Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics

Abast i tipologia de la paràfrasi. Una aproximació empírica des de la lingüística computacional

Marta Vila Rigat

Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics

Marta Vila Rigat

A Dissertation Submitted
in Partial Fulfilment of the
Requirements of the Degree of
Doctor of Philosophy
with International Mention
to the Doctoral Program in
Cognitive Science and Language,
Department of General Linguistics,
Universitat de Barcelona

under the supervision of

Dr. M. Antònia Martí Antonín
Universitat de Barcelona

Dr. Horacio Rodríguez Hontoria
Universitat Politècnica de Catalunya

April 2013

Abast i tipologia de la paràfrasi. Una aproximació empírica des de la lingüística computacional

Marta Vila Rigat

Tesi presentada
per optar al grau de **Doctor**
amb Menció Internacional
dins del Programa de Doctorat
Ciència Cognitiva i Llenguatge,
Departament de Lingüística General,
Universitat de Barcelona

sota la direcció de

Dra. M. Antònia Martí Antonín
Universitat de Barcelona

Dr. Horacio Rodríguez Hontoria
Universitat Politècnica de Catalunya

Abril de 2013

If I had to define what a language is in a few words, I would say that it is a mechanism to produce synonyms.

– Igor Mel'čuk

Interview in *UAB Divulga*, October 2007.
(The translation is ours.)

In essence, linguistics is altogether missing in contemporary natural language engineering research. [...] I want to call for the return of linguistics to computational linguistics.

– Shuly Wintner

What science underlies natural language engineering?
Computational Linguistics (2009), 35(4):641–644

To my parents, for always being there.

Abstract

Paraphrasing is generally understood as approximate sameness of meaning between snippets of text with a different wording. Paraphrases are omnipresent in natural languages demonstrating all the aspects of its multifaceted nature. The pervasiveness of paraphrasing has made it a focus of several tasks in computational linguistics; its complexity has in turn resulted in paraphrase remaining a still unresolved challenge.

Two basic issues, directly linked to the complex nature of paraphrasing, make its computational treatment particularly difficult, namely the absence of a precise and commonly accepted definition and the lack of reference corpora for paraphrasing. Based on the assumption that linguistic knowledge should underlie computational-linguistics research, this thesis aims to go a step forward in these two questions: paraphrase characterization and paraphrase-corpus building and annotation. The knowledge and resources created are then applied to natural language processing and, in concrete, to automatic plagiarism detection in order to empirically analyse their potential.

This thesis is built as an article compendium comprising six core articles divided in three blocks: (*i*) paraphrase scope and typology, (*ii*) paraphrase-corpus creation and annotation, and (*iii*) paraphrasing in automatic plagiarism detection.

In the first block, assuming that paraphrase boundaries are not fixed but depend on the field, task, and objectives, three borderline paraphrase cases are presented: paraphrases involving content loss, pragmatic knowledge, and certain grammatical features. The limits between paraphrasing and related phenomena such as coreference are also analysed. Paraphrase characterization takes on a new dimension if we look at it in extensional terms. We have built a general and linguistically-grounded paraphrase typology in line with this approach. The third issue addressed in this block is paraphrase representation, which we consider to be essential in order to formally apprehend

paraphrasing.

In the second block, the Wikipedia-based Relational Paraphrase Acquisition method (WRPA) is presented. It allows for the automatic extraction of paraphrases expressing a concrete relation from Wikipedia. Using this method, the WRPA corpus, covering different relations and two languages (English and Spanish), was built. A subset of the Spanish WRPA corpus, together with paraphrases in two English paraphrase corpora that are different in nature were annotated applying a new annotation scheme derived from our paraphrase typology. These annotations were validated applying the Inter-annotator Agreement for Paraphrase-Type Annotation measures (IAPTA), also developed in the framework of this thesis.

In the third and final block, our typology is applied to the field of automatic plagiarism detection, demonstrating that more complex paraphrase phenomena and a high density of paraphrase mechanisms make plagiarism detection more difficult, and that lexical substitutions and text-snippet additions/deletions are the most widely used paraphrase mechanisms when plagiarizing. This provides insights for future research in automatic plagiarism detection and demonstrates, through a concrete example, the value of the knowledge and data provided in this thesis to computational-linguistics research.

Resum

S'entén per paràfrasi la igualtat aproximada de significat entre fragments de text que difereixen en la forma. La paràfrasi és omnipresent en les llengües naturals, on es troba expressada de múltiples maneres. D'una banda, la ubiqüitat de la paràfrasi l'ha convertit en el centre d'interès de moltes tasques específiques dins de la lingüística computacional; de l'altra, la seva complexitat ha fet de la paràfrasi un problema que encara no té una solució definitiva.

Dues qüestions bàsiques, lligades a la naturalesa complexa de la paràfrasi, fan el seu tractament computacional particularment difícil: l'absència d'una definició precisa i comunament acceptada i la manca de corpus de paràfrasis de referència. Assumint que el coneixement lingüístic ha de ser a la base de la recerca en lingüística computacional, aquesta tesi pretén avançar en dues línies de treball: en la delimitació i comprensió del que s'entén per paràfrasi, i en la creació i anotació de corpus de paràfrasis que proporcionin dades sobre les quals fonamentar tant la recerca com futurs recursos i aplicacions. Amb l'objectiu d'avaluar empíricament el seu potencial, el coneixement i els recursos creats com a resultat d'aquest treball han estat aplicats a la detecció automàtica de plagi.

Aquesta tesi consisteix en un compendi de publicacions i comprèn sis articles principals dividits en tres blocs: (*i*) abast i tipologia de la paràfrasi, (*ii*) creació i anotació de corpus de paràfrasis i (*iii*) la paràfrasi en la detecció automàtica de plagi.

En el primer bloc, partint de la base que els límits de la paràfrasi no són fixos, sinó que depenen de l'àrea de treball, la tasca i els objectius, es presenten tres casos límit de la paràfrasi: la pèrdua de contingut, el coneixement pragmàtic i la variació en determinats trets gramaticals. La caracterització de la paràfrasi pren una nova dimensió si l'observem des d'una perspectiva extensional. En aquesta línia, s'ha construït una tipologia general de la paràfrasi lingüísticament fonamentada. La tercera qüestió tractada en

aquest bloc és la representació de la paràfrasi, essencial a l'hora de tractar-la formalment.

En el segon bloc, es presenta un mètode per a l'adquisició de paràfrasis relacionals a partir de la Wikipedia (Wikipedia-based Relational Paraphrase Acquistion, WRPA). Aquest mètode permet extreure automàticament de la Wikipedia paràfrasis que expressen una relació concreta. Utilitzant aquest mètode, s'ha creat el corpus WRPA, que cobreix diverses relacions i dues llengües (anglès i espanyol). Un subconjunt del corpus WRPA en espanyol i exemples extrets de dos corpus de paràfrasis en anglès s'han anotat amb els tipus de paràfrasis que es proposen en aquesta tesi. Aquesta anotació ha estat validada aplicant les mesures d'acord entre anotadors (Inter-annotator Agreement for Paraphrase-Type Annotation, IAPTA), també desenvolupades en el marc d'aquesta tesi.

En el tercer i últim bloc, la tipologia proposada s'ha aplicat a l'àmbit de la detecció automàtica de plagi i s'ha demostrat que els tipus de paràfrasis més complexos i l'alta concentració de mecanismes de paràfrasi fan més difícil la detecció del plagi. També s'ha demostrat que les substitucions lèxiques i l'addició/eliminació de fragments de text són els mecanismes de paràfrasi més utilitzats en el plagi. Així, es demostra el potencial del coneixement parafràstic en la detecció automàtica de plagi i en la recerca en lingüística computacional en general.

Acknowledgments

Perhaps it is because I come from a valley where two rivers meet, but reflecting on the last five years, I see myself at times climbing mountains and at times observing rivers flow. The mountains of research are somehow special; no matter how much you ascend, you never reach the top. Still, you gain something more valuable: perspective. Climbing is challenging, but it has sometimes been tougher to let difficult questions simply flow. Their release often delivers the answers.

Two experienced guides have accompanied me in this ascent: M. Antònia Martí and Horacio Rodríguez. They have been indispensable in finding the best path and they have supported me in difficult moments along the way. Toni, thank you for all you have taught me on language, on strength, and on life. Horaci, thank you for helping me with my first steps in the world of computer science. I never imagined I would go so far.

Looking for the answers, I ended up on the other side of the world, in Sydney. I would like to express my gratitude to Mark Dras for opening the doors at Macquarie University for me and for guiding my research during my stay.

My sincere appreciation to Mariona Taulé, Anabela Barreiro, and Anselmo Peñas for agreeing to be members of my dissertation committee. Thank you also to Atsushi Fujita and Eduard Hovy for evaluating this thesis from afar. And thank you again to Mariona, for always supporting me.

I am also grateful to Paolo Rosso and Robert Dale for their interest in my work and the edifying insights provided. My gratitude also to Jordi Turmo, for being part of the PhD-project acceptance committee, and to María Jesús Machuca and Joaquim Llisterri, for their helpful advice in the ClInt project.

I have also been very fortunate to encounter other climbers along the way. I am indebted to Marta Recasens, Alberto Barrón-Cedeño, and Manuel

Bertran, whose work is also present in this thesis; thank you for your constant support. I am also grateful to Aina Peris and Glòria De Valdívia for all we have shared in our thesis and beyond. My gratitude to the annotators, Oriol Borrega, Rita Zaragoza, Montse Nofre, and Patricia Fernández; it was a pleasure to work with you. I would also like to give my thanks to Edgar Gonzàlez, for opening the doors to LaTeX for me, and to David Bridgewater, for his invaluable linguistic support. Thank you also to those that have played a role in other aspects of this research: Ana Pujol, Santiago González, Esther Arias, and Cristina España; and to those whose professional company I have greatly treasured: John Roberto and Santi Reig. I am also indebted to my colleagues at Macquarie University: Yasaman Motazedi, François Lareau, Teresa Lynn, Jette Viethnen, Ben Boerschinger, Jojo Wong, Mahbub Hassan, and Suzy Howlett. Being far from home, they supported me when I was in need, both professionally and personally.

This thesis, or anything for that matter, would not have been possible without the unconditional love of my parents. I see them in me and in what I do. This reflection gives me the security to move forward in my endeavors with a firm step. This thesis is also for my sister: walking next to her in life makes me feel safe and incredibly fortunate. My thanks to the Argila team for always bringing happiness into my days, and also to my cousins Judit and Montse, who are beginning a new chapter in life. I am grateful to my grandparents for helping me to see life with perspective and to appreciate all I have. I would also like to send my appreciation to Asunción and Jorge for showing me how life can surprise you, as well as to Teresa Garnatje for her steadfast help. I have been also very blessed to have my friends Mari, Rachel, Anna, Eva, Eloi, Joan, Rosa, Mireia, Laia, Maria, Anna, and the two Blanca's by my side throughout this trip. And last, but certainly not least, I dedicate this thesis to Joel, my lord of the summits. Mountains are no longer frightening with him.

⋆ ⋆ ⋆

This work was supported by an FPU Grant (AP2008-02185) from the Spanish Ministry of Education and the FI Grant (2009FI B 00690) from the Generalitat de Catalunya.

Contents

Part I

Introduction

Chapter 1

Introduction

This thesis is about paraphrasing or the use of different wordings to express the same meaning. The omnipresence of paraphrasing in natural languages has resulted in this phenomenon becoming a focus of several tasks in computational linguistics. Paraphrase complexity, in turn, has made these tasks still unresolved challenges.

I support Wintner (2009), who calls for the return of linguistics to computational linguistics, and go a step forward in paraphrase characterization in order to provide computational linguistics with linguistic knowledge to be used in the development of tools, resources, system, and applications. This linguistic knowledge has been empirically validated through the development and annotation of paraphrase corpora, and through its exploration in a concrete computational-linguistics task, namely automatic plagiarism detection.

This thesis is built as an article compendium, preceded by this introduction, which presents the set of articles, and closed by the conclusions and results derived from them.

In this introduction, the paraphrase phenomenon is set out (Section 1.1). Section 1.2 presents a general and brief state of the art. In Section 1.3, my objectives are described. Finally, Section 1.4 sets out an outline of the thesis and the article compendium in particular.

1.1 The Paraphrase

Paraphrases are commonly defined as those linguistic expressions that, showing a different wording, hold the same meaning. Paraphrasing may take place from the smallest meaningful unit [e.g., the morpheme pair $(\beta - \beta')$ in example (1)] to a full sentence [e.g., the whole pair in (1)] and even a full document. Also, the linguistic mechanisms of paraphrasing constitute a broad and mis-

cellaneous group running from lexical substitutions ($\alpha - \alpha'$) or structural reorganizations ($\gamma - \gamma'$) to complete changes in the wording ($\delta - \delta'$).

(1) a. When we were [looking around]$_\alpha$ the Museum of Modern Art in New York, [our guide drew our attention to a [bi]$_\beta$colored Rothko work]$_\gamma$, [by which I couldn't help but feel touched]$_\delta$.

b. When we were [visiting]$_{\alpha'}$ the Museum of Modern Art in New York, [our attention was drawn by our guide to a [di]$_{\beta'}$chromatic Rothko work]$_{\gamma'}$, [which caused a deep impression on me]$_{\delta'}$.

Nevertheless, meaning preservation has been discussed at length in linguistics literature. In lexical semantics, Cruse (1986) defines absolute synonymy as an unexpected and merely transitory relationship. Sameness of meaning is also negated in paraphrase literature; Fuchs (1988) rejects the idea of paraphrasing as pure and simple identity: "the synonymy-identity myth has only given rise to sterile arguments." Therefore, with the exception of ($\beta - \beta'$), where the sameness of meaning seems to be a fact, all the paraphrases in (1) are called into question.

How can speakers consider pairs like the ones in (1) to be paraphrases if they do not hold the same meaning? The answer is that paraphrasing occurs in the field of approximation. Therefore, the above definition of paraphrasing should be reworded as follows: paraphrases are those linguistic expressions that, showing a different wording, hold *approximately* the same meaning.

The approximate and vague nature of paraphrasing seems to lead us to a dead end: standing on this shaky terrain, what are the bases, if any, that make paraphrasing a factual phenomenon adequate for systematic and formal analysis? However, what at a first sight seems to be a wasteland, is actually a rich and promising field. Numerous samples show that paraphrasing is inherent in human linguistic communication and that there is a tacit agreement between speakers justifying its existence.

Paraphrasing is at the base of the expositive and argumentative nature of any discourse. In order to make an explanation clear and in order to convince the reader or the hearer of a viewpoint, the same information is repeatedly used (giving rise to paraphrases) for clarification, expansion, or emphasis. Paraphrasing is also at the base of changes in register. Depending on the audience, the information we want to communicate will be displayed differently (again giving rise to paraphrases).

Classical biblical exegesis and training in rhetoric is also based on paraphrasing. For young orators, imitating and emulating the great authors was a way to learn their trade.[1] The other side of the story is when paraphrasing

[1]See Fuchs (1994), Chapter 1 for more reading on this topic.

is used for reprehensible purposes like plagiarism. The ability to copy others without being caught depends, indeed, on the ability to paraphrase.

Different translations of a book into the same language or different newspaper articles talking about the same event can also be considered to be paraphrases. The list is endless and shows that paraphrasing is a reality in human language communication, a reality that relies on a tacit agreement between speakers. This factual nature sets the basis that make paraphrasing adequate for systematic and formal analysis. This section is closed with a quote by Martin (1976), who also points out the "evident reality" of paraphrasing:

> The speakers [...] clearly have an intuitive understanding of paraphrasing [...]. They can recognize paraphrases [...]; they can also express the same idea in different forms [...]. Therefore, even if he is not able to say what meaning identity is, the speaker at least has a feeling for and practice with meaning-identity, so that paraphrasing has an evident reality for him. (The translation is ours.)

1.2 State of the Art

In this thesis, a state of the art is provided within each article in the compendium, when pertinent, according to its topic. In this section, a general and brief overview of works on paraphrasing from computational linguistics and their interaction with those from linguistics is provided. I focus on the main issues addressed in this thesis, namely the linguistic characterization of paraphrasing and paraphrase-corpus building and annotation.

Paraphrase knowledge is fundamental to many Natural Language Processing (NLP) applications, such as question-answering, where the wording of a question may differ to that of its answers; summarization, where paraphrase knowledge is needed to avoid redundancy in the final summary; or editing, where paraphrases offer alternative expressions that fulfill certain communicative purposes.

Research on paraphrasing in NLP has been conducted along four main lines of research: paraphrase extraction, recognition, generation, and evaluation, with promising results for NLP applications.[2] Nevertheless, despite the efforts made, paraphrasing is far from being a resolved question. Two basic problems make paraphrase-system development a challenge: the lack of

[2]See surveys by Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010) for a general overview of paraphrasing in NLP.

a precise, formalizable, and commonly accepted paraphrase characterization and the lack of standard corpora to train and evaluate paraphrase methods and systems. In this regard, Herrera et al. (2007) state that "the difficulty when working with paraphrases lies on its own definition." Chen and Dolan (2011), in turn, point out that "there are no readily available large corpora and no consistent standards for what constitutes a high-quality paraphrase." Actually, these two lines of research go hand in hand: a deeper understanding of paraphrasing would guide the compilation of standard paraphrase corpora and only the availability of paraphrase data-sets can help us to obtain a better understanding of the phenomenon.

Regarding paraphrase characterization, paraphrasing has been object of study on its own on few occasions in linguistics and, when it has, it has been in the framework of proposals that are difficult to implement, such as Meaning–Text Theory (Žolkovskij and Mel'čuk, 1965), or indirectly in other linguistic proposals that give a partial view of paraphrasing, like the idea of transformation (Harris, 1957; Chomsky, 1965). This has led people working on NLP to base their works on the vague and non-operative paraphrase definition of "approximate sameness of meaning" and to develop ad-hoc techniques to deal with the phenomenon. Underlying each of these techniques, there is a way of understanding paraphrasing that is as partial and ad-hoc as the techniques themselves. By way of illustration, Harris (1954)'s distributional hypothesis has been the basis on many works on paraphrasing (Lin and Pantel, 2001; Barzilay and McKeown, 2001; Bhagat and Ravichandran, 2008); here paraphrasing is understood as those units of text sharing contexts. Other authors (Bannard and Callison-Burch, 2005) have used another language as pivot; in this case, paraphrases are those units of text sharing the same translation into the predefined pivot language. Building paraphrase typologies has also been a productive approach to apprehend paraphrasing extensionally in NLP. Nevertheless, with the exception of a few proposals that are more complete (Dras, 1999; Fujita, 2005; Bhagat, 2009), most of these typologies are again partial and ad-hoc.

The multifaceted nature of paraphrasing prevents the creation of comprehensive paraphrase corpora, that is, paraphrase corpora covering the phenomenon as a whole. Therefore, the field lacks a general and standard data set to be used for system training and evaluation. Only corpora covering specific paraphrase types or facets, directly linked to the way paraphrases were obtained, can be created. Some paraphrase corpora in existence are the Microsoft Research Paraphrase (MSRP) corpus (Dolan and Brockett, 2005) and the Microsoft Research Video Description corpus (Chen and Dolan, 2011). The former covers paraphrases from news articles extracted by applying edit distance and an heuristic strategy pairing initial sentences in those articles.

The latter contains parallel descriptions of short videos. Obviously, the nature of the paraphrases in these corpora is substantially different. Corpora coming from paraphrase related fields may also be useful for paraphrase research, like the PAN-PC-10 corpus (Potthast et al., 2010) from the plagiarism detection domain or the Multiple-Translation Chinese corpus (MTC) from the field of machine translation.

Paraphrase annotation has generally been limited to yes/no annotations (Dolan and Brockett, 2005) or alignments at word or phrase level (Cohn et al., 2008). A special type of paraphrase corpora are those containing information about the linguistic operations underlying paraphrases, in other words, corpora with the annotation of paraphrase types. Paraphrasing presents multiple and diverse linguistic manifestations; thus, this type of corpora shows great potential. Nevertheless, while paraphrase corpora are scarce, those with type annotation are, to the best of our knowledge, almost inexistent and only some small-scale attempts exist (Bhagat, 2009; Dutrey et al., 2011).

1.3 Objectives

In the previous section, two basic issues making the paraphrase treatment in NLP a challenge have been presented, namely the absence of a precise and commonly accepted paraphrase characterization and the lack of (annotated) paraphrase corpora. This thesis aims to address both of these issues in order to provide NLP with new linguistically-grounded paraphrase knowledge and resources. In concrete, my objectives are:

- To analyse paraphrasing and its boundaries in order to provide a clearer picture of them.
- To build a general typology of paraphrasing with the aim of helping in paraphrase characterization.
- To develop an annotation scheme based on the typology in order to annotate paraphrase corpora.
- To annotate varied paraphrase corpora using the annotation infrastructure created in order to validate the typology.
- To develop inter-annotator agreement measures to check the quality of the annotated corpora.
- To study how paraphrasing should be represented.

- To develop a system to automatically extract paraphrases in order to obtain paraphrase corpora.

- To analyse the potential of paraphrase-type knowledge in NLP tasks and applications.

1.4 Thesis Outline

As already stated, this thesis is built as an article compendium, preceded by this introduction and closed by the global conclusions. In this section, we focus on the article compendium.

The results of the present investigation are comprised within six main articles at different stages of publication/acceptance in several journals. They are divided in three blocks as shown below. Moreover, appendices include previous-version publications of three of the main articles. Previous-version publications set out embryonic and earlier proposals of the topics developed further in the main articles. Moreover, previous versions are short papers and the main articles are more comprehensive. For previous-version publications, only the first page is provided. Appendices also include three collateral contributions arising from research carried out during this thesis. It is important to point out that the results of the thesis are presented in the six core articles and that the articles in the appendices are complementary. The conclusions (Section 5) are derived from the core articles. Finally, the list of the twelve publications that appears below is not organized chronologically but in a logical order within each block of the thesis.

Block I: Theoretical Approach. Paraphrase Scope and Typology

1. Vila, M., Martí, M. A., and Rodríguez, H. (submitted-b). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Lingua*

2. Recasens, M. and Vila, M. (2010). On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647

3. Vila, M. and Dras, M. (2012). Tree edit distance as a baseline approach for paraphrase representation. *Procesamiento del Lenguaje Natural*, 48:89–95

Block II: Empirical Approach. Paraphrase Corpora: Creation and Annotation

4. Vila, M., Rodríguez, H., and Martí, M. A. (submitted-c). Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus. *Natural Language Engineering*

5. Vila, M., Bertran, M., Martí, M. A., and Rodríguez, H. (submitted-a). Corpus annotation with paraphrase types. New annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation*

Block III: Applicative Approach. Paraphrasing in Automatic Plagiarism Detection

6. Barrón-Cedeño, A., Vila, M., Martí, M., and Rosso, P. (2013, to appear). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*. DOI: 10.1162/COLI_a_00153

Appendix A: Previous-Version Publications

7. Vila, M., Martí, M. A., and Rodríguez, H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90

8. Vila, M., Rodríguez, H., and Martí, M. A. (2010b). WRPA: A system for relational paraphrase acquisition from Wikipedia. *Procesamiento del Lenguaje Natural*, 45:11–19

9. Barrón-Cedeño, A., Vila, M., and Rosso, P. (2012). Detección automática de plagio: De la copia exacta a la paráfrasis. In Garayzábal, E., Jiménez, M., and Reigosa, M., editors, *Lingüística Forense: La Lingüística en el Ámbito Legal y Policial*, pages 71–101. Euphonía Ediciones, Madrid

Appendix B: Collateral Publications

10. España-Bonet, C., Vila, M., Rodríguez, H., and Martí, M. A. (2009). CoCo, a web interface for corpora compilation. *Procesamiento del Lenguaje Natural*, 43:367–368

11. Gonzàlez, E., Rodríguez, H., Turmo, J., Comas, P. R., Naderi, A., Ageno, A., Sapena, E., Vila, M., and Martí, M. A. (2013, to appear). The TALP participation at TAC-KBP 2012. In *Proceedings of the 5th Text Analysis Conference (TAC 2012)*, Gaithersburg (MD)

12. Vila, M., González, S., Martí, M. A., Llisterri, J., and Machuca, M. J. (2010a). ClInt: a biligual Spanish-Catalan spoken corpus of clinical interviews. *Procesamiento del Lenguaje Natural*, 45:105–111

Articles 2 and 6 have been published or accepted for publication in *Computational Linguistics*, which has an impact factor of 0.721 (the former article was published in the *Squibs* section of that journal); and Article 3 was published in *Procesamiento del Lenguaje Natural*, indexed in SciVerse Scopus and Repositorio Expañol de Ciencia y Tecnología (RECYT), and with the Fundación Española para la Ciencia y la Tecnología (FECYT) seal of quality. Articles 1, 4, and 5 have been submitted to three journals also included in the Journal Citation Reports (JCR). In concrete, Article 1 has been submitted to *Lingua*, with an impact factor of 0.638; Article 4 has been submitted to *Natural Language Engineering*, with an impact factor of 0.432 (in this case, the second revision process has already been completed); and Article 5 has been submitted to *Language Resources and Evaluation*, with an impact factor of 0.308.

I am the first author in five of the six articles (in Article 6, this position is shared with Alberto Barrón-Cedeño). I am the second author in Article 2. Articles 2 and 6 are the result of collaborative work. The former compares paraphrase and coreference, and was also part of Marta Recasens's thesis on coreference resolution (Recasens, 2010). The latter analyses the use of paraphrase knowledge in automatic plagiarism detection, and the information contained in the article (not the article itself) was also partially present in Alberto Barrón-Cedeño's thesis on text re-use and plagiarism detection (Barrón-Cedeño, 2012). Article 3 was the result of a research stay at Macquarie University (Sydney) under the supervision of Mark Dras.

Regarding the six articles in the appendices, most of them were published in *Procesamiento del Lenguaje Natural*; Article 9 was published in an electronic book; and Article 11 will appear in conference proceedings. I am the first author in three of them.

In what follows, an overview of the core articles is set out. Articles in the appendices are also briefly mentioned. The article compendium is divided in three blocks that correspond to different approaches to the analysis of paraphrasing: a theoretical approach, in which paraphrase characterization

is addressed; an empirical approach, where theoretical proposals are empirically validated and new paraphrase resources are created; and an applicative approach, in which the potential of our proposals for automatic plagiarism detection are discussed.

The first article in Block I was the last written: it sets out our proposals on paraphrase boundaries and typology after the analysis of the phenomenon and the empirical validations performed during the PhD research. Based on the idea that paraphrase limits depend on the field, task, and objectives, paraphrase borderline cases are presented. Paraphrase characterization takes another dimension with the typology. It was built on the basis of the state-of-the-art typologies, but goes a step forward in three fundamental issues: first, it consists of general typology of paraphrasing that leaves fine-grained linguistic mechanisms in a second term; second, it goes beyond a simple list of types and is embedded in a linguistically-based structure; finally, it was empirically validated on paraphrase corpora. Article 7 in the appendices consists of an embryonic and shorter version of some of the questions presented in Article 1. It was written before our empirical experiments.

Article 2 focusses on a concrete issue in paraphrase boundaries: the fact that paraphrase and coreference overlap considerably, which has sometimes led to confusion in the NLP community. This article provides a better understanding of the two phenomena by comparing them in the light of different features, such as their reliance on meaning or reference, and their function in discourse. The article then sets out some cases that demonstrate how the two phenomena can help each other in paraphrase extraction and coreference resolution tasks.

Article 3 consists of a first approach to paraphrase representation, which we understand as another way to characterize paraphrasing. In concrete, it analyses the performance of Tree Edit Distance (TED) as a paraphrase representation baseline. Our experiments using Edit Distance Textual Entailment Suite–EDITS (Kouylekov and Negri, 2010) show that, since TED consists of a purely syntactic approach, paraphrase alternations not based on structural reorganizations do not find an adequate representation. They also show that there is much scope for better modeling of the way trees are aligned.

In Block II, paraphrasing and our theoretical proposals are seen from an empirical perspective. In concrete, in Article 4, we focus on a specific type of paraphrase, relational paraphrasing, and present the Wikipedia-based Relational Paraphrase Acquisition method (WRPA), which automatically extracts paraphrases expressing a concrete relation from Wikipedia. Using this method, the WRPA corpus was built. It currently covers 16 relations and two languages (English and Spanish). Article 8 in the appendices presents

an early reduced version of the WRPA method; Article 11, in turn, presents the results of the TALP participation at the TAC-KBP 2012 contest, where a subset of the patterns extracted by WRPA were used.

As explained in Article 5, a subset of 1,000 paraphrases in the Spanish WRPA corpus, together with the 3,900 pairs tagged as paraphrases in the MSRP corpus, and 856 plagiarism paraphrases in the PAN-PC-10 corpus (see Section 1.2) were annotated with our typology. The annotated PAN-PC-10 corpus was called Paraphrase for Plagiarism (P4P); and the annotated versions of WRPA and MSRP, WRPA-A and MSRP-A, respectively. This article presents the results of the annotation of these diverse corpora, as well as the annotation scheme and the inter-annotator agreement measures specifically created for this task. The inter-annotator agreement measures were called Inter-Annotator Agreement for Paraphrase Type Annotation (IAPTA). Our corpus-based approach to paraphrasing also revealed two paraphrase genres: reformulative paraphrases (paraphrases taking place in reformulation frameworks) and non-reformulative paraphrases. Article 10 in the appendices presents the CoCo interface, used both in building the WRPA corpus and in paraphrase-corpus annotation.

In Block III (Article 6), we analyse the relationship between paraphrasing and plagiarism: paraphrasing is the linguistic mechanism underlying many plagiarism acts. Our experiments using the P4P corpus allowed us to detect the most frequent paraphrase types used when plagiarizing, as well as the the most difficult types to be detected by plagiarism detection systems.[3] This data offers useful insights regarding the improvement of plagiarism detection systems. Article 9 in the appendices consists of a preliminary analysis of the paraphrase–plagiarism relationship.

Finally, less tied to the present investigation, Article 12 in the appendices presents ClInt (Clinical Interview), a bilingual Spanish-Catalan spoken corpus of clinical interviews. It can be used as a source of lay-technical paraphrases.

[3]This article also presents our typology. In Article 1, the focus is on the nature and structure of the typology and Article 6 focusses on the definition of each type.

Part II

Article Compendium

Chapter 2

Paraphrase Scope and Typology

2.1 Paraphrase Boundaries and Typology

Marta Vila (Universitat de Barcelona)
M. Antònia Martí (Universitat de Barcelona)
Horacio Rodríguez (Universtat Politècnica de Catalunya)

Is this a paraphrase? What kind? Paraphrase boundaries and typology

Submitted to *Lingua*.

Journal URL `http://www.journals.elsevier.com/lingua/`
Impact Factor 0.638

Abstract A precise and commonly accepted definition of paraphrasing does not exist. This is one of the reasons that has prevented computational linguistics from a real success when dealing with this phenomenon in its systems and applications. With the aim of helping to overcome this difficulty, in this article, new insights on paraphrase characterization are provided. We first overview what has been said on paraphrasing from linguistics and the new lights shed on the phenomenon from computational linguistics. Under the light of the shortcomings observed, the paraphrase phenomenon is studied from two different perspectives. On the one hand, insights on paraphrase boundaries are set out analyzing paraphrase borderline cases and the interaction of paraphrasing with related linguistic phenomena. On the other hand, a new paraphrase typology is presented. It goes beyond a simple list of types and is embedded in a linguistically-based hierarchical structure. This typology has been empirically validated through corpus annotation and its application in the plagiarism-detection domain.

Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology

Marta Vila[a], M. Antònia Martí[a], Horacio Rodríguez[b]

[a] *CLiC, Departament de Lingüística, Universitat de Barcelona. Gran Via de les Corts Catalanes 585. 08007 Barcelona*
[b] *TALP, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. Jordi Girona Salgado 1-3. 08034 Barcelona*

Abstract

A precise and commonly accepted definition of paraphrasing does not exist. This is one of the reasons that has prevented computational linguistics from a real success when dealing with this phenomenon in its systems and applications. With the aim of helping to overcome this difficulty, in this article, new insights on paraphrase characterization are provided. We first overview what has been said on paraphrasing from linguistics and the new lights shed on the phenomenon from computational linguistics. Under the light of the shortcomings observed, the paraphrase phenomenon is studied from two different perspectives. On the one hand, insights on paraphrase boundaries are set out analyzing paraphrase borderline cases and the interaction of paraphrasing with related linguistic phenomena. On the other hand, a new paraphrase typology is presented. It goes beyond a simple list of types and is embedded in a linguistically-based hierarchical structure. This typology has been empirically validated through corpus annotation and its application in the plagiarism-detection domain.

Keywords: Paraphrasing, Paraphrase boundaries, Paraphrase typology

1. Introduction

Although the computational linguistics[1] community has been working on paraphrasing over the last decades, it continues to be a challenging and

[1] We use the terms *computational linguistics* and *natural language processing* indistinctly, because their differences are not significant in this article.

unresolved issue. One of the main reasons is found in the multifaceted and boundless nature of the phenomenon, which makes its automatic treatment complicated.

Computational linguists have looked for precise and computationally-treatable knowledge on paraphrasing in the linguistics field without reaching a definitive solution. This has led researchers to rely on vague definitions of paraphrasing, such as "expressing one thing in other words" (Shinyama et al., 2002), "alternative ways to convey the same information" (Barzilay, 2003), or "sentences or phrases that convey approximately the same meaning using different surface words" (Bhagat, 2009), and to develop techniques based on workable paraphrase notions that are partial and ad-hoc.

In this scenario, our aim is to go a step forward in paraphrase linguistic characterization in order to provide Natural Language Processing (NLP) with more solid grounds for the development of methods and systems dealing with paraphrasing. We adhere to Wintner (2009), who calls for the return of linguistics to computational linguistics: "what makes our systems special is the fact that they manipulate natural languages, and the only scientific field that can inform our work is linguistics."

In concrete, we overview what has been said about paraphrasing in linguistics, how computational linguistics has used this knowledge as a base of its systems, an what are the new insights to paraphrasing derived from them. In light of the shortcomings observed, our proposal on paraphrase characterization is set out. It aims to help in answering two questions that reflect two different approaches to the phenomenon: "is this a paraphrase?", which puts on the table where paraphrase boundaries should be drawn, and "what kind?", aiming to describe what are the paraphrase linguistic manifestations, made concrete in a typology.

Our work is not tight to any concrete theoretical framework. Moreover, it has been empirically validated through annotation with our typology of more than 5,700 paraphrase pairs from three paraphrase corpora, which are different in nature and in two languages: the PAN-PC-10 corpus (Potthast et al., 2010), the Microsoft Research Paraphrase corpus–MSRP (Dolan and Brockett, 2005), and the Wikipedia-based Relational Paraphrase Acquisition corpus–WRPA (Vila et al., submitted-b). The annotated subsets of these corpora are called, respectively, P4P, MSRP-A, and WRPA-A. P4P and MSRP-

A are in English and WRPA-A is in Spanish (Vila et al., submitted-a).[2]

In Section 2, the state of the art on paraphrasing from linguistics and computational linguistics is set out. Section 3 presents our proposals on paraphrase boundaries and typology. Finally, conclusions and future work are presented in Section 4.

2. What Has Been Said About Paraphrasing?

Paraphrasing has been conceived and apprehended from different angles in linguistics and computational linguistics. The variety of visions of paraphrasing is even larger if we consider fields like discourse analysis or psycholinguistics, which have also addressed the phenomenon. This variety is again enlarged if we adopt a diachronic view, including disciplines such as rhetoric or biblical exegesis. As can be seen, paraphrase broad and multifaceted nature has a direct reflect in the literature.

In what follows, we focus on how paraphrasing has been understood in linguistics (Section 2.1) and computational linguistics (Section 2.2).[3]

2.1. In Linguistics

In the field of linguistics, paraphrasing is at the core of two theories that set out language models focussing on production: Meaning-Text Theory (MTT) and Systemic-Functional Grammar (SFG). Their proposals are substantially different in essence, but their approaches to paraphrasing, similar: both see language production as a system of choices or alternatives, which can give rise to paraphrases.

MTT gives rise to Meaning-Text Models (MTMs). Such models incorporate a grammar organized in seven levels of representation—with semantics and phonetics at the wings—comprising six components, which contain the rules that allow for going from one level of representation to the other. The second constituent in MTMs is the Explanatory Combinational Dictionary (ECD), which governs the whole process. Lexical Functions (LF), which

[2]Annotated paraphrase corpora and the annotation guidelines used are available at `http://clic.ub.edu/corpus/en/paraphrases-en`.

[3]See Fuchs (1994), Chapters 1 and 2 for a diachronic overview on approaches to paraphrasing from linguistics and discourse analysis.

identify recurrent patterns of semantic-syntactic correspondence, are a fundamental part of the ECD. Within this framework, two paraphrase mechanisms can be identified. First, paraphrases can be produced in the transition between levels of representation: representations in one level can be projected in two or more representation in the next one. Second, paraphrases can be established through equivalence rules between representations at the same level. Paraphrasing at the deep syntax level was first described by Žolkovskij and Mel'čuk (1965), who built a paraphrasing system comprising lexical and syntactic paraphrasing rules;[4] paraphrasing at the semantic level was more recently described by Milićević (2007a,b). The axiomatic foundations and formal complexity of MTT prevent its straightforward exploitation outside the MTT framework and lead to a costly computational implementation. Nevertheless, ECD and LF in particular are useful in themselves as they encode most of the paraphrase potential in the model.

Although in a less explicit way, paraphrasing is also at the base of SFG: "the systemic theory is a theory of meaning as a choice, by which a language, or any other semiotic system, is interpreted as networks of interlocking options" (Halliday, 1994). In this framework, paraphrases are the result of making alternative choices. Obviously, not all alternants are meaning-preserving and, therefore, not all of them give rise to paraphrases.

Other linguistic proposals include elements that can be used in paraphrasing. Transformations, which are at the core of Harris (1957)'s proposal and Chomsky (1965)'s Generative Grammar, have been used as a way to represent and enumerate formal relations between sentences. Some of these transformations are paraphrastic as they preserve the meaning of sentences. Transformations take place between surface structures in Harris's approach; in Chomsky's, in contrast, they take place from deep to surface syntax structures. In the latter case, different surface representations derived from the same deep structure can be understood as paraphrases. Following Hiż (1964)'s ideas, Smaby (1971) describes a paraphrase transformational grammar which maps equivalent structures. The main interest of this work is the effort to formalize paraphrasing; nevertheless, it only deals with those paraphrases which can be formally apprehended.

With the emergence of generative semantics (Lakoff, 1971), there was a movement to a semantically-based framework. Since, in this case, the deep

[4]For a more recent reference in English, see Mel'čuk (1992).

structure is purely semantic, generative semantics appears to be a suitable means for describing paraphrasing.[5] Diathesis alternations, which stand for those alternate structures that are admitted by the same predicate, can also be viewed as paraphrases. Levin (1993) provides diathesis alternations for English, some of them, such active/passive or causative/inchoative alternations, are of general application while others are specific for English language.

There also exist works that analyse and discuss the linguistic nature of paraphrasing. Martin (1976) defines linguistic paraphrasing as logical equivalence. He also describes two mechanisms of linguistic paraphrasing: first, "semic content" identity and "actantial pattern" correspondence, which roughly corresponds to structural reorganizations, and, second, "actantial pattern" identity and "semic content" correspondence, which mainly correspond to synonymy substitutions. Fuchs (1994), in turn, describes paraphrasing in discourse and in language from a diachronic perspective. Moreover, she argues for the enunciative dimension of paraphrasing: it cannot be reduced to closed equivalence, instead it consists of a dynamic and approximate relationship. Milićević (2007a), in line with proposals within the MTT framework, analyses paraphrasing as a multifaceted and variable phenomenon focussing on the different paraphrase dimensions. Some concrete aspects discussed by these authors are taken up in subsequent sections of this article.

