojos. Decidimos restringir el corpus para ser capaces de determinar manualmente vía el uso de diccionarios y otros recursos confiables sobre el tema, el número de términos verdaderos presentes relacionados estrechamente con la temática del subcorpus y contar con una evaluación preliminar del desempeño de cada medida. Por último, seleccionamos un dominio médico debido a la disponibilidad de recursos textuales en formato digital.

Corpus de referencia

Con la meta de asignar relevancia a las palabras del dominio por medio un enfoque de contraste entre corpus, se recolectó automáticamente

un corpus de referencia de un periódico[3] online con artículos de noticias de todo el año 2014 (el tamaño del corpus es de aproximadamente de 5 millones de tokens).

Para construir el corpus se recolectaron los URLs del sitio con el módulo de Python BeautifulSoup[4]. Después, este conjunto de URLs se introdujo a la plataforma Sketch Engine[5] para recolectar automáticamente la información textual de cada página Web.

5.2 Otros recursos

Lenguajes de programación y otras herramientas

El lenguaje de programación usado para automatizar todas las tareas requeridas fue Python versión 3.4, así como el módulo NLTK version 3.0 Bird et al. (2009). Adicionalmente, el etiquetador morfosintáctico usado en este experimento fue FreeLing.

6 Análisis de resultados

En este experimento comparamos las medidas razón *log-likelihood* implementada por Gelbukh et al. (2010), la diferencia de rangos aplicada por Kit & Liu (2008), la razón de frecuencia relativa (Manning & Schütze, 1999) y la aproximación a la distribución binomial mediante el uso de la distribución normal estándar (Drouin, 2003) para medir *termhood* en palabras simples. Para las tres primeras medidas se aplica la reducción de ruido mencionada en la sección 4.3, así como la ponderación para las palabras que ocurren sólo en el dominio por medio de la frecuencia de ocurrencia. Para el caso de la medida usada en TermoStat[6] se utilizó dicha herramienta desde el sitio Web con el mismo subcorpus de dominio, sin embargo, dos factores que podrían sesgar los resultados para el caso de esta medida son el etiquetador morfosintáctico (el etiquetador usado por el sistema es TreeTagger) y el corpus de referencia, lo que afecta la comparación directa de los resultados con las tres medidas consideradas en este trabajo. Con todo, dada la cobertura altamente similar que se obtuvo del vocabulario que ocurre en ambos corpus (vocabulario común) se decidió considerarla en la comparación. Además, con el objetivo de hacer más comparables los resultados se lematizó con FreeLing el conjunto de palabras y los candidatos a término obtenidos con Termostat.

[3] http://www.lajornada.com.mx
[4] http://www.crummy.com/software/BeautifulSoup
[5] https://the.sketchengine.co.uk
[6] http://termostat.ling.umontreal.ca

Con la finalidad de comparar resultados para lograr un equilibrio entre precisión y cobertura, proponemos considerar los siguientes factores y algunas de sus combinaciones: candidatos sólo con vocabulario común (es decir, palabras que ocurren en ambos corpus), candidatos con vocabulario común y no común (incluyendo o no candidatos que ocurren una sola vez en corpus de dominio), filtro de adjetivos no relevantes de dominio y, finalmente, filtro de adjetivos no relevantes del corpus de referencia.

Vocabulario y términos del corpus

En el subcorpus analizado existen, en total, 1,842 palabras estrechamente relacionadas con el dominio y que participan en la construcción de la terminología implícita. De este subconjunto de palabras se derivan 2,253 términos unipalabra y multipalabra que cumplen con el patrón: `<NC><AQ>*`.

Ponderación del vocabulario común

El subcorpus de dominio tiene un total de 2,978 palabras que ocurren también en el corpus de referencia (vocabulario común). Estas palabras fueron ponderadas con cada una de las medidas comparadas en este trabajo: razón de frecuencia relativa (RFR), razón *log-likelihood* (RLL) y diferencia de rangos (DR). Como se mencionó en líneas anteriores, dado que el sistema Termostat considera otro corpus de referencia, el subconjunto de palabras ponderadas es diferente, en este caso específico corresponde a 1,696 palabras, por ello las celdas de los cuadros 3 y 4 en umbrales mayores que 2000 palabras no contienen datos. Estos cuadros muestran los niveles de precisión y cobertura del vocabulario común por umbrales de 500 palabras (ordenadas descendentemente por *termhood*). Así, esperaríamos que las palabras más relacionadas con el dominio se concentraran en las primeras posiciones. Por ejemplo, del primer subconjunto de 500 palabras, la cobertura obtenida por RFR, DR, TS y RLL es de 18.5 %, 18 %, 17.2 % y 14.4 %, respectivamente. Por otro lado, la precisión es de 68 %, 66.4 %, 63.2 % y 53.2 % para RFR, DR, TS y RLL, respectivamente. La cobertura total considerando sólo el vocabulario común para RFR, DR y RLL es del 43.3 % y 44.5 % para el caso de TS. Cabe señalar aquí que estos niveles de cobertura altamente similares no implican que los dos subconjuntos de palabras sean iguales, esto se debe a que consideran dos corpus de referencia diferentes.

Palabras	RLL	DF	RFR	TS
500	14.4	18	18.5	17.2
1000	24.4	29.2	30.7	30.7
1500	32.1	35.6	38.1	41.9
2000	39.7	39.2	41.9	44.5
2500	42.4	42	43.4	
3000	43.3	43.3	43.4	

Cuadro 3: Cobertura del vocabulario común.

Palabras	RLL	DF	RFR	TS
500	53.2	66.4	68	63.2
1000	45	53.8	56.6	56.5
1500	39.5	43.7	46.7	51.4
2000	36.6	36.1	38.6	41
2500	31.2	31	32	
3000	26.6	26.6	26.7	

Cuadro 4: Precisión del vocabulario común.

Ponderación del vocabulario común y no común

En este trabajo consideramos también la ponderación del vocabulario que ocurre sólo en el dominio mediante el uso de la frecuencia de ocurrencia. Por tanto asumimos que a mayor ocurrencia, mayor relevancia para el dominio. De los cuadros 5 y 6 se puede observar que se logra un aumento del 13.3 % en la cobertura del vocabulario relevante al dominio para el caso de la medida DF en los primeros 1000 candidatos, e incrementos menores para las medidas RFR y RLL. Por otro lado, la precisión más alta para las primeras 1000 palabras mejor ponderadas se obtiene con la medida DF, que alcanza un 78.3 %.

Palabras	RLL	DF	RFR	TS
500	14.7	22.1	18.7	17.2
1000	25.1	42.5	32.1	30.7
1500	35.2	62.4	42.3	41.9
2000	45.3	78.4	53.1	44.5
2500	54.6	86.4	64.1	
3000	67.5	91.6	77.9	
3500	82.4	95.4	97.2	
4000	97.3	97.1	97.9	
4500	97.8	97.8	97.9	

Cuadro 5: Cobertura del vocabulario común y no común.

Extracción y ponderación de términos sólo con vocabulario común

En los siguientes cuadros se muestran los datos de precisión y cobertura para los candidatos a término donde ocurren sólo palabras del vocabulario común, que constituyen el 43.3 % del vocabulario relevante al dominio para el caso de las medidas DR, RFR y RLL. La comparación de los datos en los cuadros 7 y 9 muestran el incremento en cobertura después de agregar los adjetivos del corpus de referencia a los extraídos del dominio mediante heurísticas lingüísticas con el objetivo de reducir ruido en los resultados. Por ejemplo, para el caso de los primeros 1000 candidatos, RLL, DF y RFR tienen un incremento por encima del 4 % en cobertura, esto debido a que se eliminan más adjetivos estrechamente relacionados con el contexto y que aparecen como modificadores de términos verdaderos (por ejemplo, *conjuntivitis rara*). Con respecto a la precisión, de los cuadros 8 y 10 se observa que hay un incremento en los primeros 1000 candidatos de 11.6 %, 11.5 %, 9.6 % para RLL, RFR y DR, respectivamente, donde las precisiones más altas y similares se encuentran en DR y RFR. Para el caso del sistema Termostat se obtuvo un conjunto de 2,695 candidatos unipalabra y multipalabra, por ello las celdas con umbrales mayores a 3000 candidatos no contienen información. La cobertura global con la reducción de ruido es de un 56.3 % sólo con el vocabulario común y después de la reducción de ruido en términos de los adjetivos tanto del corpus de dominio como los del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	54	81.6	68.8	63.2
1000	46.2	78.3	59.1	56.5
1500	43.2	76.6	51.9	51.4
2000	41.8	72.2	49.0	41.0
2500	40.2	63.7	47.2	
3000	41.4	56.2	47.8	
3500	43.4	50.2	51.2	
4000	44.8	44.7	45.1	
4500	40.0	40.0	40.1	

Cuadro 6: Precisión del vocabulario común y no común.

Palabras	RLL	DF	RFR	TS
500	13.9	14.3	14.6	7.4
1000	24.3	26.9	26.1	12.8
1500	33.4	38.3	37.9	16.4
2000	41.7	46.7	46.4	16.6
2500	49.7	51.4	52.3	16.6
3000	55.3	54.8	56.6	16.6
3500	58.0	57.2	58.1	
4000	58.5	58.5	58.5	

Cuadro 7: Cobertura en la extracción de términos sin filtrar adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	62.6	64.6	65.6	33.2
1000	54.7	60.7	58.7	28.9
1500	50.2	57.5	56.9	24.6
2000	47	52.7	52.3	18.7
2500	44.8	46.3	47.1	14.9
3000	41.5	41.1	42.5	12.4
3500	37.3	36.8	37.4	
4000	33.0	33.0	33.0	

Cuadro 8: Precisión en la extracción de términos sin filtrar adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	16.5	17.0	16.4	7.4
1000	29.4	31.2	31.2	12.8
1500	39.1	41.9	42.7	16.4
2000	47.6	48.1	49.9	16.6
2500	53.2	51.5	54.4	16.6
3000	55.7	54.8	55.8	16.6
3500	56.3	56.3	56.3	
4000	56.3	56.3	56.3	

Cuadro 9: Cobertura en la extracción de términos con filtro de adjetivos del corpus de referencia.

Extracción y ponderación de términos con vocabulario común y no común

Si consideramos el vocabulario común y no común, así como los candidatos con frecuencia mayor o igual que 1 y que solo ocurren en el corpus de dominio, se observa un incremento en cobertura de 5.2 %, 4.4 % y 4.2 % para RLL, DR y RFR en los primeros 1000 candidatos después de la reducción de ruido (véase cuadros 11 y 13). Los principales cambios en cobertura se presentan en los umbrales mayores que 2000. Con respecto a la precisión, después de la reducción de ruido se obtiene un incremento del 11.7 %, 9.9 % y 9.3 % para el caso de RLL, DR y RFR en los primeros 1000 candidatos, donde las precisiones más altas corresponden a RFR con un 72.7 % y DR con un 70.5 %.

Palabras	RLL	DF	RFR	TS
500	74.2	76.4	74	33.2
1000	66.3	70.3	70.2	28.9
1500	58.8	63.0	64.1	24.6
2000	53.6	54.2	56.2	18.7
2500	48.0	46.4	49.0	14.9
3000	41.8	41.1	41.9	12.4
3500	36.2	36.2	36.2	
4000	31.7	31.7	31.7	

Cuadro 10: Precisión en la extracción de términos con filtro de adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	13.9	14.3	15.9	7.4
1000	24.3	26.9	28.1	12.8
1500	33.4	38.4	39.2	16.4
2000	41.9	51.4	49.3	16.6
2500	49.9	65.3	57.3	16.6
3000	57.4	78.8	65.2	16.6
3500	66.6	83.1	74.1	
4000	76.9	86.5	85.4	
4500	89.9	88.8	90.1	
5000	90.1	90.1	90.1	

Cuadro 11: Cobertura en la extracción de términos sin filtro de adjetivos de corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	62.6	64.4	71.6	33.2
1000	54.7	60.6	63.4	28.9
1500	50.2	57.7	58.9	24.6
2000	47.2	57.9	55.6	18.7
2500	45.0	58.8	51.7	14.9
3000	43.1	59.2	48.9	12.4
3500	42.9	53.5	47.7	
4000	43.3	48.7	48.1	
4500	45.0	44.5	45.1	
5000	40.6	40.6	40.6	

Cuadro 12: Precisión en la extracción de términos sin filtro de adjetivos de corpus de referencia.

Por último, si eliminamos los candidatos que sólo ocurren en el corpus de dominio y con frecuencia igual a 1, la cobertura global es de un 73 % después de la reducción de ruido y la precisión para los primeros 1000 candidatos se mantiene prácticamente sin cambios respecto a los resultados incluyendo este subconjunto.

Palabras	RLL	DF	RFR	TS
500	16.5	17.0	17.5	7.4
1000	29.5	31.3	32.3	12.8
1500	39.2	43.1	44.8	16.4
2000	47.8	57.3	53.9	16.6
2500	55.6	70.8	62.8	16.6
3000	64.4	80.1	71.6	16.6
3500	75.5	83.3	82.9	
4000	87.6	86.3	87.8	
4500	87.8	87.8	87.8	
5000	87.8	87.8	87.8	

Cuadro 13: Cobertura en la extracción de términos con filtro de adjetivos del corpus de referencia.

Palabras	RLL	DF	RFR	TS
500	74.2	76.4	79	33.2
1000	66.4	70.5	72.7	28.9
1500	58.9	64.7	67.3	24.6
2000	53.9	64.5	60.7	18.7
2500	50.1	63.8	56.6	14.9
3000	48.4	60.1	53.8	12.4
3500	48.6	53.6	53.3	
4000	49.4	48.6	49.5	
4500	44.0	44.0	44.0	
5000	39.6	39.6	39.6	

Cuadro 14: Precisión en la extracción de términos con filtro de adjetivos del corpus de referencia.

7 Conclusiones

En este trabajo hemos presentado una metodología para identificar y extraer términos unipalabra y multipalabra, reconocibles en un corpus de dominio especializado. Inicialmente, para asignar relevancia a palabras simples comparamos cuatro medidas diferentes e implementamos algunas heurísticas linguísticas, para reducir *ruido* en los resultados. De las cuatro medidas comparadas, la diferencia de rangos y la razón de frecuencia relativa fueron las que lograron los mejores resultados en términos de precisión y cobertura. Además, la estrategia para reducir ruido en los resultados, que consiste en considerar heurísticas de corte lingüístico para obtener adjetivos que con frecuencia no participan en la construcción de términos, derivó en buenos resultados al permitir aumentar precisión sin dañar significativamente la cobertura. Por otro lado, la propuesta de construir *termhood* multipalabra a partir del *termhood* de los elementos constituyentes proporcionó buenos resultados ya que un elemento con una ponderación individual alta contribuirá a incrementar el *termhood* de cualquier sintagma donde se encuentre presente.

Por tanto, el enfoque de comparación de corpus resultó útil para asignar relevancia a palabras de un dominio ya que asumimos que las palabras estrechamente relacionadas con el dominio analizado tendrán una mayor probabilidad de ocurrencia en el dominio que en un corpus de otro dominio diferente o de lengua general, lo que generará un *termhood* alto.

Actualmente existen muchas fuentes de información textual disponibles en la Web en varias lenguas, lo que facilita la obtención de documentos que sean útiles para implementar enfoques de comparación de corpus. En este sentido, el empleo de noticias de periódicos electrónicos es sumamente valioso, pues contienen información sobre muchos temas: cultura, política, ciencia, tecnología, etc., lo que favorece la heterogeneidad y exhaustividad del recurso textual, mejorando con ello la precisión al momento de implementar enfoques contrastivos.

Finalmente, resulta de gran interés probar la metodología propuesta en otros corpus de dominio para explorar la estabilidad de los resultados, lo que constituye parte de nuestro trabajo futuro.

Agradecimentos

Este trabajo ha sido patrocinado por la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT), del Gobierno de Chile. Números de proyectos: 3140332 y 11130565.

Referencias

Acosta, Olga, Cesar Antonio Aguilar & Gerardo Sierra. 2013. Using relational adjectives for extracting hyponyms from medical texts. En Antonio Lieto & Marco Cruciani (eds.), *Proceedings of the First International Workshop on Artificial Intelligence and Cognition*, vol. 1100 CEUR Workshop Proceedings, 33–44.

Ahmad, Khurshid, Lee Gillam & Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). En *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland.

Ananiadou, Sophia. 1994. A methodology for automatic term recognition. En *Proceedings of the 15th Conference on Computational Linguistics - Volume 2* COLING '94, 1034–1038. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ananiadou, Sophia & John Mcnaught. 2005. *Text mining for biology and biomedicine*. Norwood, MA, USA: Artech House, Inc.

Aubin, Sophie & Thierry Hamon. 2006. Improving term extraction with terminological resources. En Tapio Salakoski, Filip Ginter, Sampo Pyysalo & Tapio Pahikkala (eds.), *Advances in Natural Language Processing*, vol. 4139 Lecture Notes in Computer Science, 380–387. Springer Berlin Heidelberg.

Baeza, Ricardo & Berthier Rivera. 2011. *Modern information retrieval*. Addison Wesley.

Barrón-Cedeño, Alberto, Gerardo Sierra, Patrick Drouin & Sophia Ananiadou. 2009. An improved automatic term recognition method for

spanish. En Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 5449 Lecture Notes in Computer Science, 125–136. Springer Berlin Heidelberg.

Basili, Roberto, Gianluca De Rossi & Maria Pazienza. 1997. Inducing terminology for lexical acquisition. En C. Cardie & R. Weischedel (eds.), *Proceeding of EMNLP 97 Conference*, 125–133.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with python*. O'Reilly.

Bourigault, Didier, Isabelle Gonzalez-Mullier & Cécile Gros. 1996. LEXTER, a natural language processing tool for terminology extraction. En Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström & Catalina Röjder Papmehl (eds.), *Proceedings of the 7th EURALEX International Congress*, 771–779. Göteborg, Sweden: Novum Grafiska AB.

Buitelaar, Paul, Philipp Cimiano & Bernardo Magnini. 2005. *Ontology learning from text: Methods, evaluation and applications*, vol. 123 Frontiers in Artificial Intelligence and Applications Series. Amsterdam: IOS Press.

Carreras, Xavier, Isaac Chao, Lluís Padró & Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. En *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, European Language Resources Association (ELRA).

Chung, Teresa. 2003. A corpus comparison approach for terminology extraction. *Terminology* 9(26). 221—-246.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.* 16(1). 22–29.

Collier, Nigel, Chikashi Nobata & Jun-ichi Tsujii. 2000. Extracting the names of genes and gene products with a hidden markov model. En *Proceedings of the 18th Conference on Computational Linguistics - Volume 1* COLING '00, 201–207. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daille, Béatrice, Éric Gaussier & Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. En *Proceedings of the 15th Conference on Computational Linguistics - Volume 1* COLING '94, 515–521. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daille, Béatrice. 1996. Study and implementation of combined techniques for automatic extraction of terminology. En Judith L. Klavans & Philip Resnik (eds.), *The balancing act: Combining symbolic and statistical approaches to language*, 49–66. MIT Press.

Daille, Béatrice, Benoît Habert, Christian Jacquemin & Jean Royauté. 1996. Empirical observation of term variations and principles for their description. *Terminology* 3(2). 197–257.

Demonte, Violeta. 1999. El adjetivo. clases y usos. la posición del adjetivo en el sintagma nominal. En Ignacio Bosque & Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*, 129–215. Espasa.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1). 99–115.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19(1). 61–74.

Enguehard, Chantal & Laurent Pantera. 1995. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics* 2(1). 27–32.

Evans, David A. & Robert G. Lefferts. 1995. CLARIT-TREC experiments. En *Proceedings of the Second Conference on Text Retrieval Conference* TREC-2, 385–395. Elmsford, NY, USA: Pergamon Press, Inc.

Frantzi, Katerina, Sophia Ananiadou & Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries* 3(2). 115–130.

Fábregas, Antonio. 2007. The internal syntactic structure of relational adjectives. *Probus* 19(1). 1–36.