Some of the works mentioned above include lists of paraphrase types. Mel'čuk (1992) enumerates 54 lexical and 29 syntactic paraphrasing rules within the MTT. Milićević (2007a) defines a set of MTT semantic-paraphrase rules and also classifies paraphrases from five different perspectives, such as accuracy of the paraphrase link (exact and approximate) or paraphrase-relationship depth (semantic, lexico-syntactic, syntactic, and morphological paraphrases). Lists of transformations (Harris, 1957) or diathesis alternations (Levin, 1993) can also be seen as typologies of potential paraphrases. The latter sets out around 60 diatheses organized in 8 main classes. Martin (1976), in turn, sets out varied paraphrase mechanisms, focussing on paraphrasing by connotative variation, double-negation or double-inversion paraphrasing, and paraphrasing by synonymy substitution.

[5]See Bagha (2011) to read more about this topic.

2.2. In Computational Linguistics

We analyse the paraphrase characterization in computational linguistics from two different perspectives. In Section 2.2.1, we analyse the notions of paraphrase which underlie NLP paraphrase techniques. In Section 2.2.2, we overview paraphrase typologies built in this field.

2.2.1. Paraphrase Notions Underlying NLP Methods

While linguistic analysis approaches paraphrasing with the aim of exploring, explaining, and formalizing it, NLP researchers focus on developing methods and techniques to deal with the phenomenon in their systems and applications.[6] Each method applied subsumes a way of understanding paraphrasing and paraphrases addressed with such a technique are of a particular nature. Sometimes these methods have their roots in linguistics; on other occasions, they were born within NLP.

A number of authors have applied MTT proposals. Boyer and Lapalme (1985) developed a paraphrase generation system based on the ECD and the lexical transformations of the model. Lareau (2002), in turn, presents an automatic text synthesis prototype system, Sentence Garden, which aimed to produce not only one sentence, but all possible sentences that express a given meaning (although the prototype only implemented the semantics–deep syntax interface).

The idea of transformation between surface structures has also been used in NLP. McKeown (1983), for example, sets out a paraphrase component for a question-answering system, where a transformational grammar is used to generate paraphrases. Romano et al. (2006) use transformation rules in their paraphrase-based approach to relation extraction.

Harris (1954)'s distributional hypothesis, which states that words occurring in the same contexts tend to have similar meanings, has been widely applied, directly or indirectly, more or less strictly, and under different forms: "sentences which appear in similar contexts are paraphrases" (Barzilay and McKeown, 2001), "if two paths [in a dependency tree] tend to occur in similar contexts, the meanings of the paths tend to be similar" (Lin and Pantel, 2001),[7] "named entities are preserved across paraphrases" (Shinyama et al.,

[6] See surveys by Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010) for a complete overview of paraphrase methods in NLP.

[7] This work and Kouylekov and Magnini (2005) below focus on entailment relations, which include paraphrases. See Section 3.1.

2002), "the meaning of the text around the source and target entities [in a concrete relation] will be similar throughout their different occurrences" (Vila et al., submitted-b).

Other authors establish the paraphrase link through a third vertex. In Rinaldi et al. (2003)'s question-answering system, paraphrases are those linguistic units mapping to the same logical representation. Bannard and Callison-Burch (2005), in turn, start out from the assumption of similar meaning when multiple phrases map onto a single foreign language phrase. The third vertex is a logical meaning representation in the first case and a sentence in another language in the second.

Similarity measures have also been used to address paraphrasing in NLP. In this framework, paraphrases are those text snippets with a high level of overlapping or a low distance. Similarity can be calculated at word level using, for example, string edit distance or ngrams overlapping (Dolan and Brockett, 2005); at syntax level, applying tree edit distance (Kouylekov and Magnini, 2005); and at semantic level, taking advantage of semantic roles in PropBank or FrameNet frames, using a semantic space such as WordNet or Wikipedia, or using distributed representations of co-occurrences, usually vector-based (Baroni and Lenci, 2010).[8] The latter approach is currently a very active research area. Semantic similarity has also been addressed in the Semantic Textual Similarity task in Semeval 2012, where paraphrases are ranked according to their similarity level.[9]

To conclude, each NLP technique applied addresses a concrete paraphrase facet, which is generally partial and ad-hoc. In this regard, a major distinction can be made. In methods relying on the formal mapping of the paraphrase members (transformations and formal similarity measures), paraphrases addressed must be similar in form. This is not the case of those methods where no formal mapping is necessarily assumed (MTT, distributional hypothesis, semantic similarity measures, and third vertex).

2.2.2. Paraphrase Typologies

Many NLP researchers have found in typology building a way to apprehend paraphrasing. Early works on paraphrase typologies are Culicover (1968) and Honeck (1971). They set out similar typologies in the sense that

[8] See Androutsopoulos and Malakasiotis (2010) for further reading on this topic.

[9] `http://www.cs.york.ac.uk/semeval-2012/task6/`

both divide their paraphrase types into formalizable and non-formalizable ones, leaving the latter group outside the scope of their work. This has been a general tendency in NLP and paraphrases where no formal mapping can be established have hardly been addressed. In concrete, Culicover (1968) presents a paraphrase typology of five types: transformational, attenuated, lexical, derivational, and real-world, and carries out a formalization attempt through the definition of some structural and semantic conditions to be fulfilled by each of the paraphrase types. He makes a division between computationally "accessible" and "inaccessible paraphrase relationships" and focusses on the accessible ones, leaving those inaccessible (most real-world paraphrases) under-explained. Honeck (1971), in the psychology field, offers a taxonomy of three types of paraphrases, including transformational, lexical and formalexic (combination of the two); however, he isolates two extra types of paraphrases that are outside the scope of his study: parasyntactic (unavailable for formal treatment) and syndetic (combination between the other types), where no formal correspondences can be established.

More recently, some typologies in NLP consist of lists of the most common types found in a corpus (Barzilay et al., 1999; Dutrey et al., 2011; Dolan et al., 2004), lists of the paraphrases they address (Dorr et al., 2004; Kozlowski et al., 2003; Boonthum, 2004), or simply lists of typical paraphrases with illustrative purposes (Rinaldi et al., 2003). In general, they are specific-work oriented and far from being comprehensive.

Sometimes paraphrasing is classified in a very generic way setting out only a few types, such as in Shimohata (2004, pp. 15–18) or Barreiro (2008, pp. 29–33). This types generally stand for the type of linguistic units or the level of language where paraphrases take place. There also exist typologies that focus on concrete paraphrase cases, such as paraphrases involving support-verb constructions (Barreiro, 2008, pp. 73–81), and typologies that come from paraphrase related fields, such as text reuse (Clough, 2003, p. 100) or editing (Faigley and Witte, 1981).

There also exist exhaustive paraphrase typologies focussing on concrete paraphrase facets, such as syntactic (Dras, 1999) or lexical mechanisms (Bhagat, 2009), or covering paraphrasing in a more comprehensive way (Fujita, 2005). More specifically, Dras (1999) sets out 54 types expressed in terms of syntactic transformations and groups them into five classes standing for paraphrase effects: change of perspective, change of emphasis, change of relation, deletion, and clause movement. Bhagat (2009), in turn, classifies paraphrases according to the lexical changes involved (e.g. actor/action substitution or

noun/adjective conversion) and links each of these types to the structural modifications accompaining them (substitution, addition/deletion, and/or permutation). Finally, Fujita (2005) presents a general classification of lexical and structural paraphrases[10] setting out 24 paraphrase types grouped into six classes including paraphrases of single content words, function-expressional paraphrases, paraphrases of compound expressions, clause-structural paraphrases, multi-clausal paraphrases, and paraphrases of idiosyncratic expressions.

Approaches to paraphrase characterization from NLP are generally partial and ad-hoc, but have opened new windows onto the paraphrase phenomenon understanding. In this section, we have shown how can computational linguistics "shed[s] new light on phenomena that traditional approaches fail to account for [and] bring refreshing insights and new points of view to al branches of linguistics" (Wintner, 2009).

3. Paraphrase Characterization

As shown in Section 2, a precise and commonly accepted definition of paraphrasing does not exist. From the perspective of linguistics and computational linguistics, the definition of "approximate sameness of meaning" is generally assumed, but it is vague (to what extent can it be "approximate"?) and actually shifts the problem to another location (what is "meaning"?)

In this article, we adopt a different approach to paraphrase characterization. Instead of focussing on the definition of paraphrasing itself, we address the questions of where to draw the boundaries between paraphrases and non-paraphrases (Section 3.1) and what phenomena fall under paraphrasing (Section 3.2). Although we are aware that paraphrase fuzziness is also present in both boundary drawing and typology building, and that they are simply another approach to the same problem, they allow us to be more precise without abandoning a general perspective on paraphrasing.

[10] This work is based on Japanese language; English and other examples can be found at `http://paraphrasing.org/paraphrase.html`. See also Atsushi Fujita's slides for the invited talk at CBA 2010 at `http://paraphrasing.org/~fujita/publications/fujita-CBA2010-slides.pdf`.

3.1. Paraphrase Boundaries

Meaning preservation has been discussed at length in the literature. In lexical semantics, Cruse (1986) defines absolute synonymy as an unexpected and merely transitory relationship. Sameness of meaning is also negated in paraphrase literature; Fuchs (1988) rejects the idea of paraphrasing as pure and simple identity: "the synonymy-identity myth has only given rise to sterile arguments." Therefore, paraphrasing must be situated in the field of the approximation, opening the path to different semantic similarity or degrees of *paraphrasability*. Paraphrasing takes place in a continuum that goes from absolute identity to the absence of semantic similarity. In this scenario, a question arises: where to draw the boundaries between paraphrases and non-paraphrases.

We consider that fixed and precise paraphrase boundaries do not exist, instead they depend on the task and objectives: sometimes a wide understanding of paraphrasing will be required, on other occasions, a more restrictive view will be necessary. Fuchs (1994) points out that a linguistic unit is a paraphrase of another one if the latter can be considered within the bounds of acceptable deformability or "distortion threshold" with respect to the former. This threshold is variable as "it depends on different parameters constituting the discursive activity: tolerance to deformation is greater or lesser depending on the subjects and situations."

In this section, we set out three cases of borderline paraphrases that are derived from our analysis of the state of the art of paraphrasing and related areas, and our experience in paraphrase-type annotation: loss of content, pragmatic knowledge, and changes in some grammatical features. These borderline paraphrases are placed in the continuum between paraphrases and non-paraphrases, in which authors can position their own paraphrase border according to their objectives. Moreover, for each of these cases, we mention the approach we adopted, which is reflected in our typology (Section 3.2). The section is closed with a comparison between paraphrasing and two related phenomena, namely coreference and textual entailment, which often lead to confusion in NLP.

Content Loss. Many paraphrase boundaries cases are due to some kind of content loss. Content loss may be due to deletion [*my favourite* in (1)] or generalization [from *pilot* to *commander* in (2)].

(1) a. Yesterday I went to the beach

b. Yesterday I went to *my favorite* beach

(2) a. The *pilot* was having breakfast
b. The *commander* was having breakfast

Depending on the quantity and relevance of the missing content, different degrees of paraphrasability are possible. In this sense, the level of paraphrasability of the sentences in (3) is lower than those in (1).

(3) a. Yesterday I went to the beach
b. Yesterday I went to the beach *which used to be my favorite when I was a child*

Moreover, the missing content can sometimes be recovered by means of implicit lexical knowledge in the context. The Generative Lexicon (Pustejovsky, 1995), though not addressing paraphrasing directly, offers useful insights in this regard. Setting out from the idea that the meaning of words reflects the deeper conceptual structures in the cognitive system, the qualia structure specifies four aspects of word meanings: formal (distinction within a larger domain), constitutive (relation between an object and its constituent parts), telic (purpose and function), and agentive (factors involved in its origin). In (4), the information contained in the qualia's telic of *book* allows for the recoverability of the deleted content (*reading*). In contrast, in (1), we have no way to recover the missing content. Therefore, the level of paraphrasability is higher in (4). Moreover, the pair in (5) shows a higher degree of paraphrasability than the pair in (2), as the context of *taking off* in the former clarifies that this *commander* is, actually, a *pilot*. In (2), we only rely on the hypernym relationship between *pilot* and *commander*.

(4) a. John began *reading* a book
b. John began a book

(5) a. The *pilot* was ready to take off
b. The *commander* was ready to take off

Depending on the task and objectives it is necessary to consider the above examples to be paraphrases or not. Many paraphrase types in our typology involve different degrees of semantic loss.[11] The ADDITION/DELETION type,

[11]Dras (1999, pp. 79–86) addresses the loss of meaning in paraphrasing regarding the

exemplified in Table 2, is a clear example of this. Although the missing content cannot always be recovered in our types, this is sometimes possible: in "light/generic element addition/deletion" within the SYNTHETIC/ANALYTIC SUBSTITUTION type (Table 3), the content of the deleted element is embedded in the one that remains, as the latter is an hyponym of the former. As shown in Vila et al. (submitted-a), ADDITION/DELETION is one of the most frequent types in the annotated corpora, demonstrating its accessibility when paraphrasing.

Pragmatic Knowledge. Examples like the ones in (6) to (10) are treated by several authors, both in linguistics and computational linguistics, as special types of paraphrases that go beyond pure semantic similarity to fall within the field of pragmatics.

(6) a. Close the door please
b. There is air flow

(7) a. *Penelope* was waiting for Ulysses return
b. *The Ithaca queen* was waiting for Ulysses return

(8) a. *Here*, life is good
b. *In Paris*, life is good

(9) a. They got married *last year*.
b. They got married *in 2004*.

(10) a. The U.S.-led *invasion* of Iraq
b. The U.S.-led *liberation* of Iraq

Martin (1976) contrasts "linguistic" to "pragmatic paraphrases", the latter standing for pairs that, in a given situation, refer to the same intention (6) or refer to the same facts and events (7).[12] Milićević (2007a), in turn, contrasts "language" to "cognitive paraphrases", the latter comprising paraphrases exploiting pragmatic data, such as (6), (8), and (9), and paraphrases exploiting encyclopedic knowledge, such as (7).[13] Fujita (2005) talks about "pragmatic paraphrases" (6) and 'referential paraphrases" (9). Dorr et al.

paraphrase classes in his typology.

[12]Martin (1976) presents a third type of pragmatic paraphrase relying on implication and coreference. We address coreference in the last part of this section.

[13]Milićević (2007a) includes a third type of cognitive paraphrases called paraphrases exploiting logic capacities, which also involves encyclopedic knowledge.

(2004) mention "viewpoint variation paraphrases" (10), also cited by Hirst (2003). Finally, Fuchs (1994) considers cases like the one in (7) to be outside the boundaries of paraphrasing.

The way to present and conceptualize all these examples varies according to the author, but all of them put forward the idea that paraphrasing may rely on something beyond pure semantic similarity. We distinguish between two main types of knowledge that can give rise to pragmatic paraphrases, namely encyclopedic knowledge [(7) and (10)] and situational knowledge (the remaining examples). These two types of knowledge are usually called *common-sense knowledge* in NLP. As Milićević (2007a) point out, we can also draw a continuum here: "between those clear and unambiguous cases, there is a gray area populated by paraphrases that can be called quasi-linguistic."

If we stick to the paraphrase definition of sameness of meaning, these examples are outside paraphrase limits. However, under certain circumstances, it may be necessary to consider these cases as a special type of paraphrase linked to the situational context. Because our typology relies on semantic content, those cases fall outside our proposal.

Grammatical Features. With the generic concept of "grammatical features", we refer to changes in person, number, and time. They generally lead to deep changes in meaning, though, on occasions, they may give rise to paraphrases.

The example in (11) is clearly nearer paraphrasing than (12), as, in (11), the first person plural includes the first person singular. In (13), the change in number is not relevant: *street* does not refer to a concrete one, but to the general sense of 'outdoors'; in (14), the change in number gains relevance as me move from the idea of 'liking a concrete cake' to 'liking cakes in general'. In (15), both tenses overlap to a high degree, which is not the case of (16), standing for different moments in time.

(11) a. *We* love flowers
b. *I* love flowers

(12) a. *She* is my collaborator
b. *He* is my collaborator

(13) a. I got lost in the *street*
b. I got lost in the *streets*

(14) a. I like *the cake*

b. I like *cakes*

(15) a. The plane *takes off* at 6:30
b. The plane *is taking off* at 6:30

(16) a. She *lives* in Barcelona
b. She *had lived* in Barcelona

Only examples (11), (13), and (15) are considered to be paraphrases in our approach. They are included in the INFLECTIONAL CHANGE type in our typology (Table 1). Contrary to content loss and pragmatic knowledge, which are language independent, this group includes phenomena that are closely related to how languages encode morpho-semantic content. In English, this is reflected in the inflection.

Paraphrase, Coreference, and Textual Entailment. Paraphrasing overlaps with coreference and textual entailment leading to recurrent confusions. In what follows, the main difference and similarities between these two phenomena and paraphrasing are presented.

Paraphrasing and coreference overlap considerably, but they differ in essence: paraphrasing is concerned with meaning, whereas coreference is about discourse referents (Recasens and Vila, 2010). In example (17), a paraphrase relationship exists between *shop assistant* and *sales person*; but, the former acts as a nominal predicate, which is not referential and cannot be part of coreference relationships. In contrast, in (18), we can establish a coreference relationship between the noun phrases in italics, but they do not hold the same meaning and, therefore, are not paraphrases. Finally, in (19), paraphrase and coreference overlap in *the coast/the seashore*.

(17) She is a *shop assistant* in that shop, but the *sales person* that assisted me was not her.

(18) – Are you a family member of *the patient in room 235*?
– Yes, *my cousin* is in that room.

(19) Yesterday I was walking along *the coast*. *The seashore* is what I really love in this area.

Paraphrases can also be seen as bidirectional entailment relations: "text A is a paraphrase of text B if and only if A entails B and B entails A" (Rus et al., 2009). Limiting paraphrasing to bidirectional entailment reduces it to very few cases; therefore, some unidirectional-entailment cases are generally

considered to be paraphrases. Dorr et al. (2004), for example, present "inference" as a paraphrase type. Kotlerman et al. (2010), in turn, introduce the concept of "directional similarity". Once again, we situate paraphrasing in a continuum with strict bidirectional entailment at one extreme and strict unidirectional entailment at the other. Where to put the boundaries between paraphrases and non-paraphrases depends again on the task and objectives.

The relationship between textual entailment and paraphrasing is intimately linked to the question of content loss mentioned above, as all paraphrases exhibiting content loss are cases of unidirectional entailment. In our typology, this is illustrated by ADDITION/DELETION (Table 2). Moreover, our typology includes types categorized as "paraphrase extremes" including IDENTICAL and NON-PARAPHRASE, which are clear paraphrase limits, and ENTAILMENT, that is, those cases of non-paraphrase that are closer to the paraphrase domain (Table 2). In the annotation task, it is worthwhile isolating these cases of entailment for researchers interested in broadening the scope of their work (Vila et al., submitted-a).

3.2. Paraphrase Typology

In this section, we focus on the characterization of paraphrasing through the description of its possible linguistic manifestations or types. Our typology is not a proposal started from scratch, but has been built on the basis of state-of-the-art typologies, which have provided ours with insights on structure and types. Actually, our typology aims to cover all the phenomena described in these typologies.[14]

A set of characteristics make our typology a step forward with respect to the state of the art. First, it consists of a comprehensive typology of paraphrasing that focusses on general paraphrase phenomena, leaving fine-grained linguistic mechanisms in a second term. Second, it goes beyond a simple list of types: it has a hierarchical structure, which is linguistically-based and uniform throughout, and it is accompanied by a linguistic reflection describing and justifying its nature. Finally, as previously mentioned, it has been empirically validated on paraphrase corpora.

The typology is displayed in Tables 1 and 2. It consists of a three-level typology of 24 paraphrase types (third column) grouped in 5 classes (first

[14]See Section 2 in this article and the appendices in the annotation guidelines (footnote 2) for a complete list of the consulted typologies.

column), two of them having two sub-classes each (second column).[15] In what follows, an overview of our typology is set out. In concrete, we describe (*i*) its scope, (*ii*) the type of units it classifies, (*iii*) its structure, (*iv*) and its types.

Scope of the typology. It is a general typology of paraphrasing in the sense that it comprehends the paraphrase phenomenon as a whole and covers all its possible manifestations, from elementary modifications like the INFLECTIONAL CHANGE type in Table 1 to deep reorganizations like SEMANTICS-BASED CHANGES in Table 2. Also, it covers paraphrases from the word to the discourse level. It should be noted that, since our typology relies on semantic content, pragmatic paraphrase fall outside our proposal (Section 3.1).

Unit of classification. The units classified according to our typology are what we call *atomic paraphrase phenomena* (*paraphrase phenomena* onwards), that is, autonomous paraphrase reorganizations consisting of a set of dependent linguistic mechanisms. The DERIVATIONAL CHANGE in Table 1, for example, comprises a change from a verb to an adjective form, as well as an involved structural modification. Among the dependent linguistic mechanisms, one of them is the trigger. In the previous example, it is the change of category or derivational change. As can be seen, paraphrase-type names stand for the linguistic mechanism triggering the paraphrase phenomenon.

Paraphrase phenomena can take place isolated or combined, giving rise to *complex paraphrase pairs*. In the pair containing a DERIVATIONAL CHANGE mentioned above, other paraphrase phenomena can be observed, such as a SAME-POLARITY SUBSTITUTION (or synonymy substitution) between *things* and *accounts*.

Typology structure: classes, subclasses, and types. Types are grouped in classes according to the nature of the trigger linguistic mechanism: (*i*) The morpholexicon-based change class comprises those types in which the paraphrase phenomenon is triggered at the morpholexicon level; (*ii*) the structure-based change class comprises those types that are the result of a different structural organization; and (*iii*) the semantic-based change class contains those types arising at the semantic level. An example of (*i*) are DERIVA-

[15]The typology was first presented (with some slight differences) in Barrón-Cedeño et al. (2013, to appear). The present article focusses on the nature and structure of the typology; Barrón-Cedeño et al. (2013, to appear), in contrast, focusses on the definition of each type.

Morpholexicon-based changes	**Morphology-based**	**Inflectional changes**	(a) it was with difficulty that the course of *streets* could be followed (b) You couldn't even follow the path of the *street*
		Modal-verb changes	(a) I [. . .] was still lost in conjectures who they *might be* (b) I was pondering who they *could be*
		Derivational changes	(a) I have heard many accounts of him [. . .] all *differing* from each other (b) I have heard many *different* things about him
	Lexicon-based	**Spelling changes**	(a) the foodservice pie business *doesn't* fit the company's long-term growth strategy (b) The foodservice pie business *does not* fit our long-term growth strategy
		Same-polarity substitutions	(a) *a teaspoonful of* vanilla (b) *very little* vanilla
		Synthetic/analytic substitutions	(a) A sequence of ideas (b) ideas
		Opposite-polarity substitutions	(a) Leicester [. . .] *failed* in both enterprises (b) he *did not succeed* in either case
		Converse substitutions	(a) the Geological Society of London in 1855 *awarded to* him the Wollaston medal (b) resulted in him *receiving* the Wollaston medal *from* the Geological Society in London in 1855
Structure-based changes	**Syntax-based**	**Diathesis alternations**	(a) the guide drew our attention to a gloomy little dungeon (b) ou[r] attention was drawn by our guide to a little dungeon
		Negation switching	(a) In order to move us, it needs *no* reference to any recognized original (b) One *does not* need to recognize a tangible object to be moved by its artistic representation
		Ellipsis	(a) In the scenes with Iago *he* equaled Salvini, yet did not in any one point surpass him (b) *He* equaled Salvini, in the scenes with Iago, but *he* did not in any point surpass him or imitate him
		Coordination changes	(a) It is estimated that he spent nearly £10,000 on these works. In addition he published a large number of separate papers (b) Altogether these works cost him almost £10,000 *and* he wrote a lot of small papers as well

Table 1: Paraphrase typology (1). Classes appear in the first column, subclasses in the second, and types in the third. Most of the examples come from the P4P corpus and also appear in Barrón-Cedeño et al. (2013, to appear). Spelling, punctuation, format, and paraphrase extremes are extracted from the MSRP-A corpus.

Structure-based changes	**(cont.)**	**Subordination-and-nesting changes**	(a) the Russian law, which limits the percentage of Jewish pupils in any school, barred his admission (b) the Russian law had limits for Jewish students so they barred his admission
	Discourse-based	**Punctuation changes**	(a) Swartz repaid it in full, *with interest,* according to his lawyer, Charles Stillman (b) Swartz fully repaid it *with interest*, according to his lawyer, Charles Stillman
		Direct/indirect-style alternations	(a) "She is mine," said the Great Spirit (b) The Great Spirit said that she is her[s]
		Sentence-modality changes	(a) The real question is, will it pay? will it please Theophilus P. Polk or vex Harriman Q. Kunz? (b) He do it just for earning money or to please Theophilus P. Polk or vex Hariman Q. Kunz
		Syntax/discourse-structure changes	(a) How he would stare! (b) He would surely stare!
Semantics-based changes			(a) The scenery was altogether more tropical (b) which added to the tropical appearance
Miscellaneous changes		**Change of format**	(a) fell 1.5% (b) fell 1.5 percent
		Change of order	(a) *First* we came to the tall palm trees (b) We got to some rather biggish palm trees *first*
		Addition/deletion	(a) *One day* she took a hot flat-iron, removed my clothes, and held it on my naked back until I howled with pain (b) As a proof of bed treatment, she took a hot flat-iron and put it on my back after removing my clothes
Paraphrase extremes		**Identical**	(a) But he added *group performance would improve in the second half of the year and beyond* (b) De Sole said in the results statement that *group performance would improve in the second half of the year and beyond*
		Entailment	(a) [...] it *was acquiring* the "intellectual property and technology assets" of GeCAD (b) [...] it *intends to acquire* the intellectual property and technology assets of Romanian antivirus firm GeCAD Software Srl
		Non-paraphrase	(a) The report was found Oct. 23, tucked inside *an old three-ring binder not related to the investigation* (b) The report was found last week tucked inside *a training manual that belonged to Hicks*

Table 2: Paraphrase typology (2)

TIONAL CHANGES, where the trigger consists of the change of category, which implies structural reorganizations. Regarding (*ii*) , a DIATHESIS ALTERNATION like the one in Table 1 involves a change of voice of the verb among others changes, but the trigger is syntactic. Finally, paraphrases in the semantics class (*iii*) are based on a different distribution of semantic content across the lexical units involving multiple and varied formal changes (Table 2).

There are two more classes in our typology: miscellaneous changes and paraphrase extremes (Table 2). The former comprises types not directly related to one single language level. The latter comprises those phenomena that are at the limits or outside the limits of paraphrasing (Section 3.1). Finally, the sub-classes follow the classical organization in formal linguistic levels from morphology to discourse and simply establish an intermediate grouping between some classes and their types.

Two main kinds of paraphrase structural reorganizations can be inferred from the previous explanation: those that are triggered by a lexical substitution (morpholexicon-based changes), and those that are not (structure-based changes). The idea of lexical trigger has its basis in the lexical projection rules put forward by Chomsky (1986) and their further reformulations.

This organization in classes and the idea of trigger determined the methodology applied to annotate the scope in Vila et al. (submitted-a).

The types.[16] Types in our typology correspond to general and contrastive categories: they stand for coarse-grained categories of paraphrase phenomena that are substantially different from each other, e.g., SAME-POLARITY SUBSTITUTION vs. PUNCTUATION CHANGE. Even types closer in nature clearly contrast. For example, the linguistic mechanisms involved in OPPOSITE-POLARITY and CONVERSE SUBSTITUTIONS are similar (both can involve a change in the order of the arguments); however, the linguistic mechanism triggering the paraphrase phenomenon (the opposite-polarity or converse lexical unit) makes them different.

An important consideration regarding the nomenclature used for the types has to be pointed out. Some paraphrase-type names refer to paraphrase relationships by default, e.g., all DERIVATIONAL CHANGES give rise to paraphrase relationships as changes of category do not affect the core

[16]See Barrón-Cedeño et al. (2013, to appear) for a detailed description and exemplification of each type.

meaning of the sentence. Other paraphrase-type names refer to linguistic mechanisms that do not necessarily give rise to paraphrases, e.g., INFLECTIONAL CHANGES may change the core meaning of the sentences. Therefore, cases like the INFLECTIONAL CHANGE type have to be understood as *meaning-preserving* changes in inflection, and not as changes in inflection as a whole (Section 3.1).

Each type is realized by a set of more fine-grained prototypes, that is, those patterns that characterize the linguistic mechanisms underlying the paraphrase. Defining a complete list of prototypes for each type is not the objective of this work. Nevertheless, while not aiming to be exhaustive, we exemplify prototypes taking SYNTHETIC/ANALYTIC SUBSTITUTIONS as an example.[17] In this case, we identified the five prototypes shown in Table 3: (*i*) compounding/decomposition, (*ii*) alternations affecting genitives and possessives, (*iii*) synthetic/analytic-superlative alternation, (*iv*) light/generic element addition/deletion, and (*v*) specifier addition/deletion.

Martin (1976) analyses in detail what he calls "double-negation" and "double inversion paraphrasing", which correspond roughly to our OPPOSITE-POLARITY and CONVERSE SUBSTITUTIONS. The equivalence rules he defines for French can be seen as a list of prototypes for these types. Barreiro (2008, pp. 73–81)'s typology involving support-verb constructions and, at a smaller scale, Peñas and Ovchinnikova (2012, pp. 399–400)'s noun-compound and genitive paraphrases can also be seen as potential lists of prototypes for the SYNTHETIC/ANALYTIC SUBSTITUTION type.

Types and prototypes differ in that types are stable and prototypes are an open class. Types represent general paraphrase phenomena covering paraphrasing as a whole. Their comprehensiveness has been tested through corpus annotation in two languages (English and Spanish). Prototypes, in contrast, are concrete linguistic mechanisms or patterns of realization for which a complete list is not necessarily provided in this work. They are more language dependent than types.

4. Conclusions and Future Work

This article has offered an overview on what has been said about paraphrasing in linguistics, how computational linguistics has used this knowledge

[17] Examples of prototypes for different types can be see in our annotations guidelines. See footnote 2.

Compounding/decomposition	(1) a.	wildlife television documentaries
	b.	television documentaries about wildlife
	(2) a.	chemical life-cycles
	b.	life-cycles for chemistry
	(3) a.	physiography
	b.	physical geography
Alternations affecting genitives and possessives	(1) a.	Tina's birthday
	b.	the birthday of Tina
	(2) a.	his reflection
	b.	the reflection of his own features
	(3) a.	the Met show
	b.	the Met's show
	(4) a.	Russia's Foreign Ministry
	b.	the Russian Foreign Ministry
Synthetic/analytic superlative alternation	(1) a.	smarter than everybody else
	b.	the smartest
Light/generic element addition/deletion	(1) a.	boast
	b.	speak boastfully
	(2) a.	cheerfully
	b.	in a cheerful way
Specifier addition/deletion	(1) a.	fog
	b.	wall of fog
	(2) a.	5
	b.	5 o'clock

Table 3: Prototypes for SYNTHETIC-ANALYTIC SUBSTITUTIONS.

as a base for its systems, and new insights on paraphrase characterization derived from computational-linguistics methods. This analysis has shown that, given the vague and multifaceted nature of paraphrasing, a precise and commonly accepted definition of the phenomenon does not exist. This has complicated paraphrase tasks in NLP on many occasions: "the difficulty when working with paraphrases lies on its own definition" (Herrera et al., 2007).

The aim of this article is to move forward in paraphrase characterization in order to provide NLP with more rigorous paraphrase knowledge. We addressed this problem from two directions. First, based on the idea that paraphrase boundaries are not fixed and depend on the task and objectives, we have presented three areas where boundary-paraphrases are placed. Second, paraphrase characterization has been addressed through the construction of a new paraphrase typology. Types in our typology are comprehensive, general, and stable. The prototypes they contain, in contrast, constitute an open and flexible group where new linguistic mechanisms can be described. This typology has been empirically validated through the annotation of more than 5,700 paraphrase pairs from three corpora that are different in nature and in two languages (Vila et al., submitted-a). Moreover, our typology proposal has already been tested in the automatic plagiarism detection field with promising results (Barrón-Cedeño et al., 2013, to appear).

Finally, this article opens a number of lines for future research, such as (*i*) further analyzing paraphrase boundaries with the aim of defining unseen borderline areas, (*ii*) the in-depth study of the idea of prototype and prototype definition, and (*iii*) seeing whether the most coarse-grained types in our typology (SYNTAX&DISCOURSE STRUCTURE and SEMANTICS-BASED CHANGES) accept a more fine-grained classification.

References

Androutsopoulos, I., Malakasiotis, P., 2010. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research 38 (1), 135–187.

Bagha, K. N., 2011. Generative semantics. English Language Teaching 4 (3), 223–231.

Bannard, C., Callison-Burch, C., 2005. Paraphrasing with bilingual parallel

corpora. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). Ann Arbor (MI), pp. 597–604.

Baroni, M., Lenci, A., 2010. Distributional memory: A general framework for corpus-based semantics. Computational Linguistics 36 (4), 673–721.

Barreiro, A., 2008. Make it Simple with Paraphrases. Automated Paraphrasing for Authoring Aids and Machine Translation. Ph.D. thesis, Universidade do Porto, Porto.

Barrón-Cedeño, A., Vila, M., Martí, M. A., Rosso, P., 2013, to appear. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics. DOI: 10.1162/COLI_a_00153.

Barzilay, R., 2003. Information Fusion for Multidocument Summarization: Paraphrasing and Generation. Ph.D. thesis, Columbia University, New York.

Barzilay, R., McKeown, K., 2001. Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001). Toulouse, pp. 50–57.

Barzilay, R., McKeown, K., Elhadad, M., 1999. Information fusion in the context of multi-document summarization. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL 1999). College Park (MD), pp. 550–557.

Bhagat, R., 2009. Learning Paraphrases from Text. Ph.D. thesis, University of Southern California, Los Angeles.

Boonthum, C., 2004. iSTART: Paraphase recognition. In: Proceedings of the ACL 2004 Student Research Workshop. Barcelona, pp. 31–36.

Boyer, M., Lapalme, G., 1985. Generating paraphrases from meaning-text semantic networks. Computational Intelligence 1 (1), 103–117.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge.

Chomsky, N., 1986. Knowledge of Language: Its Nature, Origin, and Use. Praeger Publishers, New York.

Clough, P., 2003. Measuring Text Reuse. Ph.D. thesis, University of Sheffield, Sheffield.

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

Culicover, P., 1968. Paraphrase generation and information retrieval from stored text. Mechanical Translation and Computational Linguistics 11 (1,2), 78–88.

Dolan, B., Quirk, C., Brockett, C., 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). Geneva, pp. 350–356.

Dolan, W. B., Brockett, C., 2005. Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005). Jeju Island, pp. 9–16.

Dorr, B. J., Green, R., Levin, L., Rambow, O., Farwell, D., Habash, N., Helmreich, S., Hovy, E., Miller, K. J., Mitamura, T., Reeder, F., Siddharthan, A., 2004. Semantic annotation and lexico-syntactic paraphrase. In: Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora. Lisbon, pp. 47–52.

Dras, M., 1999. Tree Adjoining Grammar and the Reluctant Paraphrasing of Text. Ph.D. thesis, Macquarie University, Sydney.

Dutrey, C., Bernhard, D., Bouamor, H., Max, A., 2011. Local modifications and paraphrases in Wikipedia's revision history. Procesamiento del Lenguaje Natural 46, 51–58.

Faigley, L., Witte, S., 1981. Analyzing revision. College Composition and Communication 32 (4), 400–414.

Fuchs, C., 1988. Paraphrases prédicatives et contraintes énonciatives. In: G. Bès, G., Fuchs, C. (Eds.), Lexique et paraphrase. No. 6 in Lexique. Presses Universitaires de Lille, Villeneuve d'Ascq, pp. 157–171.

Fuchs, C., 1994. Paraphrase et énonciation. Ophrys, Paris.

Fujita, A., 2005. Automatic Generation of Syntactically Well-Formed and Semantically Appropriate Paraphrases. Ph.D. thesis, Nara Institute of Science and Technology, Nara.

Halliday, M., 1994. An Introduction to Functional Grammar, 2nd Edition. Edward Arnold, New York.

Harris, Z., 1954. Distributional Structure. Word 10 (2–3), 146–162.

Harris, Z., 1957. Co-occurence and transformation in linguistic structure. Language 33 (3), 283–340.

Herrera, J., Peñas, A., Verdejo, F., 2007. Paraphrase extraction from validated question answering corpora in Spanish. Procesamiento del Lenguaje Natural 39, 37–44.

Hirst, G., 2003. Paraphrasing paraphrased. Keynote address for the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP 2003), Sapporo.

Hiż, H., 1964. The role of paraphrase in grammar. Monograph Series on Language and Linguistics 17, 97–104.

Honeck, R. P., 1971. A study of paraphrases. Journal of Verbal Learning and Verbal Behavior 10, 367–381.

Kotlerman, L., Dagan, I., Szpektor, I., Zhitomirsky-Geffet, M., 2010. Directional distributional similarity for lexical inference. Special Issue on Distributional Lexical Semantics. Natural Language Engineering 16 (4), 359–389.

Kouylekov, M., Magnini, B., 2005. Recognizing textual entailment with tree edit distance. In: Proceedings of the 1st PASCAL Recognising Textual Entailment Challenge (RTE I). pp. 17–20.

Kozlowski, R., McCoy, K. F., Shanker, V. K., 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In: Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP 2003). Sapporo, pp. 1–8.

Lakoff, G., 1971. On generative semantics. In: Steinberg, D. D., Jakobovits, L. A. (Eds.), Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology. Cambridge University Press, Cambridge, pp. 232–296.

Lareau, F., 2002. La synthèse automatique de paraphrases comme outil de vérification des dictionnaires et grammaires de type sens-texte. Master's thesis, Université de Montréal, Montreal.

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

Lin, D., Pantel, P., 2001. DIRT-Discovery of Inference Rules from Text. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2001). San Francisco (CA), pp. 323–328.

Madnani, N., Dorr, B. J., 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. Computational Linguistics 36 (3), 341–387.

Martin, R., 1976. Inférence, antonymie et paraphrase. Librarie C. Klincksieck, Paris.

McKeown, K., 1983. Paraphrasing questions using given and new information. American Journal of Computational Linguistics 9 (1).

Mel'čuk, I. A., 1992. Paraphrase et lexique: La théorie Sens-Texte et le Dictionnaire Explicatif et Combinatoire. In: Mel'čuk, I. A., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S. (Eds.), Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches Lexico-sémantiques III. Les Presses de l'Université de Montréal, Montreal, pp. 9–58.

Milićević, J., 2007a. La Paraphrase. Modélisation de la paraphrase langagière. Peter Lang, Bern.

Milićević, J., 2007b. Semantic equivalence rules in meaning-text paraphrasing. In: Wanner, L. (Ed.), Selected Lexical and Grammatical Issues in the Meaning-Text Theory. John Benjamins, Amsterdam/Philadelphia, pp. 267–296.

Peñas, A., Ovchinnikova, E., 2012. Unsupervised acquisition of axioms to paraphrase noun compounds and genitives. In: Gelbukh, A. F. (Ed.), CICLing 2012, Part I, LNCS 7181. Springer-Verlag, pp. 388–401.

Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P., 2010. An evaluation framework for plagiarism detection. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Beijing, pp. 997–1005.

Pustejovsky, J., 1995. The Generative Lexicon. MIT Press, Cambridge.

Recasens, M., Vila, M., 2010. On paraphrase and coreference. Computational Linguistics 36 (4), 639–647.

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., Mollá, D., 2003. Exploiting paraphrases in a question answering system. In: Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP 2003). Sapporo, pp. 25–32.

Romano, L., Kouylekov, M., Szpektor, I., Dagan, I., Lavelli, A., 2006. Investigating a generic paraphrase-based approach for relations extraction. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006). Trento, pp. 409–416.

Rus, V., McCarthy, P. M., C. Graesser, A., Danielle, S. M., 2009. Identification of sentence-to-sentence relations using a textual entailer. Research on Language and Computation 7 (2–4), 209–229.

Shimohata, M., 2004. Acquiring Paraphrases from Corpora and Its Application to Machine Translation. Ph.D. thesis, Nara Institute of Science and Technology, Nara.

Shinyama, Y., Sekine, S., Sudo, K., 2002. Automatic paraphrase acquisition from news articles. In: Proceedings of the 2nd International Conference on Human Language Technology Research (HLT 2002). San Francisco (CA), pp. 313–318.

Smaby, R. M., 1971. Paraphrase Grammars. Vol. 2 of Formal Linguistics Series. D. Reidel Publishing Company, Dordrecht.

Vila, M., Bertran, M., Martí, M. A., Rodríguez, H., submitted-a. Corpus annotation with paraphrase types. New annotation scheme and inter-annotator agreement measures.

Vila, M., Rodríguez, H., Martí, M. A., submitted-b. Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus.

Žolkovskij, A., Mel'čuk, I., 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza. Naučno-texničeskaja informacija 5, 23–28.

Wintner, S., 2009. What science underlies natural language engineering? Computational Linguistics 35 (4), 641–644.

2.2 Paraphrase and Coreference

Marta Recasens (Universitat de Barcelona)
Marta Vila (Universitat de Barcelona)

(2010)

On paraphrase and coreference

Computational Linguistics, 36(4):639–647

Journal URL `http://www.mitpressjournals.org/loi/coli`
Impact Factor 0.721

Abstract By providing a better understanding of paraphrase and coreference in terms of similarities and differences in their linguistic nature, this article delimits what the focus of paraphrase extraction and coreference resolution tasks should be, and to what extent they can help each other. We argue for the relevance of this discussion to Natural Language Processing.

Squibs

On Paraphrase and Coreference

Marta Recasens*
University of Barcelona

Marta Vila**
University of Barcelona

By providing a better understanding of paraphrase and coreference in terms of similarities and differences in their linguistic nature, this article delimits what the focus of paraphrase extraction and coreference resolution tasks should be, and to what extent they can help each other. We argue for the relevance of this discussion to Natural Language Processing.

1. Introduction

Paraphrase extraction[1] and coreference resolution have applications in Question Answering, Information Extraction, Machine Translation, and so forth. Paraphrase pairs might be coreferential, and coreference relations are sometimes paraphrases. The two overlap considerably (Hirst 1981), but their definitions make them significantly different in essence: Paraphrasing concerns meaning, whereas coreference is about discourse referents. Thus, they do not always coincide. In the following example, *b* and *d* are both coreferent and paraphrastic, whereas *a*, *c*, *e*, *f*, and *h* are coreferent but not paraphrastic, and *g* and *i* are paraphrastic but not coreferent.

(1) [Tony]$_a$ went to see [the ophthalmologist]$_b$ and got [his]$_c$ eyes checked. [The eye doctor]$_d$ told [him]$_e$ that [his]$_f$ [cataracts]$_g$ were getting worse. [His]$_h$ mother also suffered from [cloudy vision]$_i$.

The discourse model built for Example (1) contains six entities (i.e., Tony, the eye doctor, Tony's eyes, Tony's cataracts, Tony's mother, cataracts). Because *a*, *c*, *e*, *f*, and *h* all point to Tony, we say that they are coreferent. In contrast, in paraphrasing, we do not need to build a discourse entity to state that *g* and *i* are paraphrase pairs; we restrict ourselves to semantic content and this is why we check for sameness of meaning between *cataracts* and *cloudy vision* alone, regardless of whether they are a referential unit in a discourse. Despite the differences, it is possible for paraphrasing and coreference to co-occur, as in the case of *b* and *d*.

NLP components dealing with paraphrasing and coreference seem to have great potential to improve understanding and generation systems. As a result, they have been the focus of a large amount of work in the past couple of decades (see the surveys by

* CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: mrecasens@ub.edu.
** CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: marta.vila@ub.edu.

1 Recognition, extraction, and generation are all paraphrase-related tasks. We will center ourselves on paraphrase extraction, as this is the task in which paraphrase and coreference resolution mainly overlap.

Submission received: 3 March 2010; accepted for publication: 1 June 2010.

Androutsopoulos and Malakasiotis [2010], Madnani and Dorr [2010], Ng [2010], and Poesio and Versley [2009]). Before computational linguistics, coreference had not been studied on its own from a purely linguistic perspective but was indirectly mentioned in the study of pronouns. Although there have been some linguistic works that consider paraphrasing, they do not fully respond to the needs of paraphrasing from a computational perspective.

This article discusses the similarities between paraphrase and coreference in order to point out the distinguishing factors that make paraphrase extraction and coreference resolution two separate yet related tasks. This is illustrated with examples extracted/adapted from different sources (Dras 1999; Doddington et al. 2004; Dolan, Brockett, and Quirk 2005; Recasens and Martí 2010; Vila et al. 2010) and our own. Apart from providing a better understanding of these tasks, we point out ways in which they can mutually benefit, which can shed light on future research.

2. Converging and Diverging Points

This section explores the overlapping relationship between paraphrase and coreference, highlighting the most relevant aspects that they have in common as well as those that distinguish them. They are both sameness relations (Section 2.2), but one is between meanings and the other between referents (Section 2.1). In terms of linguistic units, coreference is mainly restricted to noun phrases (NPs), whereas paraphrasing goes beyond and includes word-, phrase- and sentence-level expressions (Section 2.3). One final diverging point is the role they (might) play in discourse (Section 2.4).

2.1 Meaning and Reference

The two dimensions that are the focus of paraphrasing and coreference are meaning and reference, respectively. Traditionally, paraphrase is defined as the relation between two expressions that have the same *meaning* (i.e., they evoke the same mental concept), whereas coreference is defined as the relation between two expressions that have the same *referent* in the discourse (i.e., they point to the same entity). We follow Karttunen (1976) and talk of "discourse referents" instead of "real-world referents."

In Table 1, the italicized pairs in cells (1,1) and (2,1) are both paraphrastic but they only corefer in (1,1). We cannot decide on (non-)coreference in (2,1) as we need a discourse to first assign a referent. In contrast, we can make paraphrasing judgments

Table 1
Paraphrase–coreference matrix.

		Paraphrase ✔	Paraphrase ✗
Coreference	✔	(1,1) Tony went to see *the ophthalmologist* and got his eyes checked. *The eye doctor* told him …	(1,2) *Tony* went to see the ophthalmologist and got *his* eyes checked.
	✗	(2,1) *ophthalmologist* *eye doctor*	(2,2) *His cataracts* were getting worse. *His mother* also suffered from cloudy vision.

without taking discourse into consideration. Pairs like the one in cell (1,2) are only coreferent but not paraphrases because the proper noun *Tony* and the pronoun *his* have reference but no meaning. Lastly, neither phenomenon is observed in cell (2,2).

2.2 Sameness

Paraphrasing and coreference are usually defined as sameness relations: Two expressions that have the *same meaning* are paraphrastic, and two expressions that refer to the *same entity* in a discourse are coreferent. The concept of *sameness* is usually taken for granted and left unexplained, but establishing sameness is not straightforward. A strict interpretation of the concept makes sameness relations only possible in logic and mathematics, whereas a sloppy interpretation makes the definition too vague. In paraphrasing, if the loss of *at the city* in Example (2b) is not considered to be relevant, Examples (2a) and (2b) are paraphrases; but if it is considered to be relevant, then they are not. It depends on where we draw the boundaries of what is accepted as the "same" meaning.

(2) a. The waterlogged conditions that ruled out play yesterday still prevailed *at the city* this morning.

b. The waterlogged conditions that ruled out play yesterday still prevailed this morning.

(3) On homecoming night *Postville* feels like Hometown, USA . . . For those who prefer *the old Postville*, Mayor John Hyman has a simple answer.

Similarly, with respect to coreference (3), whether *Postville* and *the old Postville* in Example 3 are or are not the same entity depends on the granularity of the discourse. On a sloppy reading, one can assume that because Postville refers to the same spatial coordinates, it is the same town. On a strict reading, in contrast, drawing a distinction between the town as it was at two different moments in time results in two different entities: the old Postville versus the present-day Postville. They are not the same in that features have changed from the former to the latter.

The concept of sameness in paraphrasing has been questioned on many occasions. If we understood "same meaning" in the strictest sense, a large number of paraphrases would be ruled out. Thus, some authors argue for a looser definition of paraphrasing. Bhagat (2009), for instance, talks about "quasi-paraphrases" as "sentences or phrases that convey approximately the same meaning." Milićević (2007) draws a distinction between "exact" and "approximate" paraphrases. Finally, Fuchs (1994) prefers to use the notion of "equivalence" to "identity" on the grounds that the former allows for the existence of some semantic differences between the paraphrase pairs. The concept of identity in coreference, however, has hardly been questioned, as prototypical examples appear to be straightforward (e.g., *Barack Obama* and *Obama* and *he*). Only recently have Recasens, Hovy, and Martí (2010) pointed out the need for talking about "near-identity" relations in order to account for cases such as Example (3), proposing a typology of such relations.

2.3 Linguistic Units

Another axis of comparison between paraphrase and coreference concerns the types of linguistic units involved in each relation. Paraphrase can hold between different

linguistic units, from morphemes to full texts, although the most attention has been paid to word-level paraphrase (*kid* and *child* in Example (4)), phrase-level paraphrase (*cried* and *burst into tears* in Example (4)), and sentence-level paraphrase (the two sentences in Example (4)).

(4) a. The kid cried.

b. The child burst into tears.

In contrast, coreference is more restricted in that the majority of relations occur at the phrasal level, especially between NPs. This explains why this has been the largest focus so far, although prepositional and adverbial phrases are also possible yet less frequent, as well as clauses or sentences. Coreference relations occur indistinctively between pronouns, proper nouns, and full NPs that are *referential*, namely, that have discourse referents. For this reason, pleonastic pronouns, nominal predicates, and appositives cannot enter into coreference relations. The first do not refer to any entity but are syntactically required; the last two express properties of an entity rather than introduce a new one. But this is an issue ignored by the corpora annotated for the MUC and ACE programs (Hirschman and Chinchor 1997; Doddington et al. 2004), hence the criticism by van Deemter and Kibble (2000).

In the case of paraphrasing, it is linguistic expressions that lack meaning (i.e., pronouns and proper nouns) that should not be treated as members of a paraphrase pair on their own (Example (5a)) because paraphrase is only possible between meaningful units. This issue, however, takes on another dimension when seen at the sentence level. The sentences in Example (5b) can be said to be paraphrases because they themselves contain the antecedent of the pronouns *I* and *he*.

(5) a. (i) A. Jiménez
(ii) I

b. (i) The Atlético de Madrid goalkeeper, A. Jiménez, yesterday realized one of his dreams by defeating Barcelona: "I had never beaten Barcelona."
(ii) The Atlético de Madrid goalkeeper, A. Jiménez, yesterday realized one of his dreams by defeating Barcelona, and said that he had never beaten Barcelona.

In Example (5b), *A. Jiménez* and *I*/*he* continue not being paraphrastic. Polysemic, underspecified, and metaphoric words show a slightly different behavior. It is not possible to establish paraphrase between them when they are deprived of context (Callison-Burch 2007, Chapter 4). In Example (6a), *police officers* could be patrol police officers, and *investigators* could be university researchers. However, once they are embedded in a disambiguating context that fills them semantically, as in Example (6b), then paraphrase can be established between *police officers* and *investigators*.

(6) a. (i) Police officers
(ii) Investigators

b. (i) *Police officers* searched 11 stores in Barcelona.
(ii) The *investigators* conducted numerous interviews with the victim.

As a final remark, and in accordance with the approach by Fuchs (1994), we consider Example (7)–like paraphrases that Fujita (2005) and Milićević (2007) call, respectively,

"referential" and "cognitive" to be best treated as coreference rather than paraphrase, because they only rely on referential identity in a discourse.

(7) a. They got married *last year*.

b. They got married *in 2004*.

2.4 Discourse Function

A further difference between paraphrasing and coreference concerns their degree of dependency on discourse. Given that coreference establishes sameness relations between the entities that populate a discourse (i.e., discourse referents), it is a linguistic phenomenon whose dependency on discourse is much stronger than paraphrasing. Thus, the latter can be approached from a discursive or a non-discursive perspective, which in turn allows for a distinction between reformulative paraphrasing (Example (8)) and non-reformulative paraphrasing (Example (9)).

(8) Speaker 1: Then they also diagnosed *a hemolytic–uremic syndrome*.
Speaker 2: What's that?
Speaker 1: *Renal insufficiency, in the kidneys*.

(9) a. X wrote Y.

b. X is the author of Y.

Reformulative paraphrasing occurs in a reformulation context when a rewording of a previously expressed content is added for discursive reasons, such as emphasis, correction, or clarification. Non-reformulative paraphrasing does not consider the role that paraphrasing plays in discourse. Reformulative paraphrase pairs have to be extracted from a single piece of discourse; non-reformulative paraphrase pairs can be extracted—each member of the pair on its own—from different discourse pieces. The reformulation in the third utterance in Example (8) gives an explanation in a language less technical than that in the first utterance; whereas Examples (9a) and (9b) are simply two alternative ways of expressing an authorship relation.

The strong discourse dependency of coreference explains the major role it plays in terms of cohesion. Being such a cohesive device, it follows that intra-document coreference, which takes place within a single discourse unit (or across a collection of documents linked by topic), is the most primary. Cross-document coreference, on the other hand, constitutes a task on its own in NLP but falls beyond the scope of linguistic coreference due to the lack of a common universe of discourse. The assumption behind cross-document coreference is that there is an underlying global discourse that enables various documents to be treated as a single macro-document.

Despite the differences, the discourse function of reformulative paraphrasing brings it close to coreference in the sense that they both contribute to the cohesion and development of discourse.

3. Mutual Benefits

Both paraphrase extraction and coreference resolution are complex tasks far from being solved at present, and we believe that there could be improvements in performance

if researchers on each side paid attention to the others. The similarities (i.e., relations of sameness, relations between NPs) allow for mutual collaboration, whereas the differences (i.e., focus on either meaning or reference) allow for resorting to either paraphrase or coreference to solve the other. In general, the greatest benefits come for cases in which either paraphrase or coreference are especially difficult to detect automatically. More specifically, we see direct mutual benefits when both phenomena occur either in the same expression or in neighboring expressions.

For pairs of linguistic expressions that show both relations, we can hypothesize paraphrasing relationships between NPs for which coreference is easier to detect. For instance, coreference between the two NPs in Example (10) is very likely given that they have the same head, head match being one of the most successful features in coreference resolution (Haghighi and Klein 2009). In contrast, deciding on paraphrase would be hard due to the difficulty of matching the modifiers of the two NPs.

(10) a. The director of a multinational with huge profits.

b. The director of a solvent company with headquarters in many countries.

In the opposite direction, we can hypothesize coreference links between NPs for which paraphrasing can be recognized with considerable ease (Example (11)). Light elements (e.g., *fact*), for instance, are normally taken into account in paraphrasing—but not in coreference resolution—as their addition or deletion does not involve a significant change in meaning.

(11) a. The creation of a company.

b. The fact of creating a company.

By neighboring expressions, we mean two parallel structures each containing a coreferent mention of the same entity next to a member of the same paraphrase pair. Note that the coreferent expressions in the following examples are printed in *italics* and the paraphrase units are printed in **bold**. If a resolution module identifies the coreferent pairs in Example (12), then these can function as two anchor points, *X* and *Y*, to infer that the text between them is paraphrastic: *X complained today before Y*, and *X is formulating the corresponding complaint to Y*.

(12) a. *Argentina*$_X$ **complained today before** *the British Government*$_Y$ about the violation of the air space of this South American country.

b. *This Chancellorship*$_X$ **is formulating the corresponding complaint to** *the British Government*$_Y$ for this violation of the Argentinian air space.

Some authors have already used coreference resolution in their paraphrasing systems in a similar way to the examples herein. Shinyama and Sekine (2003) benefit from the fact that a single event can be reported in more than one newspaper article in different ways, keeping certain kinds of NPs such as names, dates, and numbers unchanged. Thus, these can behave as anchor points for paraphrase extraction. Their system uses coreference resolution to find anchors which refer to the same entity.

Conversely, knowing that a stretch of text next to an NP paraphrases another stretch of text next to another NP helps to identify a coreference link between the two NPs, as shown by Example (13), where two diction verbs are easily detected as a paraphrase and thus their subjects can be hypothesized to corefer. If the paraphrase system

identifies the mapping between the indirect speech in Example (13a) and the direct speech in Example (13b), the coreference relation between the subjects is corroborated. Another difficult coreference link that can be detected with the help of paraphrasing is Example (14): If the predicates are recognized as paraphrases, then the subjects are likely to corefer.

(13) a. *The trainer of the Cuban athlete Sotomayor* **said** that the world record holder is in a fit state to win the Games in Sydney.

b. "The record holder is in a fit state to win the Olympic Games," **explained** *De la Torre*.

(14) a. *Police officers* **searched 11 stores in Barcelona**.

b. *The investigators* **carried out 11 searches in stores in the center of Barcelona**.

Taking this idea one step further, new coreference resolution strategies can be developed with the aid of shallow paraphrasing techniques. A two-step process for coreference resolution might consist of hypothesizing first sentence-level paraphrases via *n*-gram or named-entity overlapping, aligning phrases that are (possible) paraphrases, and hypothesizing that they corefer. Second, a coreference module can act as a filter and provide a second classification. Such a procedure could be successful for the cases exemplified in Examples (12) to (14).

This strategy reverses the tacit assumption that coreference is solved before sentence-level paraphrasing. Meaning alone does not make it possible to state that the two pairs in Example (5b), repeated in Example (15), or the two pairs in Example (16) are paraphrases without first solving the coreference relations.

(15) a. *The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona: "*I* had never beaten Barcelona."

b. *The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona, and said that *he* had never beaten Barcelona.

(16) a. Secretary of State Colin Powell last week ruled out *a non-aggression treaty*.

b. But Secretary of State Colin Powell brushed off *this possibility*.

However, cooperative work between paraphrasing and coreference is not always possible, and it is harder if neither of the two can be detected by means of widely used strategies. In other cases, cooperation can even be misleading. In Example (17), the two bold phrases are paraphrases, but their subjects do not corefer. The detection of words like *another* (Example (17b)) gives a key to help to prevent this kind of error.

(17) a. A total of 26 Cuban citizens remain in the police station of the airport of Barajas **after requesting political asylum**.

b. Another three Cubans **requested political asylum**.

On the basis of these various examples, we claim that a full understanding of both the similarities and disparities will enable fruitful collaboration between researchers working on paraphrasing and those working on coreference. Even more importantly,

our main claim is that such an understanding about the fundamental linguistic issues is a prerequisite for building paraphrase and coreference systems not lacking in linguistic rigor. In brief, we call for the return of linguistics to paraphrasing and coreference automatic applications, as well as to NLP in general, adhering to the call by Wintner (2009: 643), who cites examples that demonstrate "what computational linguistics can achieve when it is backed up and informed by linguistic theory" (page 643).

Acknowledgments

We are grateful to Eduard Hovy, M. Antònia Martí, Horacio Rodríguez, and Mariona Taulé for their helpful advice as experienced researchers. We would also like to express our gratitude to the three anonymous reviewers for their suggestions to improve this article.

This work was partly supported by FPU Grants AP2006-00994 and AP2008-02185 from the Spanish Ministry of Education, and Project TEXT-MESS 2.0 (TIN2009-13391-C04-04).

References

Androutsopoulos, Ion and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Bhagat, Rahul. 2009. *Learning Paraphrases from Text*. Ph.D. thesis, University of Southern California, Los Angeles, CA.

Callison-Burch, Chris. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh.

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program—Tasks, data, and evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon.

Dolan, Bill, Chris Brockett, and Chris Quirk. 2005. README file included in the Microsoft Research Paraphrase Corpus, March, Redmond, WA.

Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Sydney.

Fuchs, Catherine. 1994. *Paraphrase et énonciation. Modélisation de la paraphrase langagière*. Ophrys, Paris.

Fujita, Atsushi. 2005. *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Ph.D. thesis, Nara Institute of Science and Technology, Ikoma, Nara.

Haghighi, Aria and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1152–1161, Singapore.

Hirschman, Lynette and Nancy Chinchor. 1997. MUC-7 Coreference task definition–Version 3.0. In *Proceedings of the Message Understanding Conference-7 (MUC-7)*, Washington, DC.

Hirst, Graeme J. 1981. *Anaphora in Natural Language Understanding: A Survey*. Springer-Verlag, Berlin.

Karttunen, Lauri. 1976. Discourse referents. In J. McCawley, editor, *Syntax and Semantics*, volume 7. Academic Press, New York, pages 363–385.

Madnani, Nitin and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36, pages 341–387.

Milićević, Jasmina. 2007. *La paraphrase*. Peter Lang, Berne.

Ng, Vincent. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, pages 1396–1411.

Poesio, Massimo and Yannick Versley. 2009. Computational models for the interpretation of anaphora: A survey. Notes from the ACL-2009 Tutorial on State-of-the-art NLP Approaches to Coreference Resolution, Singapore.

Recasens, Marta, Eduard Hovy, and M. Antònia Martí. 2010. A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 149–156, Valletta.

Recasens, Marta and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated

corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345. doi: 10.1007/s10579-009-9108-x.

Shinyama, Yusuke and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the ACL 2nd International Workshop on Paraphrasing (IWP 2003)*, pages 65–71, Sapporo.

van Deemter, Kees and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

Vila, Marta, Santiago González, M. Antònia Martí, Joaquim Llisterri, and M. Jesús Machuca. 2010. ClInt: A bilingual Spanish-Catalan spoken corpus of clinical interviews. *Procesamiento del Lenguaje Natural*, 45, 105–111.

Wintner, Shuly. 2009. What science underlies Natural Language Engineering? *Computational Linguistics*, 35(4):641–644.

2.3 Paraphrase Representation

Marta Vila (Universitat de Barcelona)
Mark Dras (Macquarie University)

(2012)

Tree edit distance as a baseline approach for paraphrase representation

Procesamiento del Lenguaje Natural, 48:89–95

Journal URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln`

Abstract Finding an adequate paraphrase representation formalism is a challenging issue in Natural Language Processing. In this paper, we analyse the performance of Tree Edit Distance as a paraphrase representation baseline. Our experiments using Edit Distance Textual Entailment Suite show that, as Tree Edit Distance consists of a purely syntactic approach, paraphrase alternations not based on structural reorganizations do not find an adequate representation. They also show that there is much scope for better modelling of the way trees are aligned.

Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 89-95 recibido 27-01-12 revisado 09-03-12 aceptado 18-03-12

Tree Edit Distance as a Baseline Approach for Paraphrase Representation*

Distancia de edición de árboles como caso base para la representación de la paráfrasis

Marta Vila
Universitat de Barcelona
Gran Via 585
08007 Barcelona
marta.vila@ub.edu

Mark Dras
Macquarie University
Herring Rd, North Ryde
NSW 2109
mark.dras@mq.edu.au

Resumen: Encontrar un formalismo adecuado para representar la paráfrasis constituye un reto para el Procesamiento del Lenguaje Natural. En este artículo, se analiza la distancia de edición de árboles como caso base para dicha representación. Los experimentos realizados utilizando Edit Distance Textual Entailment Suite muestran que, dado que la distancia de edición de árboles es una aproximación puramente sintáctica, las paráfrasis no basadas en reorganizaciones estructurales no encuentran una representación adecuada. Asimismo, muestran la necesidad de mejorar la forma como los árboles se alinean.
Palabras clave: Paráfrasis, distancia de edición de árboles, alineación de árboles.

Abstract: Finding an adequate paraphrase representation formalism is a challenging issue in Natural Language Processing. In this paper, we analyse the performance of Tree Edit Distance as a paraphrase representation baseline. Our experiments using Edit Distance Textual Entailment Suite show that, as Tree Edit Distance consists of a purely syntactic approach, paraphrase alternations not based on structural reorganizations do not find an adequate representation. They also show that there is much scope for better modelling of the way trees are aligned.
Keywords: Paraphrasing, tree edit distance, tree alignment.

1 Introduction

In paraphrasing, different wordings express same meaning. For example, an active/passive voice alternation occurs in the paraphrase pair in (1).[1]

(1) a. The guide drew our attention to a [...] dungeon
b. Our attention was drawn by [the] guide to a [...] dungeon

String pairs like the one in (1) are obviously not very general. Formally representing paraphrasing, i.e., transforming paraphrase strings into paraphrase patterns by enriching them with linguistic knowledge and, at the same time, making them more general, makes paraphrase knowledge more efficient and scalable to various Natural Language Processing (NLP) tasks and applications. In (2), a representation of the active/passive alternation in (1) along the lines of the original Transformational Grammar representation of Chomsky (1957) can be observed. All linguistic units but prepositions have been substituted by the corresponding morpho-syntactic categories, which are mapped from one member of the pair to the other.

(2) a. NP_1 V_{active} NP_2 *to* NP_3.
b. NP_2 $V_{passive}$ *by* NP_1 *to* NP_3.

* We are grateful to M. Antònia Martí and Horacio Rodríguez for their helpful advice as experienced researchers. We would also like to express our gratitude to the anonymous reviewers for their suggestions to improve this article. This research work is supported by the TEXT-Knowledge 2.0 (TIN2009-13391-C04-04) MICINN project. Also, the work of the first author is financed by the FPU AP2008-02185 MECD grant and, within it, the funding for a 6-month stay in Macquarie University.

[1] Example from the P4P corpus. `http://clic.ub.edu/corpus/en/p4p`.

ISSN 1135-5948

Paraphrasing is a complex phenomenon, where many linguistic mechanisms—shallow or deep, formal or conceptual—can be displayed. Contrary to (1), in the example in (3),[2] a formal structural mapping between the two members of the pair in italics cannot be established.

(3) a. Michael Mitchell [...] *did not answer his phone* Wednesday afternoon
b. Michael Mitchell [...] *was not available for comment*

In this paper, we want to capture two things with respect to paraphrase representation. Primarily, we are interested in how well a representation can capture the mapping of structures (typically as instantiated by tree alignment) that occur in paraphrasing. By way of illustration, if the structural representation of (4-a) maps to the structural representation of (4-b),[3] and, in the former, *estimated* is the head of the dependent noun *people*, while the reverse is true in the latter (i.e., *people* is the head of *estimated*), the paraphrase representation must be able to capture that. That is, the trees should be aligned in a way that maps corresponding nodes to each other.

(4) a. It is estimated that 200,000 people are left behind
b. An estimated 200,000 people left behind

Secondarily, we want a representation approach capable of dealing with paraphrase complexity at a reasonable computational cost. The intrinsic variety of paraphrasing demands a highly expressive representation. Nevertheless, high expressive capacity generally entails low computational efficiency, as, in general, there is a trade-off between the two. Thus, finding an adequate balance is needed.

Given all of this, our first objective is to build a paraphrase representation baseline (in terms of expressiveness) to evaluate its level of coverage of the paraphrase phenomenon and the potential drawbacks that it presents in alignment. This baseline will be the basis for a further analysis of more expressive approaches.

Wu (2010) presents a general framework for considering alignment, including tree alignment, which is useful for considering a range of possible representations. In this paper, we work with Tree Edit Distance (TED). Given two dependency trees, TED allows for establishing the distance between them according to the number of edit operations on tree fragments (insertion, deletion, substitution) required to go from one to the other. It can thus convert any tree into any other arbitrary tree (e.g., by deleting all nodes and then inserting new nodes). We take these two dependency trees and the edit operations mapping between them as a paraphrase representation baseline, and investigate its performance regarding paraphrasing and how well the mapping of items that should be aligned is preserved.

We use the Edit Distance Textual Entailment Suite (EDITS),[4] a software package aimed at Recognizing Textual Entailment (RTE) relations between two portions of text, which embeds an implementation of the TED algorithm described in Shasha and Zhang (1990). Given that paraphrasing can be considered a bidirectional entailment, we hypothesize that such a package can also be used in the paraphrase domain. As will be seen, we do not use EDITS as a textual entailment or paraphrase classifier, and we focus instead on the edit operations between trees it provides.

Using EDITS, we represented a set of paraphrase pairs with the TED approach. First, we analyzed the coverage of this approach within the paraphrase phenomenon and the drawbacks that presents. We then proceeded to our main question of interest, how well the structures are mapped.

In what follows, after presenting a brief state of the art on paraphrase representation (Section 2), we set out our experiments and results (Section 3). Conclusions and future lines of research appear in Sections 4 and 5, respectively.

2 State of the Art

Choosing a paraphrase representation formalism implies seeking balance between expressivity and computational cost. The least

[2] Example from the MSRP corpus. `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/`.

[3] Example from Wan (2010).

[4] `http://edits.fbk.eu/`

expressive way consists in simply stating the paraphrase nature of a pair of strings. An example of this approach is the Microsoft Research Paraphrase Corpus (MSRP) (Dolan and Brockett, 2005),[5] which consists of a set of sentence pairs with a yes/no paraphrase judgement.

Expressivity may be increased transforming pairs of strings into pairs of regular expression patterns, which can be synchronized using Finite State Transducers (FST) or their probabilistic version, namely, Stochastic Finite State Transducers (SFST). Casacuberta and Vidal (2007), for instance, learn finite-state models for Machine Translation (MT). Drawbacks facing FSTs are that they are too constrained to model paraphrase mapping complexity because they only compare regular expression strings (not deeper representations), and that their expressive power is limited to Regular Grammars (RG).

A further step in expressivity, but with a higher computational cost, may be the use of the bilingual version of Context Free Grammars, namely, Synchronous Context Free Grammars (SCFG) (Aho and Ullman, 1969), which simultaneously produce strings in two languages. Dekai Wu, starting in Wu (1997), proposed several subclasses of SCFGs pruning their expressivity for reducing their computational cost, e.g., Inversion Transduction Grammars (ITG), Bracketting ITG (BITG), Linear ITG (LITG), together with their probabilistic versions. Some of these formalisms are proved to be expressive enough for learning and parsing.

A richer level of expresivity may be found in the family of the tree transducers, within the framework of Mildly Context Sensitive Grammars (MCSG). Some examples are Quasi-Synchronous Grammars (QSG), which were proposed by Smith and Eisner (2006); Synchronous Tree Adjoining Grammars (STAG), which Dras (1999) applies to syntactic paraphrasing; or Synchronous Tree-Substitution Grammar (STSG) (Eisner, 2003), a restricted version of STAGs. All these proposals have been mainly applied to MT (translation between different languages), but they may also be applied to paraphrasing (understood as translation within the same language).

TED is the approach chosen in this paper as a baseline for paraphrase representation. There is work in the related field of RTE (Kouylekov and Magnini, 2005) and also in paraphrasing (Heilman and Smith, 2010), but there all that is of interest is optimising the mapping between strings, not between structures. That is, how the trees are transformed is unimportant in those applications, as long as the transformation of the string is carried out with a minimum cost. In contrast, our interest is precisely on the tree mapping.

3 Experiments and Results

We performed two different experiments aiming at the analysis of TED performance for paraphrase representation (Section 3.1) and the analysis of the problem of tree alignments (Section 3.2). In both of them, we used EDITS.

EDITS presents a modular structure, whose main components are: algorithms used to compute a distance score; cost schemes defining the cost for each edit operation; a cost optimizer, which adapts cost schemes to specific datasets; and rules providing linguistic knowledge. Using as a starting point these modules, plus a training corpus with sentence pairs annotated with yes/no textual entailment or paraphase judgement, EDITS builds a model, which will be subsequently used to classify unseen sentence pairs.

The EDITS output which we are interested in (considering we have selected, among the possible algorithms, the TED one) consists of a file with the dependency trees of each member of the pair,[6] and a file with the edit operations between them, a score and an entailment/paraphrase judgement. Our focus is on how the trees are transformed, i.e., the trees and edit operations, not on the final score nor classification.

3.1 The Performance of TED for Paraphrase Representation

Our objective here is analyzing a set of paraphrase pairs with EDITS to see the coverage of TED regarding the paraphrase phenomenon and the potential problems arisen. The corpus used is the MSRP,[7] because it is a reference corpus fulfilling EDITS requirements: it contains a large quantity of data

[5] See note 2.

[6] The Stanford parser is the one used by EDITS. `http://nlp.stanford.edu/software/lex-parser.shtml`

[7] See Section 2 for references.

(5,800 English sentence pairs) with manual annotations indicating whether they are paraphrases (67%) or not (33%). It is already divided into training (70%) and test (30%).

We carried out a series of experiments with several EDITS configurations (always using the TED algorithm). Two considerations arise from the analysis of the output files. First, as it consists of a purely syntactic representation, some lexical and morphological paraphrases, and especially the semantic ones,[8] do not find an adequate representation. Moreover, paraphrase mechanisms based on pure changes of order are not reflected in the output, as word order is generally not taken into account in dependency analysis.

Second, the tree alignment is, on many occasions, inadequate. In Figure 1 on the left, we see how the 'technologies' node, present in both trees, is not aligned, because it does not occupy the same (or similar) position in the tree.[9] The expected alignment from the paraphrase point of view appears at the right hand representation. Such alignment problems do not have a straightforward solution in the EDITS framework, because they arise from the TED algorithm itself: it is derived from the image recognition literature and it tends to match structure more than content. Once this problem was identified, the next step was to quantify it to evaluate its scope.

3.2 The Tree Alignment Problem

Here we reach our main question of interest: the alignment problem. We compare EDITS with gold standard alignments.

We use Cohn, Callison-Burch, and Lapata (2008)'s paraphrase corpus,[10] as it contains, among other data, 370 positive pairs from the MSRP corpus with manual word or phrase alignments by two annotators (A and C). These annotations constitute the gold standard in our experiments.

In order to be able to carry out the mapping equitatively, we analyzed this same set with EDITS. As the number of pairs is small, we performed 5-fold cross validation.

[8] See Vila, Martí, and Rodríguez (2011) for the paraphrase typology we are referring to. Example (3) above is an example of a semantics based paraphrase.

[9] We understand the nodes connected with the substitution operation as aligned nodes, and the deleted or inserted nodes as non-aligned nodes.

[10] http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html

Figure 2 shows an example of gold standard alignment. Horizontal and vertical axes show the sentences of the aligned paraphrase pair. Shadowed squares represent those alignments that the annotators considered Sure (S), in black, and Possible (P), in grey. As can be seen, some of the words remain unaligned ("then" and "Texas" in the vertical sentence) because they do not have a counterpart, and others are aligned in block ("at the FAA" and "lost-aircraft" in the vertical sentence) as there is not a one-to-one word alignment. In Figure 2, we can also see EDITS alignments for the same pair of sentences (E). As can be seen, not all gold standard alignments are covered by EDITS.

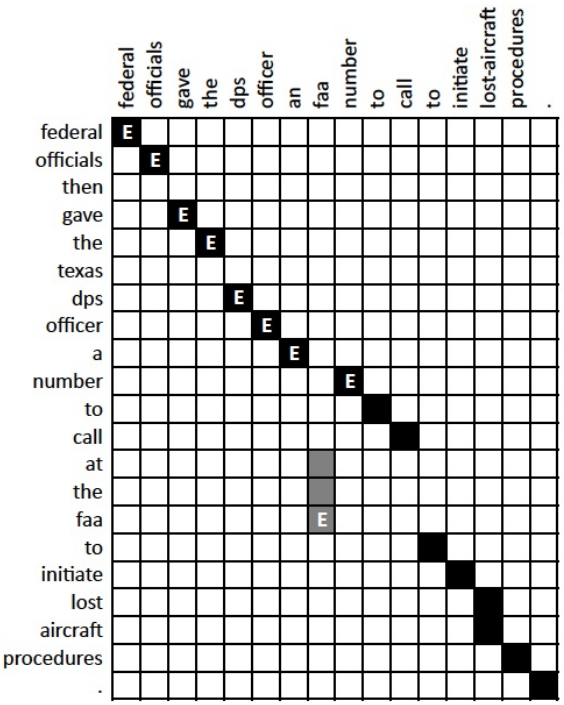

Figure 2: EDITS (E) and gold standard alignments corresponding to annotator C (black, (S)ure; grey (P)ossible) for the sentence paraphrase pair "Federal officials gave the DPS officer an FAA number to call to initiate lost-aircraft procedure" (horizontal axis) and "Federal officials then gave the Texas DPS officer a number to call at the FAA to initiate lost aircraft procedures" (vertical axis).

We performed the mapping between EDITS and gold standard alignments automatically, and computed precision, recall and F1. As can be seen in the first column in Table 1, precision is high (around 0.87) and the recall is low (around 0.50). EDITS only covers a

half of the expected alignments, but the ones that carries out are mainly correct.

We performed a second calculation of the results applying a series of filters in order to get more precise results. We did not consider cases that were erroneously penalized by EDITS in the first calculation. Specifically, we filtered block or discontinuous alignments, which refer to those cases in the gold standard in which a group of (discontinuous) words is aligned to a word or another group of (discontinuous) words. An example of this can be seen in *FAA/at the FAA* and *lost aircraft/lost-aircraft* in Figure 2. This situation cannot take place in EDITS, as the alignment is performed between nodes. We also filtered prepositions, conjunctions and punctuation marks. These elements are aligned in the gold standard, but do not appear as nodes and, thus, are not aligned in our analysis with EDITS. In the case of block alignments, the filter has a linguistic motivation as well: when the gold standard annotators use a block, it is because a word by word alignment is not possible. This case corresponds, on many occasions, to the semantic paraphrases, that cannot be treated with the TED approach (see Section 3.1). As can be seen in the second column in Table 1, once the filters applied, the precision and especially the recall rise (0.1 and more than 0.25 points, respectively).

We also analyzed the cases annotated as S in the gold standard separately. Although the recall rises again, the precision is lower. The reason for this decrease is that some EDITS alignments coincide with P alignments in the gold standard. When we do not take into consideration P alignments, these EDITS alignments are still there, which causes a decrease in the precision.

	− **Filters**		+ **Filters**		+ **Filters Only S**	
	A	*C*	*A*	*C*	*A*	*C*
Precision	0.86	0.88	0.87	0.89	0.85	0.87
Recall	0.50	0.49	0.77	0.78	0.78	0.80
F1	0.63	0.63	0.81	0.83	8.82	0.83

Table 1: EDITS alignment results classified according to the mapping with annotations by annotators A and C in the gold standard.

4 Conclusions

In this paper, we analyzed TED as a baseline approach for paraphrase representation, which may be used as the basis for further work on other approaches to paraphrase representation. As it consists of a purely syntactic approach, paraphrase alternations not based on syntactic reorganizations do no find an adequate representation. Moreover, further work needs to be done in order to improve tree alignments.

We showed that the EDITS suite, initially developped for RTE, can also be applied to the paraphrase task. As a result of the experiments, we obtained the MSRP corpus and a fragment of the Cohn, Callison-Burch, and Lapata (2008)'s corpus processed with EDITS, as well as a mapping between EDITS and Cohn, Callison-Burch, and Lapata (2008) alignments.

5 Future Work

A possible future line of research is the exploration of Tree Alignment Distance (Bille, 2003) and/or (Fanout) Weighted Tree Edit Distance (Augsten, Böhlen, and Gamper, 2010) algorithms, as we hypothesize that they can do better in terms of tree alignment.

Moreover, we plan to work on an approach dealing with paraphrase complexity in a more comprehensive way. Our objective is setting a paraphrasability measure based on the combination of relatedness measures associated to different types of the paraphrase typology by Vila, Martí, and Rodríguez (2011). EDITS would be used to build one of these dimensions.

References

Aho, Alfred V. and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56.

Augsten, Nikolaus, Michael Böhlen, and Johann Gamper. 2010. The pq-gram distance between ordered labeled trees. *ACM Transactions on Database Systems*, 35(1):art. 4.

Bille, Philip. 2003. Tree edit distance, alignment distance and inclusion. IT University Technical Report Series.

Casacuberta, Francisco and Enrique Vidal. 2007. Learning finite-state models for

machine translation. *Machine Learning*, 66:69–91.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague/Paris.

Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–614.

Dolan, William B. and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, pages 9–16.

Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Australia.

Eisner, Jason. 2003. Learning non-isomorphic tree mappings for machine translation. In *The Companion Volume to the Proceedings of the ACL*, pages 205–208.