Gelbukh, Alexander, Grigori Sidorov, Eduardo Lavin-Villa & Liliana Chanona-Hernandez. 2010. Automatic term extraction using log-likelihood based comparison with general reference corpus. En ChristinaJ. Hopfe, Yacine Rezgui, Elisabeth Métais, Alun Preece & Haijiang Li (eds.), *Natural Language Processing and Information Systems*, vol. 6177 Lecture Notes in Computer Science, 248–255. Springer Berlin Heidelberg.

Heid, Ulrich. 1998. A linguistic bootstrapping approach to the extraction of term candidates from germ text. *Terminology* 5(2). 161–181.

Jacquemin, Christian. 1997. Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.

Jacquemin, Christian. 2001. *Spotting and discovering terms through natural language processing*. Cambridge: MIT Press.

Justeson, John S. & Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1. 9–27.

Kageura, Kyo & Bin Umino. 1996. Methods of automatic term recognition: a review. *Terminology* 3(2). 259–289.

Kazama, Jun'ichi, Takaki Makino, Yoshihiro Ohta & Jun'ichi Tsujii. 2002. Tuning support vector machines for biomedical named entity recognition. En *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3* BioMed '02, 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kit, Chunyu & Xiaoyue Liu. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology* 14(2). 204–229.

Kockaert, Hendrik & Frieda Steurs. 2015. *Handbook of terminology*, vol. 1. John Benjamins.

Lopez, Patrice & Laurent Romary. 2010. HUMB: automatic key term extraction from scientific articles in GROBID. En *Proceedings of the 5th International Workshop on Semantic Evaluation* SemEval '10, 248–251. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lossio-Ventura, JuanAntonio, Clement Jonquet, Mathieu Roche & Maguelonne Teisseire. 2014. Yet another ranking function for automatic multiword term extraction. En Adam Przepiórkowski & Maciej Ogrodniczuk (eds.), *Advances in Natural Language Processing*, vol. 8686 Lecture Notes in Computer Science, 52–64. Springer International Publishing.

L'Homme, Marie-Claude. 2004. *La terminologie: príncipes et techniques*. Les Presses de l'Université de Montréal.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Marchis, Mihaela. 2010. *Relational adjectives at the syntax/morphology interface in Romanian and Spanish*: Institut für Linguistik/Anglistik, Universität Stuttgart. Tesis Doctoral.

Matsuo, Yutaka & Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* 13(1). 157–169.

Maynard, Diana & Sophia Ananiadou. 1999. Identifying contextual information for multiword term extraction. En Peter Sandrini (ed.), *Proceedings of the TKE '99 International Congress on Terminology and Knowledge Engineering*, 212—-221. Vienna, Austria.

Medelyan, Olena & Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. En *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* JCDL '06, 296–297. New York, NY, USA: ACM.

Medina, Alfonso, Gerardo Sierra, Gabriel Garduño, Carlos Méndez & Roberto Saldaña. 2004. CLI. an open linguistic corpus for engineering. En Guillermo de Ita, Olac Fuentes & Mauricio Osorio (eds.), *Memorias del IX Congreso Iberoamericano de Inteligencia Artificial IBERAMIA 2004*, 203—-208. Puebla, México.

Pantel, Patrick & Dekang Lin. 2001. A statistical corpus-based term extractor. En Eleni Stroulia & Stan Matwin (eds.), *Advances in Artificial Intelligence*, 36–46. Springer.

Pazienza, Maria Teresa. 1998. A domain-specific terminology-extraction system. *Terminology* 5(2). 183–201.

Pazienza, MariaTeresa, Marco Pennacchiotti & FabioMassimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. En Spiros Sirmakessis (ed.), *Knowledge Mining*, vol. 185 Studies in Fuzziness and Soft Computing, 255–279. Springer Berlin Heidelberg.

Poesio, Massimo. 2005. Domain modelling and nlp: Formal ontologies? lexica? or a bit of both? *Applied Ontologies* 1(1). 27–33.

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

Spasić, Irena, Mark Greenwood, Alun Preece, Nick Francis & Glyn Elwyn. 2013. FlexiTerm: a flexible term recognition method. *Journal of Biomedical Semantics* 4(1).

Vivaldi, Jordi. 2004. *Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas*. Barcelona: IULA-UPF. Tesis Doctoral.

Vivaldi, Jordi, Lluís Màrquez & Horacio Rodríguez. 2001. Improving term extraction by system combination using boosting. En Luc De Raedt & Peter Flach (eds.), *Machine Learning: ECML 2001*, vol. 2167 Lecture Notes in Computer Science, 515–526. Springer Berlin Heidelberg.

Vivaldi, Jorge & Horacio Rodríguez. 2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13(2). 225–248.

Vivanco, Verónica. 2006. *El español de la ciencia y la tecnología*. Madrid: Arco Libros.

Wermter, Joachim & Udo Hahn. 2005. Paradigmatic modifiability statistics for the extraction of complex multi-word terms. En *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 843–850. Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Wong, Wilson. 2008. Determination of unithood and termhood for term recognition. En Min Song & Yi-Fang Brook Wu (eds.), *Handbook of research on text and web mining technologies*, 500–529. Hershey, New York!: IGI Global.

Yamamoto, Kaoru, Taku Kudo, Akihiko Konagaya & Yuji Matsumoto. 2003. Protein name tagging for biomedical annotation in text. En *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 65–72. Sapporo, Japan: Association for Computational Linguistics.

A Apéndice

Un pequeño conjunto de términos candidatos antes y después de reducir ruido.

Previo a la reducción de ruido

enfermedad/NC ocular/AQ alérgico/AQ severo/AQ
vaso/NC sanguíneo/AQ anormal/AQ
catarata/NC congénito/AQ hereditario/AQ
degeneración/NC macular/AQ húmedo/AQ
ganglio/NC linfático/AQ sensible/AQ
vaso/NC sanguíneo/AQ permeable/AQ
vaso/NC sanguíneo/AQ retiniano/AQ
resequedad/NC ocular/AQ serio/AQ
vaso/NC sanguíneo/AQ defectuoso/AQ
gota/NC lubricante/AQ ocular/AQ
degeneración/NC macular/AQ seco/AQ
gota/NC oftálmico/AQ antibiótico/AQ
nervio/NC óptico/AQ saludable/AQ
anteojo/NC protector/AQ
cirujano/NC oftalmológico/AQ
vaso/NC sanguíneo/AQ cerebral/AQ
secreción/NC ocular/AQ seco/AQ
irritación/NC ocular/AQ leve/AQ
anteojo/NC común/AQ
degeneración/NC macular/AQ senil/AQ
examen/NC ocular/AQ estándar/AQ
trastorno/NC ocular/AQ común/AQ
ganglio/NC linfático/AQ circundante/AQ
músculo/NC ocular/AQ externo/AQ
movimiento/NC ocular/AQ anormal/AQ
enfermedad/NC ocular/AQ específico/AQ
órgano/NC digestivo/AQ
degeneración/NC macular/AQ temprano/AQ
vaso/NC sanguíneo/AQ frágil/AQ
oclusión/NC arterial/AQ retiniano/AQ
ganglio/NC linfático/AQ cercano/AQ
característica/NC facial/AQ anormal/AQ
anteojo/NC oscuro/AQ
afección/NC ocular/AQ grave/AQ
degeneración/NC macular/AQ intermedio/AQ
lente/NC intraocular/AQ artificial/AQ
tejido/NC corneal/AQ subyacente/AQ
coágulo/NC sanguíneo/AQ frecuente/AQ
cristalino/NC intraocular/AQ artificial/AQ
ceguera/NC nocturno/AQ congénito/AQ
enfermedad/NC intestinal/AQ inflamatorio/AQ
enfermedad/NC inflamatorio/AQ intestinal/AQ
dolor/NC ocular/AQ severo/AQ
inflamación/NC ocular/AQ
reflejo/NC nervioso/AQ anormal/AQ
ciclosporina/NC líquido/AQ
examen/NC ocular/AQ completo/AQ
dolor/NC ocular/AQ excesivo/AQ
nervio/NC craneal/AQ
examen/NC ocular/AQ minucioso/AQ
parálisis/NC facial/AQ periférico/AQ idiopático/AQ
ganglio/NC linfático/AQ
ganglio/NC linfático/AQ justo/AQ
inclinación/NC palpebral/AQ anormal/AQ
trastorno/NC neurológico/AQ agudo/AQ
examen/NC oftálmico/AQ estándar/AQ
conteo/NC sanguíneo/AQ completo/AQ
vaso/NC sanguíneo/AQ débil/AQ
músculo/NC ocular/AQ
órgano/NC hueco/AQ
presión/NC sanguíneo/AQ normal/AQ
esteroide/NC oftálmico/AQ suave/AQ
molestia/NC gastrointestinal/AQ leve/AQ
nervio/NC óptico/AQ

Después de la reducción de ruido

vaso/NC sanguíneo/AQ permeable/AQ
vaso/NC sanguíneo/AQ retiniano/AQ

cirujano/NC oftalmológico/AQ
vaso/NC sanguíneo/AQ cerebral/AQ
degeneración/NC macular/AQ senil/AQ
ganglio/NC linfático/AQ circundante/AQ
órgano/NC digestivo/AQ
oclusión/NC arterial/AQ retiniano/AQ
lente/NC intraocular/AQ artificial/AQ
cristalino/NC intraocular/AQ artificial/AQ
enfermedad/NC intestinal/AQ inflamatorio/AQ
enfermedad/NC inflamatorio/AQ intestinal/AQ
nervio/NC craneal/AQ
parálisis/NC facial/AQ periférico/AQ idiopático/AQ
ganglio/NC linfático/AQ
examen/NC oftálmico/AQ estándar/AQ
nervio/NC óptico/AQ
inflamación/NC articular/AQ
vaso/NC sanguíneo/AQ
órgano/NC abdominal/AQ
enfermedad/NC vascular/AQ cerebral/AQ
nervio/NC facial/AQ
cirujano/NC experto/AQ
examen/NC oftalmológico/AQ estándar/AQ
anteojo/NC
cirujano/NC especialista/AQ
músculo/NC facial/AQ
quiasma/NC óptico/AQ
degeneración/NC macular/AQ
gota/NC oftálmico/AQ antimicótico/AQ
neuritis/NC óptico/AQ autoinmunitario/AQ
enfermedad/NC gastrointestinal/AQ
cirujano/NC
presión/NC sanguíneo/AQ diastólico/AQ
retinopatía/NC diabético/AQ
alergia/NC nasal/AQ
cáncer/NC sanguíneo/AQ
traumatismo/NC craneal/AQ
torrente/NC sanguíneo/AQ
medicamento/NC tópico/AQ
cirugía/NC facial/AQ
resonancia/NC magnético/AQ cerebral/AQ
célula/NC sanguíneo/AQ
enfermedad/NC diabético/AQ
enfermedad/NC digestivo/AQ
gota/NC oftálmico/AQ profiláctico/AQ
gota/NC oftálmico/AQ antinflamatorias/AQ
gota/NC oftálmico/AQ homeopático/AQ
órgano/NC corporal/AQ
trastorno/NC articular/AQ
imagen/NC visual/AQ
lente/NC oftálmico/AQ
conteo/NC sanguíneo/AQ
presión/NC sanguíneo/AQ
hipertensión/NC arterial/AQ
trastorno/NC digestivo/AQ
tic/NC facial/AQ
trastorno/NC visual/AQ
hinchazón/NC facial/AQ
enfermedad/NC intestinal/AQ
inflamación/NC viral/AQ
infección/NC gastrointestinal/AQ
neuropatía/NC óptico/AQ isquémico/AQ
coágulo/NC sanguíneo/AQ

Adjetivos obtenidos del corpus de dominio.

severo, común, cierto, probable, mejor, distinto, contagioso, posible, propio, bueno, conveniente, viejo, tratable, temprano, siguiente, susceptible, alto, cuidadoso, diverso, agresivo, mayor, pequeño, efectivo, bajo, importante, intenso, visible, presunto, reseco, diferente, útil, complejo, único, largo, gran, mismo, miope, borroso, corto, redondo, saludable, rojo, abundante, inconsciente, simple, atento, peligroso, principal, nuevo, suficiente, grande, excesivo, húmedo, obvio, necesario, suceptivas, disponible, caliente, molesto, menor, usual, grave, evitable, calmo, serio, fino, eficaz, solo, cercano, excelente, pendiente, activo, frecuente, rápido, transparente, profesional, national, opaco, difícil, brillante, seguro, leve, raro, frágil, mediano, lento, potente, fácil, resistente, amplio, delgado, fuerte, american, ocular, específico, numeroso, precoz, completo, doloroso, sensible, infecto, sano, profundo, particular, especial, lleno, genial, increíble, grueso, pegajoso, oscuro, próximo, ciego, habitual, diminuto, joven, hereditario, presente, terapéutico, sólido, breve, incierto, dudoso, aceptable, claro, professional, antecedente, normal, estricto, mortal, duro, valioso, esencial, múltiple, regular, muscular, eficiente, desastroso, dañino, sensitivo, prominente, frustrante, típico, costoso, significativo, reciente, accesible, notorio, espeso, quieto, diario, delicado, vulnerable, constante, medio, proclive, estrecho, verdadero, evidente, clásico, enfermo, seco, denso, fijo, igual, menudo, distante, despierto, derecho, gafo, total, extenso, vertical, incómodo, agudo, tremendo, impresionante, inevitable, flexible, capaz, contento, general, soñoliento, característico, peor, preciso, international, riesgoso, rosado, invisible, ligero, suave, frío, tóxico, invasivo, variable, existente, pálido, futuro, débil, propenso, responsable, alérgico, nítido, graso, antiguo, prematuro, anormal, ácido, áspero, anciano, blanco, doble, letárgico, amarillo, flojo, rígido, tardío, extraño, cómodo, junto, poderoso, asimétrico, viscoso, orgulloso, moderno, lens, triste, famoso, posterior, difuso, sencillo, esférico, correcto, interno, externo, congénito, máximo, corriente, degenerativo, listo, problemático, enorme, explorador, malo, extremo, afecto

Uso de uma Ferramenta de Processamento de Linguagem Natural como Auxílio à Coleta de Exemplos para o Estudo de Propriedades Sintático-Semânticas de Verbos

Using a Natural Language Processing Tool to Assist the Collection of Samples for the Study of Syntactic-Semantic Properties of Verbs

Larissa Picoli
Universidade Federal do Espírito Santo
larissa_picoli@hotmail.com

Juliana Campos Pirovani
Universidade Federal do Espírito Santo
juliana.campos@ufes.br

Elias de Oliveira
Universidade Federal do Espírito Santo
elias@lcad.inf.ufes.br

Éric Laporte
Université Paris-Est
eric.laporte@univ-paris-est.fr

Resumo

A análise e descrição de propriedades sintático-semânticas de verbos são importantes para a compreensão do funcionamento de uma língua e fundamentais para o processamento automático de linguagem natural, uma vez que a codificação dessa descrição pode ser explorada por ferramentas que realizam esse tipo de processamento. Esse trabalho experimenta o uso do Unitex, uma ferramenta de processamento de linguagem natural, para coletar uma lista de verbos que podem ser analisados e descritos por um linguista. Isso contribui significativamente para esse tipo de estudo linguístico, diminuindo o esforço manual humano na busca de verbos. Foi realizado um estudo de caso para automatizar parcialmente a coleta de verbos de base adjetiva com sufixo *-ecer* em um *corpus* de 47 milhões de palavras. A abordagem proposta é comparada com a coleta manual e a extração a partir de um dicionário para o PLN.

Palavras chave

Unitex, Lista de verbos, Propriedades sintático-semânticas

Abstract

The analysis and description of syntactic-semantic properties of verbs are fundamental to both the knowledge of the grammar of a language and to the automatic processing of natural language, as an encoded form of this description can be exploited by automatic tools. This paper experiments with the use of Unitex, a natural language processing tool, to collect a list of verbs that can be analysed and described by a linguist. This work contributes significantly to linguistics, by decreasing the human manual effort in the search for verbs. A case study is performed to partially automate the collection of verbs in *-ecer* with adjectival bases in a *corpus* of 47 million words. The proposed approach is compared with manual collection and with extraction from an NLP dictionary.

Keywords

Unitex, List of verbs, Syntactic-semantic properties

1 Introdução

O Processamento de Linguagem Natural (PLN) é uma área interdisciplinar que estuda a geração, representação e compreensão automática de fala e textos em línguas naturais. As aplicações do PLN incluem tradução automática, reconhecimento automático de voz, geração automática de resumos, recuperação de informação, correção ortográfica e outras ferramentas que auxiliam a escrita. De acordo com Vieira & Lima (2001), o PLN busca a construção de programas capazes de interpretar e/ou gerar informação fornecida em linguagem natural. Contudo, Laporte (2009) destaca que dicionários (léxicos) e gramáticas para o PLN, que podem ser construídos artesanalmente por linguistas, são também fundamentais para a implementação de ferramentas de qualidade. Os linguistas são responsáveis pela escolha de modelos de análise oriundos da teoria linguística, pela análise e descrição da língua e pela construção e atualização dos dicionários e gramáticas.

Em prática, nesse processo, são coletadas palavras e expressões, por exemplo expressões multipalavra em geral (Ranchhod, 2005), nomes de

This work is licensed under a Creative Commons Attribution 3.0 License

pessoas (Bayraktar & Temizel, 2008), entidades nomeadas (Traboulsi, 2009), expressões jurídicas (Chieze et al., 2010), expressões de sentimento (Duran & Ramisch, 2011), expressões metafóricas (Müller, 2014), expressões com verbo-suporte[1] (*Vsup*) (Barros, 2014; Rassi et al., 2015). Os linguistas analisam os itens encontrados, tendo em vista uma descrição sintático-semântica em dicionários para o PLN.

Este artigo compara vários métodos de coleta de palavras e expressões. O contexto da coleta é uma descrição das propriedades sintático-semânticas de verbos de base adjetiva com os sufixos *-ecer* e *-izar* (Smarsaro & Picoli, 2013; Picoli, 2015).[2]

Três abordagens são focadas: a coleta manual por introspeção ou com auxílio da *web*, como no estudo inicial de Picoli (2015); a extração a partir de um dicionário para PLN existente; e a extração a partir de um *corpus* de textos com o Unitex (Paumier, 2015), uma ferramenta de PLN.

A descrição de Picoli (2015) visa as derivações que apresentam equivalência semântica entre a frase de base (1) e a frase transformada (2):

(1) *O sol aqueceu a areia*
Fórmula: N_0 *Adj-v* N_1

(2) *O sol tornou a areia quente*
Fórmula: N_0 *tornar* N_1 *Adj*

A frase de base (1) contém o verbo com o sufixo *-ecer*, no caso *aquecer*. Na fórmula sintática correspondente, N_0 significa nome ou grupo nominal que ocupa a posição de sujeito na frase base, *Adj-v* denota verbo de base adjetiva e N_1 significa nome ou grupo nominal que ocupa a posição de complemento do predicado na frase base. A frase transformada (2) contém o verbo *tornar* e o adjetivo, no caso *quente*.

A autora selecionou, para a descrição sintático-semântica, apenas verbos que admitem a correspondência semântica, como em (1) e (2). Foi construída uma lista de 88 verbos de base adjetiva com sufixo *-ecer*. A partir dessa seleção de verbos, a autora analisou e descreveu suas propriedades sintático-semânticas formais (estruturais, distribucionais e transformacionais), que são aquelas relacionadas à natureza dos argumentos admitidos pelo verbo e às transformações que o verbo pode sofrer. As propriedades foram formalizadas em uma tabela do Léxico-gramática, segundo o formato proposto por Gross (1975).

Essa tabela, que foi obtida observando o comportamento sintático-semântico dos verbos numa estrutura frasal e descreve construções sintáticas, pode ser importante para aplicações como a tradução de um texto em português para outra língua. Em português, por exemplo, *brutalizar* pode equivaler semanticamente a *tornar brutal*, mas em francês, a tradução de *tornar brutal*, que é *rendre brutal*, não equivale semanticamente ao verbo *brutaliser* "tratar com brutalidade". Com os resultados desse estudo, o verbo *brutalizar* poderia ser traduzido por *rendre brutal*. Além das aplicações para o PLN, esse tipo de base de dados permite um melhor conhecimento do uso de frases e de suas transformações, contribuindo para o ensino da língua.

Esse artigo está estruturado em 6 seções. Na Seção 2, são apresentados alguns trabalhos correlatos que examinam métodos de coleta de listas de palavras e expressões. A Seção 3 apresenta a metodologia utilizada no desenvolvimento desse trabalho. Os grafos construídos no Unitex para coleta dos verbos e exemplos desejados são apresentados na Seção 4. Na Seção 5 são apresentados e discutidos os resultados do trabalho realizado e a Seção 6 apresenta as conclusões e trabalhos futuros.

2 Revisão de literatura

A literatura científica descreve as três abordagens de coleta focadas neste artigo, mas não encontramos publicações que comparassem várias abordagens.

Em várias pesquisas linguísticas recentes na área da descrição lexical, a coleta de itens lexicais e/ou exemplos para análise é realizada com auxílio da *web* e/ou por introspecção (Rodrigues, 2009; Davel, 2009; Pacheco & Éric Laporte, 2013; Picoli, 2015).

Listas de palavras ou expressões são também extraídas de dicionários preexistentes. Por exemplo, Barros (2014) extrai do trabalho de Chacoto (2005) uma parte de sua lista de predicados nominais com o verbo-suporte *fazer*, e Rassi et al. (2014) aproveitam as listas de Vaza (1988) e Bap-

[1] Uma expressão com verbo-suporte comporta um verbo, mas o papel de núcleo do predicado e a seleção dos argumentos não são cumpridos pelo verbo, e sim por um item de outra categoria gramatical, geralmente um nome ou um adjetivo, como em *ter inveja* ou *ser invejoso* (Neves, 1999). Sendo essa definição semântica pouco precisa, critérios sintáticos foram estabelecidos para cada idioma, sempre verificando a existência de uma construção em que o verbo-suporte pode ser removido sem mudança de sentido, como em *a inveja que João tem/a inveja de João* (Gross, 1981; Langer, 2005; Rassi et al., 2015). As construções com verbo-suporte são um dos principais tipos de expressões multipalavra.

[2] O objetivo da coleta é a construção de um dicionário, mas o objetivo do artigo é a comparação de métodos de coleta.

tista (1997) para constituir uma lista de predicados nominais com os verbos-suporte *ter* e *dar*.

A terceira abordagem consiste no uso de uma ferramenta de PLN, o Unitex ou outra, para coletar itens lexicais ou exemplos automaticamente em um *corpus* de textos. Essas ferramentas realizam processamentos que incluem a segmentação em frases, a segmentação em palavras ou tokenização, a classificação gramatical de palavras e a busca em textos. Dessa forma, os linguistas podem se beneficiar do PLN por meio das ferramentas construídas pelos profissionais da computação, da mesma forma em que, simetricamente, a qualidade do PLN pode depender da descrição da língua pelos linguistas.

Assim, Ranchhod (2005), Bayraktar & Temizel (2008), Traboulsi (2009), Chieze et al. (2010), Barros (2014), Rassi et al. (2015) utilizam o Unitex para coletar diversos tipos de expressões em diversos *corpora*. Rassi et al. (2015) extraem uma lista de 4.668 construções com *Vsup*, que totalizam 45 variantes de *Vsup* e 3.200 nomes diferentes. Para criar essa lista, os autores elaboram grafos no Unitex e aplicam esses grafos a um corpus de 103.080 textos do jornal *Folha de São Paulo*. Müller (2014) coleta expressões metafóricas com o NooJ, uma ferramenta historicamente relacionada com o Unitex e com funcionalidades e funcionamento próximos.

O Unitex[3] (Paumier, 2015) é um sistema *open-source* para o PLN, desenvolvido inicialmente na universidade Paris-Est Marne-La-Vallée (França), disponível gratuitamente e utilizado por empresas de PLN. Os alicerces do Unitex (Silberztein, 1994) foram elaborados no Laboratoire d'Automatique Documentaire et Linguistique (LADL), dirigido por Maurice Gross, que guiou um trabalho de análise e descrição sintático-semântica do francês. O Unitex é distribuído com dicionários desenvolvidos para vários idiomas pela rede de laboratórios RELEX, incluindo um dicionário do português do Brasil (Muniz et al., 2005).

A ferramenta aplica dicionários a *corpora* não anotados e extrai informações através de expressões regulares e redes de transições recursivas representadas como grafos. Um exemplo de grafo no Unitex é apresentado na Figura 1. Esse grafo reconhece *o menino bonito* e *o garoto inteligente*. O código `<A>` reconhece um adjetivo: qualquer símbolo que aparecer entre "<" e ">" é interpretado pelo sistema como código de propriedade lexical nos dicionários ou como lema. O padrão simples da Figura 1 poderia ser representado na forma de uma expressão regular igualmente legível, mas a representação de padrões complexos por grafos é mais conveniente do que por expressões regulares para o leitor humano.[4] A função dos padrões não se limita ao reconhecimento dos próprios itens a extrair, e pode também se estender ao reconhecimento do contexto desses itens, em caso de ambiguidade.

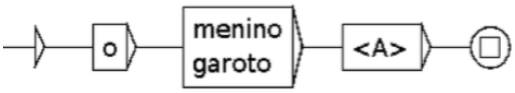


Figura 1: Exemplo de grafo no Unitex.

Além do Unitex, concebido por linguistas com experiência na descrição lexical e gramatical em grande escala, outras ferramentas são utilizadas para coleta de itens lexicais e exemplos. Por exemplo, Duran & Ramisch (2011) extraem expressões de sentimento com o auxílio do sistema mwetoolkit. Arranz et al. (2005) extraem ocorrências de expressões verbais cristalizadas morfologicamente diferentes da forma registrada no dicionário WordNet 1.6. A tecnologia dessas duas experiências tem pontos comuns com os sistemas de NLP padrão como GATE[5] e Stanford NLP Software[6], oriundos de uma inspiração voltada para a computação, e é menos adaptada à coleta do que o Unitex, por duas razões.

Primeiro, a única forma de definir padrões é a das expressões regulares, que convém para padrões simples como NA (nome seguido de adjetivo) e NAA, mas passa a ser menos legível do que os grafos quando os padrões se tornam mais complexos. Para extrair os padrões desse trabalho, as expressões se tornariam muito complexas e de difícil leitura.

A segunda razão é uma consequência da ausência de dicionários nesses sistemas. O reconhecimento das categorias gramaticais especificadas nos padrões depende da anotação das palavras do *corpus*. Essa anotação é realizada por etiquetadores automáticos, que cometem erros. Se as anotações estão revisadas, o custo da revisão limita o tamanho dos *corpora* disponíveis em português, e a diversidade dos textos. Se as anotações não foram revisadas, os erros de etiquetagem podem impedir a extração de ocorrências.

[3] http://www-igm.univ-mlv.fr/~unitex/

[4] Por exemplo, no grafo da Figura 6, as linhas que saem do nó inicial indicam visualmente quais nós correspondem ao início de cada variante do padrão. Numa expressão regular equivalente, mesmo apresentada com uma indentação cuidadosa, a identificação do início de cada variante do padrão necessita que o leitor distinga os operadores união e concatenação, o que é menos evidente visualmente do que a oposição entre a presença/ausência de uma linha entre dois nós.

[5] https://gate.ac.uk/

[6] http://nlp.stanford.edu/software/

Nos dois casos, a abrangência da extração é prejudicada. Com o Unitex, o uso de um dicionário de grande cobertura garante uma abrangência elevada.

A Sketch Engine (Kilgarriff et al., 2014) poderia ser utilizada para a coleta de palavras e expressões, com as duas dificuldades que acabamos de descrever e outras: não permite a definição de padrões, não pode ser utilizada em linha de comando e o uso é cobrado.

Cook et al. (2008) coletam expressões verbais cristalizadas no BNC, um *corpus* anotado e revisado. No caso do português, o custo da revisão limita o tamanho dos *corpora* disponíveis e a diversidade dos textos, prejudicando a abrangência da extração.

3 Metodologia

No estudo inicial, Picoli (2015) coletou os itens desejados por introspecção e manualmente, em diversos materiais como a *web* e dicionários. A descrição sintático-semântica necessitava selecionar apenas verbos que admitissem a correspondência semântica, como em (1) e (2). Essa operação, necessitada nas três abordagens, produziu uma lista de 88 verbos em *-ecer*.

A segunda abordagem, a extração dos verbos em -ecer no dicionário de lemas do Unitex (Muniz et al., 2005) com o comando *grep*, produz uma lista de 298 verbos que necessita uma revisão.

A contribuição principal deste artigo está na terceira abordagem, em que os verbos derivados com sufixo *-ecer* e os exemplos práticos reais de frases com esses verbos são tirados de um *corpus* de textos publicados.

As fórmulas (1) e (2) podem representar frases com outros verbos. Além dos verbos de base adjetiva com sufixo *-ecer*, como *enriquecer*, os verbos de base adjetiva com sufixo *-izar* também podem ser inseridos em frases com as mesmas fórmulas, por exemplo:

(3) *O adubo fertilizou a terra*
Fórmula: N_0 *Adj-v* N_1

(4) *O adubo tornou a terra fértil*
Fórmula: N_0 *tornar* N_1 *Adj*

Ferramentas de PLN podem ser usadas para buscar automaticamente frases que possuam determinadas estruturas, como as apresentadas em (1) e (2). A ferramenta de PLN utilizada neste trabalho foi o Unitex.[7]

[7] http://www-igm.univ-mlv.fr/~unitex/

Dadas as construções sintáticas descritas em Picoli (2015), representadas pelas fórmulas (1) e (2), grafos foram construídos no Unitex para reconhecer essas estruturas. Em seguida, o Unitex foi utilizado para buscar frases a partir desses grafos no *corpus* da *Tribuna*, um *corpus* de 45.908 textos jornalísticos escritos em português e publicados pelo jornal do Espírito Santo *A Tribuna*[8]. Os arquivos do *corpus* possuem, em média, 1.032 palavras. Esse *corpus*[9] possui textos publicados nos anos de 2002 a 2006. Os textos abordam assuntos diversos como Economia, Política, Família, Ciência e Tecnologia, Concursos, TV, etc.

A aplicação de grafos do Unitex a cada arquivo do *corpus* gera um arquivo de concordância que lista as ocorrências de frases identificadas pelos grafos. O Unitex permite que *tags* sejam adicionadas aos arquivos de concordância. Assim, os verbos foram colocados entre as *tags* `<verbo>` e `</verbo>` e os adjetivos entre as *tags* `<adj>` e `</adj>`. Os arquivos de concordância foram concatenados gerando um único arquivo com todas as ocorrências identificadas no *corpus* para cada grafo utilizado.

Após essa primeira etapa, os verbos e adjetivos identificados pelos dois grafos construídos por meio das fórmulas (1) e (2), respectivamente, são extraídos dos arquivos de concordância de cada grafo e dois novos arquivos são criados contendo esses verbos e adjetivos, sem repetição. Assim, para cada grafo, dois arquivos são gerados para análise de um especialista: um arquivo com os itens lexicais (`Verbos.txt` ou `Adjetivos.txt`) identificados e um arquivo com todas as frases do *corpus* utilizado onde esses itens aparecem (`ExemplosVerbos.txt` e `ExemplosAdjetivos.txt`).

Apenas a construção dos grafos foi realizada manualmente. Todas as etapas seguintes foram realizadas por um programa de computador[10], implementado em *shell script*, com essa finalidade, que utiliza as ferramentas do Unitex. O pseudocódigo do programa é apresentado na Figura 2.

As linhas de 2 a 5 mostram a aplicação de um grafo G a cada arquivo de entrada D gerando a concordância C. A linha 4 concatena o conteúdo do arquivo C gerado em um arquivo chamado `ConcordGeral.txt`. Nas linhas de 7 a 22, esse arquivo é lido linha a linha e, depen-

[8] http://www.redetribuna.com.br/jornal

[9] http://www.inf.ufes.br/~elias/dataSets/aTribuna-21dir.tar.gz

[10] http://www.inf.ufes.br/~elias/codes/linguamatica2015.tar.gz

$Grafos = \{G_1, G_2\}$ um repositório de grafos onde G_1 é o grafo construído para fórmula (1) e G_2 o grafo construído para fórmula (2)

$Textos = \{D_1, D_2, \ldots, D_N\}$ um repositório com os arquivos de entrada do corpus *A Tribuna* onde $N = 45908$

```
1  for G in Grafos do
2    for D in Textos do
3      Aplique G em D gerando C
4      cat C >> ConcordGeral.txt
5    done
6
7    while read linha do
8      if [ G == G1 ] then
9        verbo = palavra entre <verbo> e </verbo>
10       vFCanonica=`grep ^$verbo dlf | cut -d"," -f2 |
11                   cut -d "." -f1 | head -1`
12       Se vFCanonica finaliza com -ecer
13         coloque verbo no arquivo Verbos.txt
14         coloque a linha no arquivo ExemplosVerbos.txt
15     fi
16
17     if [ G == G2 ] then
18       adjetivo = palavra entre <adj> e </adj>
19       coloque o adjetivo no arquivo Adjetivos.txt
20       coloque a linha no arquivo ExemplosAdjetivos.txt
21     fi
22   done < ConcordGeral.txt
23 done
```

Figura 2: Pseudocódigo do programa em shell script.

dendo do grafo que foi aplicado, é tomada uma decisão. Se o grafo aplicado é o que reconhece a fórmula (1), deve-se obter o verbo entre as *tags* <verbo> e </verbo> e verificar se o lema (ou forma canônica) dele termina com sufixo *-ecer*. Em caso afirmativo, o verbo e a linha são colocados nos arquivos correspondentes. Se o grafo é o que reconhece a fórmula (2), basta obter o adjetivo que aparece entre as *tags* <adj> e </adj> e colocar o adjetivo e a linha nos arquivos correspondentes.

O lema do verbo é obtido a partir do arquivo chamado *dlf* gerado pelo Unitex ao aplicar os dicionários durante o pré-processamento de um texto (linha 10). O *dlf* é um dicionário gerado para as palavras simples do texto. Um exemplo de linha no arquivo *dlf* é:

```
entristeceu,entristecer.V:J3s
```

O lema está em negrito no exemplo. Ele aparece após a palavra que ocorre no texto original (forma flexionada) e é separado dela por uma vírgula. A classificação gramatical da palavra aparece após um ponto que segue o lema. No exemplo, a classe verbo é representada pelo código V no dicionário do Unitex. O que segue a classificação gramatical após ":" são as informações flexionais. Nesse caso, J representa passado, 3 indica terceira pessoa e S singular.

Feito isto, o linguista realiza uma análise minuciosa dos arquivos gerados para verificar se os verbos encontrados fazem parte do grupo de verbos que se pretende descrever e se os adjetivos encontrados derivam verbos relevantes. Em uma busca para selecionar verbos com final *-ecer* por exemplo, é possível que a ferramenta encontre alguns tipos de verbos: os de base adjetiva com sufixo *-ecer*, como *apodrecer*; os de base substantiva com o sufixo *-ecer*, como *amanhecer*; e ainda os que não são formados por derivação, como *acontecer*. O trabalho do linguista é selecionar, a partir da busca feita com o Unitex, o grupo de itens lexicais que pretende descrever.

4 Descrição dos grafos construídos no Unitex

O grafo responsável por reconhecer a estrutura (1) é apresentado na Figura 3. SN é utilizado para reconhecer um sintagma nominal. O

reconhecimento de sintagmas nominais foi detalhado previamente em um outro grafo, que será apresentado posteriormente, e incluído nesse como subgrafo. Referências a subgrafos são representados em nós com fundo cinza pelo Unitex. O código <V> no nó do grafo é utilizado durante o processamento pelo Unitex para reconhecer verbos.

Figura 3: Grafo que reconhece frases que possuem a estrutura da Fórmula (1) criado no Unitex.

O Unitex permite inserir saídas (texto em negrito sob setas) no grafo. Existem três modos de utilizar as saídas ao aplicar um grafo para identificar padrões em um texto. As saídas podem ser ignoradas (opção "are not taken into account"), podem ser usadas para substituir a sequência reconhecida no arquivo de concordância (opção "REPLACE recognized sequences") ou podem ser inseridas no arquivo de concordância (opção "MERGE with input text"). Na Figura 3, saídas são utilizadas para inserir as *tags* <verbo> e </verbo> no arquivo de concordância. Assim, aplicando esse grafo no modo "MERGE with input text", o verbo identificado será apresentado entre essas *tags* no arquivo de concordância.

Esse grafo reconhece qualquer verbo e não apenas os verbos de base adjetiva derivados com sufixo *-ecer*. Um filtro foi realizado posteriormente pelo programa implementado usando os resultados da aplicação dos dicionários do Unitex, como apresentado no pseudocódigo.

A Figura 4 apresenta uma amostra do arquivo de concordância obtido com a aplicação do grafo da Figura 3. Alguns dos exemplos listados não contêm o verbo entre dois sintagmas nominais, mas foram extraídos assim mesmo porque a descrição dos sintagmas nominais no grafo SN inclui poucas restrições e por causa das homonímias entre entradas lexicais no dicionário do Unitex. O {S} é um símbolo separador de frases inserido no Unitex, durante o pré-processamento do texto.

Para reconhecer a estrutura (2), foi criado o grafo apresentado na Figura 5. O código <tornar.V:J> é utilizado durante processamento pelo Unitex para reconhecer verbos no passado que tem *tornar* como lema. O código <A> reconhece um adjetivo. As saídas <adj> e </adj> são utilizadas para *taguear* o adjetivo identificado no arquivo de concordância. O grafo SN utilizado como subgrafo nas Figuras 3 e 5 para reconhecer sintagmas nominais é apresentado na Figura 6. Esse grafo reconhece as principais estruturas de sintagmas nominais que devem ser encontradas nas frases buscadas. Observe na Figura 6 que ele também inclui um outro subgrafo, Preposicao.grf, usado para reconhecer preposições.

Código Unitex	Reconhecimento Lexical
<N>	substantivos
<PRO + Pes>	pronomes pessoais
<DET>	determinantes
<PRO + Dem>	pronomes demonstrativos
<A>	adjetivos

Tabela 1: Códigos do Unitex usados na Figura 6 e seu significado.

A Tabela 1 mostra o significado dos códigos usados nesse grafo. Observe que o símbolo + pode ser usado para reconhecer sequências de informações gramaticais ou semânticas, isto é, sequências mais específicas. Por exemplo, o código <PRO> é usado para reconhecer quaisquer pronomes. Já o código <PRO+Pes>, usado no grafo SN, reconhecerá apenas os pronomes pessoais.

Alguns exemplos de sintagmas nominais que podem ser reconhecidos pelo grafo da Figura 6 são:

a) *João.*

b) O menino bonito.

c) Aquele belo sapato.

5 Resultados e discussão

As três abordagens focadas neste artigo podem ser comparadas em termos de abrangência, esforço e tempo.

A abrangência é o critério mais importante porque mede a qualidade e a quantidade do resultado. A abordagem manual produziu 88 entradas em *-ecer* depois da revisão. A extração automática a partir do dicionário Unitex-PB forneceu 298 entradas em *-ecer* que necessitavam revisão, mas a revisão não foi feita para evitar o efeito de repetição em que a mesma operação sobre as mesmas entradas pelo mesmo linguista se torna mais fácil e rápida a cada iteração.

A extração a partir do corpus pelo grafo da Figura 3 apresentou nos arquivos de concordância 79 verbos diferentes com sufixo *-ecer* e 10.693 exemplos de frases com esses verbos. Dos 79 verbos, foi verificado manualmente por um linguista que 27 são de base adjetiva e possuem correspondência semântica quando inseridos em frases

```
{S} A empresa<verbo> oferece</verbo> cursos para formação
{S} Este recurso<verbo> fornece</verbo> som excepcional
{S} a, na Praça do Papa, e já tem gente<verbo> aquecendo</verbo> as turbinas para
{S} Eu<verbo> agradecia</verbo> a Deus 24 horas por dia principalmente,
    me<verbo> entristeceu</verbo> muito
{S} Isso me<verbo> emagreceu</verbo> e me deixou mais saudável.
{S} A pele<verbo> apodreceu</verbo> e , se a infecção se genera
```

Figura 4: Parte da concordância obtida em modo merge com o grafo da Figura 3.