Heilman, Michael and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases and answers to questions. In *Proceedings of the HLT-NAACL*, pages 1011–1019.

Kouylekov, Milen and Bernardo Magnini. 2005. Recognizing textual entailment with tree edit distance. In *Proceedings of the PASCAL RTE Challenge*, pages 17–20.

Shasha, Dennis and Kaizhong Zhang. 1990. Fast algorithm for the unit cost editing distance between trees. *Journal of Algorithms*, 11:581–621.

Smith, David and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30.

Vila, Marta, M. Antònia Martí, and Horacio Rodríguez. 2011. Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Wan, Stephen. 2010. *Sentence Augmentation: A Text-to-Text Generation Component for Summarisation*. Ph.D. thesis, Macquarie University, Australia.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Wu, Dekai. 2010. Alignment. In Nitin Indurkhya and Fred Damerau, editors, *Handbook of Natural Language Processing*. Chapman & Hall/CRC, second edition, pages 367–408.

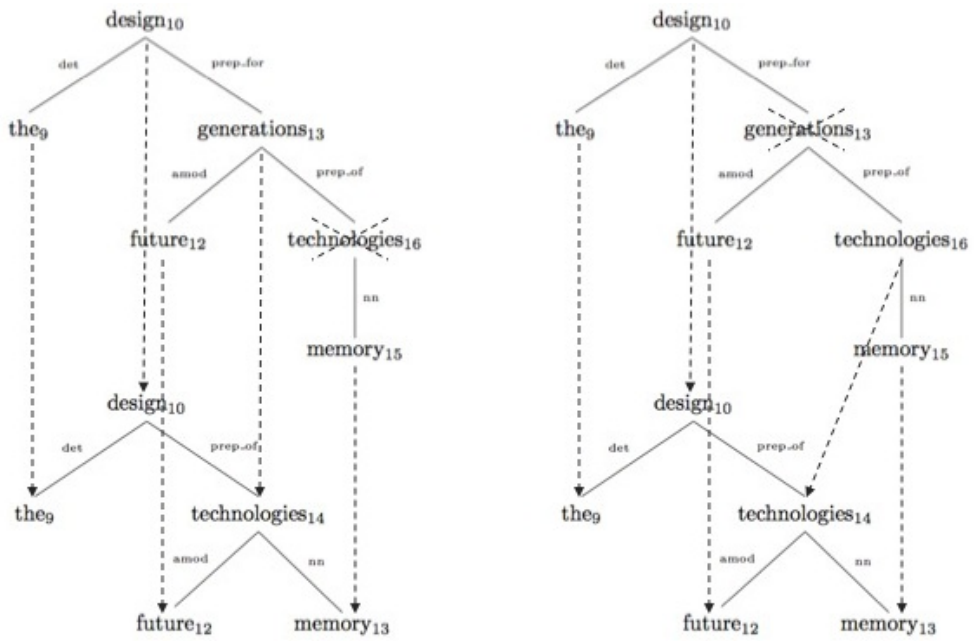

Figure 1: Representation by EDITS (left) and the expected one (right) for the MSRP corpus paraphrase pair (fragment) "the design for future generations of memory technologies" (top) and "the design of future memory technologies" (bottom). Arrow: substitution/alignment; cross: deletion.

Chapter 3

Paraphrase Corpora: Creation and Annotation

3.1 Paraphrase Corpus Building

Marta Vila (Universitat de Barcelona)
Horacio Rodríguez (Universitat Politècnica de Catalunya)
M. Antònia Martí (Universitat de Barcelona)

Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus

Submitted to *Natural Language Engineering* (second revision process completed)

Journal URL
`http://journals.cambridge.org/action/displayJournal?jid=NLE`
Impact Factor 0.432

Abstract Paraphrase corpora are an essential but scarce resource in Natural Language Processing. In this paper, we present the WRPA method, which extracts relational paraphrases from Wikipedia, and the derived WRPA paraphrase corpus. The WRPA corpus currently covers person-related and authorship relations in English and Spanish, respectively, suggesting that, given adequate Wikipedia coverage, our method is independent of the language and the relation addressed. WRPA extracts entity pairs from structured information in Wikipedia applying distant learning and, based on the distributional hypothesis, uses them as anchor points for candidate paraphrase extraction from the free text in the body of Wikipedia articles. Focussing on relational paraphrasing and taking advantage of Wikipedia structured information allows for an automatic and consistent evaluation of the results. The WRPA corpus characteristics distinguish it from other types of corpus that rely on string similarity or transformation operations. WRPA relies on distributional similarity and is the result of the free use of language outside any reformulation framework. Validation results show a high accuracy for the corpus.

Natural Language Engineering **1** (1): 1–34. Printed in the United Kingdom

Relational Paraphrase Acquisition from Wikipedia. The WRPA Method and Corpus†

M. VILA, M.A. MARTÍ
CLiC, Universitat de Barcelona
Gran Via 585, 08007 Barcelona, Spain
{marta.vila, amarti}@ub.edu

H. RODRÍGUEZ
TALP, Universitat Politècnica de Catalunya
Jordi Girona Salgado 1-3, 08034 Barcelona, Spain
horacio@lsi.upc.edu

(*Received XX; revised XX*)

Abstract

Paraphrase corpora are an essential but scarce resource in Natural Language Processing. In this paper, we present the WRPA method, which extracts relational paraphrases from Wikipedia, and the derived WRPA paraphrase corpus. The WRPA corpus currently covers person-related and authorship relations in English and Spanish, respectively, suggesting that, given adequate Wikipedia coverage, our method is independent of the language and the relation addressed. WRPA extracts entity pairs from structured information in Wikipedia applying distant learning and, based on the distributional hypothesis, uses them as anchor points for candidate paraphrase extraction from the free text in the body of Wikipedia articles. Focussing on relational paraphrasing and taking advantage of Wikipedia structured information allows for an automatic and consistent evaluation of the results. The WRPA corpus characteristics distinguish it from other types of corpus that rely on string similarity or transformation operations. WRPA relies on distributional similarity and is the result of the free use of language outside any reformulation framework. Validation results show a high accuracy for the corpus.

1 Introduction

There exists a consensus in defining paraphrases as those language expressions different in form but expressing (approximately) the same meaning. Paraphrasing is a broad and multifaceted phenomenon displaying varied linguistic mechanisms. For example, in the paraphrase pair in (1), a synonymy substitution occurs (in

† This work is supported by the MICINN projects TEXT-KNOWLEDGE 2.0 (TIN2009-13391-C04-04) and KNOW2 (TIN2009-14715-C04-04), as well as a MECD FPU grant (AP2008-02185). Also, we are grateful to Esther, Santiago, Rita and Oriol, the linguists that worked on the annotation processes.

italics); in (2), a syntactic reorganization can be observed; finally, the example in (3) shows a deeper change involving a complete rewording of the text.[1]

(1) a. But Secretary of State Colin Powell *brushed off* this possibility
b. Secretary of State Colin Powell [...] *ruled out* a non-aggression treaty

(2) a. *The company will offer songs* for 99 cents and albums for $9.95
b. *The songs are on offer* for 99 cents each, or $9.99 for an album

(3) a. Michael Mitchell [...] *did not answer his phone* Wednesday afternoon
b. Michael Mitchell [...] *was not available for comment*

The omnipresence of paraphrase processes in the ordinary use of natural languages makes a knowledge of paraphrasing essential in many Natural Language Processing (NLP) applications. By way of illustration, in question-answering, the wording of a question may differ to its possible answers. In a canonical system, the question is straightforwardly transformed into an assertion with a variable; without the help of paraphrase knowledge, only the exact occurrence of this pattern would result in an answer. Herrera et al. (2007) show the variabilty of potential answers in this field. Other examples are summarization, where paraphrase knowledge is needed to avoid redundancy in the final summary, and editing, where paraphrases offer alternative expressions that fulfill certain communicative purposes.

Quoting Dolan and Brockett (2005), Burrows et al. (2013) also pointed out that paraphrase corpora are essential, since they are necessary for evaluating and benchmarking the progress of researchers working on the foundations of paraphrasing, on new algorithms and on new tools. However, there exists a lack of paraphrase corpora. They are not created naturally or spontaneously as the Canadian Hansard corpus[2] for machine translation, which consists of parallel texts in English and Canadian French, drawn from official records of the proceedings of the Canadian Parliament. Moreover, automatically collecting paraphrases is not a straightforward task and neither is the validation of the acquired paraphrases. A more general problem is that paraphrase multifaceted nature prevents from the creation of general and comprehensive paraphrase corpora, that is, paraphrase corpora covering the phenomenon as a whole. Therefore, the field lacks a general data set that can serve as a standard against which algorithms can be trained and evaluated. Only corpora covering specific paraphrase types or facets, directly linked to the way paraphrases were obtained, may be created.

Considering all this, our objective has been building a new paraphrase corpus covering a concrete but productive paraphrase facet: relational paraphrases, that is, those paraphrases expressing the same type of relation between two entities. Examples (4) and (5), extracted from our corpus, illustrate this.[3] In the former, the

[1] Examples extracted from the Microsoft Research Paraphrase (MSRP) corpus. See footnote [13].
[2] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T20`
[3] Translations of Spanish examples appear in a smaller text font. This applies to all the examples in the paper and in the tables.

authorship relation is expressed; the latter is an example of the PERSON–PLACE OF BIRTH relation. Relational paraphrasing sets our work in a manageable framework within the broadness of paraphrasing, where processes can be automatized and results evaluated in a straightforward and consistent way.

(4) a. AUTHOR es conocido por su libro WORK
AUTHOR is known for his book WORK

b. AUTHOR es autor de los libros WORK
AUTHOR is the author of the books WORK

(5) a. PERSON was a native of PLACE

b. PERSON born in PLACE

In this paper, we present the Wikipedia-based Relational Paraphrase Acquisition (WRPA) method and describe its implementation. The WRPA method extracts relational paraphrases from Wikipedia[4] based on the distributional hypothesis (Harris, 1954) and taking advantage of Wikipedia structured information. This method is subsequently used to build the WRPA paraphrase corpus, covering some examples of relations in different languages. Building the corpus constitutes an empirical test of the usefulness of the method to produce high quality corpora and the corpus *per se* is a valuable resource for research in paraphrasing.

Several features show the value and productiveness of the WRPA corpus: (*i*) as the extraction method is based on the distributional hypothesis, paraphrases in WRPA do not necessary show a formal mapping or correspondence, as in the case of (5). This type of paraphrase presents considerable computational challenges and requires further study. (*ii*) Currently, the corpus covers 16 different relations in two languages (English and Spanish). This suggests that the WRPA method is independent of the language and the relation addressed, and that the corpus may be extended to any language and relation present in Wikipedia. Moreover, not only a variety of relations are covered, but a large number of paraphrase variants are comprised within each relation. (*iii*) The quality of the paraphrases is guaranteed by the reliance of our method on Wikipedia structured and semantically labelled data. (*iv*) WRPA corpus subsets have already been used successfully in other tasks, namely paraphrase-type annotation (Vila et al., 2013) and the Slot-Filling Task in the TAC KBP contest (Gonzàlez et al., 2013).

The rest of this paper is organized as follows. In Section 2, an overview of the WRPA method main characteristics is set out. Section 3 presents relevant related work both in paraphrase extraction and paraphrase corpora. The process of learning the WRPA components is described in Section 4; the validation of method is set out in Section 5; and its application, giving rise to the WRPA corpus, is presented in Section 6. In Section 7, the main features of the WRPA corpus are analysed. Finally, the conclusions and future work are presented in Section 8.

[4] `http://www.wikipedia.org`

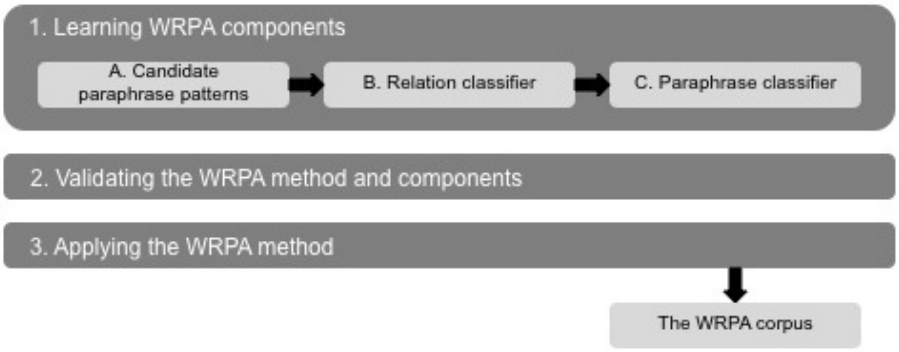

Fig. 1. WRPA method outline

2 WRPA Method Outline

The WRPA method is divided into three blocks as shown in Figure 1: learning WRPA components, validating the WRPA method and components and applying the method in order to obtain the WRPA corpus. The first block, in turn, is divided into three phases: acquiring the candidate paraphrase patterns, and learning the relation and paraphrase classifiers. In what follows, we present an overview of the whole process, organized according to the steps in Figure 1.

1A in Figure 1. We first learn a set of candidate paraphrase patterns expressing a concrete relation. Acquiring paraphrases implies ensuring sameness of meaning. In this sense, the distributional hypothesis (Harris, 1954), which states that words occurring in the same contexts tend to have similar meanings, makes up the basis of the methodology that many authors apply (1 in Figure 2). This methodology allows for the acquisition of a special types of paraphrase: those were a formal mapping is not necessary observed.

We apply this hypothesis to relational paraphrasing (2 in Figure 2). Relational paraphrases are those paraphrases in which the same type of relation is expressed in both members of the paraphrase pair. The relations we deal with are directional and binary between two entities: the SOURCE and the TARGET (i). An example of such a relation is authorship, where the SOURCE corresponds to an AUTHOR and the TARGET corresponds to a WORK (ii). SOURCE and TARGET are classes standing for sets of instances.

(i) Relation $\subset$ SOURCE $\times$ TARGET
SOURCE $\xrightarrow{Relation}$ TARGET

(ii) Authorship $\subset$ AUTHOR $\times$ WORK
AUTHOR $\xrightarrow{Authorship}$ WORK

Pairs of SOURCE and TARGET entities in a concrete relation stand for the "context" (using distributional-hypothesis terms) of our paraphrase candidates. In concrete, our approach is based on the hypotheses that the meaning of the text around the SOURCE and TARGET entities will be similar throughout their different occurrences, and that this meaning will hold the relation addressed in some way.

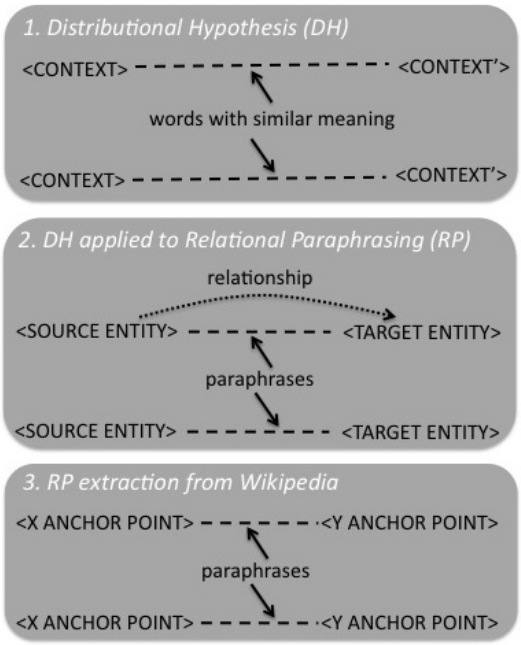

Fig. 2. Relational paraphrasing

One of the main problems facing paraphrase acquisition is disposing of an appropriate data source. Collecting paraphrases requires the availability of a corpus where different wordings for the same meaning coexist. The web is the biggest source of redundancy in existence, but it is also unstructured and unrestricted. We have therefore chosen Wikipedia. Its structure allows for an effective automatization of the task and its encyclopaedic nature allows for the acquisition of reliable and consistent examples. Imagine, for example, that we look for the WP page for "Paul Auster". Once the page corresponding to the novelist is located (sometimes after a disambiguation procedure), we are reasonably sure that this page contains information about the author, and that mentions of "Paul" or "Auster" are likely to refer to him. The adequacy of mentions is notably less reliable when accessing web pages.

WRPA obtains pairs of SOURCE and TARGET entities taking advantage of Wikipedia structured information: SOURCE entities are extracted from the titles of Wikipedia pages, and TARGET entities from infoboxes and itemized information embedded in these pages. Then WRPA uses these entities as X and Y anchor points, respectively, for the extraction of paraphrase candidate fragments in the free text of the body of Wikipedia articles (3 in Figure 2).

In more detail, the extraction of the SOURCE and TARGET entities is performed within the distant learning paradigm for relation extraction, initially proposed by Mintz et al. (2009), which uses supervised learning but with supervision not provided by manual annotation but obtained from the occurrence of positive training instances in a knowledge source or reference corpus. As we do not bootstrap the

process, the distributional hypothesis is only used for obtaining snippets of text expressing a relation between two entities and is not applied to obtain new entities. It is important to point out that, although WRPA applies relation extraction techniques, our objectives are different: while relation extraction systems are geared towards obtaining the semantic relation held by pairs of entities in a corpus, relational paraphrasing focusses on the wording used to express those relations. Our objective is to build paraphrase corpora for specific relations, not to extract as many as possible relations occurring in a text.

Finding a large number of good "contexts" is essential in our approach, since the quality and number of the obtained paraphrases depends in great measure on these contexts. The use of various Wikipedia structured sources allows for the obtention of a large number of source and target entities. Furthermore, we guarantee the correctness of these entities by extracting them directly from structured and semantically labelled data.

Once the set of SOURCE and TARGET entities has been obtained, they are used as X and Y anchor points in the body of Wikipedia articles, the sentences where they appear are extracted and the snippets within X and Y are further processed. As reducing paraphrase expressions to sequences of words has a rather limited expressive power, our method generalizes these word sequences into regular expression patterns including words, lemmas and PoS tags using a A* approach (Nilsson, 1982). Finally, a set of candidate paraphrase patterns for a concrete relation is obtained.

As reflected in Figure 2, our pairs of entities are seen from multiple perspectives in this paper. Thus, several terms are used to allude to them: when we refer to the distributional hypothesis underlying our method, we talk about *context*; when we refer to relational paraphrasing, we say *source* and *target entities*; when we refer to the technique applied within the Wikipedia framework, we use the term *anchor point*.

1B and 1C in Figure 1. Two classifiers are built: the relation classifier checks whether a paraphrase candidate is an instance of the relation addressed, while the paraphrase classifier establishes whether two candidates are actual paraphrases.

2 in Figure 1. We carried out a series of experiments in order to validate the performance of the WRPA method and components. They were applied to several relations in two languages. More specifically, our experiments are divided into two blocks: person experiments, which include the PERSON–DATE_OF_BIRTH, –DATE_OF_DEATH and –PLACE_OF_BIRTH relations in the English Wikipedia; and the authorship experiment, which focusses on the AUTHOR–WORK relation in the Spanish Wikipedia. Moreover, we performed a further validation applying our method to 13 more person relations. Therefore, the method was applied to 17 relations in total.[5]

Person relations are simple relations. Authorship is complex in the sense that it

[5] In Section 1, we stated that the WRPA corpus covers 16 relations. However, as will be explained in Section 6, one of the relations addressed was not included in the final corpus due to its low accuracy.

includes several professional or artistic activities like PAINTER–PAINTING, DIRECTOR–FILM or INVENTOR–INVENTION. It should be noted that these relations are not treated separately, but embedded in a single complex one: authorship.[6]

The person relations and the authorship one were chosen because they are frequent in biographies, for its high presence is Wikipedia structured information and for their diverse nature.

These experiments suggests that the WRPA method is independent of the language and the relation addressed and that the corpus may be extended to any language and relation present in Wikipedia.

3 in Figure 1. Finally, the application of the WRPA method resulted in the obtention of the WRPA corpus , which consists of WRPA-authorship and WRPA-person.

3 Related Work

The omnipresence and complexity of paraphrasing have given rise to a great variety of approaches to the automatic treatment of this phenomenon in NLP. Although this task is far from being solved, interesting and useful methods have been set out.[7] These methods can be grouped in three basic paraphrase tasks as follows:

Paraphrase recognition. When faced with two snippets of text, deciding whether they are paraphrases or assigning a degree of paraphrasability to the pair.

Paraphrase generation. When faced with a snippet of text, generating a (ranked) set of paraphrases of the snippet.

Paraphrase extraction (or acquisition). When faced with a corpus, extracting from the corpus a (ranked) set of paraphrases.

It should be noted that these tasks are combined or embedded in larger systems on many occasions.

Moreover, some works are devoted to building paraphrase corpora, which are generally compiled using paraphrase extraction techniques. Existing paraphrase corpora are different in nature, which basically depends on the data source and method applied to acquire them.

In this section, we set out related work in paraphrase extraction (Section 3.1) and paraphrase corpora (Section 3.2), the fields in which the WRPA method and corpus are respectively embedded. Some works on corpus building in Section 3.2 can be linked to the corresponding paraphrase extraction systems in Section 3.1. As our final objective is building paraphrase corpora, more attention is paid to this field.

[6] Due to the broad and complex nature of this relation, "creator" might be a more adequate term than "author". Nevertheless, we use "author" because it is more widely used in the field.

[7] See surveys by Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010) for a general overview of paraphrasing in NLP.

3.1 Paraphrase Extraction

Paraphrase extraction systems acquire paraphrases using a variety of data sources, operating at different levels of language description and applying a wide range of techniques. We consider the type of data source used for extraction to be a good way to organize these systems, as this selection has consequences on the methodology applied. There are five basic data sources used for paraphrase extraction: parallel, comparable, bilingual, single-monolingual and semi-structured corpora.

Parallel corpora can consist of multiple translations of the same text into the same target language. Barzilay and McKeown (2001), for example, extract paraphrases from a collection of parallel English translations of novels.[8] They use an unsupervised learning algorithm that applies a co-training procedure to decision-list classifiers for two independent sets of features: one describing the paraphrase pair itself, and the other corresponding to the contexts in which paraphrases occur. They rely on both lexical and syntactic features. Pang et al. (2003), in turn, exploit the Multiple-Translation Chinese (MTC) corpus, which consists of parallel English translations of Chinese news articles. They describe a syntax-based algorithm that builds Finite State Automata, which can be used to extract lexical and syntactic paraphrase pairs.

Comparable corpora consist of multiple texts containing approximately the same information. Barzilay and Lee (2003) extract paraphrases from articles talking about the same topic by two different news agencies. They apply multiple-sequence alignment to sentences in order to learn a set of paraphrasing patterns, which are represented by word lattice pairs. Dolan et al. (2004), starting with temporarily and topically clustered news articles, follow two different approaches to extract paraphrases: string edit distance and an heuristic strategy that pairs initial sentences from different news stories.

Bilingual corpora contain texts reporting the same information in two languages. Bannard and Callison-Burch (2005), using alignment techniques from phrase-based statistical machine translation, identify paraphrases in one language using a phrase in another language as a pivot. Going further in this line of research, Martzoukos and Monz (2012) extract paraphrases for both the source and target languages.

Single monolingual corpora are those where no explicit parallelization exists. Lin and Pantel (2001) extract paraphrases from newspapers based on the extended distributional hypothesis: if two paths in a dependency tree tend to occur in similar contexts, the meaning of the paths tends to be similar. Other systems, starting with a small set of manually collected high quality contexts (non-ambiguous pairs usually reduced to named entities), collect patterns of co-occurrence from this set and iterate this process. This bootstrapping approach has achieved notable success for simple and very specific relations such as AUTHOR_TITLE or PERSON_BIRTHDATE (Brin, 1998; Ravichandran and Hovy, 2002).[9] The TEASE algorithm (Szpektor

[8] Part of the corpus used by Barzilay and McKeown (2001) is available at `http://people.csail.mit.edu/regina/par/`.

[9] These systems focus on the paraphrase-related fields pattern and relation extraction.

et al., 2004), in turn, consists of an unsupervised learning algorithm for web-based extraction of paraphrase relations. It takes as input a verb lexicon and for each verb searches the web for related syntactic entailment templates.[10]

Semi-structured corpora are those that combine structured information and free text. Wikipedia is the paradigmatic example of this type. Methods extracting paraphrases using Wikipedia take advantage of its structure to extract paraphrases from the free text. Although there exists a significant number of NLP works on Wikipedia,[11] little research has been conducted on paraphrase extraction. In this line, Max and Wisniewski (2010) carry out a mining of Wikipedia's revision history focussing on local modifications made by human revisers and collecting paraphrases among other phenomena. Also, an earlier version of the work presented here can be found in Vila et al. (2010). Moving forward paraphrase boundaries, in Yatskar et al. (2010), edits from the Simple English Wikipedia are used to extract lexical simplifications, which overlap with paraphrasing on many occasions.

WRPA follows the line of those methods that rely on the Wikipedia semi-structured corpus. Also, our paraphrase patterns can be compared to the ones in other works performing pattern acquisition, such as Brin (1998) or Ravichandran and Hovy (2002) (see Section 5).

3.2 Paraphrase Corpora

Although paraphrase corpora are rather few in number, this set is enlarged if we take into account corpora coming from paraphrase-related fields. We divide corpora presented in this section in two groups: those created within the paraphrase field and those coming from paraphrase-related fields. In the latter case, data collections are either actual paraphrase corpora or corpora that (partially) overlap with paraphrasing.[12]

Within the first group, the Microsoft Research Paraphrase (MSRP) corpus (Dolan and Brockett, 2005)[13] contains 5,801 English sentence pairs from news articles hand-labelled with a binary judgement indicating whether human raters considered them to be paraphrases (67%) or not (33%). Chen and Dolan (2011), in turn, present the Microsoft Research Video Description (MSRVD) corpus.[14] It was collected via the crowd-sourcing platform Mechanical Turk, where participants had to watch a short video clip and then summarize it in a single sentence. They used 2,089 videos and 41 descriptions were created per video in average. As users could use a language of their choice, both paraphrase and bilingual alternations exist between different descriptions of the same video.

[10] Lin and Pantel (2001) and Szpektor et al. (2004) broaden the paraphrasing scope and are geared towards finding textual entailments.
[11] See Medelyan et al. (2009) for a survey covering information mining in Wikipedia.
[12] Some corpora in this section can be accessed from the corpus websites `https://github.com/STS-NTNU/STS13/wiki/Corpora` and `http://www.semtracks.org/web/index.php?id=Corpora%20Directory`.
[13] `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/`
[14] `http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/default.aspx`

Cohn et al. (2008)[15] present a corpus of 900 paraphrase sentence pairs aligned at the word or phrase level. The pairs were compiled from three different corpora: (1) equivalent sentence pairs from the MSRP corpus, (2) the MTC corpus, and (3) the monolingual parallel corpus used by Barzilay and McKeown (2001). Max and Wisniewski (2010) built the Wikipedia Correction and Paraphrase Corpus (WiCoPaCo)[16] from Wikipedia revision history. Apart from paraphrases, the corpus includes spelling corrections and other local text transformations. In the paper, the authors set out a typology of these revisions and classify them as meaning-preserving or meaning-altering. Wubben et al. (2010) present the ILK Headline Paraphrase Corpus.[17] It consists of 7,400,144 pairwise alignments of 1,025,605 unique headlines acquired automatically from Google News.

Still in the first group, we include works whose focus is an extraction or generation system, but which end up building a paraphrase collection. The work Barzilay and Lee (2003)[18] results in a corpus consisting of a set of 6,534 domain-dependent paraphrase pattern pairs in English. The patterns were then applied to the rewriting of new sentences. A sample of pattern pairs and sentences generated by them were manually evaluated with paraphrase judgements. Another example is the knowledge collection acquired by Lin and Pantel (2001).[19] They extracted 7 million paths from parse trees (231,000 unique) from which paraphrases were acquired. Fujita and Inui (2005) built a gold-standard corpus that is to be used to evaluate paraphrase generation models. It consists of 2,031 sentence pairs in Japanese with human judgment indicating whether the paraphrase is correct or not.

Moving to the second group, five NLP fields among others provide paraphrasing with useful data sources: plagiarism, text simplification, text compression, machine translation and textual entailment.

Plagiarism. Paraphrasing is the linguistic phenomenon underlying most plagiarism acts. PAN-PC-10 (Potthast et al., 2010)[20] is a corpus containing cases of plagiarism, where 60% involve some kind of paraphrasing. Most of the paraphrase cases were generated automatically and 6%, manually. From this 6%, the Paraphrase for Plagiarism (P4P) corpus (Barrón-Cedeño et al., 2013)[21] was created. It is composed of 847 paraphrase-plagiarism cases manually annotated with the paraphrase phenomena they contain. The Webis Crowd Paraphrase Corpus 2011 (Webis-CPC-11) (Burrows et al., 2013)[22] comprises 4,067 text samples and their corresponding paraphrases, created by human editors again via Mechanical Turk.

[15] http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html
[16] http://wicopaco.limsi.fr/
[17] http://ilk.uvt.nl/~swubben/resources.html
[18] Paraphrase patterns and derived sentence pairs, together with the manual evaluations, are available at http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html.
[19] This corpus should be requested from the authors directly.
[20] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-10.html
[21] http://clic.ub.edu/corpus/en/paraphrases-en
[22] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-webis-cpc-11.html

Clough and Stevenson (2011), in turn, present the Corpus of Plagiarized Short Answers,[23] consisting of 95 answers (200-300 words) to computer science questions, in which plagiarism has been simulated taking five Wikipedia articles as source.

Text simplification. Simplifications are a specific type of paraphrases where the complexity of the text is reduced. Coster and Kauchak (2011) set out 137K aligned sentences pairs extracted by pairing the Simple English Wikipedia with the English Wikipedia.[24] This data set contains simplification operations including rewording, reordering, insertion and deletion. Zhu et al. (2010) present a similar resource, Parallel Wikipedia Corpus (PWKP),[25] which contains more that 108K pairs of aligned sentences again from Wikipedia and the Simple Wikipedia.

Text compression. The objective of text compression is obtaining a summary paraphrase. Cohn and Lapata (2008) present the Abstractive Compression Corpus,[26] which consists of 575 pairs created by manually compressing sentences from newspaper articles. Rewriting operations as word deletion, insertion, substitution or reordering were used. Knight and Marcu (2002)'s corpus[27] contains 1,067 pairs consisting of a sentence that occurred in a newspaper article and a manually compressed version of it. Contrary to Cohn and Lapata (2008), the pairs in this corpus show a single rewriting operation, namely word deletion.

Machine translation. Paraphrase collections have been widely used in this field. Developed for machine translation evaluation, the MTC corpus[28] contains 105 news stories from journalistic Mandarin Chinese text translated into English by 11 translation agencies (human translations) and 6 machine translation systems. These parallel translations constitute a rich source of paraphrases. Buzek et al. (2010), with the aim of obtaining paraphrases for the sentence snippets that are predicted to be problematic for a translation system, create 4,821 paraphrases of 1,780 phrases using Mechanical Turk.

Textual entailment. Paraphrasing can be considered to be a bidirectional entailment. Training and test corpora in the recognizing textual entailment competitions are also available.[29] These corpora contain both positive and negative examples of entailment pairs with a relatively high number of paraphrases. The problem in entailment corpora, as in others above, is that pure paraphrases are not distinguished from the whole entailment pair set.

The corpus presented in this paper is a corpus of relational paraphrases extracted from Wikipedia. Section 7 gives an overview of the WRPA corpus main features with comparisons to other paraphrase corpora.

[23] `http://ir.shef.ac.uk/cloughie/resources/plagiarism_corpus.html`
[24] `http://www.cs.pomona.edu/~dkauchak/simplification/`
[25] `http://www.ukp.tu-darmstadt.de/data/sentence-simplification/simple-complex-sentence-pairs/`
[26] `http://staffwww.dcs.shef.ac.uk/people/T.Cohn/t3/`, under "Corpus".
[27] This corpus should be requested from the authors directly.
[28] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T01`
[29] `http://www.nist.gov/tac/2011/RTE/index.html`

4 Learning WRPA Components

WRPA performance is based on the learning and application of a set of patterns associated with each relation addressed. In this section, we set out the learning process (1 in Figure 1). It is performed in three steps, grouped in two sections: learning the set of candidate paraphrase patterns (Section 4.1), and learning the relation and paraphrase classifiers (Section 4.2).

4.1 Learning the Set of Candidate Paraphrase Patterns

Learning the set of candidate paraphrase patterns (1A in Figure 1) proceeds in four steps, reflected in Figure 3. First, the working corpus is set up (Section 4.1.1). Second, collections of anchor points are extracted (Section 4.1.2). Anchor points are then used for sentence selection in the text, and relevant snippets are subsequently extracted from these sentences (Section 4.1.3). Finally, candidate paraphrase patterns are obtained from these snippets applying additional processes (generalization and ngram modeling) when needed (Section 4.1.4).

4.1.1 Working Corpus Set Up

Our corpus consists of the Spanish Wikipedia and the English Wikipedia. We downloaded the versions of February 2009 into an MySQL database from Wikipedia dumps. We used the JWPL software (Zesch et al., 2008)[30] to access Wikipedia. Although the use of Wikipedia dumps prevents us from working on the latest version of Wikipedia, it allows us to perform the experiments on a static version, avoiding the problems arising from content instability.

WRPA extracts SOURCE–TARGET pairs and, from them, our paraphrase candidates, taking advantage of the following Wikipedia components:

- *Source pages*: Wikipedia pages describing SOURCE entities.
- *Redirection pages* that point to source pages.
- *Infoboxes* that are present in the source pages containing attributes corresponding to TARGET entities.
- *Target sections*: Itemized sections in the source pages containing TARGET entities.
- *Target pages*: Pages linked to source pages containing an itemized list of TARGET entities.
- *Category-page links*, which assign categories to articles, and the *category graph*.

For each relation addressed, the first step is to collect Wikipedia articles corresponding to the SOURCE (e.g., PERSON and AUTHOR). To do so, we first select Wikipedia categories that correspond to the SOURCE. This process can sometimes be done straightforwardly: when the SOURCE corresponds to an existing Wikipedia

[30] `http://www.ukp.tu-darmstadt.de/software/jwpl/`

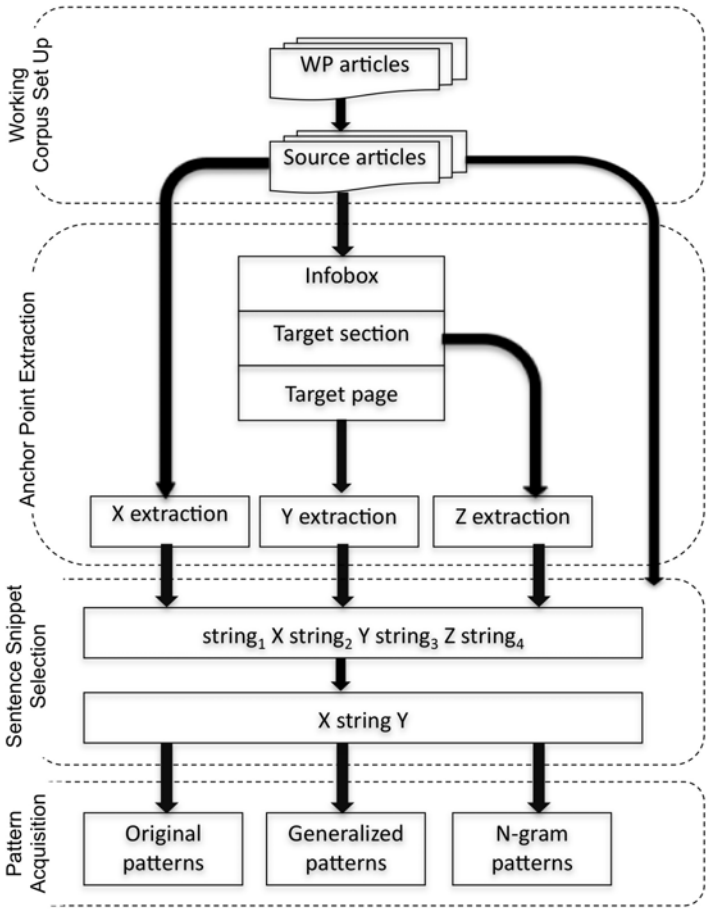

Fig. 3. The process of learning the set of candidate paraphrase patterns

category ('person' in the English Wikipedia), this category, which we call top category, is selected. When such a correspondence does not exist ('author' in the Spanish Wikipedia), a manual inspection is needed in order to get top Wikipedia categories related to the SOURCE.[31] A set of top categories is selected in this case. From this top category or top category set, the category graph of Wikipedia is traversed top-down following the subcategory links. Once the set of relevant categories and subcategories for a SOURCE has been established, we collect the set of pages under them using the category-page links. These pages constitute the working corpus for each relation addressed. The working corpora are then cleaned and segmented into sentences.

The second step is to collect TARGET-related infobox attributes (e.g., DATE_OF_BIRTH and WORK). Again, this task can sometimes be done straight-

[31] Manual interventions only apply when Wikipedia's lack of consistency prevents a fully automatic approach. They are minor interventions.

forwardly and sometimes it requires further work. For the person experiments, we took advantage of the set of generic attributes and mappings to specific ones provided in the framework of the Slot-Filling Task in the TAC KBP contest.[32] Thus, we only had to select the generic attributes corresponding to our TARGET entities. For the authorship experiment, we had to follow a less direct approach: we collected all the infoboxes with their attributes occurring in the working corpus and we manually selected the ones containing an attribute referring to works. Once the set of relevant infobox attributes was built, the pages in the working corpus containing any of these attributes were selected. They constituted what we call the learning corpus.

In our work, we talk about different partitions of the same corpus, namely Wikipedia: 1) the whole English Wikipedia and Spanish Wikipedia, our point of departure data sources; 2) the *working corpus*, which consists of the SOURCE pages; and 3) the *learning corpus*, which consists of the SOURCE pages containing a TARGET entity in the structured information, i.e., infoboxes and/or TARGET sections, and the TARGET pages (see Section 4.1.2).

4.1.2 Anchor Point Extraction

In this step, the set of SOURCE and TARGET entities is extracted. These entities will be used as X and Y anchor points, respectively, in the candidate paraphrase extraction process.

As shown in Figure 4, we extract X anchor points from the titles of source pages; Y anchor points, in turn, are extracted from infoboxes, target sections and target pages. We sometimes extract a third anchor, Z, which consists of additional information complementing Y. The relation holds between X and Y; Z simply makes Y more reliable. Z, when present, is extracted together with Y. In our experiments, Z is only obtained for authorship and corresponds to the work date.

X extraction is simpler than Y extraction. Once the source pages have been selected (Section 4.1.1), X extraction simply consists in looking for it in the title of each of these pages. Name variants for X, when they exist, are also extracted from redirection pages linked to the source page and/or from infobox attributes like 'alternate names' or 'nicknames'.

In the case of Y, a first set of anchor points is extracted from infoboxes. As information encoded in the attribute values shows variability (e.g., the variable format of the dates in Figure 5), we learned a set of grammars for extracting these values: DATE, DATE+PLACE, PLACE and WORK grammars in our experiments. Grammars were inferred using the ALERGIA system (Carrasco and Oncina, 1994), an efficient stochastic regular grammar inference engine that learns only with positive examples. Probabilities in ALERGIA generated grammars were not taken into account.

In the case of person experiments, information extracted from infoboxes is enough

[32] `http://www.nist.gov/tac/2012/KBP/index.html` It is important to point out that the objective of this task is different from ours: we aim to build paraphrase corpora and their aim is to learn slot values for an entity.