Figura 5: Grafo que reconhece frases que possuem a estrutura da Fórmula (2) criado no Unitex.

como (1) e (2): vários dos verbos com sufixo *-ecer* não eram de base adjetiva e não foram selecionados, como o verbo *oferecer* apresentado na Figura 4. Comparando a lista obtida com a de Picoli (2015), foi observado que um dos verbos identificados que possui correspondência semântica, *enraivecer*, não consta na lista de verbos descritos pela autora.

Já o grafo da Figura 5 deu 177 adjetivos e 234 exemplos de frases com esses adjetivos. Dos 177 adjetivos, foi verificado manualmente que 9 têm derivados em *-ecer* que admitem a correspondência semântica em frases como (1) e (2). Nessa lista de adjetivos, apenas 3 formam verbos que não foram coletados a partir do grafo da Figura 3. Assim, utilizando a extração a partir do *corpus*, foi construída uma lista final com 30 verbos (27 coletados pelo grafo da Figura 3 e 4 outros coletados a partir do grafo da Figura 5).[11] Esse resultado é comparável com os 88 verbos apresentados em Picoli (2015). Assim, a extração a partir de um grande *corpus* homogêneo é eficiente para coleta de verbos e adjetivos, mas menos do que a coleta manual, que leva em conta as várias comunidades em que o pesquisador está inserido.

Todavia, os dois métodos de coleta são complementares. O Unitex extraiu do *corpus* um grande número de exemplos de frases com esses verbos que podem ser utilizados no estudo e na atestação de suas propriedades. Analisando exemplos de frases com o mesmo verbo, podem-se observar propriedades distintas. Tal riqueza de exemplos é complementar à análise por introspecção, indispensável para identificar, entre outras, as propriedades que os verbos não possuem.

Enfim, a comparação entre a abordagem por dicionário e as duas outras sugere que o dicionário contém entradas pouco usadas, que podem dificilmente ser submetidas a um estudo aprofundado de suas propriedades sintático-semânticas, pois tal estudo pressupõe que o linguista domine o uso das entradas em todas suas construções.

O segundo critério, o do esforço humano desempenhado, é difícil de avaliar, mas durante esta experiência, achamos mais trabalhosa a busca lexical sem ferramentas computacionais.

O terceiro critério é a duração do processo de coleta.

No estudo inicial, Picoli (2015) não registrou uma estimação do tempo gasto para a busca manual de itens lexicais, nem para a seleção dos itens que possuem correspondência semântica quando inseridos em frases como (1) e (2). A busca manual não consistiu em ler grandes quantidades de textos; contudo, considerando que aproximadamente 4 minutos são necessários, em média, para ler um dos arquivos do *corpus*, seria necessário aproximadamente 3.000 horas para apenas ler todos os 45.908 arquivos desse *corpus*.

Na segunda abordagem, a resposta do sistema é imediata.

Na terceira abordagem, o programa implementado para coleta dos verbos e adjetivos foi executado em um computador com as seguintes características: processador Intel core i5, memória de 4GB, sistema operacional Ubuntu 14.04. O tempo de execução para coleta dos adjetivos foi 1h35min e para coleta dos verbos 3h54min. Essa diferença de tempo é consequência do tempo adicional necessário para gerar outro arquivo de concordância mantendo apenas os verbos com sufixo *-ecer*. Portanto, esses verbos foram identificados em poucas horas, e o processo pode ser agilizado, concatenando todos os textos do *corpus* em um único arquivo e apli-

[11] Esse mesmo grafo poderia ser usado para coletar adjetivos que derivam verbos com outros sufixos, como *-izar*. Analisando rapidamente as listas, verificou-se que 13 adjetivos extraídos do corpus têm derivados em *-izar* e 272 verbos têm o sufixo *-izar*. Portanto, esse sufixo é mais frequente e mais produtivo neste corpus do que *-ecer*.

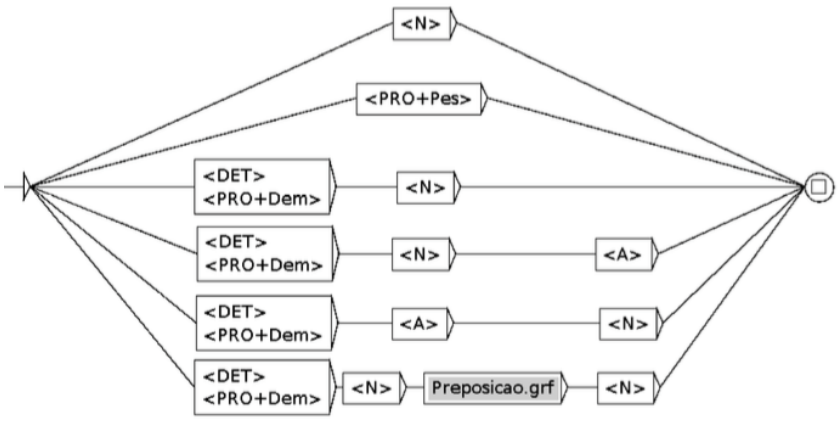


Figura 6: Subgrafo que reconhece sintagmas nominais (SN.grf) criado no Unitex.

cando cada grafo ao *corpus* completo. Para verificar se os verbos são de base adjetiva e se possuem correspondência semântica em frases como (1) e (2), analisando cada verbo e adjetivo extraído para verificar se eles ocorrem na estrutura buscada e se possuem as propriedades de interesse, o linguista gastou cerca de 2 minutos por verbo.

Assim, com a utilização de uma ferramenta de PLN, o linguista pode, em um curto período de tempo, extrair uma parte do seu objeto de análise, que são neste estudo os verbos que aparecem na estrutura sintática de interesse.

6 Conclusão

Nesse trabalho foi utilizada uma ferramenta de processamento de linguagem natural, o Unitex, para auxiliar o recenseamento de verbos de base adjetiva com sufixo *-ecer* e de exemplos de frases com estes verbos, para análise e descrição de propriedades sintático-semânticas.

Foram construídos dois grafos no Unitex: um para coleta de frases contendo verbos e outro para coleta de frases contendo adjetivos que podem derivar os verbos de interesse. Esses grafos foram construídos a partir das fórmulas (1) e (2) que descrevem as estruturas de frase cuja equivalência semântica poderia ser estudada como em Picoli (2015). Foi implementado um programa que aplica esses grafos no *corpus* da *Tribuna* e gera arquivos para análise de um linguista.

O linguista deve selecionar para descrição apenas aqueles verbos com as características desejadas: base adjetiva e correspondência semântica em frases como (1) e (2). Foram identificados 30 verbos com essas características, sendo que um deles não fora identificado pelo estudo de Picoli (2015).

A coleta dos verbos por este método produz resultados menos abrangentes, mas também é menos trabalhosa para os linguistas que buscam descrever a língua portuguesa. O linguista pode ainda poupar esforço e tempo com o Unitex por meio de outros métodos: por exemplo, extraindo todos os 298 verbos com sufixo *-ecer* do dicionário de lemas do Unitex, antes de revisar a lista obtida.

Como trabalho futuro, pretende-se realizar o mesmo experimento em outros *corpus*, maiores e de gêneros textuais diferentes do apresentado no jornal *A Tribuna* e realizar a coleta de itens em dicionários existentes. Pretende-se também repetir a experiência com versões mais completas dos grafos. Por exemplo, pode ser incluído no grafo da Figura 3 o reconhecimento de frases com o verbo na primeira pessoa do singular, sem pronome sujeito explícito, como em *Fortaleço meus braços*. Da mesma forma, é possível relaxar as restrições do grafo da Figura 5, para aceitar o verbo em outros tempos que não o perfeito (*O sol torna a areia quente*), ou aceitar os complementos na ordem inversa (*O sol tornou quente a parte superficial*). Grafos adicionais podem ser construídos para representar construções frequentes associadas a (1), por exemplo,

$$N_1 \; Adj\text{-}v \leftarrow A \; areia \; aqueceu.$$

Além disso, pretende-se também criar uma interface *web* onde o linguista poderá informar as fórmulas sintáticas de interesse, tais como as apresentadas nas fórmulas (1) e (2). Essas estruturas serão convertidas automaticamente em grafos do Unitex que serão usados para buscá-las em alguns corpora internos predefinidos. Assim, uma lista de verbos será criada automaticamente para que o linguista continue seu trabalho.

Referências

Arranz, Victoria, Jordi Atserias & Mauro Castillo. 2005. Multiwords and word sense disambiguation. Em Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 3406 Lecture Notes in Computer Science, 250–262. Springer Berlin Heidelberg.

Baptista, Jorge. 1997. Sermão, tareia e facada. uma classificação das construções conversas dar-levar. Em *Seminários de Linguística*, vol. 1, 5–37.

Barros, Cláudia D. 2014. *Descrição de classificação de predicados nominais com verbo-suporte fazer: especificidades do Português do Brasil*: Universidade Federal de São Carlos. Tese de Doutoramento.

Bayraktar, Özkan & Tuğba Taşkaya Temizel. 2008. Person name extraction from Turkish financial news text using local grammar-based approach. Em *23rd International Symposium on Computer and Information Sciences*, 1–4.

Chacoto, Lucília Maria Vieira Gonçalves. 2005. *O verbo 'fazer' em construções nominais predicativas*: Universidade de Algarve. Tese de Doutoramento.

Chieze, Emmanuel, Atefeh Farzindar & Guy Lapalme. 2010. An automatic system for summarization and information extraction of legal information. Em Enrico Francesconi, Simonetta Montemagni, Wim Peters & Daniela Tiscornia (eds.), *Semantic Processing of Legal Texts*, 216–234. Springer-Verlag.

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2008. The VNC-Tokens dataset. Em *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE)*, 19–22.

Davel, Alzira da P. C. 2009. *Um estudo sobre o verbo-suporte na construção Dar+SN*: UFES. Tese de Mestrado.

Duran, Sanches & Carlos Ramisch. 2011. How do you feel? Investigating lexical-syntactic patterns in sentiment expression. Em *Proceedings of Corpus Linguistics*, online.

Gross, Maurice. 1975. *Méthodes en syntaxe. régime des constructions complétives.* Hermann.

Gross, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 15(63). 7–52.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch engine: ten years on. *Lexicography* 1(1). 7–36.

Langer, Stefan. 2005. A linguistic test battery for support verb constructions. *Lingvisticae Investigationes* 27(2). 171–184.

Laporte, Éric. 2009. Lexicons and grammar for language processing: industrial or handcrafted products? Em L. Rezende, B. Dias da Silva & J. B. Barbosa (eds.), *Léxico e gramática: dos sentidos à descrição da significação*, 51–84. Cultura Acadêmica.

Müller, Ralph. 2014. NooJ as concordancer in computer-assisted textual analysis. the case of the German module. Em *Formalising Natural Languages with NooJ 2013: Selected papers*, 203–214.

Muniz, Marcelo C., Maria das Graças Volpe Nunex & Éric Laporte. 2005. UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. Em *Proceedings of the Workshop on Technology on Information and Human Language (TIL)*, 2059–2068.

Neves, Maria H. M. 1999. *Gramática de usos do português.* UNESP.

Pacheco, Wagner L. & Éric Laporte. 2013. Descrição do verbo cortar para o processamento automático de linguagem natural. Em *Dialogar é preciso. Linguística para o processamento de línguas*, 165–175. PPGEL/UFES.

Paumier, Sébastien. 2015. Unitex 3.1 user manual. Disponível em http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf.

Picoli, Larissa. 2015. *Descrição de verbos de base adjetiva derivados com os sufixos -ecer e -izar, para o processamento automático de linguagem natural*: UFES. Tese de Mestrado.

Ranchhod, Elisabete. 2005. Using corpora to increase Portuguese MWU dictionaries. Tagging MWU in a Portuguese corpus. Em Pernilla Danielsson & Martijn Wagenmakers (eds.), *Proceedings from the Corpus Linguistics Conference Series*, online.

Rassi, Amanda, Cristina Santos-Turati, Jorge Baptista, Nuno Mamede & Oto Vale. 2014. The fuzzy boundaries of operator verb and support verb constructions with dar "give" and ter "have" in Brazilian Portuguese. Em *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), COLING 2014*, 92–101.

Rassi, Amanda Pontes, Jorge Baptista & Oto Araújo Vale. 2015. Um corpus anotado de construções com verbo-suporte em português. *Gragoatá* 38. 207–230.

RELEX. 2015. The RELEX network. Disponível em http://infolingu.univ-mlv.fr/Relex/introduction.html. Acesso em: 13/07/2015.

Rodrigues, Carlos R. de S. 2009. *Descrição e formalização de estruturas com verbos de ação-processo para a elaboração de um parser*: UFES. Tese de Mestrado.

Silberztein, Max D. 1994. INTEX: a corpus processing system. Em *COLING 94*, vol. 1, 579–583.

Smarsaro, Aucione & Larissa Picoli. 2013. Propriedades sintático-semânticas de verbos adjecer. *Cadernos do CNLF (CiFEFil)* 17(2).

Traboulsi, Hayssam. 2009. Arabic named entity extraction: A local grammar-based approach. Em *International Multiconference on Computer Science and Information Technology*, vol. 4, 139–143.

Vaza, Aldina. 1988. *Estruturas com nomes predicativos e o verbo-suporte dar*: Faculdade de Letras, Universidade de Lisboa. Tese de Mestrado.

Vieira, Renata & Vera L. S. Lima. 2001. Linguística computacional: princípios e aplicações. Em Luciana Nedel (ed.), *IX Escola de Informática da SBC-Sul*, SBC-Sul.

El Test de Turing para la evaluación de resumen automático de texto

The Turing Test for Automatic Text Summarization Evaluation

Alejandro Molina[1], Juan-Manuel Torres-Moreno[1,2]

[1]Laboratoire Informatique d'Avignon - UAPV, France
[2]Ecole Polytechnique de Montréal, Québec

Resumen

Actualmente existen varios métodos para producir resúmenes de texto de manera automática, pero la evaluación de los mismos continua siendo un tema desafiante. En este artículo estudiamos la evaluación de la calidad de resúmenes producidos de manera automática mediante un método de compresión de frases. Abordamos la problemática que supone el uso de métricas automáticas, las cuales no toman en cuenta ni la gramática ni la validez de las oraciones. Nuestra propuesta de evaluación está basada en el test de Turing, en el cual varios jueces humanos deben identificar el origen, humano o automático, de una serie de resúmenes. También explicamos como validar las respuestas de los jueces por medio del test estadístico de Fisher.

Palabras clave

Evaluación de resumen automático, Compresión de frases, Test de Turing, Crowdsourcing

Abstract

Currently there are several methods to produce summaries of text automatically, but the evaluation of these remains a challenging issue. In this paper, we study the quality assessment of automatically generated abstracts. We deal with one of the major drawbacks of automatic metrics, which do not take into account either the grammar or the validity of sentences. Our proposal is based on the Turing test, in which a human judges must identify the source of a series of summaries. We propose how statistically validate the judgements using the Fisher's exact test.

Keywords

Text Summarization Evaluation, Sentence Compression, Turing Test, Crowdsourcing

1 Introducción

La compresión de frases consiste en eliminar las partes menos importantes de una oración o de una frase de manera automática. Aunque es un tema relativamente reciente, ya se han propuesto diversas aplicaciones para su uso, por ejemplo en los dispositivos móviles que cuentan con pantallas reducidas en tamaño y donde el número de caracteres mostrados es limitado. La compresión de frases permitiría reducir la extensión del texto mostrado y, de esta manera, incluir más información en un espacio reducido.

Otra aplicación es la de traducción automática de subtítulos. Un modulo de traducción automática puede estar acoplado con módulo de compresión de frases de manera que se garantice una longitud específica del texto traducido. La compresión de frases también puede servir para ayudar a las personas con problemas visuales. Grefenstette (1998) presenta un método de reducción de textos que tiene por objetivo disminuir el tiempo de lectura de un sintetizador para ciegos.

Si bien es cierto que el tema de resumen automático continúa siendo de mucho interés, la evaluación presenta aún muchos retos que necesitan ser considerados y estudiados. Por ejemplo, se sabe poco acerca de la subjetividad en los criterios para calificar un resumen. En este artículo abordaremos metódicamente el tema de la calidad mínima esperada para un resumen. Esto es, dejando de lado que el resumen sea bueno o malo, verificar que éste cumple con las expectativas mínimas esperadas. Dicho de otra manera, que no se pueda distinguir si el resumen ha sido producido por una máquina o por una persona.

En la sección 2 pondremos en evidencia la necesidad de considerar cierta calidad mínima en la evaluación. Después, en la sección 3 discutiremos algunos métodos de evaluación de resúmenes y puntualizaremos por qué algunos de ellos no resultan adecuados para evaluar resúmenes por

This work is licensed under a
Creative Commons Attribution 3.0 License

compresión. En las secciones siguientes propondremos un método de evaluación basado en el test de Turing cuyos resultados pueden ser validados estadísticamente con una prueba de hipótesis.

2 Resumen por compresión de frases

La compresión de frases fue definida por Knight & Marcu (2000) como un método de reducción de oraciones. Los autores proponen algoritmos para eliminar palabras de una frase, sin cambiar el orden, de manera que la secuencia resultante, considerada como una compresión de la original, puede o no ser una oración válida en Inglés.

En (Molina, 2013) se plantea usar la compresión de frases como un método para generar resúmenes de manera automática. La idea es eliminar ciertos elementos de las frases de un texto pero considerando su contexto original en lugar de comprimir las frases aisladas. Para esto, se propone dividir la oración en segmentos discursivos y luego, mediante un algoritmo basado en aprendizaje de máquina, se decide cuáles de los segmentos se pueden eliminar. Los criterios para generar el resumen son que éste sea más corto, informativo y gramaticalmente correcto.

Sin embargo, la evaluación de un resumen por compresión de frases es un tema que merece ser tratado cuidadosamente. A diferencia del método de resumen automático por extracción, el resumen por compresión de frases puede modificar la estructura gramatical de las oraciones.

Por ejemplo, considere el Cuadro 2 que presenta las 2^3 compresiones posibles de una frase con 3 segmentos. Sea φ = [En casa es útil tener un termómetro]$\_s_1$ [para saber con precisión]$\_s_2$ [si alguien de la familia tiene fiebre.]$\_s_3$. Note que las compresiones $\widetilde{\varphi}_3$, $\widetilde{\varphi}_4$, $\widetilde{\varphi}_6$ y $\widetilde{\varphi}_7$ no son gramaticalmente correctas o cambian el sentido original de la frase. Esto nos lleva a concluir que la evaluación de un resumen por compresión de frases debe considerar la validez de las frases resultantes. En la sección 3 discutiremos más en detalle algunos métodos de evaluación de resúmenes.

3 La evaluación de resúmenes

La evaluación del resumen automático ha sido una cuestión compleja, que ha propiciado el surgimiento de varios enfoques. En (Amigó et al., 2005) se discuten ampliamente varios métodos de evaluación pero nuestro interés principal es con respecto a la fuente de validación de los resúmenes que puede ser: manual o automática.

3.1 Evaluación manual

La evaluación manual consiste básicamente en la lectura y comparación de los resúmenes automáticos con respecto a los resúmenes producidos por humanos (Edmundson, 1969). Su principal ventaja es que el criterio humano es garantía de validez y pertinencia. Su principal desventaja es que a partir de un mismo texto se puede producir una infinidad de resúmenes válidos y esto puede provocar que los evaluadores no muestren acuerdo.

Como parte de los trabajos representativos de la evaluación manual esta el de Mani et al. (1999) en el que los autores proponen dar a los anotadores tanto los resúmenes producidos con métodos automáticos como los documentos originales. La hipótesis es que los originales, contienen frases-clave que determinan la temática del texto y por lo tanto éstas deben estar incluidas en un resumen. Los anotadores deben comparar los documentos con los resúmenes y verificar que estos últimos en efecto contengan dichas frases-clave.

Saggion & Lapalme (2000) propone otro método en la misma línea pero aplicado a resúmenes mono-documento. La variante es que en lugar de frases, se entrega a los anotadores una lista de conceptos-clave que deben ser mencionados en los resúmenes automáticos.

Orasan & Hasler (2007) propone evaluar la calidad de un resumen con un test comparativo. Los anotadores deben elegir el mejor resumen de entre un par tal que uno de ellos fue elaborado con una herramienta de resumen asistido por computadora (*Computer-Aided Summarization*) y el otro sin esta herramienta. Su hipótesis es que no existe diferencia, estadísticamente significativa, entre ambos tipos y por lo tanto, los anotadores son incapaces de distinguirlos.

3.2 Evaluación automática

La evaluación automática consiste en que un programa evalúe los resúmenes. Su principal ventaja es que permite tratar cantidades masivas de documentos. Sin embargo, muchos métodos automáticos, no consideran ni la coherencia ni la validez gramatical ni la sucesión retórica de las ideas. Es decir que para muchos de estos métodos no importa el orden lógico de las palabras sino simplemente si aparecen o no. En las subsecciones siguientes, mencionaremos las características de algunos métodos automáticos.

$\widetilde{\varphi}_1$	(s_1, s_2, s_3)	En casa es útil tener un termómetro para saber con precisión si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_2$	(s_1, s_3)	En casa es útil tener un termómetro si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_3$	(s_1, s_2)	En casa es útil tener un termómetro para saber con precisión.
$\widetilde{\varphi}_4$	(s_2, s_3)	Para saber con precisión si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_5$	(s_1)	En casa es útil tener un termómetro.
$\widetilde{\varphi}_6$	(s_2)	Para saber con precisión.
$\widetilde{\varphi}_7$	(s_3)	Si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_8$	$()$	

Cuadro 1: Ejemplo de las compresiones posibles de una frase.

3.2.1 Evaluación usando referencias: ROUGE

Uno de los métodos de evaluación automática más utilizados es ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Éste se utiliza incluso durante las campañas internacionales de *Document Understanding Conferences - Text Analysis Conference* desde el año 2008. La idea general es comparar un resumen candidato (automático) con varios resúmenes elaborados por expertos, llamados resúmenes modelo o referencias. Así, la métrica de evaluación se basa en la coocurrencia de n-gramas entre el resumen candidato y los de referencia. La ecuación (1) corresponde a la fórmula para calcular la métrica ROUGE-N (R_n), donde n es el tamaño del n-grama y $Cnt_m(gram_n)$ es el número máximo de n-gramas que aparecen tanto en el resumen candidato como en las referencias (refs).

$$R_n = \frac{\sum_{S\epsilon\{\text{refs}\}} \sum_{gram_n \epsilon S} Cnt_m(gram_n)}{\sum_{S\epsilon\{\text{refs}\}} \sum_{gram_n \epsilon S} Cnt(gram_n)} \quad (1)$$

Si bien ROUGE ha tenido gran aceptación entre la comunidad, no resulta adecuado para evaluar resúmenes con frases comprimidas. Note que la métrica se basa en la cobertura entre el resumen candidato y el conjunto de los resúmenes de referencia. Dado que los resúmenes con frases comprimidas contienen menos palabras, estos son penalizados aunque su contenido sea pertinente. Por esta razón, ROUGE no es adecuado para evaluar la calidad de un resumen producido mediante la compresión de frases.

3.2.2 Evaluación inspirada en la traducción automática: BLEU

En (Molina et al., 2010), se estudia la evaluación de frases comprimidas mediante una métrica semiautomática, desarollada originalmente por IBM para la tarea de la traducción automática, llamada Bleu (Papineni et al., 2002). Inspirada por ROUGE, esta métrica está basada en la precisión entre los n-gramas de una frase candidata y un conjunto de frases de referencia. La idea general es calcular la proporción de n-gramas de la frase candidata presentes en las referencias y el número total de n-gramas de la frase candidata. En la ecuación (2), C corresponde a la frase candidata y Cnt_{clip}(n) es el número máximo de veces que un n-grama de la frase candidata fue encontrado en alguna de las referencias. Una penalización con respecto a la longitud (*Brevity Penalty*, BP) se impone a las frases demasiado largas o demasiado cortas en la ecuación (3). Cuanto más corta sea la frase mayor será su penalización. En consecuencia, la frases comprimidas obtienen un score bajo. Por lo tanto, Bleu no es adecuado para evaluar resúmenes producidos por compresión de frases.

$$P_n = \frac{\sum_{C\epsilon\{\text{Cands}\}} \sum_{\text{n-gram}\epsilon C} Cnt_{clip}(\text{n})}{\sum_{C\epsilon\{\text{Cands}\}} \sum_{\text{n-gram}\epsilon C} Cnt(\text{n})} \quad (2)$$

$$\text{Bleu} = BP \times e^{\left(\sum_{n=1}^{N} \frac{1}{N} \mathbf{log}(P_n)\right)} \quad (3)$$

3.2.3 Evaluación sin referencias: FRESA

En (Molina et al., 2012), se intenta evaluar resúmenes con frases comprimidas utilzando la métrica FRESA (Torres-Moreno et al., 2010; Saggion et al., 2010; Torres-Moreno, 2014), la cual no requiere resúmenes de referencia dado que solamente utiliza el documento de origen. La idea es calcular las divergencias entre las distribuciones de frecuencias de términos entre el resumen que se quiere evaluar y el texto de origen. Estas divergencias corresponden a las de Kullback-Leibler (KL) y Jensen-Shannon (JS) como se describe en (Louis & Nenkova, 2008).

Sea T el conjunto de términos contenidos en el documento de origen. Para cada $t \in T$, C_t^T es el número de apariciones de t en el documento de origen y C_t^S es el número de apariciones de t

en el resumen que se quiere evaluar. En la ecuación (4), se calcula la diferencia absoluta entre las divergencias de dichas distribuciones (en el espacio *log*). Los valores altos (poca divergencia) están asociados a la similitud entre el resumen y el texto de origen mientras que los balores bajos (alta divergencia) implican disimilitud entre ellos.

$$\mathcal{D} = \sum_{t \in T} \left| \log\left(\frac{C_t^T}{|T|} + 1\right) - \log\left(\frac{C_t^S}{|S|} + 1\right) \right| \quad (4)$$

La interpretación de esta métrica no es trivial ni intuitiva. La única conclusión que se puede sacar a partir de los valores de la métrica es que el valor de divergencia entre un texto y su resumen es elevado, pero esto siempre es aplicable a las frases comprimidas como se muestra en los experimentos de (Molina et al., 2012). Así, FRESA asocia valores de alta divergencia independientemente de la estrategia utilizada, incluyendo la compresión aleatoria. Por lo tanto, tampoco resulta una manera adeacuada de evaluación de resúmenes por compresión de frases.

Después de presentar tres diferentes medidas automáticas en esta sección, queda claro que ninguna de ellas toma en cuenta la estructura gramatical de las frases comprimidas. En efecto, una de las principales desventajas de las evaluaciones automáticas es que no consideran ni la gramática ni la retórica ya que se basan solamente en las apariciones de elementos léxicos como los n-gramas.

En la siguiente sección, proponemos afrontar la problemática de la evaluación de otra manera, usando el test de Turing.

4 El juego de la imitación

Las ideas que tuvo Alan Turing, acerca de las máquinas y el pensamiento, siguen generando polémica. Sin embargo, vamos a explorar cómo pueden resultar ventajosas para la evaluación de algunas tareas del Procesamiento de Lenguaje Natural (PLN) y en concreto del resumen automático.

Nos referimos al famoso test de Turing descrito en el artículo (Turing, 1950) en el cual se discute la cuestión: "¿Y si las máquinas pudieran pensar?"

Para evitar la complicación de tener que definir qué significa pensar, Turing estableció el juego de la imitación, hoy conocido como el test de Turing. En el juego hay dos jugadores y un juez. El primer jugador es un ser humano (A) y el segundo es una máquina (B). Otra persona que funge como el juez (C) debe adivinar la identidad de cada uno de los jugadores sin verlos. Unicamente se permite que los jugadores interactúen con el juez mediante una terminal. Por ejemplo, el juez escribe preguntas con la ayuda de un teclado y lee las respuestas de los jugadores en una pantalla. Al final del juego, el juez debe indicar quién es la máquina y quién es el humano a partir de las respuestas obtenidas durante el intercambio.

Por supuesto, el objetivo de este experimento hipotético propuesto por Turing no era el de engañar a alguien en particular, sino el de plantear cuestiones filosóficas en torno al pensamiento. Concretamente, sobre la posibilidad de recrear artificialmente las funciones cognitivas del cerebro humano y sobre la posibilidad de evaluar si dichas funciones corresponden a lo que podríamos esperar de "algo"que piensa.

Para nuestro estudio, hemos rescatado algunos aspectos del protocolo del test que nos parecen aplicables a la evaluación de una tarea compleja de procesamiento del lenguaje y para la cual no se ha propuesto ningún método eficaz. Concordamos con Harnad (2000) sobre el hecho de que Turing privilegió, en el test, la comunicación por medio del lenguaje natural. ¿No es acaso la lengua uno de los principales medios para vehicular el pensamiento? No obstante, las cuestiones filosóficas del test de Turing no conciernen el presente estudio. Nuestro objetivo es simplemente validar, mediante una adaptación del test de Turing, un tipo específico de función lingüística: la generación de resúmenes.

En el test de Turing el juez no tiene el derecho de ver a los jugadores. Con esta restricción, Turing puso de manifiesto que son los aspectos funcionales y no los aspectos físicos los que deben ser juzgados. Nos parece entonces natural utilizar este test para evaluar tareas del PLN cuyo objetivo es simular la habilidad de los humanos. A este respecto, consideramos una variante del test de Turing destinada a la evaluación de resúmenes automáticos.

5 El test de Turing para evaluar resúmenes automáticos

Supongamos que un ser humano (A) y una máquina (B) producen respectivamente dos resúmenes a partir del mismo documento. (A) y (B) deben respetar las mismas reglas para que las producciones sean homogénas y, en consecuencia, comparables. Un juez humano (C), debe determinar cuál de los resúmenes fue elaborado por (A) y cual fue elaborado por (B). Para esto,

el juez debe revelar la identidad de cada jugador apoyándose unicamente en la lectura de sus resúmenes.

5.1 Protocolo experimental

Para nuestro experimento, hemos convocado 54 jueces quienes leyeron y evaluaron exactamente los mismos resúmenes. Como la evaluación requiere la lectura directa, se eligieron solamente seis resúmenes humanos (A) y seis resúmenes automáticos (B) producidos, en ambos casos, mediante el mismo algoritmo (Algoritmo 1), descrito en (Molina, 2013).

El Algoritmo 1 toma como argumentos un umbral de probabilidad ($\alpha \in [0, 1]$) y el documento a resumir (Doc). Primero, el documento original es segmentado en frases y luego en segmentos discursivos. Posteriormente, se decide para cada segmento si éste debe ser eliminado según un valor de probabilidad. El algoritmo termina cuando se han procesado todas las frases del documento y se ha producido un resumen. La computadora utilizó un modelo de regresión lineal que calcula la probabilidad de eliminar un segmento basándose en el aprendizaje de 60 844 segmentos anotados manualmente. Para el caso de los humanos, la desición está basada simplemente en su criterio para decidir si algún segmento es importante o no lo es.

Algoritmo 1 Generación de resúmenes por eliminación de segmentos.

Argumentos:(α, Doc)
$Segmentar_{\varphi}(Doc)$ //En frases.
$Segmentar_{s}(Doc)$ //En segmentos discursivos.
para todo φ en Doc **hacer**
  para todo s en φ **hacer**
    si ($\mathbf{P_{elim}}(s, \varphi) > \alpha$) **entonces**
      $Eliminar(s)$ de φ
    fin si
  fin para
fin para
  devolver resumen // Doc con φs modificadas.

Las frases fueron segmentadas mediante dos métodos distintos, un segmentador retórico para el español llamado DiSeg (da Cunha et al., 2012) y un segmentador adaptado a la compresión de frases (Molina, 2013). Para cada segmentador seleccionamos tres categorías de resumen de acuerdo con la tasa de compresión τ (Cuadro 2): poca compresión ($\tau < 50\,\%$), compresión media ($\tau \approx 50\,\%$) y mucha compresión ($\tau > 50\,\%$). Para el test, conservamos los que tenían mejores scores en gramática para cada una de las categorías.

Los 54 jueces (C), todos hispanohablantes con nivel de estudios de 4 o más años en la universidad, ignoraban toda la información respecto al juego de la imitación. Unicamente se les otorgaron los doce resúmenes y se les dio una sola instrucción: determinar para cada resumen si éste había sido producido por un humano o por una máquina. El Anexo 1, al final de este artículo, se muestra una copia del documento entregado a los jueces.

	Pals. origen	**Pals. resumen**	τ (%)	
$\tau < 50\,\%$	303	49	16.1	DiSeg
$\tau \approx 50\,\%$	209	104	49.6	DiSeg
$\tau > 50\,\%$	156	119	76.3	DiSeg
$\tau < 50\,\%$	217	57	26.2	CoSeg
$\tau \approx 50\,\%$	165	76	43.4	CoSeg
$\tau > 50\,\%$	234	186	79.4	CoSeg

Cuadro 2: Criterios de selección para la evaluación de resúmenes con un test de Turing.

6 La catadora de té

Para validar estadísticamente nuestros resultados, nos inspiramos en el experimento de la "dama del té", descrito en (Agresti, 2002), por medio del cual Ronald A. Fisher desarrolló un test estadístico exacto. Una dama (Muriel Bristol) se jactaba de ser capaz de distinguir si una taza de té con leche había sido servida primero con la leche o primero con el té. Para examinar su pretensión, Fisher le pidió probar 8 tazas de té con leche. En 4 de ellas se sirvió el té sobre la leche y en las otras 4 se sirvió la leche sobre el té. El test estadístico propuesto por Fisher se basa sobre el conteo del número de buenas y malas respuestas mediante una tabla de contingencia como la del Cuadro 3.

	Respuesta correcta	
Respuesta dama	*primero leche*	*primero té*
primero leche	a =3	b =1
primero té	c =1	d =3

Cuadro 3: Tabla de contingencia de las respuestas de la catadora de té.

Fisher mostró que la probabilidad de obtener una tabla de contingencia como la del del Cuadro 3 esta dada por la ley hipergeométrica (ecuación 5). Donde $\binom{l}{k}$ es el coeficiente binomial y n es la suma de todas las celdas de la tabla. Usando las respuestas del Cuadro 3, se tiene que $p = 0{,}229$ es la probabilidad de obtener los resultados de esa tabla por mera coincidencia. Se sigue que las

respuestas de la catadora no establecen prueba de su supuesta habilidad.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \quad (5)$$

7 Validación de resultados del juego de la imitación con el test exacto de Fisher

En nuestra evaluación, le pedimos a los 54 jueces que distinguieran para cada uno de los 12 resúmenes si había sido creado por un humano o por una máquina y se crearon tablas de contingencia, como la mostrada en el Cuadro 4. Finalmente se aplicó el test de Fisher a las respuestas obtenidas.

	Origen del resumen	
Respuesta juez	*Humano*	*Máquina*
Humano	a	b
Máquina	c	d

Cuadro 4: Tabla de contingencia para la evaluación de resúmenes con el test exacto de Fisher.

La hipótesis nula de nuestros tests (H_0) es que no existe asociación entre las respuestas y el origen del resumen. La hipótesis alternativa (H_1) es que sí existe una asociación positiva. Utilizamos la función *fisher.test* del lenguaje de programación R para calcular los valores de p. En nuestros experimentos utilizamos la configuración estándar del test: dos cola con un intervalo de confianza del 95 %.

El Cuadro 9 muestra las tablas de contingencia así como los valores de p de la evaluación del origen de los resúmenes para los 54 jueces del experimento.

Una sola persona (el Juez 1 en el Cuadro 9) presenta un resultado estadísticamente significativo, de haber distinguido entre los dos tipos de resúmenes. Respecto a los 53 jueces restantes, no podemos decir que el resultado sea significativo en la distinción del origen verdadero de los resúmenes. Nos inclinamos a afirmar entonces que los 53 jueces restantes encontraron la misma calidad en los resúmenes manuales que en los automáticos.

8 Evaluación de resúmenes según el tipo de segmentación y el tamaño

También utilizamos el test exacto de Fisher para verificar si se observan diferentes resultados según el segmentador automático empleado. El Cuadro 5 muestra el número de veces que los jueces identificaron correcta o incorrectamente los resúmenes según el segmentador. Para afirmar con significación estadística que una segmentación en particular permite identificar más facilmente el origen de los resúmenes, la hipótesis nula es, en este caso, que el grado de identificación es independiente del tipo de segmentación. Los resultados dan un valor $p = 0{,}4965$ al 95 % con un intervalo de confianza de $[0{,}63; 2{,}76]$. Se sigue que como $p > 0{,}05$, entonces aceptamos H_0: el hecho de que un resumen haya sido segmentado con DiSeg o CoSeg no influye en la identificación realizada por los jueces.

	Origen correctamente identificado	**Origen erroneamente identificado**
DiSeg	45	63
CoSeg	19	35

Cuadro 5: Evaluación de la influencia del tipo de segmentación para la identificación de los resúmenes.

	Correctamente identificado (*observado*)	**Correctamente identificado** (*esperado*)
$\tau < 50\,\%$	27	25
$\tau \approx 50\,\%$	30	25
$\tau > 50\,\%$	18	25

Cuadro 6: Valores esperados y observados en la identificación correcta del origen de los resúmenes.

	Erróneamente identificado (*observado*)	**Erróneamente identificado** (*esperado*)
$\tau < 50\,\%$	27	29
$\tau \approx 50\,\%$	24	29
$\tau > 50\,\%$	36	29

Cuadro 7: Valores esperados y observados en la identificación errónea del origen de los resúmenes.

	Correctamente identificado	**Erróneamente identificado**
$\tau < 50\,\%$	0.668	-0.668
$\tau \approx 50\,\%$	1.671	-1.671
$\tau > 50\,\%$	-2.339	2.339

Cuadro 8: Desviación estándar de la varianza residual con respecto a la tasa de compresión τ en la identificación de resúmenes automáticos.

Juez id	Contingencia	p	◦	Juez id	Contingencia	p	◦	Juez id	Contingencia	p	◦
Juez 1	4	0.030	falso	Juez 2	3	0.500	verdadero	Juez 3	5	0.772	verdadero
	0				2				5		
	2				3				1		
	6				4				1		
Juez 4	1	0.998	verdadero	Juez 5	3	0.716	verdadero	Juez 6	4	0.727	verdadero
	5				3				4		
	5				3				2		
	1				3				2		
Juez 7	5	0.772	verdadero	Juez 8	4	0.283	verdadero	Juez 9	5	0.272	verdadero
	5				2				3		
	1				4				1		
	1				2				3		
Juez 10	3	0.969	verdadero	Juez 11	3	0.878	verdadero	Juez 12	4	0.283	verdadero
	1				4				2		
	3				3				2		
	5				2				4		
Juez 13	3	0.716	verdadero	Juez 14	3	0.716	verdadero	Juez 15	2	0.878	verdadero
	3				3				3		
	3				3				4		
	3				3				3		
Juez 16	3	0.969	verdadero	Juez 17	1	0.969	verdadero	Juez 18	3	0.500	verdadero
	5				3				2		
	3				5				3		
	1				3				4		
Juez 19	3	0.500	verdadero	Juez 20	2	0.960	verdadero	Juez 21	3	0.716	verdadero
	2				4				3		
	3				2				3		
	4				4				3		
Juez 22	3	0.878	verdadero	Juez 23	2	1.000	verdadero	Juez 24	1	0.992	verdadero
	4				6				4		
	3				4				5		
	2				0				2		
Juez 25	2	0.727	verdadero	Juez 26	4	0.283	verdadero	Juez 27	2	0.960	verdadero
	2				2				4		
	4				4				2		
	4				4				4		
Juez 28	4	0.727	verdadero	Juez 29	1	0.969	verdadero	Juez 30	3	0.878	verdadero
	4				3				4		
	2				5				3		
	2				3				2		
Juez 31	5	0.772	verdadero	Juez 32	2	0.960	verdadero	Juez 33	3	0.878	verdadero
	5				4				4		
	1				2				3		
	1				4				2		
Juez 34	2	0.960	verdadero	Juez 35	3	0.878	verdadero	Juez 36	4	0.960	verdadero
	4				4				2		
	4				2				4		
	2				3				2		
Juez 37	4	0.909	verdadero	Juez 38	3	0.878	verdadero	Juez 39	4	0.727	verdadero
	5				4				4		
	1				2				2		
	2				3				2		
Juez 40	4	0.283	verdadero	Juez 41	4	0.727	verdadero	Juez 42	1	0.998	verdadero
	2				4				5		
	4				2				5		
	2				2				1		
Juez 43	3	0.878	verdadero	Juez 44	2	1.000	verdadero	Juez 45	5	1.000	verdadero
	3				6				6		
	3				4				1		
	3				0				0		
Juez 46	4	0.500	verdadero	Juez 47	2	0.878	verdadero	Juez 48	4	0.283	verdadero
	3				3				2		
	3				4				2		
	2				3				4		
Juez 49	3	0.716	verdadero	Juez 50	2	0.878	verdadero	Juez 51	4	0.500	verdadero
	3				3				3		
	3				4				2		
	3				3				3		
Juez 52	2	0.992	verdadero	Juez 53	2	0.992	verdadero	Juez 54	3	0.716	verdadero
	5				5				3		
	2				2				3		
	1				1				3		
	4				4				3		

Cuadro 9: Resultados del test de Turing en evaluación de resumen automático aplicado a 54 jueces.