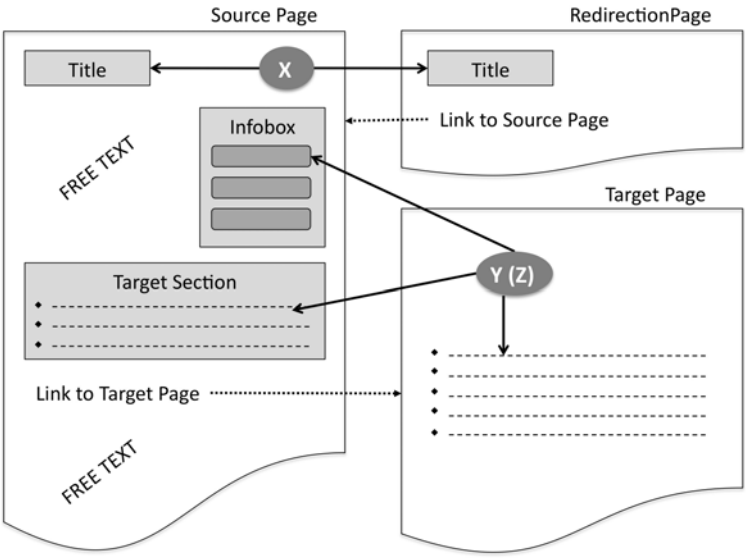

Fig. 4. Anchor point extraction from Wikipedia

birth_date :	[[August 15]], [[1974]]
birthdate :	1960|8|24
date of death :	2 November , 1916

Fig. 5. Infobox attributes and their values

for Y anchor point extraction, because, as can be seen in Table 1, i) the English Wikipedia is extensive enough, ii) there is a relatively high number of person pages with an infobox, iii) many of the target attributes appear in person infoboxes and iv) attributes are univalued, allowing for an easier and more accurate extraction. However, in the authorship experiment, the coverage of Ys in infoboxes is low, because i) the Spanish Wikipedia is smaller, ii) few author pages have an infobox, iii) few infoboxes have a work attribute and iv) this attribute is multivalued consisting of an unconstrained string with possible noise.

In order to improve the recall of the TARGET entities in the authorship experiment, we take advantage of, when they exist, the target sections and the links to target pages. Target sections present the difficulty involved in obtaining the section limits (heading and ending), a lack of homogeneity in the format and the presence of complementary material which results in noise. By way of illustration of the productivity of target sections and pages, Paul Auster's infobox lacks the work attribute. In contrast, 33 works (25 were correct) were extracted from the target section and 18 (all correct) were added from the target page. Table 2 contains some anchor points extracted by WRPA.

Table 1. *Counts for relevant Wikipedia pages for authorship and person experiments*

	Person experiments (English Wikipedia)	Authorship experiments (Spanish Wikipedia)
Content pages	1,660,067	484,550
Categories	20,741	4,054
Source pages	418,352	76,653
Source pages with infobox	142,452 (34%)	8,343 (10.9%)
Source pages with target infobox attributes	36,958 (8.8%)	305 (0.4%)

Table 2. *Anchor point examples*

X	**Y**	**Z**
	Authorship	
Canaletto	La Riva degli Schiavoni	(1730-31)
Edgar Allan Poe	Eureka	
Luis Eduardo Aute	Templo de carne	(1986)
	Date of birth	
David Kaye	October 14, 1964	
Sara Rue	1979|01|26	
Joan of Arc	c. 1412	
	Date of death	
Menachem Begin	9 March 1992 (aged 78)	
Sharon Tate	August 9, 1969 (aged 26)	
Diana Dors	4 May 1984 (aged 52)	
	Place of birth	
Giovanni Branca	San Angelo in Lizzola, Pesaro	
Grigore Preoteasa	Bucharest, Romania	
Tomas Plekanec	Kladno, Czech Republic|CZE	

4.1.3 Sentence Snippet Selection

Sentence snippet selection consists of first gathering the sentences that contain the anchor points in the body of the articles, and second selecting relevant snippets within these sentences.

Sentences are extracted from the free text of the same pages where the corresponding pair of anchor points has been extracted from the structured information. Sentences extracted by our method present the following structure:

$$\text{string}_1 \text{ X } \text{string}_2 \text{ Y } \text{string}_3 \text{ Z } \text{string}_4$$

X, Y and Z stand for our entities or anchor points. They may occur in any order (what is shown in the above formula is the canonical order) and only Y is mandatory.

Table 3. *Sentence snippet counts for authorship.*

XY	XYZ	YX	_Y	Y_
4,338	1,257	1,866	55,770	50,873

$String_i$ stands for a snippet of text and only $string_1$ or $string_2$ are mandatory. Table 3 shows frequency information for different orderings.

In order to improve the recall, we look for all the variant forms in which X and Y can be expressed. In order to expand X, apart from using the variants obtained in the anchor point extraction step, we apply a person-name grammar (Arévalo et al., 2004) to each entity. In order to expand Y, the grammars used for the extraction of infobox attribute values are also used for the generation of Y variants. One of WRPA's strengths for learning a large amount of patterns is the use of anchor point variants instead of restricting the search to the original ones alone. By way of illustration, in the authorship experiment, the average number of variants for X is 5.2. Using only the initial anchor point instead of the whole set of variants obtains only 28% of the total of patterns.

Relevant sentence snippets are subsequently extracted. The criteria for selecting them depends on its frequency and the presence and order of X, Y and Z. In the following, we will concentrate on XY and XYZ snippets. We hypothesize that the most relevant information expressing the relation can be found between a SOURCE and a TARGET in that order, as the former is likely to correspond to the subject of the sentence and the latter, to the object, and the canonical sentence order is <S V O>. Also, this subset is common in our learning corpus, as can be seen in Table 3 for the authorship experiment. The set of snippets only containing Y is bigger because subject (normally the X) elision is extremely frequent in Spanish. We decided not to work with this subset because the absence of X makes it much less reliable. Moreover, including YX patterns might result on a set of snippets different in nature to XY ones. However, again, this would result in an increase of noise. We are leaving these issues for future work.

Table 4 shows examples of XY snippets extracted by our method. In the case of proper names, numbers and dates, we replaced specific words with their generic tags to reduce sparseness (e.g., <LOCATION> or <DATE>).

4.1.4 Pattern Acquisition

Pattern acquisition consists of transforming sentence snippets into candidate paraphrase patterns. The process is more demanding in the case of complex relations like authorship, due to the high variability of the sentence snippets and their low relative frequency. In simple relations like date of birth, the sentence snippets are less variable and more frequent and, thus, the process is more straightforward.

For simple relations, the obtained snippets are simply transformed into regular expression patterns. For instance, the snippet in (f) in Table 4 is transformed into

Table 4. *Sentence snippet examples.*

	Authorship
(a)	X sigue escribiendo la novela Y X continues writing the novel Y
(b)	X lanzó su álbum debut Y X released his debut album Y
(c)	X dirigirá Y X will direct Y
	Date of birth
(d)	X was born in <LOCATION> in about Y
(e)	X was born in Y
(f)	X was born in <LOCATION> on Y
	Date of death
(g)	X died in Y
(h)	X died peacefully in his sleep on Y
(i)	X was executed on Y
	Place of birth
(j)	X born on <DATE> in Y
(k)	X was born free in Y
(l)	X was born to a family of <ENTITY> missionaries in Y

the regular expression X +was +born +in +<LOCATION> +on +(.+).*. The underlined group in the regular expression must be recognized by the grammar that corresponds to the TARGET (in this example, the DATE grammar). This consistency check allows for the removal of many spurious candidates extracted by the pattern matching mechanism. For complex relations, the direct use of the snippets as patterns leads to severe over-fitting and further work is therefore needed. We have developed two approaches: generalization and ngram modeling.

Generalization aims to extract patterns accounting for a larger number of texts from our initial sentence snippets. It implies going from a representation of the snippet as a sequence of words to its representation as a sequence of tokens with word forms, lemmas and PoS tags. Additionally, each token can be mandatory, skippable or omitted. The generalized pattern will be subsequently transformed into a regular expression formula.

In concrete, we first perform PoS tagging and named-entity recognition and classification on sentence snippets using the Freeling toolbox (Padró et al., 2010).[33] Then, we represent each token in the snippet as a <word, lemma, PoS, matching condition> tuple as shown in Table 5. Conditions state which part of the token has to be matched (w for word, l for lemma, or p for PoS) and its mandatory, skippable (?) or omitted (-) nature. Before generalization, all tokens fulfill the w condition. Snippets longer than 10 tokens are removed.

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

Table 5. *Generalization pre-process corresponding to (a) in Table 4.*

Word	sigue	escribiendo	la	novela
Lemma	seguir	escribir	el	novela
PoS	VB	VB	DT	NN
Matching condition	w	w	w	w

Table 6. *Generalization process corresponding to Table 5.*

Initial state
<sigue:w> <escribiendo:w> <la:w> <novela:w>
Snippet matching the pattern
sigue escribiendo la novela
Generalized pattern
<sigue:l?> <escribiendo:l> <la:w?> <novela:l>
Snippets matching the pattern
sigue escribiendo la novela (continues writing the novel) siguió escribiendo novelas (continued writing novels) escribió la novela (wrote the novel) escribía novelas (wrote novels)

Then, generalization is performed using an A* approach (Nilsson, 1982). The search is controlled by 3 parameters: N) minimum number of matches with other snippets, M) maximum number of states to be traversed, B) maximum number of generalized patterns to be returned.[34] The search proceeds until either of the following conditions is fulfilled: i) a state matching at least N other snippets is reached, ii) the number of traversed states reaches M. Once either of these stopping criteria is fulfilled, the B best scored patterns are returned. Note that the algorithm returns an optimum value only when the first halting criterion (N matches) holds, the second condition (M states) only prevents an expensive and unsuccessful search. In each state, snippets are represented as a sequence of <word, lemma, PoS, matching condition> tuples. The operators allow for moving from w to $w?$, from $w?$ to l and so on. An example of the generalization process can be seen in Table 6.

As the number of applicable operators in each state is huge, we perform a clustering of the candidates using an agglomerative approach with Levenstein distance as measure in order to reduce the search space. In this way, pattern generalization is carried out within each cluster and a set of generalized patterns is obtained from each cluster independently. Table 7 shows the contents of one of these clusters.

A final filtering sets the number of tokens fulfilling the w condition in a pattern from 1 to 4 (one of them being a common noun or a main verb) and the number

[34] In our experiments, N is set to 10, M is set to 100,000, and B is set to 100.

Table 7. *Cluster example*

comenzó a grabar su álbum debut, started to record his debut album,
lanzó su sexto álbum de estudio, released his sixth studio album,
lanzó su primer álbum solista <ENTITY> y released his first solo album <ENTITY> and
lanzó su primer álbum en <DATE>, released his first album on <DATE>,
lanzó dos álbumes más, released two more albums,
lanzó su primer álbum solista released his first solo album
lanzó su álbum debut released his debut album
registró su debut recorded his debut
edita el álbum produces the album
lanzó el álbum released the album

Table 8. *Generalized pattern and ngram examples*

Generalized patterns
<lanzó:w> <su:w> <álbum:w> <debut:w?> <lanzó:l> <su:p?> <álbum:w> <debut:p?> released his debut album
Simple ngrams
a trabajar en (to work in) todos los manuscritos del (all the manuscripts by)
Double ngrams
junto con su comedia [...] esta obra porque el together with his comedy [...] this work because the
junto con su comedia [...] título ha llegado hasta together with his comedy [...] title has reached

of words fulfilling the p? condition from 0 to 4. The reason for establishing such restrictive constraints is that we are looking for patterns with lexical content. Not establishing these constraints would result in paraphrase candidates with no text between X and Y, or only functional words or punctuation marks between them, which are not relevant for our purpose of building paraphrase corpora. Examples of generalized patterns can be found in Table 8 (they are derived from the seventh example in Table 7).

The alternative approach to generalization is the use of ngram fragments between the involved anchor points. This can be considered to be an intermediate approach between generalization and a pure bag of words clustering, which is not appropriate for our purposes because word order is pertinent. In this approach, all word ngrams

of any length are extracted from all the sentence snippets of the form XY[Z]. This set of ngrams is filtered to remove those not including any main verb or common noun. Moreover, a frequency threshold is used in order to discard infrequent ngrams. Finally, a regular expression is built for each remaining ngram, including the ngram within the anchor points and possibly a snippet of text (string_i) with a limited length (LL, in tokens) between both X and the ngram and between the ngram and Y ($|\text{string}_i| \leqslant \text{LL}$).

$$\text{X } \{\text{string}_1\} \text{ ngram } \{\text{string}_2\} \text{ Y}$$

We also build patterns including a pair of ngrams between X and Y:

$$\text{X } \{\text{string}_1\} \text{ ngram}_1 \ \{\text{string}_2\} \text{ ngram}_2 \ \{\text{string}_3\} \text{ Y}$$

Examples of the so called simple and double ngrams can be found in Table 8.

4.2 Learning the Classifiers

The next steps in WRPA are learning a relation classifier, which classifies a specific pattern as an example or not of the relationship of interest (1B in Figure 1), and learning a paraphrase classifier, which decides, given a pair of patterns for a relation, whether or not they are paraphrases (1C in Figure 1). In both cases, we used a supervised approach for learning decision trees in the WEKA toolbox (Hall et al., 2009).[35] In the experiments reported in this paper, we built these two classifiers for the authorship relation only. The simpler nature of person experiments made this process unnecessary.

4.2.1 Learning the Relation Classifier

In order to build the classifier, we first created a learning data collection. With this aim, the set of candidate paraphrase patterns was applied to the body of the articles in the learning corpus and 977 instances that matched different patterns were randomly selected. They were then manually annotated using the CoCo annotation tool (España-Bonet et al., 2009).[36] The annotation consisted in deciding whether the pattern instances expressed an authorship relation or not. By way of illustration, the instance (a) in Table 9 is a positive example and (b) is a negative one.

The classifier was learned using a total of 5,231 features extracted from this set of instances. These features consisted of the most frequent nouns, verbs and their lemmas, the most frequent word ngrams (Section 4.1.4), and the most frequent subpatterns containing a verb or a noun. The relation classifier resulted in 32 rules corresponding to branches of the decision tree. We finally applied this classifier to the whole set of candidate paraphrase patterns and selected the positive ones.

[35] `http://www.cs.waikato.ac.nz/ml/weka/`
[36] `http://www.lsi.upc.edu/~textmess/`

Table 9. *Relationship annotation*

(a)	✔	AUTHOR se embarca en la creación una de sus obras más famosas: WORK AUTHOR embarks on the creation of one of his most famous works: WORK
(b)	✗	AUTHOR vivió la mayor parte de su vida en su casa WORK AUTHOR lived most of his life in his house WORK

4.2.2 Learning the Paraphrase Classifier

In order to create the learning data set, we collected the pairwise cross product of all the patterns considered as positive by the relation classifier. We then applied them to the learning corpus and 8,746 instance pairs that matched different patterns were randomly selected for manual annotation using the CoCo interface. The manual annotation consisted of determining whether the pairs were paraphrases or not. Our point of departure is that, while what we call general authorship expressions can be applied to any type of authorship (e.g. *is the author of* can be applied to painting, writing, inventing, etc.), concrete authorship expressions can only be applied to one or some types (e.g. *wrote* can only be applied to writing). In the light of these ideas, the basic criteria were to annotate as paraphrases those pairs where both members expressed general authorship ((a) in Table 10) and those pairs where one member expressed general authorship and the other, concrete authorship (b). In the latter case, the meaning of the general case includes the meaning of concrete one. The specific-specific authorship relations were annotated as paraphrases only when they belonged to the same type(s) (c). When they did not belong to the same type(s), they were mutually exclusive and thus annotated as non-paraphrases (d).

The classifier was learned using a total of 234 features measuring different types of overlapping, including the most frequent words, word ngrams, lemma ngrams and tree edit distance information calculated with the EDITS suite (Kouylekov and Negri, 2010)[37]. The paraphrase classifier resulted in 55 rules.

Although the classifiers are only built for authorship and for Spanish in this work, they can be built for other relations and languages as the features, i.e., the most frequent nouns, verbs, etc. are not fixed but extracted from the set of instances corresponding to the relation in question.

5 Validating the WRPA Method and Components

WRPA validation consists in evaluating the set of candidate paraphrase patterns and the two classifiers (2 in Figure 1). Candidate patterns are validated according to their ability to match a snippet of text ending in a correct Y. Following this approach, we applied the set of patterns to the body of the articles in the learning corpus: we looked for the occurrences of any variant of X in the sentences in the

[37] `http://edits.fbk.eu/`

Table 10. *Paraphrase annotation. G(eneral), S(pecific). α and β are authorship types.*

(a)	**G** **G**	✔	AUTHOR en su obra WORK AUTHOR in his work WORK AUTHOR planeaba la realización de WORK AUTHOR was planning the execution of WORK
(b)	**G** **S**	✔	AUTHOR es autor de la obra WORK AUTHOR is the author of the work WORK AUTHOR publicó más de 30 libros, entre ellos: WORK AUTHOR published more than 30 books, among them: WORK
(c)	**S** α **S** α	✔	AUTHOR es autor de numerosos libros, entre los que pueden contarse WORK AUTHOR is the author of numerous books, among which we may include WORK AUTHOR publicó el libro WORK AUTHOR published the book WORK
(d)	**S** α **S** β	✗	AUTHOR publicó su primer manga, WORK AUTHOR published his first manga, WORK AUTHOR decepcionó con la película WORK AUTHOR disappointed with the film WORK

learning corpus (X is determined by the author in the title of the source page) and applied the pattern in order to obtain all possible Ys (corresponding to that X). We then computed their precision, recall and F1. The values for frequency, i.e., number of occurrences of each pattern, were also obtained.

In the case of precision, if the resulting Y is a variant or the original Y, the agreement is positive. For estimating recall, in the case of relations linked to univalued attributes, i.e., person experiment, we make the conservative assumption that all the pages contain at least a Y corresponding to each relation in the free text. Obviously, this is not always the case, because many pages lack these relations or some of them are simply not pertinent (e.g., a date of death is not adequate for living people). For each relation, recall is thus the ratio between the number of Ys recovered by the obtained patterns and the total number of pages in the learning corpus. In the case of multivalued attributes, i.e., the authorship experiment, it is not easy to hypothesize how many works occur in Wikipedia pages, so there are no clear ways to measure recall automatically. As an approximate alternative, we select the author pages containing links to their corresponding work pages and assume that all the works by an author can be obtained from these work pages. This resulted in 46 works per author on average. We then select from these works the ones appearing in the author page free text, which resulted in 13 works per author on average. We then compute the recall as the ratio between the number of Ys recovered by our patterns and the total number of Ys in the work pages occurring in the author pages. Thus, recall is only measured using author pages containing a link to a work page.

Table 11 presents the number of occurrences, precision, recall and F1 for the five most frequent patterns (individual and overall values) and the results for all

Table 11. *Person experiment results*

		Frequency	Precision	Recall	F1
Date of birth	X born Y	19,447	0.95	0.70	0.81
	X <LOCATION> Y	1,699	1.00	0.06	0.11
	X on Y	972	0.98	0.04	0.07
	X was born on Y	695	1.00	0.03	0.05
	X was born in Y	528	1.00	0.02	0.04
	Top 5 patterns	**23,341**	**0.96**	**0.84**	**0.90**
	All patterns	**25,624**	**0.95**	**0.96**	**0.96**
Date of death	X <DATE> - Y	1,188	0.86	0.35	0.50
	X died on Y	194	1.00	0.06	0.11
	X died in Y	161	1.00	0.05	0.09
	X died Y	111	0.96	0.03	0.06
	X <LOCATION> <DATE> Y	104	0.96	0.03	0.06
	Top 5 patterns	**1,758**	**0.90**	**0.52**	**0.66**
	All patterns	**2,094**	**0.91**	**0.69**	**0.78**
Place of birth	X was born in Y	1,352	1.00	0.25	0.40
	X born Y	1,188	0.70	0.22	0.34
	X born <DATE> in Y	657	0.98	0.12	0.22
	X on Y	187	0.63	0.04	0.07
	X <LOCATION> Y	163	0.65	0.03	0.06
	Top 5 patterns	**3,547**	**0.86**	**0.66**	**0.75**
	All patterns	**4,406**	**0.85**	**0.95**	**0.90**

the patterns in each person relation. As can be seen, the frequency is considerably higher in the first or first two patterns. Precision is very high in all cases, due to the relatively low variability of the patterns, and the fact that patterns have not undergone generalization or ngram processes. Date of birth resulted in a high recall, basically due to the high recall of the first and most frequent pattern. In the case of date of death, the recall is lower due to the conservative criteria used for defining it.

The results for authorship are shown in Table 12. We set out the results of the generalization, ngram and double-ngram approaches individually (rows 1 to 7). After removing the patterns showing a precision under 0.5, we split the remaining patterns in each approach into three groups: those having a precision of 1, those having a precision of between 0.5 and 1, and those having a precision of 0.5, respectively. The third group is empty for some sources. Results are shown in a cumulative way. The evaluation of several combinations between the top sets is also presented (rows 8 to 13).

Patterns coming from generalization are more precise than the patterns obtained from ngrams and double ngrams, though the latter show better recall. As patterns within each approach are sorted decreasingly by precision, increasing the number of patterns results in an improvement in recall at a cost of a drop in precision. The results of the different combinations, in turn, are significant since they demonstrate that the generalization and ngram approaches are complementary for pattern building, which allows for the acquisition of non-overlapping data. As an illustrative example, the overlapping between the works extracted by the pattern subsets

Table 12. *Authorship experiment results. Pat.=number of patterns; Freq.=frequency; P=precision; R=recall.*

			Pat.	Freq.	P	R	F1
Generalization	1	$P = 1$	15	386	1.000	0.006	0.012
	2	$1 > P > 0.5$	267	2,273	0.980	0.026	0.050
Ngram	3	$P = 1$	50	6,166	1.000	0.097	0.176
	4	$1 > P > 0.5$	100	6,770	0.910	0.150	0.257
	5	$P = 0.5$	152	6,962	0.734	0.188	0.299
Double ngram	6	$P = 1$	170	1,020	1.000	0.051	0.098
	7	$1 > P > 0.5$	204	1,127	0.755	0.062	0.115
Combination	8	2+3+6	487	9,459	1.000	0.145	0.254
	9	2+3+7	521	9,566	1.000	0.154	0.258
	10	2+4+6	537	10,063	0.989	0.188	0.315
	11	2+4+7	571	10,170	0.786	0.197	0.315
	12	2+5+6	589	10,255	0.900	0.221	0.358
	13	2+5+7	623	10,362	0.765	0.228	0.352

shown in rows 2 and 4 is only 6%. The relatively low recall in all cases is due to the greater difficulty of the task and the constraints we have imposed, namely, that they must contain at least a common noun or a main verb. Limiting ourselves to XY patterns also contributes to a drop in recall.

As our objective is to build paraphrase corpora, our main interest is to obtain precise and lexically rich paraphrases rather than a high recall. Actually, we observed that some patterns with a low recall are, in some sense, more interesting from the paraphrase point of view. By way of illustration, patterns like *X escribió Y* ('X wrote Y') show a high recall and the recall is lower in cases like *X había por fin publicado Y* ('X finally published Y'). However, the second pattern is of greater interest for the analysis of the linguistic mechanisms underlying paraphrasing.

Regarding the classifiers, they were validated using 10-fold cross-validation. The relation classifier achieved an accuracy of 85% and the paraphrase classifier an accuracy of 76%.

A direct comparison of the whole task of relational paraphrase corpus building to related approaches is not possible because no comparable corpora exist. However, we can perform a partial comparison of the subtask of learning the set of candidate paraphrase patterns to works on pattern acquisition.

One of the relations used in this paper to illustrate our approach, namely, date of birth, is also explored by Ravichandran and Hovy (2002). Ravichandran and Hovy (2002)'s patterns differ from ours in four main aspects:

- They extract patterns from the web while we extract them from Wikipedia.
- They accept XY and YX patterns while we limit ourselves to the former.
- They extract patterns split into prefix, infix and suffix strings while we limit ourselves to infixes.
- We apply a consistency check (grammar application) of the obtained Ys to ensure grammaticality, while they apply a simple matching procedure.

In their paper, a list of the 10 most precise patterns for this relation is presented.

Three of them are comparable to ours (they consist of XY infixes) and also appear in the top positions in our set: *X was born on Y*, *X was born in Y*, *X was born Y*. In Ravichandran and Hovy (2002), these patterns have a precision of 0.85, 0.6 and 0.59, respectively. In our work, these patterns have a precision of 1.0 in all three cases (Table 11 shows the first two cases). Our better results in precision are due to the structured and more limited nature of our corpus, and the consistency check performed. Although no recall information is provided by Ravichandran and Hovy (2002), we assume that our recall is lower, due to our less expressive patterns (only infixes and XY patterns are extracted) and our smaller corpus.

Regarding authorship, a comparison to the DIPRE system (Brin, 1998) can be made, although DIPRE and WRPA differ two aspects. First, DIPRE is applied to English and WRPA to Spanish; second, DIPRE is only applied to the AUTHOR–TITLE (books) relation and WRPA, to authorship in general. DIPRE patterns are again more expressive (they accept YX patterns, as well as prefixes and suffixes) than ours and come from a greater but less reliable source (24 million web pages). DIPRE does not provide values for precision, a qualitative analysis of the results is performed instead. It acquires 3,972 occurrences of books after convergence. Even though this result does not correspond to the whole authorship relation, it is comparable to our snippet counts in Table 3 (4,338 for XY and 1,866 for YX).

6 Applying the WRPA Method to Build the WRPA Corpus

The process of construction of the WRPA corpus (3 in Figure 1) consists of candidate paraphrase pattern set application and filtering with the two classifiers. Contrary to the validation process, here the application of the patterns was performed on the free text of the whole working corpus, rather than on the learning one.

In the building of the WRPA-person corpus, the whole set of patterns was applied to the working corpus to obtain the number of occurrences. The patterns were then organized in three groups according to the relation involved. All patterns within each group are considered to be paraphrases of each other in the sense that they all express the same type of relation. This massive grouping is possible due to the low variability of person patterns. The WRPA-person corpus consists of 355 patterns with 27,581 occurrences for date of birth, 433 patterns with 3,261 occurrences for date of death, and 942 patterns with 15,208 occurrences for place of birth.

For the building of the WRPA-authorship corpus, we selected the group of patterns with a higher F1 (row 12 in Table 12) and, from them, we selected the ones accepted by the authorship classifier. Then we obtained all the occurrences of these patterns from the working corpus, cross-paired them, computed the Levenstein distance between all pairs, sorted them by distance and discarded the most distant half. We finally applied the paraphrase classifier and collected the instance pairs accepted as paraphrases. The WRPA-authorship corpus consists of 81,101 pairs. The size of the WRPA-person and WRPA-authorship are not comparable. The reason is, as already stated, that the person corpus is not expanded to all possible pairs, but presented as a single massive grouping.

In order to check the quality of the final WRPA-person and WRPA-authorship

Table 13. *WRPA-person and WRPA-person-2 validation. Pat.=number of patterns in the corpus; Acc.=accuracy.*

Subcorpus	**Pat.**	**Acc.**	**Subcorpus**	**Pat.**	**Acc.**
date of birth	355	0.87	employee of	233	0.95
date of death	433	0.89	member of	375	0.99
place of birth	942	0.84	origin	555	0.69
age	155	0.04	parent	40	1
alternate name	55	1	religion	62	0.81
charge	40	0.9	school attended	94	0.92
child	54	1	spouse	413	0.95
residence	238	0.72	title	532	0.99

corpora, subsets for each were manually revised. WRPA-person validation consisted in manually revising 25% of the cases in each subcorpora to determine whether they were an expression of the corresponding relation. Besides using this revision for evaluation, the rejected patterns were removed from the corpus. Table 13 shows the final volume of each WRPA-person subcorpora and its accuracy. Moreover, a further validation was performed: the WRPA method was applied to 13 more person relations[38] and the results were revised in the same way. The figures for these new relations can also be seen in Table 13. They constitute the WRPA-person-2 corpus.

The accuracy for *age* shows a major decrease. *Age*, understood as a relation, is very unstable and it normally appears associated to concrete events in the life of a person in Wikipedia. This makes the value of the corresponding attribute in infoboxes highly variable. Due to its low accuracy, we decided not to include the *age* relation in the WRPA-person-2 corpus. The remaining relations show a high accuracy, only locations (place of birth, residence, and origin) are slightly lower. A reason for that is that a same location may appear in a text expressing different relations. Also, locations sometimes do not appear in isolation, but combined in infoboxes (e.g., *California, USA*).

In the case of WRPA-authorship, the manual validation comprises two steps. The first step was previous to building the paraphrase pairs and consisted of a manual validation of 922 instances corresponding to different patterns. The objective was to reject those instances not expressing an authorship relation. The global accuracy obtained was 0.63. The second step was performed after building the paraphrase pairs and consisted in manually validating 1,000 of these pairs stating whether they were paraphrases or not. The criteria used was the same as in Section 4.2.2. In this case, the accuracy obtained was 0.91.

[38] All the person relation in our paper are also present in the Slot-Filling Task in the TAC KBP contest.

7 WRPA Corpus Main Features

The WRPA corpus consists of a collection of relational paraphrases extracted from Wikipedia.[39] In what follows, WRPA corpus main features are set out. In concrete, the nature of paraphrases in WRPA is discussed and compared to the ones in other paraphrase corpora. Also, we analyse the quantity, variety and quality of WRPA paraphrases. Then, some actual uses and applications of WRPA are presented. This section finishes with an analysis of the most common errors found in the corpus.

Paraphrase nature. Paraphrase multifaceted nature prevents from the creation of general and comprehensive paraphrase corpora, that is, paraphrase corpora covering the phenomenon as a whole. Only corpora covering specific paraphrase facets, directly linked to the way paraphrases were obtained, may be created. In this sense, some paraphrase corpora rely on the formal similarity between the paraphrase members in order to establish paraphrasability. Part of the MSRP, for example, was built using string edit distance. Other corpora apply transformations to one member of the pair in order to obtain the other. The P4P corpus was built in this way. Despite following different routes, both string similarity and transformation techniques generally cover paraphrases with some kind of formal mapping. In contrast, other corpora do not necessarily show formal similarity between the pairs. This absence of formal mapping may be the result of different gathering techniques. In the case of the MSRVD corpus, paraphrases are the result of using videos as stimulus. WRPA, in turn, follows the line of those works that rely on the distributional hypothesis. In this case, as paraphrasability is established through context, what may be found in between is not predictable. By way of illustration, together with paraphrases with a high level of formal similarity (6), our corpus also contains paraphrases that are completely different in form (7).

(6) a. AUTHOR realizó su primer álbum en solitario, WORK
AUTHOR made his first solo album, WORK

b. AUTHOR publicó su primer álbum en solitario, WORK
AUTHOR released his first solo album, WORK

(7) a. AUTHOR corresponden los títulos WORK
the titles WORK correspond to AUTHOR

b. AUTHOR es autor de los libros WORK
AUTHOR is author of the books WORK

These ideas have been demonstrated in Vila et al. (2013), which presents the results of annotating 1,000 pairs of the WRPA-authorship corpus, together with the 856 pairs in P4P and 3,900 pairs of MSRP, with our typology of paraphrases (Barrón-Cedeño et al., 2013), comprising 24 paraphrase types.[40] The SAME-POLARITY type, which includes cases of lexical-unit substitution, synonymy among

[39] The corpus is available at `http://clic.ub.edu/corpus/en/paraphrases-en`.

[40] Annotated corpora are available at `http://clic.ub.edu/corpus/en/paraphrases-en` as a search interface and as a package to download. Annotation guidelines can also be accessed.

them, is one of the most frequent types in the three corpora (25-50% of the cases approximately). ADDITION-DELETION is the other most frequent type in P4P and MSRP (12.91% and 25.94%, respectively); in contrast, the second most frequent type in WRPA is precisely SEMANTIC (16.22%), which includes those cases of semantic similarity where a formal mapping is not possible.

The higher presence of SEMANTIC paraphrases in WRPA makes it a valuable resource in NLP, as this type of paraphrase presents considerable computational challenges and requires further study. Barrón-Cedeño et al. (2013) illustrates this in the field of automatic plagiarism detection: cases of plagiarism comprising SEMANTIC paraphrases are the most difficult to detect by plagiarism detection systems.

Paraphrase quantity. The WRPA corpus currently covers 16 relations in two languages (English and Spanish). Moreover, it can be extended to any relation and language with sufficient coverage in Wikipedia by applying our acquisition method. Also, for each relation, a great number of paraphrases is provided. We consider that a total of 4,421 patterns for person and 81,101 pairs for authorship to be significant numbers. It is important to note that a direct numerical comparison between existing paraphrase corpora is not feasible due to their different nature.

Paraphrase variety. Through the variety of relations covered (simple and complex, and with univalued and multivalued targets), we observed that there is a spectrum between two extremes within relational paraphrasing. Very complex relations like authorship show a high variablity with respect to semantic content, that is why paraphrases cannot be combined in a single paraphrase grouping. Very simple relations like date of birth show a low variablity, so a single group is possible (Section 6). Between these two extremes there is a whole spectrum of possibilities. What all these relations share is the underlying paraphrase concept, which is wider than in corpora like MSRP or P4P: in WRPA, paraphrases are those units expressing the same kind of relation, even though the semantic content is not always the same, as in example (8) for the *alternate name* relation.

(8) a. PERSON also writes under the name PERSON
b. PERSON changed his name to PERSON

Paraphrase quality. Taking advantage of Wikipedia structured information and semantically labelled data to obtain our anchor points, as well as extracting our paraphrases from the body of the articles of the same encyclopaedia, allowed for the acquisition of reliable and consistent examples. The precision of the paraphrase pattern sets is good and their quality is further tested through the use of the two classifiers, whose accuracy is also reasonably high (Section 5). The same can be stated for the results of the manual validation of the corpus (Section 6).

Corpus applicability. WRPA is not fully accurate because it was automatically created without a final and complete human revision. Therefore, this corpus cannot be used as gold standard for evaluating paraphrase systems. However, it is useful in other NLP tasks. By way of illustration, patterns in WRPA-person were used in the Slot Filling Task in the TAC-KBP-2012 contest Gonzàlez et al. (2013). Also,

Table 14. *Patterns not expressing the expected relation*

(a)	place of birth	PERSON was born outside of PLACE
(b)	authorship	AUTHOR tema de la película homónima, WORK AUTHOR topic of the homonymous film, WORK
(c)	spouse	PERSON saw and fell in love with PERSON
(d)	age	PERSON DATE – DATE NUMBER
(e)	origin	PERSON was born in LOCATION on LOCATION

paraphrase corpora annotated with paraphrase types show a great potential for the study of the phenomenon of paraphrasing and the development of NLP tools involving paraphrasing. As already explained, a subset of WRPA-authorship was annotated with our paraphrase typology (Vila et al., 2013).

Due to the complexity of the task and the diversity of the paraphrase phenomenon, the method applied is not error free. In what follows, an analysis of the drawbacks present in the WRPA corpus is performed. According to Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010), extraction methods based on the distributional hypothesis can produce pairs of templates that are not true paraphrase pairs, even though they share the same context. In fact, pairs involving antonyms are frequent. Lin and Pantel (2001) find *X solves Y* to be very similar to *X worsens Y*, and the same problem has been reported by Bhagat and Ravichandran (2008). WRPA is not different in this regard. It extracted cases of antonymy, such as (a) in Table 14 for place of birth; cases that do not express the relation in question, such as (b) for authorship; cases in which the relation does not necessarily hold, such as (c) for spouse; cases in which is difficult to determine the adequacy of the pattern, such as (d) for age; and cases where there exists a confusion with other relations, such as (e) for origin, where the preposition *on* clearly introduces a date. These errors have been removed from the final corpus when detected in our manual revisions.[41]

Other issues causing error in WRPA, specially regarding the authorship classifier, can be seen in Table 15. The example in (a) ends by a list of Ys instead of a single Y. The example in (b) introduces a series of alternatives to the name of the work (*better known as* or *which can be translated as*). In (c), a lot of complementary information appears (*revolutionized the field of theoretical linguistics* and *based on his doctoral thesis*). These cases make the task more difficult.

8 Conclusions and Future Work

In this paper, we set out the WRPA method and corpus: a method to build paraphrase corpora and a new corpus created accordingly. The WRPA method extracts

[41] Although the final-corpus revision is not a component of the WRPA method and it is only a partial revision, we decided to remove the detected errors from the corpora while maintaining the figures of global accuracy.

Table 15. *Troubling instances*

(a)	AUTHOR escribió numerosas obras especializadas, como WORK (1931), WORK (1933), WORK AUTHOR wrote numerous specialized works, such as WORK (1931), WORK (1933), WORK
(b)	AUTHOR publica su obra maestra, WORK, más conocido como WORK, que se puede traducir como WORK o WORK AUTHOR publishes his masterpiece, WORK, better known as WORK, which can be translated as WORK or WORK
(c)	AUTHOR revolucionó el campo de la lingüística teórica con la publicación de la obra WORK, basada en su tesis doctoral – WORK AUTHOR revolutionized the field of theoretical linguistics with the publication of the work WORK, based on his doctoral thesis – WORK

relational paraphrases, that is, paraphrases expressing a relation between two entities, from Wikipedia. In concrete, it extracts entity pairs from structured information in Wikipedia applying distant learning and, based on the distributional hypothesis, uses them as anchor points for candidate paraphrase extraction from the free text in the body of Wikipedia articles. Some of the extracted paraphrases are then generalized following two complementary approaches, namely generalization and ngram processes. Also, two classifiers that take the candidate paraphrases as input were built: a relation and a paraphrase classifier. Due to the vastness and diversity of paraphrasing, the evaluation of such a task is a challenging issue in NLP. Focussing on the relational paraphrase framework and the use of Wikipedia structured information allowed for a direct and straightforward evaluation of the results.

Paraphrase multifaceted nature prevents from the creation of general and comprehensive paraphrase corpora and only corpora covering specific paraphrase types or facets may be created. In this sense, the WRPA corpus is one more piece in the paraphrase puzzle, a piece that unveils paraphrase features not present in other types of corpora, making the words by Fillmore (1992) true: "every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way". Contrary to corpora that rely on string similarity or transformation operations, WRPA relies on coincident anchor points. The form of the obtained paraphrases is the result of the free use of language and not the result of operations of transformation and reformulation.

The fact that the corpus currently covers both simple and complex relations, and two languages suggest that the WRPA method is independent of the language and relation addressed and that the corpus may be enlarged, assuming that there exist both Wikipedia and shallow linguistic tools for that particular language, and that the relation appears in Wikipedia structured information.

Also, finding numerous good entities, i.e., our SOURCE and TARGET entities, is essential in our approach, since the quality and number of the obtained paraphrases depends in great measure on these entities. First, we guarantee their correctness by

extracting them directly from structured and semantically labelled data. Second, the TARGET entity is not only extracted from infoboxes but also from itemized information in or outside the source page. Also, several entity variant generation mechanisms are used both for the SOURCE and the TARGET. Thus, this method allows for the acquisition of a large number of quality paraphrase variants, which makes the application of bootstrapping techniques unnecessary, thereby avoiding data degradation.

Finally, the usefulness of the corpus has also been discussed: WRPA corpus subsets have already been used successfully in paraphrase-type annotation (Vila et al., 2013) and the Slot-Filling Task in the TAC KBP contest (Gonzàlez et al., 2013).

Potential lines for future work are (*i*) applying our method to other types of sentence snippets, such as <Y string X> or <string Y>, and to other relations and languages, (*ii*) evaluation on non-Wikipedia texts and (*iii*) studying the dependence between paraphrase extraction performance and the number of available Wikipedia pages, infoboxes and filled attribute values.

References

Androutsopoulos, I. and P. Malakasiotis (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research 38*, 135–187.

Arévalo, M., M. Civit, and M. A. Martí (2004). MICE: A module for named entity recognition and classification. *International Journal of Corpus Linguistics 9*(1), 53–68.

Bannard, C. and C. Callison-Burch (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL 2005*, pp. 597–604.

Barrón-Cedeño, A., M. Vila, M. A. Martí, and P. Rosso (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. To appear in *Computational Linguistics*. DOI: 10.1162/COLI_a_00153.

Barzilay, R. and L. Lee (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL 2003*, pp. 16–23.

Barzilay, R. and K. McKeown (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of ACL 2001*, pp. 50–57.

Bhagat, R. and D. Ravichandran (2008). Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of HLT/ACL 2008*, pp. 674–682.