Para verificar la influencia de la tasa de compresión de un resumen en la elección de los jueces, utilizamos el test de χ^2. En este caso, no podemos utilizar el test exacto de Fisher porque la tabla de contingencia asociada es de 3×2 (Cuadros 6 y 7). Los resultados del test de χ^2 dan un valor $p = 0{,}0547$, apenas superior al valor crítico lo que nos llevó a comparar los valores esperados bajo la hipótesis nula en ambos cuadros, confirmandose que, para los resúmenes con $\tau > 50\,\%$, resultó más difícil identificar el origen artificial.

Este hecho puede confirmarse en el Cuadro 8 a partir de las varianzas residuales donde la desviación estándar para los resúmenes fijando $\tau > 50\,\%$ es más de dos veces superior a la media. Para los jueces resultó muchos más complicado identificar correctamente un resumen automático cuando había sido menos comprimido.

9 Conclusiones

En este trabajo hemos abordado la evaluación de resúmenes de documentos textuales producidos con métodos automáticos. La motivacion principal de este trabajo es que, a pesar que existen métodos efectivos de evaluación para resúmenes por extracción, estos resultan inadecuados para evaluar resúmenes por compresión de frases, porque que no toman en cuenta la gramática.

Ante este panorama, hemos propuesto un método basado en el test de Turing en el que los jueces humanos deben develar el origen (automático o manual) de varios resúmenes, y por medio del test exacto de Fisher, se calcula la fiabilidad de las respuestas de dichos jueces.

Aunque hemos aplicado la evaluación al área de resumen automático, encontramos que la metodología resulta lo bastante general para ser aplicada a cualquier otra área del procesamiento del lenguaje natural.

Agradecimentos

A todos los voluntarios que realizaron el test. A Mariana Tello Signoret por su ayuda con las traducciones. A Eric SanJuan, Gerardo Sierra y Carlos Mendez por la asesoría para la realización de este trabajo.

Referencias

Agresti, Alan. 2002. *Categorical data analysis*, vol. 359. Wiley interscience.

Amigó, Enrique, Julio Gonzalo, Anselmo Peñas & Felisa Verdejo. 2005. Qarla: a framework for the evaluation of text summarization systems. En *43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 280–289. Ann Arbor, MI, Etats-Unis: ACL.

da Cunha, Iria, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes & Irene Castellón. 2012. Diseg 1.0: The first system for spanish discourse segmentation. *Expert Systems with Applications* 39(2). 1671–1678.

Edmundson, H. P. 1969. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery* 16(2). 264–285.

Grefenstette, G. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. En *AAAI Spring Symposium on Intelligent Text summarization (Working notes)*, 111–118. Stanford University, CA, Etats-Unis.

Harnad, Stevan. 2000. Minds, machines and turing. *Journal of Logic, Language and Information* 9(4). 425–445.

Knight, Kevin & Daniel Marcu. 2000. Statistics-based summarization – step one: Sentence compression. En *17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, 703–710. Austin, TX, Etats-Unis.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. En Marie-Francine Moens & Stan Szpakowicz (eds.), *Workshop Text Summarization Branches Out (ACL'04)*, 74–81. Barcelone, Espagne: ACL.

Louis, Annie & Ani Nenkova. 2008. Automatic Summary Evaluation without Human Models. En *First Text Analysis Conference (TAC'08)*, Gaithersburg, MD, Etats-Unis.

Mani, Inderjeet, David House, Gary Klein, Lynette Hirschman, Therese Firmin & Beth Sundheim. 1999. The tipster summac text summarization evaluation. En *ninth conference on European chapter of the Association for Computational Linguistics*, 77–85. ACL.

Molina, Alejandro. 2013. Compresión automática de frases: un estudio hacia la generación de resúmenes en espanol. *Inteligencia Artificial* 16(51). 41–62.

Molina, Alejandro, Iria da Cunha, Juan-Manuel Torres-Moreno & Patricia Velazquez-Morales. 2010. La compresión de frases: un recurso para la optimización de resumen automático de documentos. *Linguamática* 2(3). 13–27.

Molina, Alejandro, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan & Gerardo Sierra. 2012. Sentence compression in spanish driven by discourse segmentation and language models. *Cornell University, Computation and Language (cs.CL), Information Retrieval (cs.IR)* arXiv:1212.3493.

Orasan, Constantin & Laura Hasler. 2007. Computer-aided summarisation: how much does it really help. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)* 437–444.

Papineni, K., S. Roukos, T. Ward, & W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. En *40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 311–318. Philadelphie, PA, Etats-Unis: ACL.

Saggion, H. & G. Lapalme. 2000. Concept identification and presentation in the context of technical text summarization. En *ANLP/NAACL Workshop on Automatic Summarization*, 1–10. Seattle, WA, Etats-Unis: ACL.

Saggion, Horacio, Juan-Manuel Torres-Moreno, Iria da Cunha & Eric SanJuan. 2010. Multilingual summarization evaluation without human models. En *23rd International Conference on Computational Linguistics: Posters (COLING'10)*, 1059–1067. Beijing, Chine: ACL.

Torres-Moreno, Juan-Manuel. 2014. *Automatic text summarization*. Wiley and Sons.

Torres-Moreno, Juan-Manuel, Horacio Saggion, Iria da Cunha & Eric SanJuan. 2010. Summary Evaluation With and Without References. *Polibits: Research journal on Computer science and computer engineering with applications* 42. 13–19.

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 59(236). 433–460.

A Anexo 1: Test de evaluación

Algunos de los siguientes doce resúmenes que se muestran a continuación han sido creados de manera automática por un programa y otros han sido creados por humanos. Determine cuáles.

La persona con el cociente intelectual más alto del mundo
Su nombre es Marilyn vos Savant y nació en San Louis (Missouri) el 11 de agosto de 1946. Marilyn vos Savant está considerada como la persona con el cociente intelectual más alto del mundo. Hoy en día es una más que reputada columnista, escritora, conferenciante y dramaturga. En 1986 comenzó una columna dominical llamada Pregunta a Marilyn (Ask Marilyn) en la revista Parade, donde responde preguntas de los lectores acerca de diversos temas. Su mayor aspiración era el convertirse en escritora. Durante su juventud trabajó en la tienda de ultramarinos de su padre. Cursó varios seminarios de filosofía en la universidad. En la actualidad está casada con el prestigioso cardiólogo Robert Jarvik. A Marilyn se le asocia con el famoso problema de Monty Hall, o bien le fue planteado a ella a través de una consulta en su columna Ask Marilyn.

Cuadro 10: Resumen de *La persona con el cociente intelectual más alto del mundo* (tipo de segmentación: DiSeg, origen del resumen : Humano, τ =51.83 % del contenido original).

El Pulque
El Pulque o Neutle se obtiene de la fermentación de la savia azucarada o aguamiel, concentrados en el corazón de el maguey, antes de que salga el pedúnculo de la inflorescencia del maguey por el proceso conocido como raspado, que consiste en quitar el centro de la planta donde crecen las hojas tiernas dejando una oquedad que se tapa con una penca del maguey. El interior es entonces raspado con una especie de cuchara, lo que provoca que el maguey suelte un jugo el cual se concentra en el hueco. Este es, luego, a intervalos de uno o dos días absorbido hacia un cuenco hueco (llamado acococte, fruto de una cucurbitácea) y depositado en un recipiente llamado odre. Este proceso lo lleva a cabo el Tlachiquero o raspador, y el jugo se recolecta durante dos meses como máximo. Después es depositado en barriles de pino o, en cubas de acero inoxidable, donde se fermenta con la bacteria Zymomonas mobilis durante uno o dos días obteniendose un líquido blanco de aspecto lechoso con un 5 % de alcohol. Se debe beber inmediatamente ya que al seguirse fermentando adquiere un gusto muy fuerte.

Cuadro 11: Resumen de *El Pulque* (tipo de segmentación: CoSeg, origen del resumen: Humano, τ =90.90 % del contenido original).

La música en el antiguo Egipto
La Música en el antiguo Egipto se empleaba en varias actividades, pero su desarrollo principal fue en los templos, donde era usada durante los ritos dedicados a los diferentes dioses y era utilizada como remedio terapéutico. Como en otros pueblos, también se consideraba un medio de comunicación con los difuntos y los músicos alcanzaban una categoría tal que algunos están enterrados en las necrópolis reales. No se conoce cómo era realmente ya que no desarrollaron un sistema para representarla, se transmitía de maestro a alumno. También arrojan luz sobre este tema los instrumentos conservados en los museos y la representación en bajorrelieves y pinturas de instrumentos y bailarines, además de lo conservado por tradición oral por los cantores coptos.

Cuadro 12: Resumen de *La música en el antiguo Egipto* (tipo de segmentación: DiSeg, origen del resumen: Máquina, τ =76.28 % del contenido original).

Confirman en Veracruz caso de influenza en niño de 5 años
El gobierno de Veracruz confirmó este domingo un caso de influenza porcina de la cepa H1N1 en un niño de cinco años originario de el poblado La Gloria. El subdirector de prevención y control de enfermedades de la Secretaría de Salud estatal dijo que el menor de nombre Edgar Hernández Hernández superó el cuadro de infección pulmonar.

Cuadro 13: Resumen de *Confirman en Veracruz caso de influenza en niño de 5 años* (tipo de segmentación: CoSeg, origen del resumen: Máquina, τ =26.26 % del contenido original).

Efectos de la LSD
Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, el entorno en que se use la droga, la pureza de ésta, la personalidad, el estado de ánimo y las expectativas del usuario. Algunos consumidores de LSD experimentan una sensación de euforia, mientras que otros viven la experiencia en clave terrorífica. Cuando la experiencia tienen un tono general desagradable, suele hablarse de mal viaje. Cuando la sustancia se administra por vía oral, los efectos tardan en manifestar se entre 30 minutos y una hora y, según la dosis, pueden durar entre 8 y 10 horas. Entre los efectos fisiológicos recurrentes están los siguientes: contracciones uterinas, fiebre, erizamiento del vello, aumento de la frecuencia cardíaca, transpiración, pupilas dilatadas, insomnio, hiperreflexia y temblores.

Cuadro 14: Resumen de *Efectos de la LSD* (tipo de segmentación: CoSeg, origen del resumen: Humano, τ =90.30 % del contenido original).

Introducción a las matemáticas
Cada vez que vas a la tienda, juegas en la computadora o en la consola de video juegos; cuando sigues las incidencias de un juego de béisbol o fútbol americano, cuando llevas el ritmo de una canción, estás utilizando relaciones numéricas y en tu mente realizas una serie de operaciones que tienen que ver con el lenguaje matemático. En este sentido, podemos afirmar que el pensamiento matemático está presente en la mayoría de nuestras actividades, desde las más sencillas hasta las más especializadas. Sin embargo, no siempre estamos conscientes de los conceptos, reglas, modelos, procedimientos y operaciones matemáticas que realizamos mentalmente a diario. A lo largo de esta unidad, mediante la adquisición de distintos conocimientos y la resolución de una serie de problemas y ejercicios, descubriremos cómo representar y formalizar algunas de las operaciones que mencionamos. Los cursos de matemáticas que llevaste con anterioridad, te han familiarizado con la utilización de ciertas operaciones básicas. Con ello podríamos decir que posees los conocimientos básicos para manejar algoritmos elementales. Así que reconocerás diferentes tipos de números como los naturales, los enteros, los fraccionarios (racionales) y los irracionales que son temas de esta unidad. Gracias al conocimiento de los distintos tipos de números construirás y aplicarás modelos matemáticos, los cuales trabajarás con razones y proporciones, así como con series y sucesiones, que te ayudarán a resolver diferentes situaciones de la vida cotidiana. Todos estos aprendizajes te servirán en las siguientes unidades para identificar, resolver, plantear, interpretar y aplicar diferentes procedimientos (algoritmos) con un distinto nivel de complejidad, en variedad de situaciones.

Cuadro 15: Resumen de *Introducción a las matemáticas* (tipo de segmentación: DiSeg, origen del resumen: Humano, τ =84.31 % del contenido original).

Ética de robots
Existe la preocupación de que los robots puedan desplazar o competir con los humanos. Las leyes o reglas que pudieran o debieran ser aplicadas a los robots u otros entes autónomos en cooperación o competencia con humanos han estimulado las investigaciones macroeconómicas de este tipo de competencia, notablemente por Alessandro Acquisti basandose en un trabajo anterior de John von Neumann. Actualmente, no es posible aplicar las Tres leyes de la robótica, dado que los robots no tienen capacidad para comprender su significado. Entender y aplicar las Tres leyes de la robótica, requeriría verdadera inteligencia y consciencia del medio circundante, así como de sí mismo, por parte del robot.

Cuadro 16: Resumen de *Ética de robots* (tipo de segmentación: CoSeg, origen del resumen: Humano, τ =63.52 % del contenido original).

Por qué el embarazo de las elefantas es tan largo
El período de gestación, que se prolonga por casi dos años, es una de esas rarezas de la biología que le permite al feto desarrollar suficientemente su cerebro. Los resultados de este estudio servirán para mejorar los programas de reproducción de elefantes en los zoológicos y podrían también contribuir al desarrollo de un anticonceptivo. Los elefantes son mamíferos muy sociales con un alto grado de inteligencia, similar a la de los homínidos y los delfines. Son, además, los que tienen el período de gestación más largo, que puede extenderse hasta por 680 días. Los elefantes nacen con un nivel avanzado de desarrollo cerebral, que utilizan para alimentarse mediante sus habilidosas trompas. Hasta ahora, los científicos no habían logrado entender en profundidad los procesos biológicos del maratónico embarazo de las elefantas. Pero gracias a los avances de las técnicas de ultrasonido, los veterinarios pudieron utilizar nuevas herramientas.

Cuadro 17: Resumen de *Por qué el embarazo de las elefantas es tan largo* (tipo de segmentación: DiSeg, origen del resumen: Humano, τ =69.85 % del contenido original).

Hallan genes asociados a migraña
Investigadores europeos y australianos indicaron el domingo que habían localizado cuatro nuevos genes asociados con la forma más común de la migraña. Las variantes genéticas fueron detectadas en el genoma de 4800 pacientes de migraña sin aura, la forma que asumen tres de cada cuatro ataques de migraña. Estas estas variantes genéticas no fueron halladas, sin embargo, en el grupo testigo de 7000 personas libres de la enfermedad, dijeron los investigadores. El estudio también confirmó la existencia de otros dos genes de predisposición, en un trío de genes ya identificados en un trabajo anterior. La migraña afecta a aproximadamente una de cada seis mujeres y a uno de cada ocho hombres. Los nuevos genes identificados en este estudio refuerzan el argumento según el cual la disfunción de las moléculas responsables de la transmisión de señales entre las células nerviosas, contribuye a la aparición de la migraña. Además, dos de estos genes refuerzan la hipótesis de un posible papel de las venas. La investigación, publicada en la revista especializada Nature Genetics, fue realizada por un consorcio internacional dedicado a la investigación sobre la genética de la migraña.

Cuadro 18: Resumen de *Hallan genes asociados a migraña* (tipo de segmentación: CoSeg, origen del resumen: Máquina, τ =79.48 % del contenido original).

Problemas globales
Hoy se reconoce que existen problemas que denominamos globales. Estos problemas se presentan fundamentalmente por la carga de contaminantes liberados hacia la atmósfera terrestre. Por su magnitud y complejidad constituyen un grave problema que requiere medidas muy drásticas para su solución. La composición química de la atmósfera es muy inestable: cambia a través del tiempo y en función de diversas reacciones e interacciones de sus componentes. Hoy sabemos que además de los numerosos gases que la componen, existe una compleja interrelación de los gases. Esta interacción se manifiesta en el hecho de que la radiación solar aporta la energía necesaria para que se realicen las reacciones químicas que modifican la composición de la atmósfera. El diálogo entre la atmósfera y la radiación solar ha sido alterado por el hombre.

Cuadro 19: Resumen de *Problemas globales* (tipo de segmentación: DiSeg, origen del resumen: Máquina, τ =59.44 % del contenido original).

Descubrimiento de mamut emociona a científicos
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. Al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses.

Cuadro 20: Resumen de *Descubrimiento de mamut emociona a científicos* (tipo de segmentación: CoSeg, origen del resumen: Máquina, τ =43.42 % del contenido original).

La Tundra
El ambiente de la tundra está caracterizado por una sequía prolongada. Las especies más típicas de la flora son los arbustos enanos, líquenes y musgos. Algunas especies, particularmente de aves, sólo pasan el verano en la tundra, época en la que anidan. Existen pocas especies de anfibios y reptiles.

Cuadro 21: Resumen de *La Tundra* (tipo de segmentación: DiSeg, origen del resumen: Máquina, τ =16.17 % del contenido original).

Projetos, Apresentam-Se!

ASinEs: Prolegómenos de un atlas de la variación sintáctica del español

ASinEs: Prolegomena to an atlas of syntactic variation in Spanish

Alba Cerrudo
Universitat Autònoma de Barcelona
alba.cerrudo@gmail.com

Ángel J. Gallego
Universitat Autònoma de Barcelona
angel.gallego@uab.cat

Anna Pineda
Universitat Autònoma de Barcelona
anna.pineda@uab.cat

Francesc Roca
Universitat de Girona
francesc.roca@udg.edu

Resumen

En este artículo se presenta el ASinEs[1], una aplicación con formato de atlas dedicada al estudio sincrónico de la variación sintáctica de los geolectos del español. Este proyecto es innovador, ya que no existe ningún atlas dedicado exclusivamente a investigar la variación geolectal de la sintaxis de esta lengua. La versatilidad del ASinEs permite también el estudio de geolectos de otros estadios del español, así como los de otras lenguas con las que está actualmente en contacto. Todo ello proporciona una potente herramienta para la investigación en el campo de la variación de las lenguas románicas y no románicas (vasco, inglés, lenguas amerindias, etc.).

El desarrollo de este proyecto cuenta con la colaboración del *Centre de Lingüística Teòrica* (Universitat Autònoma de Barcelona), el *Centro IKER* con sede en Bayona (Francia) y la *Real Academia Española*.

Palabras clave

atlas, español, geolectos, sintaxis, variación

Abstract

This paper introduces the ASinEs[1], an atlas-based application devoted to the study of the syntactic variation of Spanish geolects. This project is groundbreaking, as there is no other atlas exclusively devoted to study the geolectal variation of geolectal variants of Spanish. Although ASinEs was originally conceived to explore the current geolects of Spanish, its flexibility allows it to study both the geolects of previous stages and the geolects of other close-by languages. This provides us with a po-werful tool to study variation of both Romance and non-Romance languages (Basque, English, Amerindi-an languages, etc.).

This project is being developed in collaboration with the *Centre de Lingüística Teòrica* (Universitat Autònoma de Barcelona), the *IKER Center* at Bayonne (France), and the *Real Academia Española*.

[1] http://www.asines.org

Keywords

atlas, Spanish, geolects, syntax, variation

1 Introducción

Tradicionalmente, los estudios de dialectología (o geolingüística) y variación lingüística se han centrado en fenómenos que pertenecen a los niveles léxico, fonético o morfológico (cf. Chambers & Trudgill (1980); Chambers & Schilling-Estes (2013); Labov (1994, 2001); Labov et al. (2006); Petyt (1980), entre otros), utilizando a tal efecto técnicas cuantitativas (estadísticas), de reconstrucción (diacrónica), y comparativas (cf. Campbell (2001) y referencias allí citadas).

Muchos de esos trabajos tienen en cuenta factores sociales y geográficos para explicar el cambio / variación, y han dado lugar a avances importantes en nuestra comprensión de fenómenos socio-lingüísticos como la diglosia, los *continuum* geolectales o las áreas de transición (el llamado "*Sprachbund*").

Otro de los resultados de esta línea de acción fue una caracterización adecuada de unidades como el "fonema", el "morfema" o el "rasgo distintivo" (introducidos por el estructuralismo europeo), lo cual permitió y facilitó la investigación basada en el trabajo de campo y condujo a estudios tipológicos como los de Greenberg (1963).

En este apartado, revisamos brevemente los antecedentes en los estudios de geolingüística del español, así como los avances que ha habido en las teorías sintácticas y de variación en las últimas décadas, los cuales han permitido plantear un cambio de tendencia en la investigación geolectal.

This work is licensed under a Creative Commons Attribution 3.0 License

1.1 Antecedentes en el estudio del español

Los estudios sobre la variación geolectal, en el caso de lenguas como el español, se han centrado en los mismos dominios mencionados al inicio: el léxico, la fonética y la morfología (cf. Alvar (1996b,a); Fernández-Ordóñez (2011); García Mouton (1994); Kany (1945), entre otros). El énfasis puesto en dichos fenómenos puede verse en los diferentes atlas lingüísticos (p.ej., ALPI, ALEANR, ALBI, ALEA, ALECAN, ALCyL), que suelen representar los siguientes tipos de variación:

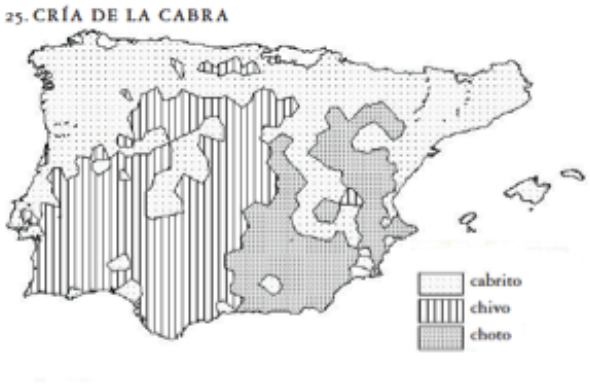
- **Léxica:** reflejado en Figura 1a, donde aparecen recogidas las respuestas extraídas del ALPI a la pregunta sobre el nombre de "la cría de la cabra".
- **Morfológica:** podemos ver muestras en Figura 1b, donde se observan las variantes del pronombre de segunda persona del singular.
- **Fonética:** tenemos ilustraciones en Figura 1c, donde se ofrecen las soluciones fonéticas del sustantivo *tejón*, que proviene del latín TAXU.

La misma situación puede observarse en los estudios diacrónicos, donde son mayoritarios los trabajos sobre el léxico y la morfofonología (con alguna excepción; cf. Company Company (2006, 2009, 2014); Lapesa (2000). La siguiente cita, extraída de Sánchez Lobato (1994), recoge la idea —extendida— de que la sintaxis apenas presenta variación geolectal:

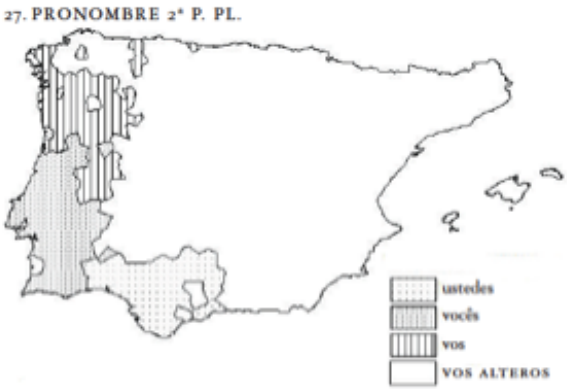
> Los rasgos lingüísticos, pues, más característicos del español americano —frente a la subnorma castellana— se encuentran en esa nueva coiné surgida de Andalucía. Lo esencialmente autóctono del español de América se encuentra en su aliento, en su voz, es decir, en la entonación, en el ritmo y en el léxico, no en la morfología. *En la sintaxis no hay diferencias notables.*
>
> [Sánchez Lobato (1994, pg. 560), énfasis nuestro]

La opinión de Sánchez Lobato no constituye, en absoluto, un caso aislado. Otros autores enfatizan el predominio del interés por el léxico y la morfo-fonología en los estudios geolectales del español:

> La falta de atención a los problemas de sintaxis dialectal por parte de la dialectología tradicional, el estructuralismo y la gramática generativa está dejando un vacío en nuestro conocimiento de la realidad viva de las lenguas que tardará muchos años en llenarse. Son varias las razones de esta situación contra la que se viene clamando desde hace algunos años, sin que los intentos de ponerle remedio pasen de ser esfuerzos aislados. La primera causa me parece enteramente imputable a la dialectología que, desde sus orígenes, ha tendido a obviar este tipo de asuntos, posiblemente



(a) Variación Léxica



(b) Variación Morfológica

(c) Variación Fonética

Figura 1: Ejemplos de fenómenos estudiados en atlas del español (casos tomados del ALPI y discutidos en Fernández-Ordóñez (2011, 2014))

> porque los métodos que fue perfilando para la recolección de materiales fueran poco aptos para obtener información sobre la sintaxis de las hablas estudiadas. Por eso es totalmente acertada la observación de Gregorio Salvador de que «*la sintaxis nunca ha sido hasta ahora ocupación seria de dialectólogos, sino de filólogos*».
>
> [Morillo-Velarde Pérez (1992, pgs. 219–220), énfasis nuestro]

Pueden leerse palabras similares en relación al dominio diacrónico:

> Es casi un lugar común la afirmación de que los estudios de sintaxis histórica del español presentan un desarrollo muy limitado. Y aunque se ha publicado una bibliografía de sintaxis histórica con 548 títulos (Narbona 1984–1985), lo cierto es que, si se comparan los estudios de sintaxis con los de fonética o morfología históricas, son obvios tanto el retraso metodológico como la escasez de los trabajos sintácticos.
>
> [Ridruejo (1992, pg. 587), énfasis nuestro]

> Me atrevo a decir que la sintaxis ha pasado de ser el patito feo de la lingüística histórica, romance y general, con una escasez notoria de estudios y estudiosos hace cincuenta años —si la comparamos con la fonología, y en buena parte, la morfología históricas— a ser el cisne de las subdisciplinas diacrónicas hoy en día [...] Por décadas se dio una escisión tajante entre lingüistas sincrónicos y lingüistas diacrónicos, que conllevó el retraimiento de los estudios diacrónicos, muy especialmente los de sintaxis histórica. Subyacen a esta escisión varias razones teóricas.
>
> [Company Company (2005, pgs. 144–146), énfasis nuestro]

En las últimas décadas se han puesto en marcha numerosas líneas de investigación que pretenden hacer una transición entre los estudios de corte tradicional de la gramática del español (notablemente, las gramáticas de Andrés Bello, Salvador Fernández Ramírez, o la misma Real Academia Española) y los estudios más actuales. Por lo general, esos esfuerzos han adoptado la forma de monografías, manuales, tesis doctorales inéditas, artículos publicados e incontables actas de congresos.

Puede afirmarse, no obstante, que hay un punto de inflexión en dicha tendencia con la aparición de la *Gramática Descriptiva de la Lengua Española* (Bosque & Demonte, 1999) y, especialmente, con la *Nueva Gramática de la Lengua Española* (RAE-ASALE, 2009), obras en las que encontramos capítulos dedicados a fenómenos sintácticos en los que se recogen diferentes casos de variación. Es interesante observar, aun así, que la expresión "variación sintáctica" solo se encuentra en tres epígrafes de la NGLE (§§ 34.11.d, 34.11.e y 41.12o), lo cual da a entender la escasa popularidad que ese tipo de variación ha experimentado. Esfuerzos similares a los de RAE-ASALE (2009) se han llevado a cabo para otras lenguas, como el catalán (con la *Gramàtica del Català Contemporani*, 2002), el italiano (con la *Grande Grammatica Italiana di Consultazione*, 1991), el inglés (con la *Cambridge Grammar of the English Language*, 2002), y el portugués (con la *Gramática do Português Contemporâneo*, 2014).

1.2 Otras herramientas actuales

La ausencia de estudios de variación sintáctica en el caso del español contrasta con la situación de otras lenguas cercanas, que sí poseen atlas (o bases de datos) sintácticos. Hay diferentes ejemplos de ello:

1. El *Dynamic Syntactic Atlas of the Dutch dialects* [DynaSAND] http://www.meertens.knaw.nl/sand/zoeken/index.php
2. El *Syntax-oriented Corpus of Portuguese Dialects* [CORDIAL-SIN] http://www.clul.ul.pt/en/resources/212-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects
3. El *Atlas Linguístico-Etnográfico de Portugal e da Galiza* [ALEPG] http://www.clul.ul.pt/en/resources/205-linguistic-and-ethnographic-atlas-of-portugal-and-galicia-alepg?showall=1
4. El *Atlante Sintattico d'Italia* [ASIt] http://asit.maldura.unipd.it/
5. La *Base de Datos de la Sintaxis Vasca* [BASYQUE] http://ixa2.si.ehu.es/atlas2/index.php?lang=es

Junto a estos atlas y bases de datos, otras herramientas en línea han sido desarrolladas en los

últimos años. Lo más relevante es que todas ellas incorporan fenomenología sintáctica:

1. El *World Atlas of Linguistic Structures* [WALS] http://wals.info/
2. El proyecto *TERRALING* http://www.terraling.com/
3. El *Syntactic Structures of the World's Languages* [SSWL] http://sswl.railsplayground.net/
4. El proyecto *Symila* http://blogs.univ-tlse2.fr/symila/en/

Como hemos dicho, no existe, al menos de forma exclusiva, ninguna herramienta análoga en el caso del español. Visto con cierta perspectiva, dicha ausencia probablemente se debe a la conjunción de diferentes factores. Uno de ellos es la falta de herramientas formales suficientes para reflejar adecuadamente la variación sintáctica que presentan no solo lenguas tipológicamente diferenciadas (p.ej., español / cingalés), sino también lenguas cercanas (p.ej., español / catalán) o geolectos de estas (p.ej., español de Santander / español de Buenos Aires). Como veremos en la siguiente sección, tales herramientas existen dentro de marcos teóricos formales como el de la Gramática Generativa.

1.3 La variación sintáctica: estado de la cuestión

En el contexto que estamos comentando, es importante destacar la contribución a la variación morfo-sintáctica de los enfoques formales, especialmente desde la llamada Teoría de Principios y Parámetros (TPP; cf. Chomsky (1981, 1986)), en la que el lenguaje es concebido como una facultad biológica que consta de un estado inicial (E_I) y un estado final (E_F), que se corresponden con lo que se ha dado en llamar Gramática Universal (GU) y Gramática Particular (GP) respectivamente. La gran aportación del modelo TPP fue, precisamente, intentar reconciliar la supuesta inmutabilidad de la GU (los "principios") con la obvia variabilidad de las diferentes GPs (los "parámetros"). En particular, la TPP permitió aliviar la tensión existente entre una descripción robusta de la variación lingüística (que pretendía caracterizar la variedad de cada lengua) y la explicación de la adquisición lingüística (un proceso aparentemente simple y sin instrucción explícita, incompatible con la proliferación de reglas que debían ser memorizadas por el niño que adquiere su lengua) mediante la sustitución de los sistemas de reglas por principios universales que se fijaban mediante la experiencia lingüística. Desde tal perspectiva, un niño lo único que tendría que hacer es fijar un principio como el de (1), famoso en la bibliografía, de manera positiva (en el caso del inglés o el francés) o negativa (en el caso del español o el vasco):

(1) Parámetro del Sujeto Nulo (PSN)
El sujeto de una oración debe manifestarse fonéticamente

Como puede verse, (1) únicamente habla de la posibilidad de que un sujeto se exprese fonéticamente (e.g., *los estudiantes han aprobado*) o no (e.g., *∅ han aprobado*). Sin embargo, pronto se observó que las lenguas que fijaban negativamente (1), desplegaban otras características (un "racimo" de propiedades; cf. Chomsky (1981, 1986)):

(2) Efectos de racimo del PSN
   a. Sujetos nulos:
   ∅ han aprobado
   b. Sujetos postverbales:
   Han aprobado ellos
   c. Extracción larga de sujeto:
   ¿Quién dices que vino?
   d. Pronombres expletivos nulos:
   ∅ Llueve

Este tipo de correlaciones guardan ciertas similitudes con los "universales implicativos" propuestos por Joseph Greenberg (cf. Greenberg (1963); Mairal & Gil (2006), y referencias allí citadas), que tienen el formato "Si una lengua L tiene un rasgo x, entonces L tiene un rasgo y". En (3) ofrecemos una muestra de ese tipo de universales:

(3) Universales implicativos
   a. **Universal 5.** Si una lengua tiene el orden dominante SOV y el complemento genitivo aparece después del sustantivo, entonces el adjetivo también aparece después del sustantivo
   b. **Universal 7.** Si en una lengua con orden dominante SOV, no hay orden alternativo básico, o solo OSV es una alternativa, entonces todos los modificadores adverbiales del verbo también preceden al verbo
   c. **Universal 13.** Si el objeto nominal siempre precede al verbo, entonces las formas verbales subordinadas al verbo principal también lo preceden [tomados de Greenberg (1963, pgs. 79, 80, 64), traducción nuestra]

Acabamos de ver un parámetro cuya fijación tiene consecuencias para otros rincones de la gramática (los efectos de racimo), pero no hemos dicho mucho sobre dónde ni cómo se codifican los "puntos de variación". Borer (1984) fue el primer trabajo en el que se planteó la pregunta de dónde se encuentran los parámetros —es decir, en qué componente de la gramática. Tal y como acabamos de ver, la TPP concibe los parámetros como principios moldeados mediante la experiencia, por lo que podríamos pensar que aquellos están en cualquiera de los módulos que posee una gramática: léxico, morfología, fonética, sintaxis, etc. Junto con Borer (1984), autores como Fukui (1986), Fukui & Speas (1986), Kayne (2000, 2005), Ouhalla (1991) y Webelhuth (1992) replantearon la propuesta de Chomsky (1986) al suponer que los parámetros se encontraban en el léxico. Borer (1984) expuso esta idea de manera clara al afirmar que "no hay elecciones de cada lengua específica con respecto a los principios y procesos universales. En vez de eso, la variación entre lenguas debería estar restringida a las propiedades idiosincrásicas de las unidades léxicas" (Borer (1984, pg. 2), traducción nuestra). Adoptando el término acuñado por Baker (2008), esta hipótesis se ha conocido como Conjetura Borer-Chomsky (CBC):

(4) Conjetura Borer-Chomsky (CBC)
Todos los parámetros de variación son atribuibles a diferencias en los rasgos de ítems particulares (p.ej., núcleos funcionales) en el léxico [tomado de Baker (2008, pg. 353), traducción nuestra]

La CBC (y sus diversas manifestaciones) marcó un punto de inflexión para el estudio de la variación lingüística, especialmente al permitir distinguir dos tipos de parámetros: (i) Microparámetros y (ii) Macroparámetros. El PSN es un ejemplo clásico de macroparámetro: se trata de un parámetro anclado a la sintaxis con consecuencias a gran escala una vez fijado y suele servir para distinguir lenguas tipológicamente distantes. Los microparámetros, por su lado, son parámetros compatibles con la CBC, con consecuencias limitadas a aquellas propiedades que puedan codificarse en las unidades léxicas, y con efectos visibles en lenguas cercanas (o geolectos de estas, lo cual es particularmente relevante para el estudio desarrollado en el ASinEs). La caracterización que de unos y otros ofrece Baker (2008) es esta:

> La visión microparamétrica estándar es que las diferencias primitivas científicamente significativas entre lenguas [o dialectos] son siempre diferencias pequeñas, típicamente asociadas a (como mucho) unas cuantas construcciones relacionadas [...] Las grandes diferencias entre lenguas siempre se reducen a muchas de las diferencias pequeñas [...] Por el contrario, la visión macroparamétrica es que hay al menos unos pocos parámetros (no compuestos) que definen tipológicamente diferentes tipos de lenguas.
>
> [Baker (2008, pgs. 255–256), traducción nuestra]

Las principales asimetrías entre el punto de vista macroparamétrico y microparamétrico pueden resumirse como en el cuadro 1.

	Macro Parámetros	Micro Parámetros
DÓNDE	gramática (o sintaxis)	léxico
CÓMO	efectos masivos	efectos limitados
QUÉ	lenguas no relacionadas filogenéticamente	lenguas relacionadas filogenéticamente

Cuadro 1: Principales asimetrías entre el punto de vista macroparamétrico y microparamétrico

Es importante observar, para los propósitos del presente proyecto, que existen determinados fenómenos sintácticos del español a los que sí se ha prestado atención en los estudios geolectales (cf. Bosque (1999); Brucart (1994); Demonte (2000); Demonte & Fernández-Soriano (2005); García Mouton (1994); Gómez Torrego (1999); Fernández-Ordóñez (1993, 1999); Ordóñez & Olarrea (2006)). Son los siguientes:

(5)
a. Dequeísmo:
Me dijo de que vendría tarde
b. Queísmo:
La idea que no la volveré a ver más
c. Relativas enfáticas:
No tienes idea de las cosas que dice
d. Perífrasis (focales) de relativo:
Es en Boston que lo vi
e. Leísmo:
Le criticaron duramente (de persona)
Ese libro, no le he leído (de cosa)
f. Laísmo:
La dije la verdad a María
g. Loísmo:
Lo di el libro, a Juan
h. Duplicación pronominal:
La toqué a la sonata
i. Distinción indicativo/subjuntivo:
No sé qué te diga

Aunque, como acabamos de decir, los fenómenos de (5) tienden a presentarse como sintácticos, creemos que, salvo los casos de (5-c) y (5-d), los demás podrían considerarse morfológicos (sobre todo si se adopta el marco teórico de la "Morfología Distribuida"; cf. Halle & Marantz (1993)). Esto plantea la cuestión de si existen fenómenos genuinamente sintácticos que manifiesten variación en el español (y en cualquier otra lengua; cf. Picallo (2014)). Sea cual sea la respuesta a esa pregunta, parece evidente que existe una asimetría entre los estudios sintácticos y los puramente morfo-fonéticos de la variación geolectal. Creemos, como decíamos hace un momento, que ello es el resultado de más de un factor. Además de la ausencia de un marco teórico que permita analizar las sutilidades de algunos casos de variación, la influencia del estructuralismo también dificultó que los estudios de variación abordasen datos sintácticos. En un trabajo reciente, Noam Chomsky observa lo siguiente:

> La publicación de lo que fue la fundación de la lingüística estructuralista americana, *Métodos de lingüística estructural*, de Zellig Harris, se llamó "métodos" porque parecía haber poco que decir sobre el lenguaje más allá de los métodos que había para reducir los datos de lenguas que variaban sin límite a una forma organizada. El estructuralismo europeo fue esencialmente idéntico. La introducción clásica de Nikolai Trubezkoy al análisis fonológico tenía una concepción similar. De manera más general, las investigaciones estructurales se centraron casi exclusivamente en la fonología y la morfología, las áreas en las que las lenguas parecen diferir ampliamente y de manera compleja.
>
> [apud Chomsky (2008, pgs. 2-3), traducción nuestra]

Asumiendo que el primero de estos factores ha sido corregido (en la actualidad sí contamos con una teoría sintáctica que nos permite desarrollar análisis detallados), el proyecto ASinEs pretende paliar el vacío existente en los estudios de sintaxis geolectal del español mediante la elaboración de un atlas que tenga una base de datos asociada.