Brin, S. (1998). Extracting patterns and relations from the World Wide Web. In *Proceeding of the 1st International Workshop on the World Wide Web and Databases (WebDB 1998)*, pp. 172–183.

Burrows, S., M. Potthast, and B. Stein (2013). Paraphrase acquistion via crowdsourcing and machine learning. To appear in *ACM Transactions on Intelligent Systems and Technology*.

Buzek, O., P. Resnik, and B. B. Bederson (2010). Error driven paraphrase annotation using Mechanical Turk. In *Proceedings of the HLT/NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pp. 217–221.

Carrasco, R. C. and J. Oncina (1994). Learning stochastic regular grammars by means of a state merging method. In R. C. Carrasco and J. Oncina (Eds.), *Grammatical Inference and Applications. Proceedings of the Second International Colloquium (ICGI 1994)*, pp. 139–152. Springer-Verlag.

Chen, D. L. and W. B. Dolan (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of HLT/ACL 2011*, Volume 1, pp. 190–200.

Clough, P. and M. Stevenson (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation 45*(1), 5–24.

Cohn, T., C. Callison-Burch, and M. Lapata (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics 34*(4), 597–614.

Cohn, T. and M. Lapata (2008). Sentence compression beyond word deletion. In *Proceedings of COLING 2008*, pp. 137–144.

Coster, W. and D. Kauchak (2011). Simple English Wikipedia: A new text simplification task. In *Proceeding of HLT/ACL 2011*, pp. 665–669.

Dolan, B., C. Quirk, and C. Brockett (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING 2004*, pp. 350–356.

Dolan, W. B. and C. Brockett (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, pp. 9–16.

España-Bonet, C., M. Vila, H. Rodríguez, and M. A. Martí (2009). CoCo, a web interface for corpora compilation. *Procesamiento del Lenguaje Natural 43*, 367–368.

Fillmore, C. J. (1992). "Corpus linguistics" or "computer-aided armchair linguistics". In J. Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, pp. 35–60. Mouton de Gruyter.

Fujita, A. and K. Inui (2005). A class-oriented approach to building a paraphrase corpus. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2003)*, pp. 25–32.

Gonzàlez, E., H. Rodríguez, J. Turmo, P. R. Comas, A. Naderi, A. Ageno, E. Sapena, M. Vila, and M. A. Martí (2013). The TALP participation at TAC-KBP 2012. In *Proceedings of TAC 2012* (to appear).

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter 11*(1), 10–18.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Herrera, J., A. Peñas, and F. Verdejo (2007). Paraphrase extraction from validated question answering corpora in Spanish. *Procesamiento del Lenguaje Natural 39*, 37–44.

Knight, K. and D. Marcu (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence 139*, 91–107.

Kouylekov, M. and M. Negri (2010). An open-source package for recognizing textual entailment. In *Proceedings of the ACL System Demonstrations (ACLDemos 2010)*, pp. 42–47.

Lin, D. and P. Pantel (2001). DIRT-Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD 2001*, pp. 323–328.

Madnani, N. and B. J. Dorr (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics 36*(3), 341–387.

Martzoukos, S. and C. Monz (2012). Power-law distributions for paraphrases extracted from bilingual corpora. In *Proceedings of EACL 2012*, pp. 2–11.

Max, A. and G. Wisniewski (2010). Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of LREC 2010*, pp. 3143–3148.

Medelyan, O., D. Milne, C. Legg, and I. H. Witten (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies 67*(9), 716–754.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extraction without labeled data. In *In Proceedings of ACL 2009*, pp. 1003–1011.

Nilsson, N. J. (1982). *Principles of Artificial Intelligence*. Springer-Verlag.

Padró, L., M. Collado, S. Reese, M. Lloberes, and I. Castellón (2010). Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of LREC 2010*, pp. 931–936.

Pang, B., K. Knight, and D. Marcu (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL 2003*, pp. 102–109.

Potthast, M., B. Stein, A. Barrón-Cedeño, and P. Rosso (2010). An evaluation framework for plagiarism detection. In *Proceedings of COLING 2010*, pp. 997–1005.

Ravichandran, D. and E. Hovy (2002). Learning surface text patterns for a question answering system. In *Proceeding of ACL 2002*, pp. 41–47.

Szpektor, I., H. Tanev, I. Dagan, and B. Coppola (2004). Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP 2004*, pp. 41–48.

Vila, M., M. Bertran, M. A. Martí, and H. Rodríguez (2013). Corpus annotation with paraphrase types. New annotation scheme and inter-annotator agreement measures (manuscript under submission).

Vila, M., H. Rodríguez, and M. A. Martí (2010). WRPA: A system for relational paraphrase acquisition from Wikipedia. *Procesamiento del Lenguaje Natural 45*, 11–19.

Wubben, S., A. van den Bosch, and E. Krahmer (2010). Paraphrase generation as monolingual translation: Data and evaluation. In *Proceedings of INLG 2010*, pp. 203–207.

Yatskar, M., B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee (2010). For the sake of simplicity: Unsupervised extraction of lexical simplications from Wikipedia. In *Proceedings of NAACL 2010*, pp. 365–368.

Zesch, T., C. Müller, and I. Gurevych (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of LREC 2008*, pp. 1646–1652.

Zhu, Z., D. Bernhard, and I. Gurevych (2010). A monolingual tree-based translation method for sentence simplification. In *Proceedings of COLING 2010*, pp. 1353–1361.

3.2 Paraphrase-Type Annotation

Marta Vila (Universitat de Barcelona)
Manuel Bertran (Universitat de Barcelona)
M. Antònia Martí (Universitat de Barcelona)
Horacio Rodríguez (Universitat Politècnica de Catalunya)

Corpus annotation with paraphrase types. New annotation scheme and inter-annotator agreement measures.

Submitted to *Language Resources and Evaluation*.

Journal URL
`http://www.springer.com/education+%26+language/linguistics/journal/10579`
Impact Factor 0.308

Abstract Paraphrase corpora annotated with the types of paraphrases they contain constitute an essential resource for the understanding of the phenomenon of paraphrasing and the improvement of paraphrase-related systems in Natural Language Processing. In this article, a new annotation scheme for paraphrase-type annotation is set out, together with newly created measures for the computation of inter-annotator agreement. Three corpora diferent in nature and in two languages have been annotated using this infrastructure. The annotation results and the inter-annotator agreement scores for these corpora are prove of the adequacy and robustness of our proposal.

Noname manuscript No.
(will be inserted by the editor)

Corpus Annotation With Paraphrase Types

New Annotation Scheme and Inter-annotator Agreement Measures

Marta Vila · Manuel Bertran · M. Antònia Martí · Horacio Rodríguez

Received: date / Accepted: date

Abstract Paraphrase corpora annotated with the types of paraphrases they contain constitute an essential resource for the understanding of the phenomenon of paraphrasing and the improvement of paraphrase-related systems in Natural Language Processing. In this article, a new annotation scheme for paraphrase-type annotation is set out, together with newly created measures for the computation of inter-annotator agreement. Three corpora different in nature and in two languages have been annotated using this infrastructure. The annotation results and the inter-annotator agreement scores for these corpora are prove of the adequacy and robustness of our proposal.

Keywords Paraphrasing, Paraphrase typology, Corpus annotation, Inter-annotator agreement.

1 Introduction

Paraphrasing, which stands for different wordings expressing (approximately) the same meaning, is omnipresent in the ordinary use of natural languages. This pervasiveness makes paraphrase knowledge indispensable in many Natural Language Processing (NLP) systems. However, paraphrasing is a complex phenomenon and, although it has been an object of study in NLP over the last few decades, it sometimes gives the sense of a still unexplored field.

This is quite evident at the sphere of paraphrase corpora, where there is a lack of standard or reference corpora, due to the difficulties in compiling large,

Marta Vila, Manuel Bertran, M. Antònia Martí
CLiC, Universitat de Barcelona
Gran Via 585, 08007 Barcelona, Spain
E-mail: {marta.vila, manu.bertran, amarti}@ub.edu

Horacio Rodríguez
TALP, Universitat Politècnica de Catalunya
Jordi Girona Salgado 1-3, 08034 Barcelona, Spain
E-mail: horacio@lsi.upc.edu

general, and accurate datasets.[1] Without these resources, researchers have resorted to developing their own small, ad hoc datasets. This situation has complicated and sometimes impeded progress in the field (Chen and Dolan, 2011). Paraphrase corpora are essential as they allow for a better understanding of the linguistic nature of paraphrasing, as well as the development of paraphrase tools based on real data and their subsequent evaluation.

A special type of paraphrase corpora are those containing information about the linguistic operations underlying paraphrases, in other words, corpora with the annotation of paraphrase types. Paraphrasing presents multiple and diverse linguistic manifestations; thus, this type of corpora show a great potential in order to go a step further in solving the puzzle of paraphrasing in NLP. Nevertheless, if paraphrase corpora are few, those with type annotation are, to the best of our knowledge, almost inexistent. Building annotation schemes is inherently difficult and the task is even more complicated for phenomena that are still not well understood (Zaenen, 2006), as is the case of paraphrasing. A great variety of linguistic operations give rise to paraphrases, a single paraphrase may include multiple combined paraphrase phenomena, and determining the scope of each phenomenon is not a straightforward task. This scenario makes the creation of such corpora a complex, costly, time-consuming challenge.

The development of these corpora involves building a powerful infrastructure backed by solid linguistic bases. In this article, we present such an infrastructure, as well as three corpora annotated by applying it. Firstly, we set out a new annotation scheme based on our paraphrase typology (Barrón-Cedeño et al, 2013, to appear). This scheme comprises a set of 24 paraphrase-type tags, as well as instructions to detect and annotate the scope of each of these tags within the paraphrases. Secondly, we set out new measures for inter-annotator agreement in order to guarantee the quality of these annotations. We finally present three corpora annotated with our infrastructure: the Paraphrase for Plagiarism corpus (P4P), the Microsoft Research Paraphrase corpus-Annotated (MSRP-A), and the Wikipedia-based Relational Paraphrase Acquisition corpus-Annotated (WRPA-A). The latter is in Spanish; the other two, in English. The annotation of such diverse corpora is prove of the adequacy and robustness of our proposal.

Section 2 sets out the state of the art on corpora or small datasets with some kind of paraphrase-related annotation. Sections 3 and 4 describe the two components of our annotation infrastructure: the annotation scheme and the inter-annotator agreement measures, respectively. Section 5 sets out the figures for the three annotated corpora, as well as a discussion and error analysis. Finally, conclusions and future work appear in Section 6.

[1] See Madnani and Dorr (2010), Section 5 for a discussion on this topic.

2 State of the Art on Paraphrase-related Annotations

Besides corpora containing paraphrase pairs sometimes with manual yes/no annotations, such as the Microsoft Research Paraphrase corpus – MSRP (Dolan and Brockett, 2005),[2] or manual alignments at word or phrase level, such as the corpus developed by Cohn et al (2008),[3] there exist some works that have gone further in paraphrase or paraphrase-related annotations. In this section, we focus on such works.[4]

Bhagat (2009) presents a paraphrase typology of 25 lexical changes (e.g., actor/action substitution or noun/adjective conversion) and 3 structural modifications that can accompany them (substitution, addition/deletion, and permutation). He empirically quantifies the distributions of the types annotating a small dataset: 30 sentences from the MSRP corpus and 30 sentences from the Multiple-Translation Chinese corpus (MTC).[5] Regarding the lexical changes, he took advantage of the alignments by Cohn et al (2008) and broke the sentences into 145 and 210 phrases, respectively. These phrases were the units used for annotation. Regarding the structural changes, he annotated the entire sentences allowing more than one phenomena per sentence.

Using a set of 60 paraphrases in French created by 60 people reformulating the same sentence, Fuchs (1988) analyses the formal mechanisms in paraphrases, as well as the changes in thematization and referential values. She states that paraphrases are the result of four formal operations that can be combined: substitution, deletion, movement, and addition. Note that these basic operations coincide with the structural changes in Bhagat (2009) above. Moreover, from a different framework and pursuing different objectives, Vila and Dras (2012), after representing the MSRP corpus using dependency trees and tree edit distance operations, show that explaining paraphrasing using only substitution, addition, and deletion operations is too simplistic to account for paraphrase complexity.

Dutrey et al (2011) define a typology of local modifications which are present in Wikipedia Correction and Paraphrase Corpus (WiCoPaCo), a corpus of rewritings extracted from the revision history in the French Wikipedia (Max and Wisniewski, 2010).[6] Although it is not a paraphrase typology, it accounts for the degree of semantic variation of the types and includes rephrasings, which roughly correspond to paraphrases. The authors present the results of the manual annotation of 200 pairs of modification segments from WiCoPaCo. The annotation scheme consisted of four main classes based on the typology: surface corrections, rephrasings, strong semantic variations, and misalignments. Each annotation had to cover the entire segment and it was possible to assign several labels to the same segment. After the annotation,

[2] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

[3] http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html

[4] See Vila et al (submitted) for a more general state of the art on paraphrase corpora.

[5] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2002T01

[6] http://wicopaco.limsi.fr/

they observe that rephrasings have the largest number of occurrences, followed by strong semantic variations.

Liu et al (2010) present Paraphrase Evaluation Metric (PEM), which evaluates the quality of paraphrases and that of paraphrase generation systems. This metric is based on three criteria: adequacy (semantic similarity), fluency, and lexical dissimilarity. For validation purposes, they manually annotated 1,200 paraphrases, some of them created by humans, some of them automatically built. The MTC corpus was used as a source of paraphrases. The annotation for each paraphrase pair consisted of four scores, each given in a five-point scale: the above three criteria plus an overall score.

The dataset of the Semantic Textual Similarity task in Semeval 2012 is also relevant, although it does not contain paraphrase-type annotations but information about paraphrasability.[7] It consists of 5,250 sentence pairs coming from different sources, the MSRP corpus among them, annotated from 0 to 5 according to their degree of semantic similarity.

The corpora and datasets presented in this section are very different in nature. They do not necessarily consist of annotations with paraphrase types and, if they do, they are very small datasets and/or types are too coarse-grained. Nevertheless, all of them are small steps moving forward in the field of paraphrase-related annotations.

3 The Annotation Scheme

Our annotation infrastructure consists of an annotation scheme and inter-annotator agreement measures. In this section, we focus on the former. It comprises a set of 24 paraphrase-type tags and instructions to annotate the scope of each of these tags within the paraphrase pairs. This annotation scheme was specified in the annotations guidelines[8] and is based on our paraphrase typology (Barrón-Cedeño et al, 2013, to appear).

The development of paraphrase corpora annotated with types does not only involve building a consistent annotation scheme, it also has to be backed by solid linguistic bases: "annotations are not substitute for the understanding of a phenomenon. They are an encoding of that understanding" (Zaenen, 2006). In this sense, our work relies on our thoughts and proposals on the paraphrase phenomenon presented in Recasens and Vila (2010) and Vila et al (2011).

[7] `http://www.cs.york.ac.uk/semeval-2012/task6/`. Although Semeval organizers distinguish between semantic textual similarity and paraphrasing, being the former a sort of graded paraphrasing, this distinction is not relevant here.

[8] Annotation guidelines are available at `http://clic.ub.edu/corpus/en/paraphrases-en`.

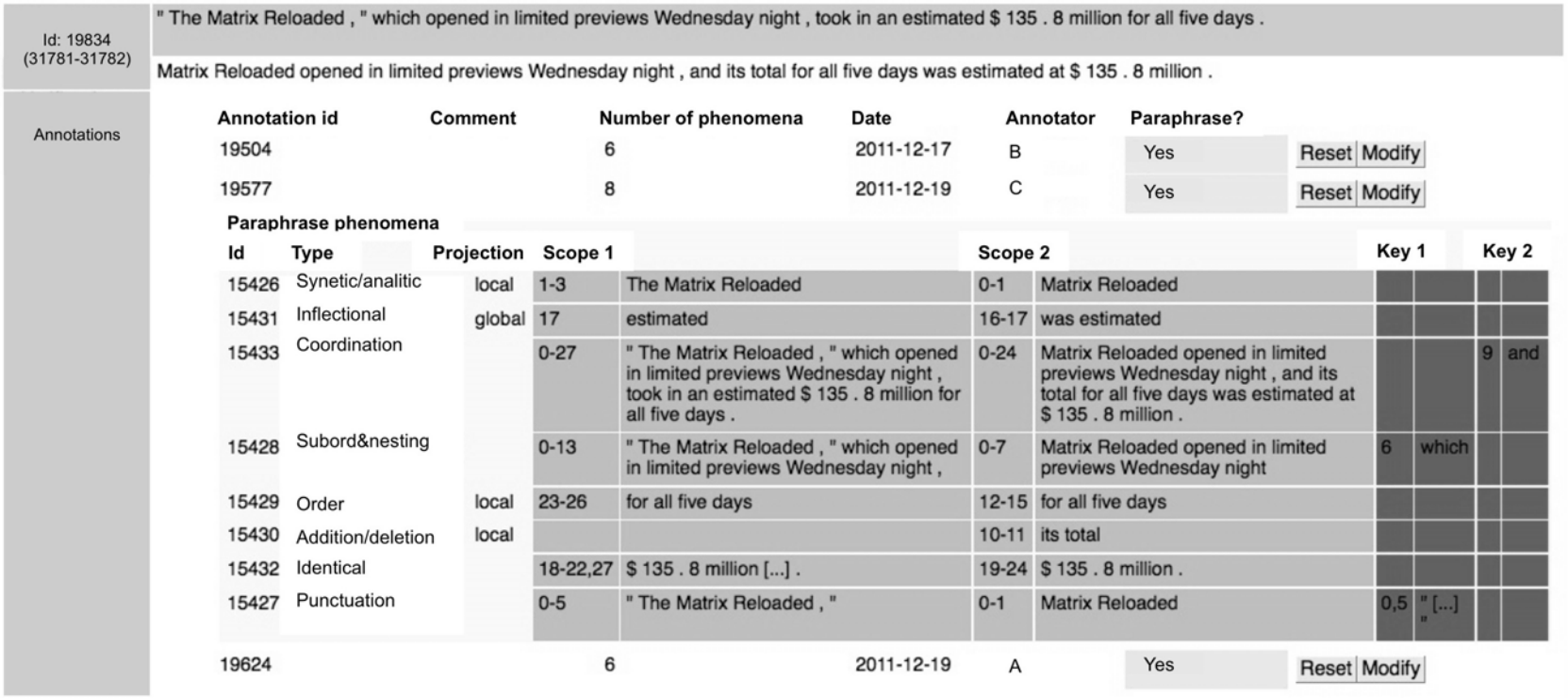

Fig. 1 Annotation example from the MSRP-A. It consists of a screenshot of the CoCo interface used for annotation (See Section 5) with some terminology adaptations for consistency in this article. The paraphrase pair appears on the top. It has been annotated by three annotators: A, B, and C, although only the annotation by C is displayed. For each annotation, the number of phenomena and a paraphrase judgement (yes/no) can be seen. In the case of C, annotated phenomena are shown, with information about the type and the projection, as well as the scope and key elements corresponding to each member of the pair.

The annotation task comprises two steps: (*i*) the classification of pairs as paraphrases and non-paraphrases and (*ii*) the annotation of paraphrase types within those pairs. Only pairs considered paraphrases in the first step will be subsequently annotated in the second. An example of an annotated paraphrase pair from the MSRP-A corpus can be seen in Figure 1. This is used to illustrate the explanation below.[9]

Regarding (*i*) the classification of the pairs as paraphrases or non-paraphrases, we consider paraphrase pairs to be those containing, at least, one paraphrase unit. We consider as paraphrase units those having the same or an equivalent propositional content: the core meaning is the same, although more peripherical aspects of meaning may vary. As can be seen, paraphrase pairs may contain only a fragment that is a paraphrase, regardless of the content of the rest of the pair. This decision was taken in order not to disregard paraphrase fragments within sentences that are not full paraphrases. The subsequent annotation with paraphrase types will make it possible to distinguish the non-paraphrase fragments within these sentences with the NON-PARAPHRASE tag (see below).[10] In the example in Figure 1, the three annotators annotated the pair as a paraphrase.

Regarding (*ii*) type annotation, the units we annotate are atomic paraphrase phenomena within possibly complex paraphrase pairs. Each paraphrase phenomenon is assigned a tag (the type) and a scope (the corresponding fragment in one member of the pair and the corresponding fragment in the other). In Figure 1, eight paraphrase phenomena within the paraphrase pair are displayed for the annotator C.

The typology on which the tagset is based consists of a three-level typology of 5 classes, 4 sub-classes, and 24 paraphrase types (light grey, dark grey, and ticked in Figure 2, respectively). Paraphrase types refer to the linguistic phenomena underlying paraphrases, classes and sub-classes group them according to the level or sphere of language where they arise.[11] The tagset used for annotation corresponds to the 24 paraphrase types.

In most of the cases, paraphrase phenomenon scopes correspond to standard linguistic units (e.g., phrase or clause), such as the nominal phrase in SYNTHETIC/ANALYTIC in Figure 1.[12] Scopes can be discontinuous, such as the case of IDENTICAL in the same figure (indicated by [...]). Also, scopes corresponding to different paraphrase phenomena can overlap: in our example, SYNTHETIC/ANALYTIC overlaps with PUNCTUATION. Finally, the scope affects the annotation task differently depending on the class:

[9] All the examples in this article are extracted from the three annotated corpora, namely P4P, MSRP-A, and WRPA-A. Typos in the original corpora have not been corrected.

[10] It should be taken into account that corpora we annotate consist of positive cases of paraphrasing; therefore, non-paraphrases or non-paraphrase fragments are a minority.

[11] See Barrón-Cedeño et al (2013, to appear) for a more detailed presentation of our paraphrase typology.

[12] We refer to the tags with small capital letters and sometimes using short names, e.g., SYNTHETIC/ANALYTIC for synthetic/analytic substitutions.

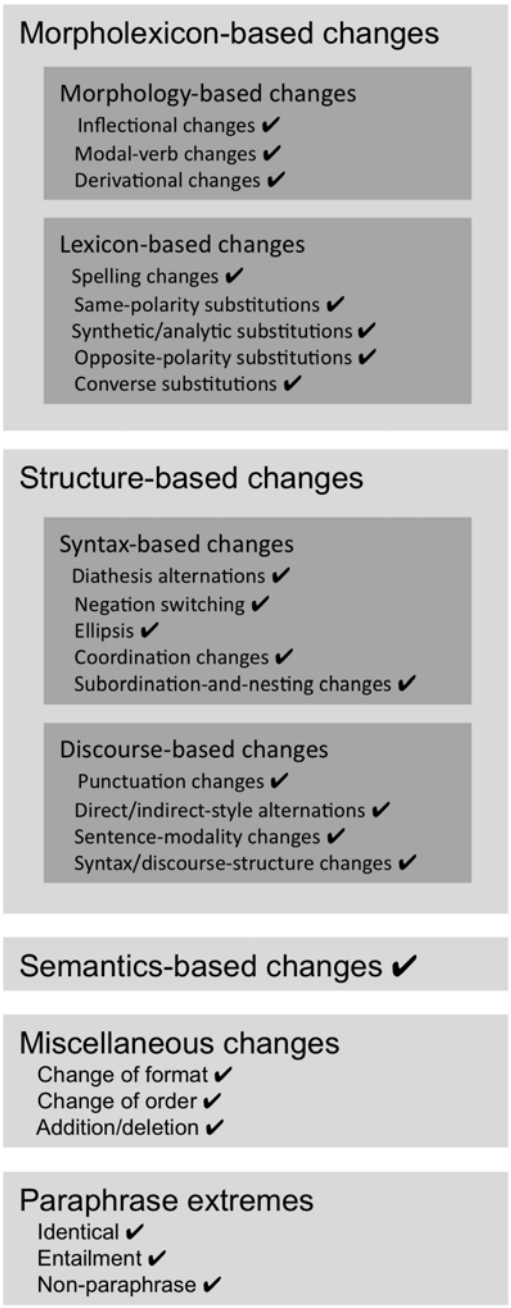

Fig. 2 Three-level paraphrase typology. Types are indicated with a tick.

Morpholexicon-based changes, semantics-based changes, and miscellaneous changes. Only the affected linguistic unit(s) is(are) tagged. As some of these changes entail other changes (mainly inflectional or structural), the annotators can choose between two different facets for each phenomenon: LOCAL, which stands for those cases not entailing other changes; and GLOBAL, which stands for those cases entailing them. For these entailed changes, neither the type of change nor the fragment undergoing the change are specified in the

annotation. We call this distinction between LOCAL/GLOBAL *projection* and it is compulsory for the tags in this group. By way of illustration, in Figure 1, a change of order (ORDER tag) without any other implication takes place, so the LOCAL attribute is used.

Structure-based changes. The whole linguistic unit undergoing the syntactic or discourse reorganisation is tagged. Moreover, most structure-based changes have a key element that gives rise to the change and/or distinguishes it from others. As the scope of structure-based changes is generally long, key elements allow for the identification of the structural change the annotator is referring to. Contrary to the projection, key elements are not compulsory. In Figure 1, the two full sentences in the pair constitute the scope of the coordination change (COORDINATION) and the conjunction *and* stands for the key element.

Paraphrase extremes. No projections or key elements are used, and only the affected fragment is tagged. A case of IDENTICAL can be seen in Figure 1.

4 Inter-annotator Agreement

Due to the complexity of the task, one of the main challenges in paraphrase-type annotation is to guarantee the quality of the resulting corpora. We measure the quality in terms of inter-annotator agreement, which corresponds to the second component in our annotation infrastructure.

Inter-annotator agreement is mostly calculated through observed agreement (Fleiss, 1981) or the kappa measure (Cohen, 1960). However, when the task is complex and the global score is the result of the combination of heterogeneous partial scores computed over smaller units, getting global agreement is rare and so these measures are close to 0. Moreover, the use of these smaller units increases the difficulty of computing agreement: they have to be carefully selected according to the task and the way to combine partial scores has to be set. These smaller units have been used in several NLP tasks, such as automatic summarization evaluation: ROUGE (Lin and Hovy, 2003), Basic Elements (Hovy et al, 2006), and the Pyramid method (Nenkova and Passonneau, 2004) have been widely used in DUC and TAC contests.[13] More recently, ORANGE (Lin and Och, 2004) and QARLA (Amigó et al, 2006) have been proposed as a way of combining heterogeneous measures and raters. Although these measures were created for evaluation, their application to inter-annotator agreement is straightforward. Another illustrative example, where the task of computing inter-annotator agreement is not a trivial one, is the case of annotations consisting of selecting a subset of tags from a set of interdependent ones. Kupper and Hafner (1989) proposed an agreement metric for these cases, derived from kappa. Cohn et al (2008) argue for the usefulness of using this metric for the case of paraphrasing.

[13] `http://www.nist.gov/tac/`

Comparing paraphrase annotations involving multiple pieces with variable type, scope, projection, and key elements is a challenge, and there are no established approaches to do it. To fill this gap, we created the Inter-annotator Agreement for Paraphrase Type Annotation measures (IAPTA). These are ranged in $[0, 1]$ and classified in three groups of increasing granularity level:

- **Number measures (N-measures)**. They compute agreement of the total number of annotated tokens or phenomena, sometimes filtered by type.
- **Total/Partial-Overlapping measures (TPO-measures)**. They compute agreement taking into account the type and the full or partial overlapping of the scope. They are based on evaluation measures in Dale and Narroway (2011).
- **Degree-Overlapping measures (DO-measures)**. They compute agreement taking into account the type, the degree of overlapping of the scope, the projection, and the key elements.

Althought the measures relevant to our work are DO-measures, because they are the most precise, we present more coarse-grained measures, namely N- and TPO-measures, because they may be useful for other approaches to paraphrase-type annotation, which are less precise in terms of scope and less costly in terms of human effort. N-measures and TPO-measures can be considered to be a baseline for DO-measures.

In what follows, we present each of these measures and illustrate them through two paraphrase pairs from the MSRP-A corpus annotated by B and C. Figure 3 shows these annotated pairs and Table 1 sets out the corresponding IAPTA-measure scores.

4.1 N-measures

N-measures are the most coarse-grained IAPTA measures. They only take into account the total number of tokens covered by the scope of paraphrase phenomena or the total number of phenomena annotated, sometimes filtered by type. Projection or key elements are not considered.

In concrete, agr_n (Equation 1) is the ratio between the number of tokens (in this case, agr_n is called agr_w)[14] or phenomena (agr_{ph}) annotated by B (n_B) and C (n_C).

$$agr_n = \frac{\min(|n_B|, |n_C|)}{\max(|n_B|, |n_C|)} \tag{1}$$

[14] We use the subindex w (words) instead of t (tokens) in order to avoid confusion with the superindex t (type) that will appear in what follows.

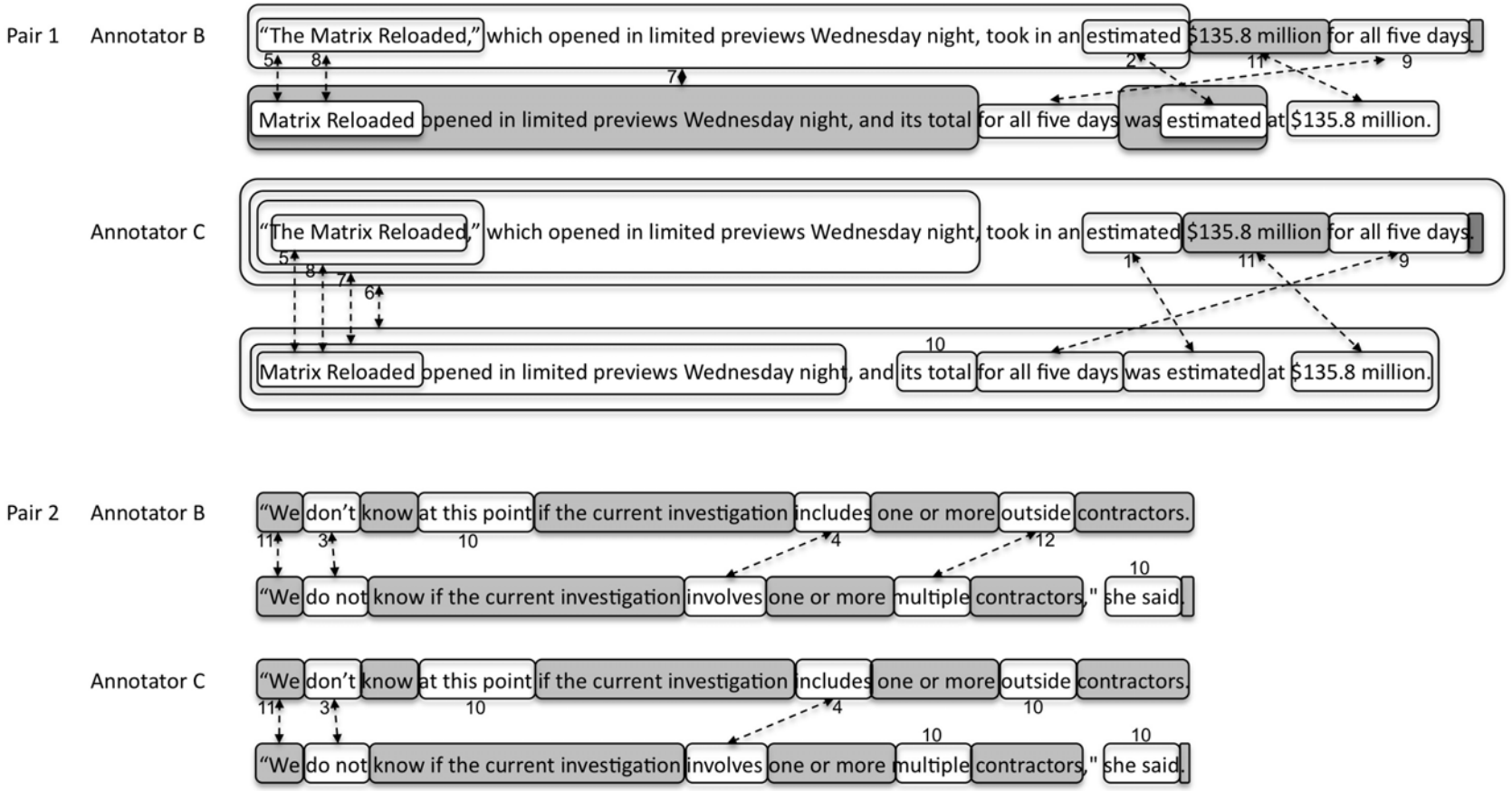

Fig. 3 Annotation examples from the MSRP-A. Pair 1 corresponds to the pair in Figure 1. The meaning of the numbers is as follows: 1. INFLECTIONAL 2. DERIVATIONAL 3. SPELLING 4. SAME-POLARITY 5. SYNTHETIC/ANALYTIC 6. COORDINATION 7. SUBORDINATION&NESTING 8. PUNCTUATION 9. ORDER 10. ADDITION/DELETION 11. IDENTICAL 12. NON-PARAPHRASE. Shadowed boxes stand for discontinuous scopes. For the sake of readability, projection and key elements are not included.

Each agr_n measure can also be computed for each paraphrase type independently (agr_n^t) and combined into a global score by averaging ($agr_n^{\overline{t}}$). Moreover, it can be computed for each paraphrase pair (agr_n^p) and then all the pairs averaged ($agr_n^{\overline{p}}$). All the possibles combinations result in 8 measures. In order to calculate $agr_n^{\overline{p},\overline{t}}$, we first compute $agr_n^{\overline{p},t}$ for each type independently and then the average of all types ($agr_n^{\overline{p},\overline{t}}$). By way of illustration, the scores corresponding to Figure 3 are set out in Table 1.

N-measures	
agr_w	0.98
$\text{agr}_w^{\overline{p}}$	0.98
$\text{agr}_w^{\overline{t}}$	0.56
$\text{agr}_w^{\overline{p},\overline{t}}$	0.57
agr_{ph}	0.81
$\text{agr}_{ph}^{\overline{p}}$	0.81
$\text{agr}_{ph}^{\overline{t}}$	0.62
$\text{agr}_{ph}^{\overline{p},\overline{t}}$	0.60

TPO-measures	
agr_o^p	0.76
agr_o^t	0.55

DO-measures	
F1	0.74

Table 1 IAPTA-measure scores for annotation examples in Figure 3.

4.2 TPO-measures

TPO-measures are based on the evaluation measures in Dale and Narroway (2011), which were used in the pilot round of the Helping Our Own (HOO) shared task.[15] HOO aims to promote the development of automated tools and techniques that can assist authors in the writing task. Systems participating in the task had to detect errors and infelicities in texts, indicate their extent and optionally a type, and correct them. For evaluation, a set of gold-standard edits were compared with the set of edits corresponding to the participating team's output. Three scoring measures were used: (*i*) detection, which indicates whether the system determined that an edit is required at some point in the text; (*ii*) recognition, which indicates whether the system correctly determined the extent of the source text that requires editing; and (*iii*) correction, which indicates whether the system offered a correction that is amongst the

[15] `http://clt.mq.edu.au/research/projects/hoo/hoo2011/index.html`. See also Dale and Kilgarriff (2011) and Dale and Narroway (2012).

corrections provided in the gold standard. For each of these measures, they computed precision, recall, and F1.

We adapted the measures relevant to our work, that is, detection and recognition, which gave rise to agr_o^p and agr_o^t, respectively. The agr_o^p measure accounts for paraphrase phenomena of the same type that at least (p)artially (o)verlap in scope; the agr_o^t measure, in turn, accounts for phenomena that (t)otally (o)verlap. Positive cases in agr_o^t are also considered in agr_o^p, but not vice versa. Projection and key elements are not considered. The scores corresponding to Figure 3 are again set out in Table 1.

Our approach differs from Dale and Narroway (2011) in these main aspects:

- They compute precision, recall, and F1 of the systems' edits compared to the gold-standard ones. Their approach is, therefore, directional. As inter-annotator agreement lacks directionality, we compute the precision, recall, and F1 of annotator B taking C as gold standard, and vice versa.
- Their unit for comparison are "fragments"; in concrete, they have 19 fragments of approximately 1,000 words in length with gold standard edits and several systems' output. Our unit of comparison are pairs of snippets annotated by different annotators (see Table 3 for the figures corresponding to each corpus). To handle this, we concatenate the two members of the paraphrase pair into a single fragment.
- Their scores are calculated on a fragment-by-fragment basis and on a dataset as a whole (computing the average across the fragments). As in the case of N-measures, we calculate the scores in a pairwise and a non-pairwise way. In this work, we use non-pairwise scores.
- Participants were not required to indicate the type of error and this feature was not evaluated in that round. Type annotation was only used, when present, to obtain scores filtered for the individual types. As we are interested in taking types into account within our inter-annotator agreement calculation, we only consider the overlapping between paraphrase phenomena of the same type. In cases where a paraphrase phenomenon only overlaps with phenomena of a different type, we consider there is no overlapping.
- Their distinction between optional and mandatory edits is not relevant to our work.
- They work at character level; we work at token level.

Dale and Narroway (2011) state that a possible improvement to their proposal would be "modifying scoring regime to give partial marks depending on the degree of overlap, rather than the current binary correct vs incorrect". This degree of overlap is considered in our DO-measures, presented in the next section.

4.3 DO-measures

The DO-measures are the most fine-grained of the IAPTA measures. For each paraphrase phenomenon annotated, they calculate the degree of overlapping

at token level with annotations of the other annotator of the same type. They do not only account for those annotations that totally or partially overlap, but determine to what extend they coincide. They also consider projection and key elements.

The rationale behind our inter-annotator agreement computation is as follows: let B and C be the set of paraphrase phenomena annotated by annotators $\mathcal{B}$ and $\mathcal{C}$ (we consider independently all the phenomena occurring in all the pairs). For a phenomenon $b \in B$, b_t refers to the type (t), b_{s_i} refers to the scope (s) in the i member of the pair, b_p refers to the projection (p), and b_{k_i} refers to the scope of the key element (k) in the i member of the pair. We define the inter-annotator agreement between B and C as:

$$F_1 = 2 \cdot \frac{K_B \cdot K_C}{K_B + K_C} \quad . \tag{2}$$

K_B is computed as:

$$K_B = \frac{\sum_{b \in B} \min\left(1, \sum_{c \in C} overlapping(b, c)\right)}{|B|} \quad , \tag{3}$$

K_C is computed accordingly. The *overlapping* measure is defined as:

$$overlapping(x, y) = \left\{ \begin{array}{l} 0 \text{ if } x_t \neq y_t; \\ \text{otherwise,} \\ \quad \alpha \cdot \pi \cdot \kappa \cdot (coverage(x_{s_1}, y_{s_1}, 0) + coverage(x_{s_2}, y_{s_2}, 0)) \end{array} \right\} . \tag{4}$$

The nucleous of Equation 4 is the sum of *coverages*. The α, π and κ factors weight this nucleous. In concrete, $\alpha = 1$ for ADDITION-DELETION phenomena and $\alpha = 0.5$ for others (in the case of ADDITION-DELETION only one text fragment, either in member 1 or 2 of the paraphrase pair, exists). The π and κ factors include the information about projection and key elements, respectively.[16] Regarding projection, $\pi = 1$ if $b_p = c_p$; otherwise, $\pi = 0.75$. Regarding key elements,

$$\kappa = \left\{ \begin{array}{l} 1 \text{ if all } b_{k_i} \text{ are empty;} \\ \text{otherwise,} \\ \quad 0.75 + 0.125 \cdot coverage(b_{k_1}, c_{k_1}, 1) + 0.125 \cdot coverage(b_{k_2}, c_{k_2}, 1) \end{array} \right\} .^{17} \tag{5}$$

$Coverage$ is defined as:

$$coverage(x, y, \chi) = \left\{ \begin{array}{l} \chi \text{ if } |x| = 0; \\ \text{otherwise,} \\ \quad \frac{|x \cap y|}{|x|} \end{array} \right\} . \tag{6}$$

The value of χ is 0 in Equation 4 and 1 in Equation 5.[18]

[16] The π and κ factors can be omitted from the calculus if they are not relevant, as in Barrón-Cedeño et al (2013, to appear).