La ausencia a la que nos estamos refiriendo es doblemente sorprendente: por un lado, porque muchas otras lenguas poseen este tipo de herramientas y, por el otro, porque las peculiaridades del español, tanto cuantitativas (es la segunda lengua más hablada del mundo, solamente superada por el chino mandarín, según los datos de Ethnologue[2]) como cualitativas (actualmente, está en contacto con lenguas románicas, germánicas, amerindias, así como con lenguas como el vasco) lo convierten en una fuente de interés tipológico para estudios de tipo geolectal. Por ello mismo, nuestro proyecto tiene un componente transversal, de "espectro amplio", que lo hace interesante para diferentes usuarios: desde geolingüistas hasta sintactistas teóricos, pasando por sociolingüistas, tipólogos y estudiantes / profesores de español.

2 Objetivos del proyecto ASinEs

El proyecto ASinEs pretende desarrollar un estudio detallado y pionero, así como una caracterización precisa, de la variación sintáctica manifestada en los diferentes geolectos del español, una lengua hablada por aproximadamente 420 millones de hablantes nativos (460 millones, si tenemos en cuenta a los hablantes que tienen el español como lengua segunda), mayoritariamente en la Península Ibérica y el continente americano.

Una iniciativa de tales características no tiene precedentes en el campo (no existe ningún estudio sistemático de la variación sintáctica del español), y mucho menos con una herramienta en línea: un atlas que pretende incorporar una amplia base de datos, creada a partir de las gramáticas de referencia del español (GDLE 1999 y NGLE 2009). De manera más general, nuestro proyecto contribuye no solo a la creación de una hoja de ruta detallada de la sintaxis del español (un objetivo con consecuencias muy provechosas para investigadores con diferentes posicionamientos teóricos, como hemos apuntado), sino también a la comprensión de un fenómeno complejo (la facultad del lenguaje y sus diferentes manifestaciones) que posee un amplio interés para otras áreas, de tipo cognitivo, histórico y sociológico (p.ej., bilingüismo, biodiversidad, antropología lingüística, diglosia, conflictos sociopolíticos, *code-switching*, *Sprachbund*, desarrollo del lenguaje, lenguas heredadas, sistemas de comunicación, psicolingüística, etc.).

El proyecto se articula alrededor de la creación de dos herramientas compactadas que ejercen de eje vertebrador:

- Una **base de datos / corpus de fenómenos y construcciones sintácticas del español**. Los datos son extraídos, principalmente, de las gramáticas de referencia del español, la GDLE (1999) y la NGLE (2009),

[2] https://www.ethnologue.com

aunque también se tienen en cuenta monografías, tesis doctorales, artículos publicados y otros materiales. En fases posteriores del proyecto, se incorporará información extraída de trabajo de campo (encuestas, entrevistas, etc.)

- Un **atlas interactivo en línea del español** (ASinEs), que se encarga de reflejar la variación sintáctica de las variantes del español, tanto americanas como europeas.

Este proyecto es original al menos por dos razones. Por un lado, a nivel teórico, el desarrollo de ASinEs conlleva, por primera vez, una investigación sistemática sobre la sintaxis geolectal del español, aprovechando para ello una base teórica y tipológica robusta, y también incorporando el reciente trabajo descriptivo de la GDLE (1999) y la NGLE (2009). Además, nuestra iniciativa aborda de manera sistemática aspectos que atañen al contacto de lenguas y la transición geolectal, incorporando a tal efecto tanto las estrategias tradicionales de investigación geolectal (la bibliografía disponible y los atlas existentes) como los mecanismos más actuales (corpus informáticos, bases de datos y atlas en línea, como el WALS, el BASYQUE, el SSWL o el TERRALING).

Por otro lado, a nivel metodológico, la investigación llevada a cabo tomando el español como objeto de estudio proporciona dos herramientas de trabajo que no tienen precedentes: una base de datos de construcciones con variación sintáctica y un atlas interactivo capaz de geolocalizar los puntos de variación. En este sentido, es importante volver a destacar el estatus especial del español en términos de cobertura empírica y geográfica, número de hablantes, y número de lenguas con las que interactúa (francés, catalán, gallego, vasco, inglés, árabe, rumano, etc.).

En resumen, tanto teórica como metodológicamente, el actual proyecto se adentra en territorio inexplorado. Los resultados de esta experiencia proporcionarán, sin duda, una visión novedosa y estimulante que servirá de referente para proyectos similares en el dominio de la lingüística, así como para otros ámbitos en los que el trabajo interdisciplinar es necesario. Y, naturalmente, el ASinEs como tal se convertirá en una herramienta para investigadores no solo de la variación sintáctica del español, sino también de la variación sintáctica de otras lenguas con las que esta esté en contacto (tanto románicas como no románicas).

3 Funcionamiento de la aplicación

La aplicación ASinEs (cuya interfaz de entrada puede verse en la Figura 2) utiliza la tecnología de Google Maps para geolocalizar los fenómenos sujetos a estudio; es decir, las construcciones sintácticas que presenten variación geolectal.

Figura 2: Interfaz principal del ASinEs.

La base de datos se compone de un sistema de fichas que contienen la siguiente información:

(6) Información fichas ASinEs
Ejemplo: La dije que no era verdad
Fuente: GDLE (§§ 23.4., 11.5), NGLE (§§ 34.1., 30.2., 4.7)
Lengua: español
Dialecto: español de Castilla
Fenómeno: laísmo
Construcción: construcción de doble objeto
Elem. gram.: pronombre, determinante, etc.
relacionado con: leísmo, loísmo
¿Se encuentra en otras lenguas? No
¿Se encuentra en otros estadios del español? No
Bibliografía: Romero (1997)
Geolocalización: puntos / áreas (predefinidas) / áreas (dibujo libre)
Sexo del informante (si es relevante)
Edad del informante (si es relevante)
Formación del informante (si es relevante)
Archivo de audio / vídeo
Comentarios:
Autor:
Fecha de creación:

A continuación, en la Figura 3, ofrecemos un ejemplo concreto de las fichas del ASinEs.

Como puede verse, el motor de búsqueda del ASinEs incluye análisis morfológicos y sintácticos. El sistema permite utilizar cualesquiera de los campos para realizar búsquedas, ya sean simples (basadas en un parámetro de las fichas) o

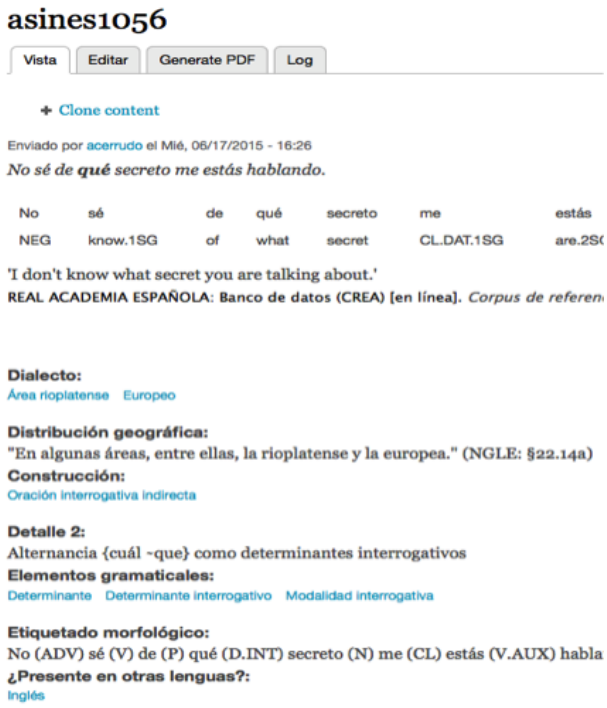


Figura 3: Ejemplo de ficha del ASinEs.

cruzadas (combinando más de un parámetro). De manera adicional, se incorpora un etiquetado morfológico y sintáctico, como se indica en (7):

(7) Ejemplo: *María dijo la verdad a Pedro*
ETIQUETADO MORFOLÓGICO
María(N) dijo(V) la(D) verdad(V) a(P) Pedro(N)
ETIQUETADO SINTÁCTICO
[(O) María [(SV) dijo [(SN) la verdad] [(SP) a Pedro]]]

Un etiquetado de doble nivel, como el de (7), ofrece la posibilidad de realizar búsquedas de símbolos terminales (cadenas markovianas, generables por máquinas de estados finitos) o secuencias que combinen símbolos terminales con símbolos no terminales. Ambos tipos de búsquedas se ilustran en (8):

(8) Búsqueda 1: `V + N` > dice cosas, etc.
Búsqueda 2: `V + SN` > dice [SN esas cosas], etc.

La interfaz de búsqueda asociada a la base de datos del ASinEs aparece en la Figura 4.

El tipo de sistema que estamos implementando permite, asimismo, desarrollar un analizador (*parser*), capaz de etiquetar automáticamente. Este analizador puede utilizar información disponible en las redes sociales (que, por defecto, contienen información geográfica) para complementar la información extraída de las gramática de referencia. Además, el sistema de etiquetado de (7) empleará una sintaxis que puede ser utilizada para generar diagramas arbóreos[3], como los que se muestran en la Figura 5.

Figura 4: Ejemplo de menú de búsqueda del ASinEs.

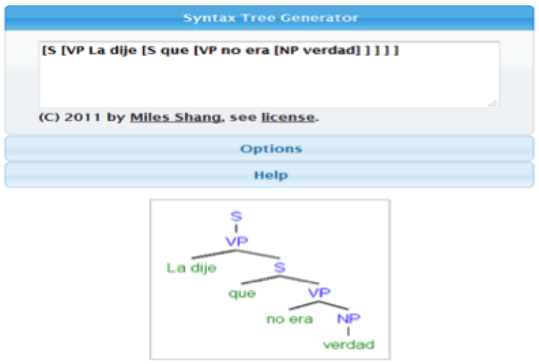


Figura 5: Ejemplo del analizador sintáctico.

Por su parte, el sistema de geolocalización permite asociar áreas geolectales concretas a las variantes del español relevantes, como se ve en las Figuras 6, 7 y 8.

Figura 6: Geolocalización (área) de Colombia.

Este tipo de información, codificable con puntos o áreas, está presente en muchas de las redes sociales, lo cual abre considerablemente el abanico de fuentes de las que extraer información para caracterizar tal o cual construcción sintáctica.

[3]Como, por ejemplo, `http://mshang.ca/syntree/`

Figura 7: Geolocalización (área) de Cataluña.

Figura 8: Geolocalización (área) de Castilla-La Mancha.

Entre estas redes se encuentran Instagram (utilizada sobre todo por jóvenes para comentar fotografías), Facebook (utilizada por usuarios de todas las edades, para conectar con amigos y conocidos), LinkedIn (utilizada por adultos por motivos profesionales), Tripadvisor (utilizada para compartir opiniones sobre restaurantes, hoteles, etc.) o Twitter (utilizada sobre todo por usuarios adultos con fines laborales o socioculturales). Toda esa información podría incorporarse al ASinEs en el futuro para complementar los datos conseguidos a través de las gramáticas de referencia, libros y manuales, tesis doctorales y artículos publicados, o información recopilada mediante trabajo de campo.

En resumen, el proyecto ASinEs ofrece una herramienta versátil para la investigación de la variación sintáctica de los geolectos del español, diseñada alrededor de una base de datos y un atlas que proporcionan diversos parámetros de búsqueda.

4 Conclusiones

El objetivo de este artículo ha sido el de presentar el Proyecto ASinEs, una aplicación que desarrolla un atlas sintáctico de la variación de los geolectos del español. Para ello, hemos revisado los antecedentes existentes (en los estudios geolectales de corte tradicional) y cuáles han sido los cambios teóricos que han permitido abordar el estudio de la variación sintáctica de manera realista.

El presente proyecto tiene, como se ha visto, objetivos ambiciosos, y pretende convertirse en una herramienta para los investigadores y estudiantes de variación sintáctica tanto del español como de otras lenguas relacionadas. En una primera fase (iniciada en enero de 2015) nos hemos centrado en la confección de una base de datos que recoja los puntos de variación presentes en las gramáticas de referencia del español, la *Gramática Descriptiva de la Lengua Española* (1999) y la *Nueva Gramática de la Lengua Española* (2009). En fases posteriores, pretendemos incorporar información de las siguientes fuentes:

- Tesis, monografías y artículos publicados
- Trabajo de campo (entrevistas, cuestionarios, etc.)
- Corpus que codifiquen información sintáctica relevante para el ASinEs (COSER, BASYQUE, etc.)
- Redes sociales (Twitter, Facebook, etc.)

El proyecto que presentamos en estas páginas es, en definitiva, de largo recorrido y puede ampliarse si se desarrollan las líneas de colaboración internacionales que existen en el Proyecto Edisyn[4].

Agradecimentos

Los autores de este artículo queremos dejar constancia de las diferentes fuentes de financiación que han permitido trabajar en el desarrollo del Atlas Sintáctico del Español (ASinEs). En primer lugar, todos agradecemos a la Fundación BBVA la concesión de una *Ayuda a Investigadores, innovadores y creadores culturales* 2014, que ha permitido poner en marcha el proyecto. En segundo lugar, agradecemos también las ayudas que provienen de la Generalitat de Catalunya (2014SGR-1013 y 2014SGR-1511) y el Ministerio de Economía y Competitividad (FFI2014-56968-C4-2-P, FFI22011-29440-C03-03, FFI2012-31415 y FFI2014-56968-C4-4-P). Gracias, por último, a los útiles y acertados comentarios de dos revisores anónimos. Si hay algún error, es nuestro.

[4] http://www.dialectsyntax.org/wiki/Main_Page

Referencias

Alvar, Manuel. 1996a. *Manual de dialectología hispánica. el español de américa*. Barcelona: Ariel.

Alvar, Manuel. 1996b. *Manual de dialectología hispánica. el español de españa*. Barcelona: Ariel.

Baker, Mark C. 2008. *The syntax of agreement and concord*. Cambridge University Press.

Borer, Hagit. 1984. *Parametric syntax: Case studies in semitic and romance languages*. Dordrecht: Foris Publications.

Bosque, Ignacio. 1999. Sobre la estructura sintáctica de una construcción focalizadora. En *En Homenaje al profesor Ambrosio Rabanales*, 207–231. BFUCh XXXVII.

Bosque, Ignacio & Violeta Demonte (eds.). 1999. *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe.

Brucart, José María. 1994. Syntactic variation and gramatical primitives in generative grammar. En A. Briz y M. Pérez-Saldanya (ed.), *Lynx. Categories and functions. A monographic series in Linguistics and world perception*, 145–176. Publicacions de la Universitat de València/University of Minnesota.

Campbell, Lyle. 2001. The history of linguistics. En Mark Aronoff & Janie Rees-Miller (eds.), *The Handbook of Linguistics*, 81–104. London: Blackwell Publishers.

Chambers, Jack K. & Natalie Schilling-Estes (eds.). 2013. *The handbook of language variation and change*, vol. 129. John Wiley & Sons.

Chambers, Jack K. & Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.

Chomsky, Noam. 1981. *Lectures on government and binding: The pisa lectures*. Dordrecht, Holland: Foris Publications.

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Chomsky, Noam. 2008. The biolinguistic program: Where does it stand today? Ms. MIT.

Company Company, Concepción. 2005. Una paradoja de la lingüística histórica romance: el florecimiento de la sintaxis histórica románica. *La Crónica* 34(1). 144–163.

Company Company, Concepción (ed.). 2006. *Sintaxis histórica de la lengua española. primera parte: La frase verbal*. México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Company Company, Concepción (ed.). 2009. *Sintaxis histórica de la lengua española. segunda parte: La frase nominal*. México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Company Company, Concepción (ed.). 2014. *Sintaxis histórica de la lengua española. tercera parte: Adverbios, preposiciones y conjunciones. relaciones interoracionales*. México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Demonte, Violeta. 2000. Gramática, variación y norma. una tipología. *Estudios Hispánicos (Revista de la Sociedad Coreana de Hispanistas)* 17(12). 3–49.

Demonte, Violeta & Olga Fernández-Soriano. 2005. Features in comp and syntactic variation: the case of '(de)queísmo'in spanish. *Lingua* 115(8). 1063–1082.

Fernández-Ordóñez, Inés. 1993. Leísmo, laísmo y loísmo: estado de la cuestión. En O. Fernández Soriano (ed.), *Los pronombres átonos*, Madrid: Taurus.

Fernández-Ordóñez, Inés. 1999. Leísmo, laísmo y loísmo. En I. Bosque & V. Demonte (eds.), *Gramática descriptiva de la lengua española*, Madrid: Espasa Calpe.

Fernández-Ordóñez, Inés. 2011. La lengua de castilla y la formación del español. Discurso leído el 13 de febrero de 2011 en su recepción pública por la Excma. Sra. D.ª Inés Fernández-Ordóñez y contestación del Excmo. Sr. D. José Antonio Pascual. Madrid.

Fernández-Ordóñez, Inés. 2014. Los dialectos del español de España. En Javier Gutiérrez Rexach (ed.), *Enciclopedia lingüística hispánica*, Routledge. En prensa.

Fukui, Naoki. 1986. *A theory of category projection and its applications*: MIT. Tesis Doctoral.

Fukui, Naoki & Margaret Speas. 1986. Specifiers and projections. *MIT Working Papers in Linguistics* 8. 128–172.

García Mouton, Pilar. 1994. *Lenguas y dialectos de españa*. Madrid: Arco/Libros.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. En Joseph H. Greenberg (ed.), *Universals of Human Language*, 73–113. Cambridge, Mass: MIT Press.

Gómez Torrego, Leonardo. 1999. La variación en las subordinadas sustantivas: dequeísmo y queísmo. En Ignacio Bosque & Violeta Demonte (eds.), *Gramática descriptiva de la lengua española*, Madrid: Espasa Calpe.

Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. En K. Hale & S.J. Keyser (eds.), *The view from Building 20: Essays in honour of Sylvain Bromberger*, 111–176. Cambridge, MA: MIT Press.

Kany, Charles Emil. 1945. *American spanish syntax*. Chicago: The University of Chicago Press. [traducción esp. *Sintaxis hispanoamericana*. Madrid: Gredos. 1970].

Kayne, Richard S. 2000. *Parameters and universals* Oxford studies in comparative syntax. Oxford University Press, USA.

Kayne, Richard S. 2005. *Movement and silence* Oxford Studies in Comparative Syntax. Oxford University Press, USA.

Labov, William. 1994. *Principles of linguistics change: Internal factors*, vol. 1. Oxford: Blackwell.

Labov, William. 2001. *Principles of linguistics change: Social factors*, vol. 2. Oxford: Blackwell.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of north american english: Phonetics, phonology and sound change*. Berlin: Mouton/de Gruyter.

Lapesa, Rafael. 2000. *Estudios de morfosintaxis histórica del español*. Madrid: Gredos.

Mairal, Ricardo & Juana Gil. 2006. *Linguistic universals*. Cambridge: Cambridge University Press.

Morillo-Velarde Pérez, Ramón. 1992. Un modelo de variación sintáctica dialectal: el demostrativo de realce en andaluz. En M. Ariza, R. Cano, J. M.a Mendoza & A. Narbona (eds.), *Actas del II Congreso Internacional de Historia de la Lengua Española*, vol. 2, 219–227. Madrid: Pabellón de España.

Ordóñez, Francisco & Antxon Olarrea. 2006. Microvariation in caribbean/non caribbean spanish interrogatives. *Probus* 18(1). 59–96.

Ouhalla, Jamal. 1991. *Functional categories and parametrization*. London: Routledge.

Petyt, Keith Malcolm. 1980. *The study of dialect: An introduction to dialectology*. London: The language library.

Picallo, M. Carme. 2014. *Linguistic variation in the minimalist framework*. Oxford: Oxford University Press.

RAE-ASALE. 2009. *Nueva gramática de la lengua española*. Madrid: Espasa.

Ridruejo, Emilio. 1992. Sintaxis histórica. En *Actas del I Congreso de la Lengua Española*, http://www.cvc.cervantes.es/obref/congresos/sevilla/unidad/ponenc_eridruejo.htm.

Sánchez Lobato, Jesús. 1994. El español en américa. En *Problemas y métodos en la enseñanza del español como lengua extranjera. Actas del IV Congreso Internacional de ASELE*, 553–570. Madrid.

Webelhuth, Gert. 1992. *Principles and parameters of syntactic saturation*. Oxford: Oxford University Press.

http://www.linguamatica.com/

linguaMATICA

Artigos de Investigação

Projetos, Apresentam-Se!