[18] Assigning a different value to χ comes from the fact that, in the case of key elements, we consider a disagreement to be more harmful than their simple omission.

In summary, we compute how $\mathcal{B}$'s annotations are covered by $\mathcal{C}$'s, and vice versa. K_B may be understood as a regression precision taking the annotation by $\mathcal{C}$ as a reference, and a regression recall taking the annotations by $\mathcal{B}$ as a reference. The inverse is true for K_C.

K_B and K_C can be computed typewise and pairwise analogously to agr_n, resulting in the four F_1 measures. However, it should be taken into account that typewise measures ($F_1^{\overline{t}}$ and $F_1^{\overline{p},\overline{t}}$) are not relevant here, as *overlapping* (Equation 4) is only computed over same-type phenomena. If we are interested in a global measure for all the types, we already have the non-typewise ones (F_1 and $F_1^{\overline{p}}$). The difference between pairwise ($F_1^{\overline{p}}$ and $F_1^{\overline{p},\overline{t}}$) and non-pairwise measures (F_1 and $F_1^{\overline{t}}$), in turn, is that pairwise variants give the same importance to each pair independently of the number of annotated phenomena; in the non-pairwise variants, all phenomena contribute equally to the final score independently of which pair they belong to. In our work, we decided to use the non-pairwise and non-typewise measure. The scores corresponding to Figure 3 are set out in Table 1.

5 The Annotated Corpora

The annotation scheme and inter-annotator agreement measures presented above were used to annotate the following corpora:[19]

- The 3,900 paraphrases in the MSRP corpus (see Section 2). This gave rise to the MSRP-A corpus.
- 847 paraphrases in the "simulated" cases of plagiarism in the PAN-PC-10 corpus (Potthast et al, 2010).[20] This gave rise to the P4P corpus, first presented in Barrón-Cedeño et al (2013, to appear).
- 1,000 paraphrases in the authorship cases in the WRPA corpus (Vila et al, submitted).[21] This gave rise to the WRPA-authorship-A corpus (simplified as WRPA-A).

The latter corpus is in Spanish and the other two, in English. These corpora were compiled from different sources and by applying diverging techniques, which make them different in nature. The MSPR corpus was built using simple string edit distance and a heuristic strategy that pairs initial (presumably summary) sentences from different news stories in the same cluster; sentences were collected from the web over a 2-year period. The PAN-PC-10 corpus was created in the plagiarism domain. The "simulated" subset contains paraphrases manually created by reformulation: people were asked to simulate cases of plagiarism by rewording a given text snippet. Finally, WRPA

[19] Annotated corpora are available at `http://clic.ub.edu/corpus/en/paraphrases-en` as a downloadable package and as a search interface.

[20] `http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-10.html`

[21] `http://clic.ub.edu/corpus/en/paraphrases-en`

P4P	(a)	Bonaparte retreated to Lausanne to prepare to go to Mount St. Bernard. The veteran Austrian general did not sufficiently prepare to fight Bonaparte's arrival, as he did not think such an expedition likely.
	(b)	Bonaparte repaired to Lausanne to prepare the expedition of Mount St. Bernard; the old Austrian general could not believe in the possibility of so bold an enterprise, and in consequence made inadequate preparations to oppose it.
MSRP-A	(a)	Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
	(b)	Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
WRPA-A	(a)	AUTHOR es autor de WORK ('is author of')
	(b)	AUTHOR lanzó su primer álbum: WORK ('released his first album')

Table 2 Paraphrase examples from the annotated corpora.

was built extracting relational paraphrases from Wikipedia applying the distributional hypothesis. The subset used for annotation contains paraphrases expressing the authorship relation, that is, the relationship between an author and his work. Paraphrase examples from P4P, MSRP-A, and WRPA-A corpora appear in Table 2.

There are two most prominent differences between the paraphrase pairs in these corpora: (*i*) the level of paraphrasability, that is, their semantic similarity and (*ii*) the level of formal correspondence, in other words, the possibility to isolate the paraphrase phenomena in them (Example (1) in Section 5.1 exemplifies this). Regarding (*i*) , the paraphrasability level in WRPA-A is considerably lower than that of P4P and MSRP-A, because paraphrases in WRPA-A are understood as pairs expressing the same kind of relation, although their semantic content sometimes differs. Regarding (*ii*) , both P4P and MSRP-A were created in reformulation frameworks to a greater o lesser degree: P4P was built precisely through manual reformulations; and MSRP-A contains news talking about the same or related topics and we assume that, in media, there exists reformulation between agencies and newspapers. In contrast, WRPA-A was built by applying the distributional hypothesis. Paraphrases created in reformulation frameworks show a clearer formal mapping than paraphrases created by applying the distributional hypothesis. Features (*i*) and (*ii*) can be seen in the examples in Table 2.

Our annotation scheme can adapt to the multifaceted nature of paraphrasing. In this sense, WRPA-A annotation required some adjustments in the line of criterion simplification. Instead of using NON-PARAPHRASE or ADDITION/DELETION to tag the semantically diverging or non-parallel fragments in the pairs (which would result in an excessive number of less informative tags), only actual paraphrase phenomena were annotated, leaving the remaining fragments without any type of annotation. Also, overlapping between tags

was not allowed and only the most representative tag was annotated on each fragment. Finally, projection and key elements were not used.

Three annotators participated in the annotation of each corpus. They were all linguists, native Spanish speakers with an advanced level of English. The annotation of each corpus was performed in three phases: annotator training, inter-annotator agreement calculation, and final annotation. In the annotator training phase, one set of 50 cases was annotated by all annotators. Then, problems and disagreements were discussed, the guidelines were better specified regarding these issues, and the 50 annotations by one of the annotators was revised to be included in the corpus. In the inter-annotator agreement phase, one set of 100 cases was again annotated by all annotators, the inter-annotator agreement was computed, and a discussion was again held. In the final annotation phase, the remaining cases in each corpus were annotated only once by one of the three annotators. The examples to be annotated in each phase (training, inter-annotator agreement, and final annotation) were randomly selected. CoCo (España-Bonet et al, 2009)[22] was the interface used for annotation.

The typology, the annotation scheme, and the inter-annotator agreement measures are independent up to a point: other tagsets can be used, some features in the annotation can be obviated, and modifications in the metrics are allowed. The annotation infrastructure can be applied to any corpora satisfying the following constraints: (*i*) units to be annotated are paraphrase pairs; (*ii*) the pair is a complex paraphrase where a set of paraphrase phenomena are annotated; (*iii*) each phenomenon is tagged with a paraphrase type from a closed tagset and eventually a scope consisting of a mapping between not necessarily contiguous spans of the two members of the pair.

In what follows, the results of the corpus annotation (Section 5.1) and the inter-annotator agreement scores (Section 5.2) are presented, together with a discussion and error analysis.

5.1 Annotation Results and Discussion

Table 3 shows the figures for the three annotated corpora. The lower results in WRPA-A for average phenomena per pair and per word can be explained by the shorter length of the pair members and the adaptation of the annotation scheme mentioned above.

Table 4 shows the details for each paraphrase type in the three corpora; in concrete, their relative frequencies and average length are displayed. Empty cells are due to different reasons. In P4P, FORMAT and ENTAILMENT are empty because these tags did not exist in that annotation process. In MSRP-A, SENTENCE MODALITY is empty because no cases of change in the modality of the sentences were found there, as sentences in news articles are generally affirmative. In WRPA-A, many tags are empty due to the adaptation of the guidelines mentioned above.

[22] `http://www.lsi.upc.edu/~textmess/`

	P4P	MSRP-A	WRPA-A
Words	83,745	186,616	20,544
Pairs	856	3,900	1,000
Paraphrase phenomena	11,420	22,105	1,332
Word average in pair members	48.92	23.93	10.27
Average phenomena per pair	13.34	5.67	1.33
Average phenomena per word	1.47	1.28	0.36

Table 3 Global figures for the annotated corpora.

Type	P4P		MSRP-A		WRPA-A	
	RF	AL	RF	AL	RF	AL
Inflectional	2.22	1.30	2.78	1.45	3.75	1.07
Modal verb	1.02	2.47	0.83	2.37	0.38	2.00
Derivational	2.29	1.03	0.85	1.05	1.73	1.00
Spelling	3.83	1.58	2.91	2.06		
Same-polarity	**44.41**	1.53	**24.81**	1.75	**53.15**	1.76
Synthetic/analytic	5.86	3.33	4.42	3.53		
Opposite-polarity	0.57	2.65	0.09	2.03		
Converse	0.29	2.09	0.20	1.95		
Diathesis	1.14	13.28	0.73	11.52		
Negation	0.29	11.73	0.09	6.88		
Ellipsis	0.76	11.08	0.30	12.88		
Coordination	1.84	26.02	0.22	14.61		
Subordination&nesting	5.23	18.83	2.14	12.31		
Punctuation	4.71	23.16	3.77	18.09		
Direct/indirect	0.32	20.40	0.30	21.06		
Sentence modality	0.31	18.37				
Syntax/discourse structure	2.74	17.36	1.39	16.33		
Semantic	2.98	9.10	1.53	6.25	**16.22**	3.35
Format			1.10	1.69		
Order	5.04	5.39	3.89	5.98	0.08	2.00
Addition/deletion	**12.91**	1.67	**25.94**	1.41		
Identical	0.88	14.20	**17.54**	13.80	**14.57**	4.15
Entailment			0.37	6.82	4.05	2.33
Non-paraphrase	0.39	15.35	3.81	2.73	6.08	9.15

Table 4 Per-type figures for the annotated corpora. RF=relative frequency (percentage); AL=average length (tokens). Figures above 10 in the RF column are in bold.

Regarding relative frequencies, SAME-POLARITY and ADDITION/DELETION are the most prominent types both in P4P and MSRP-A. This is due to the accessibility of these types in reformulation processes, as they are mechanisms that are relatively simple to apply to a text by humans: changing one lexical unit for its synonym (understanding synonymy in a general sense) and deleting a text fragment, respectively. In P4P, SAME-POLARITY clearly surpasses ADDITION/DELETION, showing the high accessibility of this mechanism in conscious human reformulations. ADDITION/DELETION slightly surpasses SAME-POLARITY in MSRP-A, pointing to the recurrence in adding or deleting certain details depending on the newspaper.

The most frequent type in WRPA-A is again SAME-POLARITY; and, at a considerable distance, we find SEMANTIC, which involves a different lexicalization of the same content units. The nature of the corpus and the adaptation to the guidelines meant that the annotators tended to use the SAME-POLARITY tag when the fragment to map was a single lexical unit and the SEMANTIC tag when it was a more complex unit.

IDENTICAL is the third most frequent type in MSRP-A and WRPA-A, but among the least frequent in P4P. This is due to a change in the way the scope of this tag was marked: in P4P, IDENTICAL was only used when the identical fragment appeared between strong punctuation marks. In the other corpora, all identical fragments in the pair were tagged as a single discontinuous tag. Almost all the sentences have some identical words; therefore, this is a frequent type in MSRP-A and WRPA-A.

Finally, distributions are clearly biased towards two or three types. This can correspond to either a real distribution of paraphrase phenomena or some inertia in the way of annotating. We are confident of the first interpretation because of the relatively high correlation of RFs, with 0.74, 0.63, and 0.86 of Pearson's correlations.

Comparing our distributions with those of Bhagat (2009) (see Section 2), all types appear in our corpora to a greater o lesser extent, which is not the case in Bhagat (2009), where many types are not present in the corpus—in part explained by its small size. Also, in Bhagat (2009)'s resulting distributions, synonymy substitutions, function word variations, and external knowledge have the highest frequency. In structural changes, substitutions and additions/deletions are more frequent than permutations. As can be seen, there are points in common with our results.

Regarding the length of the annotated fragments, the paraphrase types with the greatest average length are those in the class of structure-based changes (see Figure 2). The reason is to be found in the above distinction between the two ways to annotate the scope: in structural reorganizations, we annotate the whole linguistic unit undergoing the change.

One of the difficulties we had to deal with during annotation was that, in some pairs, the rewording made it difficult to isolate the paraphrase phenomena. Example (1) from P4P illustrates this situation. In these cases, we tried to isolate as many paraphrase phenomena as possible, assuming that other changes could remain without annotation. In (1), a SAME-POLARITY between *thought of* and *concieve* can be isolated, among others. When isolating some phenomena was not possible, the SEMANTIC tag was used, as in the sentence in square brackets in (1).

(1) a. No longer was the body *thought of* as just a vessel, it was treated with the most respect and reverence. [From that time on artists have shown the human body to be worth of royalty and utmost fidelity.]

b. Men began to *conceive* that the human body is noble in itself and worthy of patient study. [The object of the artist then became to

unite devotional feeling and respect for the sacred legend with the utmost beauty and the utmost fidelity of delineation.]

5.2 Inter-annotator Agreement Scores and Discussion

Table 5 shows the IAPTA scores for the three annotated corpora. Each column corresponds to the agreement between two annotators. Only one column appears for P4P as only two annotators participated in the inter-annotator agreement phase of this corpus.[23]

The score values are consistent between the three corpora and are in line with our expectations. In N-measures, the scores for agr_w, agr_{ph}, and their pairwised versions are the highest, almost all above 0.90. The scores decrease when analysed by type, generally not being below 0.50.

In TPO- and DO-measures, the scores are in general lower than in N-measures, because TPO- and DO-measures are more fine-grained. In TPO-measures, the scores for agr_o^p are higher (around 0.75) than agr_o^t (around 0.50). DO-measure scores are below agr_o^p and above agr_o^t (nearer agr_o^p). Both TPO- and DO-measures take into account the scope of the phenomena, but do it to different degrees: agr_o^p is the loosest, because, if there is overlapping at some point, whatever its degree, the example is considered positive; agr_o^t is the most strict measure, because it only accepts as positive a total overlapping; finally, DO-measures are not discrete but consider the degree of overlapping.

The scores for the P4P corpus tend to be lower than those of MSRP-A, as the former was more complex to annotate: the pair members were longer and there was a higher concentration of paraphrase phenomena, as shown in Table 3. Despite the lower semantic similarity and formal correspondence of the paraphrase pairs in WRPA-A, which would have made the annotation more complex, scores in WRPA-A are higher due to the simplification in the annotation scheme applied.

Regarding individual scores per type, some aspects should be pointed out.[24] Our types are (*i*) generic (e.g., NEGATION covers multiple and diverse phenomena), but, at the same time, (*ii*) precisely define which linguistic phenomenon they refer to (e.g., it stands for those paraphrases where the negation has changed its position in the sentence). However, one type in our tagset, namely SYNTAX&DISCOURSE, does not fulfill property (*ii*) : this type is a kind of "others" for the structure-based class. As a result, this is one of the types with the lowest inter-annotator agreement. The other tag with the lowest agreement is SEMANTICS, which, up to a point, is again a by-default tag standing for cases involving multiple and varied paraphrase changes. In future work, an analysis of the phenomena annotated under these tags to see whether they accept a

[23] Little differences in the P4P annotation process can be seen in Barrón-Cedeño et al (2013, to appear).

[24] For reasons of space, we do not include the per-type scores of inter-annotator agreement. Instead, we point out the most relevant issues in this respect.

	P4P	**MSRP-A**			**WRPA-A**		
	A–B	A–B	A–C	B–C	A–B	A–C	B–C
N-measures							
agr_w	0.96	0.99	0.99	1.00	0.91	0.95	0.96
$\text{agr}_w^{\overline{p}}$	0.96	0.99	0.99	1.00	0.91	0.94	0.97
$\text{agr}_w^{\overline{t}}$	0.65	0.62	0.59	0.69	0.70	0.79	0.79
$\text{agr}_w^{\overline{p},\overline{t}}$	0.65	0.63	0.60	0.69	0.70	0.77	0.78
agr_{ph}	0.98	0.98	0.99	0.99	0.97	0.99	0.96
$\text{agr}_{ph}^{\overline{p}}$	0.85	0.89	0.88	0.88	0.92	0.92	0.91
$\text{agr}_{ph}^{\overline{t}}$	0.67	0.64	0.65	0.74	0.70	0.84	0.79
$\text{agr}_{ph}^{\overline{p},\overline{t}}$	0.36	0.49	0.42	0.48	0.56	0.52	0.60
TPO-measures							
agr_o^p	0.67	0.80	0.77	0.77	0.78	0.75	0.77/0.78
agr_o^t	0.42	0.54/0.55	0.49	0.51	0.55	0.63	0.53
DO-measures							
F1	0.62	0.74	0.73	0.73	0.73	0.75	0.74

Table 5 IAPTA-measure scores for the three corpora. Each column corresponds to the agreement between two annotators. For each TPO-measure score, there are two values: one taking annotator X as gold standard, the other taking as gold standard annotator Y. The values for both directions tend to be the same, so generally only one score appears.

more fine-grained classification could be performed. The types with the highest agreement are IDENTICAL and SAME-POLARITY.

In our work, we consider the most precise and adequate measure to be F1 in DO-measures. The scores obtained for that measure (generally above 0.70) are satisfactory given the difficulty of the task: it requires thorough annotator training and even experienced annotators make errors due to the complexity in the annotation of some paraphrase pairs (long snippets, high paraphrase phenomena density) and the number of features they have to take into account (type, scope, projection, and key elements). Moreover, we cannot avoid some degree of subjectivity: on occasions, different ways to annotate the same phenomenon are acceptable depending on the perspective (see "false negatives" below). In a much simpler task, the binary decision of whether two sentences are paraphrases in the MSRP corpus, a similar agreement was obtained (Dolan and Brockett, 2005).

It should be pointed out that, at the end of the MSRP-A annotation process, we performed a new inter-annotator agreement calculation with a new set of 100 cases. The scores of F1 in DO-measures are 0.79, 0.78, and 0.78, respectively. This results are slightly higher than those corresponding to the first inter-annotator agreement phase (Table 5), which shows that the annotation guidelines succeed in its cohesive function by reducing disagreements.

Finally, we performed a manual analysis of a sample of annotated pairs in the inter-annotator agreement set. We classified the infelicities found into

two classes: false negatives and false positives, which stand for complementary situations.

False negatives are those cases considered to be disagreements in the inter-annotator agreement calculation when they should not be considered as such in absolute terms. They are due to the assumption in the inter-annotator agreement formulae that, when different-type tags by two annotators overlap, they are not referring to the same phenomenon. However, this is not always true. In the example in (2) from the P4P corpus, *regular soldiers/soldiers* was annotated by B as a change from an analytic structure to a synthetic one (SYNTHETIC/ANALYTIC tag); in contrast, A used an ADDITION/DELETION tag for *regular*. In our calculation, it is considered that B lacks an ADDITION/DELETION tag and A, a SYNTHETIC/ANALYTIC one. Nevertheless, although annotators define it differently, both tags refer to the same phenomenon and they are not contradictory. Therefore, it would be better to consider these cases as partial agreement.

(2) a. [...] *Regular soldiers* and the militia maintained order and discipline [...]
b. [...] *Soldiers* and militia kept everyone in line [...]

These cases cannot be solved straightforwardly, because different-type overlapping between annotators is also due to different phenomena that simply occur together, and these two types of overlapping are not easily automatically distinguishable. We are therefore forced to accept that there is some hidden agreement in our scores.

False positives are those cases erroneously considered as agreements. They are due to the assumption in the inter-annotator agreement formulae that, when same-type tags by two annotators overlap, they refer to the same phenomenon. This is not always true. In the example (3) from P4P, B annotates with the PUNCTUATION tag the absence of a comma before *and* in (3-a) versus its presence in (3-b) (the corresponding scope appears in curly brackets); A annotates with the same tag the change from the full stop before *taxes* in (3-a) to the comma before *those* in (3-b) (scope in square brackets). The corresponding punctuation marks are the key elements of the annotations, which allows us to detect the annotators' intention. As there exists same-tag overlapping, these cases are considered positive in the calculation when they should not be, as they are referring to different phenomena.

(3) a. [...] [He remitted the excise duties on beer, {cider and leather}$_B$. Taxes on spirits were increased.]$_A$
b. [...] [The excise duties on beer, {cider, and leather}$_B$ were now totally remitted, those on spirits being somewhat increased.]$_A$

A possible way to solve this problem would be to discard those cases with excluding key elements. However, once again, this is not a straigthforward task, due to the relatively freedom in key element annotations and their variabiliy.

Moreover, only some paraphrase types have key elements. It should be pointed out that false positives are rare by their very nature.

These infelicities have slightly biased our results. Given the nature of each of them, we assume that false negatives are more frequent. Therefore, the bias affects our results negatively. Addressing these issues is left for future work.

6 Conclusions and Future Work

In this article, we have presented a new annotation infrastructure for paraphrase-type annotation consisting of an annotation scheme and inter-annotator agreement measures, as well as three corpora annotated accordingly. The main components in the annotation scheme are a tagset and instructions on how to annotate the scope of each paraphrase phenomena; the IAPTA measures, in turn, compute agreement at different levels of granularity. The annotation of such diverse corpora as P4P, MSRP-A, and WRPA-A, which are different in nature and in two languages, has demonstrated the comprehensiveness of the annotation scheme. IAPTA measures, in turn, have shown the quality of the annotations and the adequacy of the annotation scheme to annotate new paraphrase corpora.

Paraphrasing presents multiple and diverse linguistic manifestations; therefore, this type of resource shows a great potential in order to better understand the linguistic nature of paraphrasing and to go a step further towards solving the puzzle of paraphrasing in NLP. In concrete, these corpora constitute a powerful resource for machine learning and a source for deriving new tools, such as paraphrase lexicons. These annotated corpora show which are the most frequent paraphrase types and, consequently, where to put the focus in improving NLP systems. In this sense, the P4P corpus has already been used to determine the most frequent paraphrase types in plagiarism and which types are the most difficult to detect for plagiarism detection systems (Barrón-Cedeño et al, 2013, to appear).

This is the first time that this type of annotation infrastructure and corpora have been built, which makes our work experimental. They constitute a primary step in an almost unexplored field and open the path to new proposals and improvements. In concrete, further work could be done in (*i*) seeing whether the most coarse-grained tags in our proposal (SYNTAX&DISCOURSE and SEMANTICS) accept a more fine-grained classification and (*ii*) solving the issue of false positives and false negatives in the IAPTA measures.

Acknowledgements We are grateful to the people that participated in the annotation of the corpora: Rita Zaragoza, Montse Nofre, Patricia Fernández, and Oriol Borrega. We would also like to thank Alberto Barrón-Cedeño for his help in shaping inter-annotator agreement measure formulae. This work is supported by the Spanish government through the projects DIANA (TIN2012-38603-C02-02) and SKATER (TIN2012-38584-C06-01) from Ministerio de Ciencia e Innovación, as well as a FPU grant (AP2008-02185) from Ministerio de Educación, Cultura y Deporte.

References

Amigó E, Giménez J, Gonzalo J, Màrquez L (2006) MT evaluation: Human-like vs. human acceptable. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Sydney, pp 17–24

Barrón-Cedeño A, Vila M, Martí MA, Rosso P (2013, to appear) Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics 39(4), doi: 10.1162/COLI_a_00153

Bhagat R (2009) Learning Paraphrases from Text. PhD thesis, University of Southern California, Los Angeles

Chen DL, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011), Portland, vol 1, pp 190–200

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46

Cohn T, Callison-Burch C, Lapata M (2008) Constructing corpora for the development and evaluation of paraphrase systems. Computational Linguistics 34(4):597–614

Dale R, Kilgarriff A (2011) Helping Our Own: The HOO 2011 pilot shared task. In: Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011), Nancy, pp 242–249

Dale R, Narroway G (2011) The HOO pilot data set: Notes on release 2.0. Resource document. `http://clt.mq.edu.au/research/projects/hoo/hoo2011/files/HOOReleaseNotes20110621.pdf`. Accessed 8 February 2013

Dale R, Narroway G (2012) A framework for evaluating text correction. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, pp 3015–3018

Dolan WB, Brockett C (2005) Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005), Jeju Island, pp 9–16

Dutrey C, Bernhard D, Bouamor H, Max A (2011) Local modifications and paraphrases in Wikipedia's revision history. Procesamiento del Lenguaje Natural 46:51–58

España-Bonet C, Vila M, Rodríguez H, Martí MA (2009) CoCo, a web interface for corpora compilation. Procesamiento del Lenguaje Natural 43:367–368

Fleiss JL (1981) Statistical methods for rates and proportions. John Wiley, New York

Fuchs C (1988) Paraphrases prédicatives et contraintes énonciatives. In: G Bès G, Fuchs C (eds) Lexique et Paraphrase, no. 6 in Lexique, Presses Universitaires de Lille, Villeneuve d'Ascq, pp 157–171

Hovy E, Lin CY, Zhou L, Fukumoto J (2006) Automated summarization evaluation with basic elements. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, pp 899–902

Kupper LL, Hafner KB (1989) On assessing interrater agreement for multiple attribute responses. Biometrics 45(3):957–967

Lin CY, Hovy E (2003) Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2003), Edmonton, vol 1, pp 71–78

Lin CY, Och FJ (2004) ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In: Proceedings of the 20th international conference on Computational Linguistics (COLING 2004), Geneva

Liu C, Dahlmeier D, Ng HT (2010) PEM: A paraphrase evaluation metric exploiting parallel texts. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), Cambridge, MA, pp 923–932

Madnani N, Dorr BJ (2010) Generating phrasal and sentential paraphrases: A survey of data-driven methods. Computational Linguistics 36(3):341–387

Max A, Wisniewski G (2010) Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, pp 3143–3148

Nenkova A, Passonneau R (2004) Evaluating content selection in summarization: the pyramid method. In: Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2004), Boston, pp 145–152

Potthast M, Stein B, Barrón-Cedeño A, Rosso P (2010) An evaluation framework for plagiarism detection. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, pp 997–1005

Recasens M, Vila M (2010) On paraphrase and coreference. Computational Linguistics 36(4):639–647

Vila M, Dras M (2012) Tree edit distance as a baseline approach for paraphrase representation. Procesamiento del Lenguaje Natural 48:89–95

Vila M, Martí MA, Rodríguez H (2011) Paraphrase concept and typology. A linguistically based and computationally oriented approach. Procesamiento del Lenguaje Natural 46:83–90

Vila M, Rodríguez H, Martí MA (submitted) Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus

Zaenen A (2006) Mark-up barking up the wrong tree. Computational Linguistics 32(4):577–580

Chapter 4

Paraphrasing in Automatic Plagiarism Detection

4.1 Paraphrasing in Automatic Plagiarism Detection

Alberto Barrón-Cedeño (Universitat Politècnica de València)
Marta Vila (Universitat de Barcelona)
M. Antònia Martí (Universitat de Barcelona)
Paolo Rosso (Universitat Politècnica de València)

Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection

To appear in *Computational Linguistics*, DOI: 10.1162/COLI_a_00153.

Journal URL `http://www.mitpressjournals.org/loi/coli`
Impact Factor 0.721

Abstract Although paraphrasing is the linguistic mechanism underlying many plagiarism cases, little attention has been paid to its analysis in the framework of automatic plagiarism detection. Therefore, state-of-the-art plagiarism detectors find it difficult to detect cases of paraphrase plagiarism. In this article, we analyse the relationship between paraphrasing and plagiarism, paying special attention to which paraphrase phenomena underlie acts of plagiarism and which of them are detected by plagiarism detection systems. With this aim in mind, we created the P4P corpus, a new resource which uses a paraphrase typology to annotate a subset of the PAN-PC-10 corpus for automatic plagiarism detection. The results of the Second International Competition on Plagiarism Detection were analysed in the light of this annotation.

The presented experiments show that (*i*) more complex paraphrase phenomena and a high density of paraphrase mechanisms make plagiarism detection more difficult, (*ii*) lexical substitutions are the paraphrase mechanisms used the most when plagiarising, and (*iii*) paraphrase mechanisms tend to shorten the plagiarized text. For the first time, the paraphrase mechanisms behind plagiarism have been analysed, providing critical insights for the improvement of automatic plagiarism detection systems.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Plagiarism meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection

Alberto Barrón-Cedeño [*†]
Universitat Politècnica de Catalunya

Marta Vila [** †]
Universitat de Barcelona

M. Antònia Martí [‡]
Universitat de Barcelona

Paolo Rosso [§]
Universitat Politècnica de València

Although paraphrasing is the linguistic mechanism underlying many plagiarism cases, little attention has been paid to its analysis in the framework of automatic plagiarism detection. Therefore, state-of-the-art plagiarism detectors find it difficult to detect cases of paraphrase plagiarism. In this article, we analyse the relationship between paraphrasing and plagiarism, paying special attention to which paraphrase phenomena underlie acts of plagiarism and which of them are detected by plagiarism detection systems. With this aim in mind, we created the P4P corpus, a new resource which uses a paraphrase typology to annotate a subset of the PAN-PC-10 corpus for automatic plagiarism detection. The results of the Second International Competition on Plagiarism Detection were analysed in the light of this annotation.

*The presented experiments show that (*i*) more complex paraphrase phenomena and a high density of paraphrase mechanisms make plagiarism detection more difficult, (*ii*) lexical substitutions are the paraphrase mechanisms used the most when plagiarising, and (*iii*) paraphrase mechanisms tend to shorten the plagiarized text. For the first time, the paraphrase mechanisms behind plagiarism have been analysed, providing critical insights for the improvement of automatic plagiarism detection systems.*

1. Introduction

Plagiarism is the re-use of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source (IEEE 2008). While plagiarism may occur incidentally, it is often the outcome of a conscious process. Inde-

∗ TALP Research Center, Jordi Girona Salgado 1-3, 08034 Barcelona, Spain. E-mail: albarron@lsi.upc.es
∗∗ CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: marta.vila@ub.edu
† Both authors contributed equally to this work.
‡ CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: amarti@ub.edu
§ NLE Lab-ELiRF, Department of Information Systems and Computation, Camino de Vera s/n, 46022 Valencia , Spain. E-mail: prosso@dsic.upv.es

Submission received: 13 March 2012; Revised submission received: 17 October 2012; Accepted for publication: 7 November 2012

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

pendently from the vocabulary or channel an idea is communicated through, a person that fails to provide its corresponding source is suspected of plagiarism. The amount of text available in electronic media nowadays has caused cases of plagiarism to increase. In the academic domain, some surveys estimate that around 30% of student reports include plagiarism (Association of Teachers and Lecturers 2008) and a more recent study increases this percentage to more than 40% (Comas et al. 2010). As a result, its manual detection has become infeasible. Models for automatic plagiarism detection are being developed as a countermeasure. Their main objective is assisting people in the task of detecting plagiarism—as a side effect, plagiarism is discouraged.

The linguistic phenomena underlying plagiarism have barely been analyzed in the design of these systems, which we consider to be a key issue for their improvement. Martin (2004) identifies different kinds of plagiarism: of ideas, of references, of authorship, word by word, and praphrase plagiarism. In the first case, ideas, knowledge or theories from another person are claimed without proper citation. In plagiarism of references and authorship, citations and entire documents are included without any mention of their authors. Word by word plagiarism, also known as copy–paste or verbatim copy, consists of the exact copy of a text (fragment) from a source into the plagiarized document. Regarding paraphrase plagiarism, in order to conceal the plagiarism act, a different form expressing the same content is often used. Paraphrasing, generally understood as sameness of meaning between different wordings, is the linguistic mechanism underlying many plagiarism acts and the linguistic process on which plagiarism is based.

In this article, the relationship between plagiarism and paraphrasing, which consists of a largely unexplored problem, is analyzed, and the potential of such a relationship in automatic plagiarism detection is set out. We aim not only to investigate

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

how difficult detecting paraphrase cases for state-of-the-art plagiarism detectors is, but to understand which types of paraphrases underlie plagiarism acts and which are the most difficult to detect.

For this purpose, we created the **Paraphrase for Plagiarism corpus** (P4P) annotating a portion of the PAN-PC-10 corpus for plagiarism detection (Potthast et al. 2010) on the basis of a paraphrase typology, and we mapped the annotation results with those of the Second International Competition on Plagiarism Detection (PAN-10 competition onwards).[1] The results obtained provide critical insights for the improvement of automatic plagiarism detection systems.

The rest of the article is structured as follows. Section 2 sets out the paraphrase typology used in this research work. Section 3 describes the construction of the P4P corpus. Section 4 gives an overview of the state of the art in automatic plagiarism detection; special attention is given to the systems participating in the PAN-10 competition. Section 5 discusses our experiments and the findings derived from mapping the P4P corpus and the PAN-10 competition results. Section 6 draws some conclusions and offers insights for future research.

2. Paraphrase Typology

Typologies are a precise and efficient way to draw the boundaries of a certain phenomenon, identify its different manifestations, and, in short, go into its characterization in depth. Also, typologies constitute the basis of many corpus annotation processes, which have their own effects on the typologies themselves: the annotation process tests the adequacy of the typology for the analysis of the data, and allows for the identification of new types and the revision of the existing ones. Moreover, an annotated

1 http://www.webis.de/research/events/pan-10

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

corpus following a typology is a powerful resource for the development and evaluation of Computational Linguistics systems. In this section, after setting out a brief state of the art on paraphrase typologies and the weaknesses they present, the typology used for the annotation of the P4P corpus is described.

Paraphrase typologies have been addressed in different fields, including Discourse Analysis, Linguistics, and Computational Linguistics, which has originated typologies that are very different in nature. Typologies coming from Discourse Analysis classify paraphrases according to the reformulation mechanisms or communicative intention behind them (Cheung 2009; Gülich 2003), but without focusing on the linguistic nature of paraphrases themselves, which, in contrast, is our main focus of interest. From the perspective of Linguistic Analysis, some typologies are strongly tied to concrete theoretical frameworks, as the case of Meaning–Text Theory (MTT) (Mel'čuk 1992; Milićević 2007). Still in this field, typologies of transformations and diathesis alternations can be considered indirect approaches to paraphrasing in the sense that they deal with equivalent expressions (Harris 1957; Chomsky 1957; Levin 1993). However, they do not cover paraphrasing as a whole, but focus on lexical and syntactic phenomena. Other typologies come from Linguistics-related fields like editing (Faigley and Witte 1981), which is interesting in our analysis because it is strongly tied to paraphrasing.

A number of paraphrase typologies have been built from the perspective of Computational Linguistics. Some of these typologies are simple lists of paraphrase types useful for a specific system or application, or the most common types found in a corpus. They are specific-work oriented and far from being comprehensive: Barzilay, McKeown, and Elhadad (1999), Dorr et al. (2004), and Dutrey et al. (2011), among others. Other typologies classify paraphrases in a very generic way, setting out only two or three types (Barzilay 2003; Shimohata 2004); these classifications do not reach the category of

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

typologies sensu stricto. Finally, there are more comprehensive typologies, such as the ones by Dras (1999), Fujita (2005), and Bhagat (2009). They usually take the shape of very fine-grained lists of paraphrase types grouped into bigger classes following different criteria. They generally focus on these lists of specific paraphrase mechanisms, which will always be endless.

Our paraphrase typology is based on the paraphrase concept defined in Recasens and Vila (2010) and Vila, Martí, and Rodríguez (2011), and consists of an upgraded version of the one presented in the latter. Our paraphrase concept is based on the idea that paraphrases should have the same or an equivalent propositional content, that is, the same core meaning. This conception opens the door to paraphrases sometimes disregarded in the literature, mainly focused on lexical and syntactic mechanisms.

The paraphrase typology attempts to capture the general linguistic phenomena of paraphrasing, rather than presenting a long, fine-grained, and inevitably incomplete list of concrete mechanisms. In this sense, it also attempts to be comprehensive of paraphrasing as a whole: it was contrasted with, and sometimes inspired by, state-of-the-art paraphrase typologies to cover the phenomena described in them;[2] and it was used to annotate (*i*) the plagiarism paraphrases in the P4P corpus (cf. Section 3), (*ii*) 3,900 paraphrases from the news domain in the Microsoft Research Paraphrase corpus (MSRP) (Dolan and Brockett 2005),[3] and (*iii*) 1,000 relational paraphrases (i.e., paraphrases expressing a relation between two entities) extracted from the Wikipedia-based Relational Paraphrase Acquisition corpus (WRPA) (Vila, Rodríguez, and Martí Submitted).[4] P4P and MSRP are English corpora, whereas WRPA is a Spanish one.

2 The list of the consulted typologies can be seen in the Appendix of the annotation guidelines. See footnote 9 for more information.
3 http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/
4 http://clic.ub.edu/corpus/en/paraphrases-en

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

The success in the annotation of such diverse corpora with our paraphrase typology guarantees its adequacy for general paraphrasing not only in English.

The typology is displayed in Fig. 1. It consists of a three-level typology of 20 paraphrase types grouped in 4 classes and 4 sub-classes. Paraphrase types stand for the linguistic mechanism triggering the paraphrase phenomenon. They are grouped in classes according to the nature of such trigger linguistic mechanism: (*i*) those types where the paraphrase phenomenon arises at the morpholexicon level, (*ii*) those that are the result of a different structural organization, and (*iii*) those types arising at the semantic level. Classes inform about the origin of the paraphrase phenomenon, but such paraphrase phenomenon can involve changes in other parts of the sentence. For instance, a morpholexicon-based change (derivational) like the one in (1), where the nominal form *failure* is exchanged for the verb *failed*, has obvious syntactic implications; however, the paraphrase phenomenon is triggered by the morpholexical change.[5] A structure-based change (diathesis) like the one in (2) involves an inflectional change in *heard/hear* among others, but the trigger change is syntactic. Finally, paraphrases in semantics are based on a different distribution of semantic content across the lexical units involving multiple and varied formal changes (3). Miscellaneous changes comprise types not directly related to one single class. Finally, the sub-classes follow the classical organization in formal linguistic levels from morphology to discourse and simply establish an intermediate grouping between some classes and their types.

(1) a. the comical *failure* of the head master's attempt at a "Parents' Committee"

5 All the examples in this article are extracted from the P4P corpus. In some of them, only the fragment we are referring to appears; in others, its context is also displayed (with the fragment in focus in italics). Neither the fragment set out nor italics necessarily refer to the annotated scope (cf. Section 3), although they sometimes coincide. These fragments are not complete cases of plagiarism. Refer to Table 4 to see some entire instances of plagiarism in the P4P corpus.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

CLASS	SUB-CLASS	TYPE
MORPHOLEXICON-BASED CHANGES	**Morphology-based changes**	Inflectional changes Modal verb changes Derivational changes
	Lexicon-based changes	Spelling and format changes Same-polarity substitutions Synthetic/analytic substitutions Opposite-polarity substitutions Converse substitutions
STRUCTURE-BASED CHANGES	**Syntax-based changes**	Diathesis alternations Negation switching Ellipsis Coordination changes Subordination and nesting changes
	Discourse-based changes	Punctuation and format changes Direct/indirect style alternations Sentence modality changes
		Syntax/discourse structure changes
SEMANTICS-BASED CHANGES		Semantics based changes
MISCELLANEOUS CHANGES		Change of order Addition/deletion

Figure 1
Overview of the paraphrases typology, including 4 classes, 4 sub-classes, and 20 types.

b. how the headmaster *failed* at the attempt at a "Parent's Committee"

(2) a. the report of a gun on shore was still heard at intervals

b. We were able to hear the report of a gun on shore intermittently

(3) a. I've got a hunch that we*'re not through with that game yet*

b. I'm guessing we *won't be done for some time*

Although the types in our typology are presented in isolation, they can be combined: in (4), changes of order of the subject (β) and the adverb (γ), and two same-polarity substitutions (*said/answered* [α] and *cautiously/carefully* [γ]) can be observed. A difference

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

between cases such as (4) and, for example, (1) should be noted: in (1), the derivational change implies the syntactic one, so only one single paraphrase phenomenon is considered; in (4), same-polarity substitutions and changes of order are independent and can take place in isolation, so four paraphrase phenomena are considered.

(4) a. "Yes," *[said]*$_\alpha$ *[I]*$_\beta$ *[cautiously]*$_\gamma$

b. "Yes," *[I]*$_\beta$ *[carefully]*$_\gamma$ *[answered]*$_\alpha$

In what follows, types in our typology are briefly described.

Inflectional changes consist of changing inflectional affixes of words. In (5), a plural/singular alternation (*streets/street*) can be observed.

(5) a. it was with difficulty that the course of *streets* could be followed

b. You couldn't even follow the path of the *street*

Modal verb changes are changes of modality using modal verbs, like *might* and *could* in (6).

(6) a. I [. . .] was still lost in conjectures who they *might be*

b. I was pondering who they *could be*

Derivational changes consist of changes of category with or without using derivational affixes. These changes imply a syntactic change in the sentence in which they occur. In (7), the verbal form *differing* is changed to the adjective *different*, with the consequent structural reorganization.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

(7) a. I have heard many accounts of him [...] all *differing* from each other

b. I have heard many *different* things about him

Spelling and format changes comprise changes in the spelling and format of lexical (or functional) units, such as case changes, abbreviations, or digit/letter alternations. In (8), case changes occur (*Peace/PEACE*).

(8) a. And yet they are calling for *Peace*!–*Peace*!!

b. Yet still they shout *PEACE*! *PEACE*!

Same-polarity substitutions change one lexical (or functional) unit for another with approximately the same meaning.[6] Among the linguistic mechanisms of this type, we find synonymy, general/specific substitutions, or exact/approximate alternations. In (9), *very little* is more general than *a teaspoonful of*.

(9) a. *a teaspoonful of* vanilla

b. *very little* vanilla

Synthetic/analytic substitutions consist of changing synthetic structures for analytic structures, and vice versa. This type comprises mechanisms such as compounding/decomposition, light element, or lexically emptied specifier additions/deletions, or alternations affecting genitives and possessives. In (10-b), a (lexically emptied) specifier

6 The object of study of both paraphrasing and lexical semantics fields converge in lexicon-based changes in general and same-polarity substitutions in particular. In this sense, many works and tasks in lexical semantics are also relevant for our purposes. By way of illustration, the lexical substitution task within SemEval-2007 aimed to produce a substitute word (or phrase), that is, a paraphrase, for a word in context (McCarthy and Navigli 2009).

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

(*a sequence of*) has been deleted: it did not add new content to the lexical unit, but emphasized its plural nature.

(10) a. A sequence of ideas

b. ideas

Opposite-polarity substitutions. Two phenomena are considered within this type. First, there is the case of double change of polarity, when a lexical unit is changed for its antonym or complementary and another change of polarity has to occur within the same sentence in order to maintain the same meaning. In (11), *failed* is substituted for its antonym *succeed* and a negation is added. Second, there is the case of change of polarity and argument inversion, where an adjective is changed for its antonym in comparative structures. Here an inversion of the compared elements has to occur. In (12), the adjectival phrases *far deeper* and *more general* change to the opposite-polarity ones *less serious* and *less common*. To maintain the same meaning, the order of the compared elements (i.e., what the Church considers and what is perceived by the population) has to be inverted.

(11) a. Leicester [. . .] *failed* in both enterprises

b. he *did not succeed* in either case

(12) a. the sense of scandal given by this is *far deeper* and *more general* than the Church thinks

b. the Church considers that this scandal is *less serious* and *less common* than it really is

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Converse substitutions take place when a lexical unit is changed for its converse pair. In order to maintain the same meaning, an argument inversion has to occur. In (13), *awarded to* is changed to *receiving [...] from*, and the arguments *the Geological Society in London* and *him* are inverted.

(13) a. the Geological Society of London in 1855 *awarded to* him the Wollaston medal

b. resulted in him *receiving* the Wollaston medal *from* the Geological Society in London in 1855

Diathesis alternation type gathers those diathesis alternations in which verbs can participate, such as the active/passive alternation (14).

(14) a. the guide drew our attention to a gloomy little dungeon

b. ou[r] attention was drawn by our guide to a little dungeon[7]

Negation switching consists of changing the position of the negation within a sentence. In (15), *no* changes to *does not*.

(15) a. In order to move us, it needs *no* reference to any recognized original

b. One *does not* need to recognize a tangible object to be moved by its artistic representation

7 Typos in the examples are also present in the original corpus. When there was any modification of the original, this is indicated with square brackets.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Ellipsis includes linguistic ellipsis, i.e, those cases in which the elided fragments can be recovered through linguistic mechanisms. In (16-b), the subject *he* appears in both clauses; in (16-a), it is only displayed in the first one.

(16) a. In the scenes with Iago *he* equaled Salvini, yet did not in any one point surpass him

b. *He* equaled Salvini, in the scenes with Iago, but *he* did not in any point surpass him or imitate him

Coordination changes consist of changes in which one of the members of the pair contains coordinated linguistic units, and this coordination is not present or changes its position and/or form in the other member of the pair. The juxtaposed sentences with a full stop in (17-a) are coordinated with the conjunction *and* in (17-b).

(17) a. It is estimated that he spent nearly £10,000 on these works. In addition he published a large number of separate papers

b. Altogether these works cost him almost £10,000 *and* he wrote a lot of small papers as well

Subordination and nesting changes consist of changes in which one of the members of the pair contains a subordination or nested element, which is not present, or changes its position and/or form within the other member of the pair. What is a relative clause in (18-a) (*which limits the percentage of Jewish pupils in any school*) is part of the main clause in (18-b).

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

(18) a. the Russian law, which limits the percentage of Jewish pupils in any school, barred his admission

b. the Russian law had limits for Jewish students so they barred his admission

Punctuation and format changes consist of any change in the punctuation or format of a sentence (not of a lexical unit, cf. lexicon-based changes). In (19-a), the list appears numbered and, in (19-b), it does not.

(19) a. At Victoria Station you will purchase (1) a return ticket to Streatham Common, (2) a platform ticket

b. You will purchase a return ticket to Streatham Common and a platform ticket at Victoria station

Direct/indirect style alternations consist of changing direct style for indirect style, and vice versa. The direct style can be seen in (20-a) and the indirect in (20-b).

(20) a. "She is mine," said the Great Spirit

b. The Great Spirit said that she is her[s]

Sentence modality changes are those cases in which there is a change of modality (not provoked by modal verbs, cf. modal verb changes), but the illocutive value is maintained. In (21-a), interrogative sentences can be observed; they are changed to an affirmative sentence in (21-b).

(21) a. The real question is, will it pay? will it please Theophilus P. Polk or vex Harriman Q. Kunz?

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

b. He do it just for earning money or to please Theophilus P. Polk or vex Hariman Q. Kunz

Syntax/discourse structure changes gather a wide variety of syntax/discourse reorganizations not covered by the types in the syntax and discourse sub-classes above. An example can be seen in (22).

(22) a. How he would stare!

b. He would surely stare!

Semantics-based changes are those that involve a different lexicalization of the same content units.[8] These changes affect more than one lexical unit and a clear-cut division of these units in the mapping between the two members of the paraphrase pair is not possible. In example (23), the content units TROPICAL-LIKE ASPECT (*scenery was [...] tropical/tropical appearance*) and INCREASE OF THIS ASPECT (*more/added*) are present in both fragments, but there is not a clear-cut mapping between the two.

(23) a. The scenery was altogether more tropical

b. which added to the tropical appearance

Change of order includes any type of change of order from the word level to the sentence level. In (24), *first* changes its position in the sentence.

(24) a. *First* we came to the tall palm trees

b. We got to some rather biggish palm trees *first*

8 This type is based on the ideas of Talmy (1985).

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Addition/deletion This type consists of all additions/deletions of lexical and functional units. In (25-b), *one day* is deleted.

(25) a. *One day* she took a hot flat-iron, removed my clothes, and held it on my naked back until I howled with pain

b. As a proof of bed treatment, she took a hot flat-iron and put it on my back after removing my clothes

3. Building the P4P Corpus

This section describes how P4P, a new paraphrase corpus with paraphrase type annotation, was built.[9] First, we will set out a brief state of the art on paraphrase corpora.

Paraphrase corpora in existence are rather few. One of the most widely used is the MSRP corpus (Dolan and Brockett 2005), which contains 5,801 English sentence pairs from news articles hand-labeled with a binary judgment indicating whether human raters considered them to be paraphrases (67%) or not (33%). Cohn, Callison-Burch, and Lapata (2008), in turn, built a corpus of 900 paraphrase sentence pairs aligned at word or phrase level.[10] The pairs were compiled from three different types of corpora: (*i*) sentence pairs judged equivalent from the MSRP corpus, (*ii*) the Multiple-Translation Chinese corpus (MTC), and (*iii*) the monolingual parallel corpus used by Barzilay and McKeown (2001). The WRPA corpus (Vila, Rodríguez, and Martí Submitted) is a corpus of relational paraphrases extracted from Wikipedia. It comprises paraphrases expressing relations like person–date_of_birth in English and author–work in Spanish. Moreover, Max and Wisniewski (2010) built the Wikipedia Correction and Paraphrase

9 The P4P corpus and guidelines used for its annotation are available at `http://clic.ub.edu/corpus/en/paraphrases-en`. The subsets of the MSRP and WRPA corpora annotated with the same typology are also available in this website.

10 `http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html`

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Corpus (WiCoPaCo) from the Wikipedia revision history.[11] Apart from paraphrases, the corpus includes spelling corrections and other local text transformations. In the paper, the authors set out a typology of these revisions and classify them as meaning-preserving or meaning-altering. There also exist works where the focus is not to build a paraphrase corpus, but to create a paraphrase extraction or generation system, which ends up in also building a paraphrase collection, such as Barzilay and Lee (2003).

Plagiarism detection experts are starting to turn their attention to paraphrasing. Burrows, Potthast, and Stein (2012 (to appear)) built Webis Crowd Paraphrase Corpus, by crowd-sourcing more than 4,000 manually simulated samples of paraphrase plagiarism.[12] In order to create feasible mechanisms for crowd-sourcing paraphrase acquisition, they built a classifier to reject bad instances of paraphrase plagiarism (e.g., cases of verbatim plagiarism). These crowd-sourced instances are similar to the cases of simulated plagiarism in the PAN-PC-10 corpus, and hence the P4P (see below).

P4P was built upon the PAN-PC-10 corpus, from the International Competition on Plagiarism Detection.[13] The PAN competition appeared with the aim of creating the first large-scale evaluation framework for plagiarism detection. It relies on two main resources: a corpus with cases of plagiarism and a set of evaluation measures specially suited to the problem of automatic plagiarism detection (cf. Section 4) (Potthast et al. 2010). We focus on the PAN-10 plagiarism detection competition. The corpus used in this edition, known as PAN-PC-10, was composed of a set of suspicious documents D_q that may or may not contain plagiarized fragments, together with a set of potential source documents D. In order to build it, text fragments were extracted randomly from documents $d \in D$ and inserted into some $d_q \in D_q$. The PAN-PC-10 contains circa 70,000

11 http://wicopaco.limsi.fr/
12 http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-webis-cpc-11.html
13 http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-10.html

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

cases of plagiarism; 40% of them are exact copies, while the rest involved some kind of obfuscation (paraphrasing). Most of the obfuscated cases were generated artificially, that is, rewriting operations were imitated by a computational process.[14] The rest (6%) were created by humans who aimed at simulating paraphrase cases of plagiarism. These cases were generated through Amazon Mechanical Turk, with clear instructions to rewrite text fragments to simulate the act of plagiarizing. According to Potthast et al. (2010), most of the *turkers* had attended college and 62% identified themselves as native English speakers.[15] Cases in this subset of the corpus are referred to onwards as simulated plagiarism.[16]

The P4P corpus was built using cases of simulated plagiarism in the PAN-PC-10 (plg_{sim}). They consist of pairs of source and plagiarized fragments, where the latter was manually created reformulating the former. From this set, we selected those cases containing 50 words or less ($|plg_{sim}| \leq 50$); 847 paraphrase pairs met these conditions and were selected as our working subset. The decision was taken for the sake of simplicity and efficiency, and is backed by state-of-the-art paraphrases corpora. As way of illustration, the MSRP contains 28 words per case on average and the Barzilay and Lee (2003) collection includes examples of about 20 words in length only.

The tagset and the scope. After tokenization of the working corpus, the annotation was performed by, on the one hand, tagging the paraphrase phenomena present in each source/plagiarism pair with our tagset (each pair contains multiple paraphrase

14 The strategies include: (*i*) randomly shuffling, removing, inserting, or replacing short phrases from the source to the plagiarized fragment, (*ii*) randomly substituting a word for its synonym, hyponym, or antonym, and (*iii*) randomly shuffling the words, but preserving the POS sequence of the source text (Potthast et al. 2010, 2010).
15 Turkers aimed at finishing the cases as soon as possible in order to get paid for the task, hence facing a similar time constraint to that of people tempted to take the plagiarism short-cut.
16 In contrast to *simulated plagiarism*, *paraphrase plagiarism* is a more general term referring to plagiarism based on paraphrase mechanisms.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

phenomena) and, on the other hand, indicating the scope of each of these tags (the range of the fragment affected by each paraphrase phenomenon).

Our tagset consists of our 20 paraphrase types plus identical and non-paraphrase tags. Identical refers to those text fragments in the source/plagiarism pairs that are exact copies; non-paraphrase refers to fragments in the source/target pairs that are not semantically related. The reason for adding these two tags is to see how they perform in comparison to the actual paraphrase cases.

Regarding the scope, we do not annotate strings but linguistic units (words, phrases, clauses, and sentences). In (26), although a change takes place between the fragments *brotherhood among* and *other brothers with*, the paraphrase mapping has to be established between *the brotherhood* and *the other brothers* (α), and between *among* and *with* (β), two different pairs of linguistic units, fulfilled, respectively, by nominal phrases and prepositions. They consist of two same-polarity substitutions.

(26) a. [*the brotherhood*]$_\alpha$ [*among*]$_\beta$ whom they had dwelt

b. [*the other brothers*]$_\alpha$ [*with*]$_\beta$ whom they lived

It is important to note that paraphrase tags can overlap. In example (27), a same-polarity substitution overlaps a change of order in *sagely/wisely*. Tags can also be discontinuous, such as in (28-a): *distinct [. . .] from*. The pair *distinct [. . .] from* and *unconnected to* are a same-polarity substitution.

(27) a. *sagely* shaking his head

b. shaking his head *wisely*

(28) a. But yet I imagine that the application of the term “Gothic” may be found to be quite *distinct*, in its origin, *from* the first rise of the Pointed Arch

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

b. Still, in my opinion, the use of "Gothic" might well have origins *unconnected to* the emergence of the pointed arch

The scope affects the annotation task differently regarding the classes:

Morpholexicon-based changes, semantics-based changes, and miscellaneous changes: only the linguistic unit(s) affected by the trigger change is (are) tagged. As some of these changes entail other changes, two different attributes are provided: LOCAL, which stands for those cases in which the trigger change does not entail any other change in the sentence; and GLOBAL, which stands for those cases in which the trigger change does entail other changes in the sentence. In (29), an isolated same-polarity substitution takes place, so the scope *older/aging* is annotated and the attribute LOCAL is used. In (30), the same-polarity substitution entails changes in the punctuation. In that case, only *but/however* is annotated using the attribute GLOBAL. For the entailed changes indicated by the GLOBAL attribute, neither the type of change nor the fragment suffering the change are specified in the annotation. This distinction between LOCAL / GLOBAL is called "projection" in our tag system.

(29) a. The *older* trees
b. The *aging* trees

(30) a. would not have had to endure; *but* she does not seem embittered
b. wouldn't have been. *However*, she's not too resentful

Structure-based changes: The whole linguistic unit suffering the syntactic or discourse reorganization is tagged. Moreover, most structure-based changes have a key element that gives rise to the change and/or distinguishes it from others. This key element is also

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

tagged. In (31), the coordination change affects two juxtaposed sentences in (31-a) and two coordinated clauses in (31-b), so all of them constitute the scope of the phenomenon. The conjunction *and* stands for the key element.

(31) a. They were born of the same universal fact. They are of the same Father!

b. They are the sons of the same Father *and* are born and brought up with the same plan

In the case of identical and non-paraphrases, no LOCAL/GLOBAL attributes nor key elements are used, and only the affected fragment is tagged.

The annotation process. The annotation process was carried out by three postgraduate linguists experienced in annotation and having an advanced English level. Among the annotators, there was one of the authors of the typology (annotator $\mathcal{A}$); the other two were not familiar with the typology before the annotation (annotators $\mathcal{B}$ and $\mathcal{C}$). This mixed group allowed for sharing experienced and blind knowledge regarding the typology, both necessary for the test of the paraphrasing types when applied to the P4P corpus.

The annotation was performed using the CoCo interface (España-Bonet et al. 2009)[17] in three phases: annotators' training, inter-annotator agreement, and final annotation. In the annotators' training phase, 50 cases were doubly annotated by $\mathcal{B}$ and $\mathcal{C}$ under the supervision of $\mathcal{A}$, following a preliminary version of the guidelines. Problems and disagreements were discussed. Following this discussion, some changes were made to the guidelines (see footnote 9), and the 50 annotations by one of the annotators revised to be included in the corpus. In the inter-annotator agreement phase, 100 cases were

17 http://www.lsi.upc.edu/~textmess/

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

doubly annotated by $\mathcal{B}$ and $\mathcal{C}$ and the inter-annotator agreement computed. In the final annotation phase, we annotated the remaining cases in P4P; the examples were annotated only once by $\mathcal{A}$, $\mathcal{B}$ or $\mathcal{C}$.

The examples corresponding to each phase (training, agreement, and final annotation) were randomly selected. Once the annotation process finished, we calculated the similarity between the distributions of paraphrase types in the inter-annotator subset and the rest of the corpus. We used the well known cosine measure, ranged in $[0, 1]$ with 1 implying maximum similarity. The similarity was 0.988.

Regarding the inter-annotator agreement calculation, Kappa measures (e.g., Fleiss') are not suitable for our work, because agreement by chance is almost impossible, due to the fact that we do not only annotate types but also scope: the amount of possible scope combinations in each pair is in the order of $2^{|src|+|plg|}$, where $|\cdot|$ represents the number of tokens in the source or plagiarized fragment. As an alternative, we developed a measure for inter-annotator agreement in paraphrase type annotation ranged in $[0, 1]$. For each paraphrase phenomenon, we calculate the degree of overlapping between the two annotations at token level, considering types and scope.

The rationale behind our inter-annotator agreement computation is as follows. Let B and C be the set of paraphrase phenomena annotated by $\mathcal{B}$ and $\mathcal{C}$ (we consider independently all the phenomena occurring over all the plagiarism–source pairs). We define the inter-annotator agreement between B and C as:

$$F_1 = 2 \cdot \frac{K_B \cdot K_C}{K_B + K_C} \tag{1}$$

K_B is computed as:

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

$$K_B = \frac{\sum_{b \in B} \min\left(1, \sum_{c \in C} overlapping(b, c)\right)}{|B|} \quad (2)$$

The $overlapping$ measure is defined as:

$$overlapping(b, c) = \alpha \cdot \left(\frac{|b_s \cap c_s|}{|b_s|} + \frac{|b_p \cap c_p|}{|b_p|} \right) \quad (3)$$

where s and p refer to the source and plagiarized tokens in the annotation, respectively; $\alpha = 1$ for phenomena of the type addition/deletion and $\alpha = 0.5$ for others (in the case of addition/deletion only one text fragment, either in the source or plagiarized text, exists). As expected, an overlapping between b and c exists only if the two phenomena are annotated with the same paraphrase type (otherwise, the overlapping is 0).

In summary, we compute how $\mathcal{B}$'s annotations are covered by $\mathcal{C}$'s, and vice versa. K_B may be understood as a regression precision taking the annotation by $\mathcal{C}$ as reference, and a regression recall taking the annotations by $\mathcal{B}$ as reference. K_C is computed accordingly. Thus, F_1 obtains the same value independently of what we could take as a reference annotation.

The overall inter-annotator agreement thus obtained is $F_1 = 0.63$. In a much simpler task, the binary decision of whether two sentences are paraphrases in the MSRP corpus, a similar agreement was obtained (Dolan and Brockett 2005); hence we consider this as an acceptable result. These results show the suitability of our paraphrase typology for the annotation of plagiarism examples.

Annotation Results. Paraphrase type frequencies, and total and average lengths are collected in Tables 1 and 2. Same-polarity substitutions represent the most frequent paraphrase type ($freq_{rel} = 0.46$). At a considerable distance, the second most frequent

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

type is addition/deletion ($freq_{rel} = 0.13$). We hypothesize that the way paraphrases were collected has a major impact on these results. They were created manually, asking people to simulate plagiarizing by re-writing a collection of text fragments, that is, they were originated in a reformulation framework, where a conscious reformulative intention by a speaker exists. Our hypothesis is that the most frequent paraphrase types in the P4P corpus correspond to the paraphrase mechanisms most accessible to humans when asked to reformulate or plagiarize. Same-polarity substitutions and addition/deletion are mechanisms which are relatively simple to apply to a text by humans: changing one lexical unit for its synonym (understanding synonymy in a general sense) and deleting a text fragment, respectively.

In general terms, the lengths of the annotated paraphrases in the plagiarism fragments are shorter than in the source. As a result, the entire plagiarized fragments tend to be shorter than their source (cf. top of Table 2). This means that, while reformulating (plagiarizing), people tend to use shorter expressions for same meaning, or, as already said, just delete some fragments. Finally, the paraphrase types with the largest average length are in syntax- and discourse-based change classes. The reason is to be found in the above distinction between the two ways to annotate the scope: in structural reorganizations, we annotate the whole linguistic unit suffering the change.

A question that remains open is how realistic the cases of simulated plagiarism in the PAN corpora are. In order to check this, two small collections of cases of real text re-use, RWP (Real Web Plagiarism) and sub-METER, were annotated with our typology. RWP is composed of actual cases of plagiarism reported on-line and sub-METER includes a set of re-used sentences extracted from the METER (MEasuring TExt Re-use) corpus, which contains cases of journalistic text re-use (Clough, Gaizauskas, and Piao

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Table 1
Absolute and relative frequencies of the paraphrase phenomena occurring within the 847 source–plagiarism pairs in the P4P corpus. Note that the values of the classes (in bold) are the sum of the corresponding types. In the right-hand column average of paraphrase phenomena for each pair are shown.

	$freq_{abs}$	$freq_{rel}$	$avg \pm \sigma$
Morphology-based changes	**631**	**0.057**	
Inflectional changes	254	0.023	0.30±0.60
Modal verb changes	116	0.010	0.14±0.38
Derivational changes	261	0.024	0.31±0.60
Lexicon-based changes	**6,264**	**0.564**	
Spelling and format changes	436	0.039	0.52±1.20
Same-polarity substitutions	5,056	0.456	5.99±3.58
Synthetic/analytic substitutions	658	0.059	0.79±1.00
Opposite-polarity substitutions	65	0.006	0.08±0.31
Converse substitutions	33	0.003	0.04±0.21
Syntax-based changes	**1,045**	**0.083**	
Diathesis alternations	128	0.012	0.14±0.39
Negation switching	33	0.003	0.04±0.20
Ellipsis	83	0.007	0.10±0.35
Coordination changes	188	0.017	0.25±0.52
Subordination and nesting changes	484	0.044	0.70±0.92
Discourse-based changes	**805**	**0.072**	
Punctuation and format changes	430	0.039	0.64±0.91
Direct/indirect style alternations	36	0.003	0.04±0.29
Sentence modality changes	35	0.003	0.04±0.22
Syntax/discourse structure changes	304	0.027	0.37±0.65
Semantics-based changes	**335**	**0.030**	0.40±0.74
Miscellaneous changes	**2,027**	**0.182**	
Change of order	556	0.050	0.68±0.95
Addition/deletion	1,471	0.132	1.74±1.66
Others	**136**	**0.012**	
Identical	101	0.009	0.12±0.40
Non-paraphrases	35	0.003	0.04±0.22

2002).[18] Around 150 cases of re-use were annotated with our typology. As in the P4P corpus, the most frequent paraphrase operations are: (*a*) same-polarity substitutions, with 27% (36%) in the METER (RWP) sample and (*b*) addition/deletion, with 29% (23%) in the METER (RWP) sample. The distributions of other paraphrase operations are also very similar to those in P4P (cf. Fig. 2). Regarding the lengths, the behavior is

18 http://nlp.shef.ac.uk/meter/

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Table 2
Character-level lengths of the annotated paraphrases in the P4P corpus. On top the lengths corresponding to the entire source and plagiarized fragments. Total and average lengths included (avg. lengths $\pm\sigma$).

	tot_{src}	tot_{plg}	$avg_{src} \pm \sigma$	$avg_{plg} \pm \sigma$
Entire fragments	210,311	193,715	248.30±14.41	228.71±37.50
Morphology-based changes				
Inflectional changes	1,739	1,655	6.85±3.54	6.52±2.82
Modal verb changes	1,272	1,212	10.97±6.37	10.45±5.80
Derivational changes	2,017	2,012	7.73±2.65	7.71±2.66
Lexicon-based changes				
Spelling and format changes	3,360	3,146	7.71±5.69	7.22±5.68
Same-polarity substitutions	42,984	41,497	8.50±6.01	8.21±5.24
Synthetic/analytic substitutions	12,389	11,019	18.83±12.78	16.75±12.10
Opposite-polarity substitutions	888	845	13.66±8.67	13.00±6.86
Converse substitutions	417	314	12.64±8.82	9.52±5.93
Syntax-based changes				
Diathesis alternations	8,959	8,247	69.99±45.28	64.43±37.62
Negation switching	2,022	1,864	61.27±39.84	56.48±38.98
Ellipsis	4,866	4,485	58.63±45.68	54.04±42.34
Coordination changes	25,363	23,272	134.91±76.51	123.79±71.95
Subordination and nesting changes	48,764	45,219	100.75±69.53	93.43±60.35
Discourse-based changes				
Punctuation and format changes	51,961	46,894	120.84±79.04	109.06±68.61
Direct/indirect style alternations	3,429	3,217	95.25±54.86	89.36±50.86
Sentence modality changes	3,220	2,880	92.0±67.14	82.29±57.99
Syntax/discourse structure changes	27,536	25,504	90.58±64.67	83.89±56.57
Semantics-based changes	16,811	13,467	50.18±41.85	40.20±29.36
Miscellaneous changes				
Change of order	15,725	14,406	28.28±30.89	25.91±24.65
Addition/deletion	16,132	6,919	10.97±17.10	4.70±10.79
Others				
Identical	6,297	6,313	62.35±63.54	62.50±63.60
Non-paraphrases	1,440	1,406	41.14±26.49	40.17±24.11

as observed already in the P4P corpus: the resulting re-used texts tend to be shorter. The length of a source text and its re-used counterpart has already been exploited in cross-language plagiarism detection (Barrón-Cedeño et al. 2010; Potthast et al. 2011), representing a promising factor to consider in the detection of paraphrase plagiarism.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

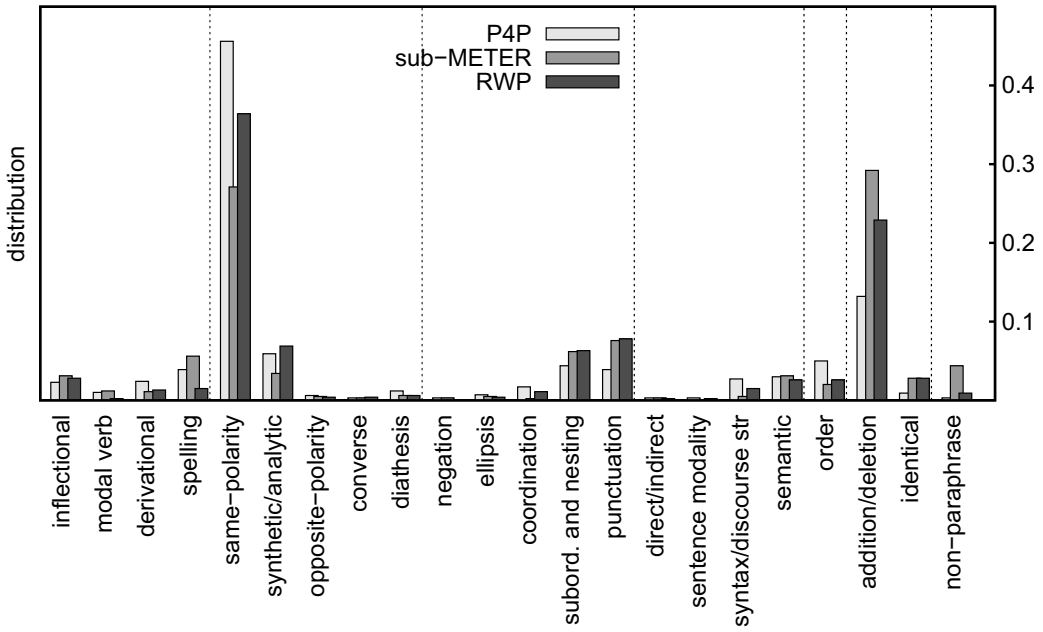

Figure 2
Overview of the paraphrase distribution in the P4P corpus with respect to the samples from the sub-METER and RWP corpora.

4. Plagiarism Detection Approaches at PAN-10

In this section, we move to the analysis and evaluation of existing systems for plagiarism detection. Generalities on models for plagiarism detection are set out, focusing on the PAN-10 competition. This information will be taken up in Section 5, where the performance of these systems when dealing with paraphrase plagiarism is analyzed by comparing it to the P4P dataset.

We consider that when a reader reviews a document d_q, there are two main factors that trigger suspicions of plagiarism: (*i*) inconsistencies or disruptive changes in terms of vocabulary, style and complexity throughout d_q; and (*ii*) the resemblance of the contents in d_q to previously consulted material. Our analysis is focused on factor (*ii*): the detection of a suspicious text fragment and its claimed source. This approach is

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

generally known as external plagiarism detection.[19] Research on paraphrasing has a direct application in this case: in order to conceal the plagiarism act, a different form expressing the same content, that is, a paraphrase, is often used.

External plagiarism detection is considered to be an information retrieval (IR) task. d_q is analyzed with respect to a collection of potential source documents D. The aim is to identify text fragments in d_q that are potential cases of plagiarism (if there are any), in conjunction with their respective source fragments from D (Potthast et al. 2009).

Here we discuss the models for plagiarism detection proposed in the framework of the PAN-10 competition.[20] As observed by Potthast et al. (2010), most of the participants' approaches to the external plagiarism detection task followed a three steps schema: (*1*) retrieval: for a suspicious document d_q, the most closely related documents $D' \subset D$ are retrieved; (*2*) detailed analysis: d_q and $d \in D'$ are compared section-wise in order to identify specific plagiarism–source candidate fragment pairs; and (*3*) post-processing: bad candidates (very short or not similar enough) are discarded and neighbor text fragments are combined. For the sake of clarity, we consider the IR pre-processing techniques applied by some participants as a preliminary step *(0)*. The pre-processing step gathers shallow linguistic processes and splitting of the source and suspicious documents in order to handle smaller text chunks. A summary of the parameters used at the PAN-10 competition for the four steps is included in Table 3. Note that this table represents a generalization of the different approaches that will be taken into account when investigating the correlation with paraphrase plagiarism detection (cf. Section 5.2).

19 We do not consider the approach related to factor (*i*): intrinsic plagiarism detection. See Stein, Lipka, and Prettenhofer (2011) and Stamatatos (2009) for further reading on this approach to plagiarism detection.
20 Refer to Clough (2000, 2003) and Maurer, Kappe, and Zaka (2006) for a general overview on approaches to plagiarism detection.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Table 3
Generalization of the modules applied by the models in the PAN-10 competition. The participant corresponds to the surname of the first member of each team. A black square appears if the participant applied a certain parameter and a number appears for values of n. Four steps are considered: pre-processing (sw=stopword, !αnum= non-alphanumeric, doc.=document, syn=synonymic), retrieval, detailed analysis, and post-processing (s=pair of plagiarism (s_q) source (s) detected fragments, $thres_k$=threshold, sim= similarity, δ= distance, $|\cdot|$=length of $\cdot$).

	Step (*0*) Pre-processing							**Step (*1*) Retrieval**		**Step (*2*) Detailed an.**				**Step (*3*) Post-processing**		
														discard s if		**merge s_1, s_2 if**
Participant	case-folding	sw removal	!αnum removal	stemming	doc. splitting	n-grams ordering	syn. normalization	word n-grams	char. n-grams	word n-grams	char. n-grams	dotplot	greedy str. tiling	$\lvert s_q\rvert < thres_1$	$sim(s_q, s) < thres_2$	$\delta(s_1, s_2) < thres_3$
Kasprzak						■		5		5						■
Zou	■			■				5				■			■	
Muhr					■			1		3				■	■	■
Grozea									16		16	■				
Oberreuter		■	■		■			3		3						
Rodriguez	■	■		■		■		3		3				■		
Corezola		■		■	■			1		1						■
Palkovskii								5		5						
Sobha								4		4						
Gottron					■	■		1		5		■		■		■
Micol		■	■					1			30			■		■
Costa-jussà	■	■		■	■			1				■				■
Nawab	■		■					5					■			■
Gupta								9		7						■
Vania		■						1		6					■	
Alzahrani		■		■			■	3		1						■

Most of the systems apply some kind of pre-processing (0) for one or both of steps (1) and (2), whereas a few of them do not.[21] Most of the pre-processing operations aim at minimizing the effect of paraphrasing, such as case-folding (spelling and format changes in our typology), n-gram ordering (change of order) and synonymic normalization (same-polarity substitutions).

During step (*1*) , retrieval, Gupta, Sameer, and Majumdar (2010) extract those non-overlapping word 9-grams with at least one named entity in order to compose the queries. The rest of the participants make a comparison on the basis of word n-grams

21 Systems such as the one of Gupta, Sameer, and Majumdar (2010) use standard information retrieval engines (e.g., Indri `http://www.lemurproject.org/`), which could incorporate some pre-processing.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

(with $n = \{1, 3, 4, 5\}$) or character 16-grams. Some of them order the n-grams' tokens alphabetically (Gottron 2010; Kasprzak and Brandejs 2010; Rodríguez Torrejón and Martín Ramos 2010).

During step (*2*), detailed analysis, several strategies are applied. Kasprzak and Brandejs (2010) and Rodríguez Torrejón and Martín Ramos (2010), as well as Gottron (2010), apply ordered n-grams. Corezola Pereira, Moreira, and Galante (2010) apply a classification system considering different features: bag-of-words cosine similarity, the similarity score assigned by an IR engine, and length deviation between the two fragments, among others. Alzahrani and Salim (2010) is the only team that, on the basis of WordNet synsets, expands the documents' vocabulary. The best systems participating in the competition were those using word n-grams (Kasprzak and Brandejs 2010; Muhr et al. 2010) as well as character n-grams (dot–plot technique) (Grozea and Popescu 2010b; Zou, jiang Long, and Ling 2010) in either one or both of steps (*1*) and (*2*) .[22]

Finally, in the post-processing step (3), models apply two different heuristics: (*i*) discarding a detected case if its length s_q is lower than a previously estimated threshold or the similarity $sim(s_q, s)$ (i.e., the similarity between the presumed plagiarism and its source) is not high enough to be considered relevant, and (*ii*) merging detected discontinuous fragments if the distance $\delta(s_1, s_2)$ between them is shorter than a given threshold (i.e., they are particularly close to each other). Probably the most interesting operation is merging. The maximum merging threshold is 5,000 characters (Costa-jussà et al. 2010).

22 In the dot–plot technique, documents are represented in an X, Y plane: d is located in X, while d_q is located in Y. The coordinates are filled with dots representing either common character n-grams, tokens, or word n-grams. As Clough (2003) points out, dot–plot provides "a visualization of matches between two sequences where diagonal lines indicate ordered matching sequences, and squares indicate unordered matches."

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

As automatic plagiarism detection is identified as an IR task, evaluation on the basis of recall and precision comes naturally. Nevertheless, plagiarism detection aims at retrieving specific (plagiarized–source) fragments rather than documents. Given a suspicious document d_q and a collection of potential source documents D, the detector should retrieve: (*a*) a specific text fragment $s_q \in d_q$, potential case of plagiarism; and (*b*) a specific text fragment $s \in d$, the claimed source for s_q. Therefore, special versions of precision and recall have been proposed that specially fit in this framework (Potthast et al. 2010). The plagiarized text fragments are treated as basic retrieval units, with $s_i \in S$ defining a query for which a plagiarism detection algorithm returns a result set $R_i \subseteq R$. The recall and precision of a plagiarism detection algorithm are defined as:

$$prec_{PDA}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(s \sqcap r)|}{|r|} \quad \text{and} \tag{4}$$

$$rec_{PDA}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(s \sqcap r)|}{|s|} \tag{5}$$

where $\sqcap$ computes the positionally overlapping characters. In both equations, S and R represent the entire set of actually plagiarized text fragments and detections, respectively.

Consider Fig. 3 for an illustrative example. $\{s_1, s_2, s_3\} \in S$ represent text sequences in the document that are known to be plagiarized. A given detector recognizes the sequences $\{r_1, r_2, r_3, r_4, r_5\} \in R$ as plagiarized. Substituting the values in Equations 4 and 5:

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

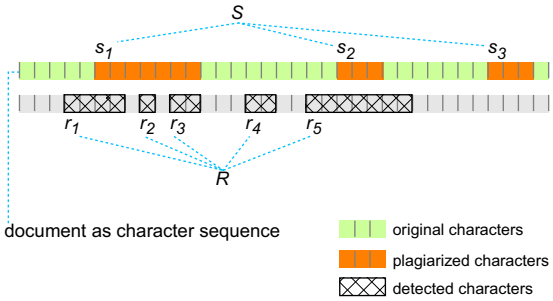

Figure 3
A document as character sequence, including plagiarized sections S and detections R returned by a plagiarism detection algorithm (used with permission of Potthast et al. [2010]).

$$prec_{PDA}(S,R) = \frac{1}{|R|} \cdot \left(\frac{|r_1 \sqcap s_1|}{|r_1|} + \frac{|r_2 \sqcap s_1|}{|r_2|} + \frac{|r_3 \sqcap s_1|}{|r_3|} + \frac{|\emptyset|}{|r_4|}^{0} + \frac{|r_5 \sqcap s_2|}{|r_5|} \right)$$
$$= \frac{1}{5} \cdot \left(\frac{2}{4} + \frac{1}{1} + \frac{2}{2} + \frac{3}{7} \right) = 0.5857 \quad \text{and}$$

$$rec_{PDA}(S,R) = \frac{1}{|S|} \cdot \left(\frac{|(s_1 \sqcap r_1) \bigcup (s_1 \sqcap r_2) \bigcup (s_1 \sqcap r_3)|}{|s_1|} + \frac{|s_2 \sqcap r_5|}{|s_2|} + \frac{\emptyset}{|s_3|}^{0} \right)$$
$$= \frac{1}{3} \cdot \left(\frac{5}{7} + \frac{3}{3} \right) = 0.5714$$

Once precision and recall are computed, they are combined into their harmonic mean (F_1-measure). In the next section, we analyze the performance of the PAN-10 plagiarism detection systems over the paraphrase-annotated cases in the P4P corpus on the basis of these measures.

5. Analysis of Paraphrase Plagiarism Detection

Paraphrase plagiarism has been identified as an open issue in plagiarism detection (Potthast et al. 2010; Stein et al. 2011). In order to figure out the limitations of current plagiarism detectors when dealing with paraphrase plagiarism, we analyze their performance

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

on the P4P corpus. Our aim is to understand what types of paraphrases make plagiarism more difficult to detect.

In Section 5.1 we group together the cases of plagiarism in the P4P corpus according to the paraphrase phenomena occurring within them. This grouping allows for the analysis of detectors' performance in Section 5.2. In order to obtain a global picture, we first analyze the detectors considering the entire PAN-PC-10 corpus. The aim is to give a general perspective of how difficult detecting cases with a high paraphrase density is respect to cases of verbatim copy and algorithmically simulated paraphrasing. Then we analyze the detectors' performance when considering the above mentioned groupings in the P4P corpus. We do so in order to identify those (combinations of) paraphrase operations that better allow a plagiarized text to go unnoticed. These analyzes open the perspective to research directions in automatic plagiarism detection that aim at detecting these kinds of borrowing.

5.1 Clustering Similar Cases of Plagiarism in the P4P Corpus

Paraphrase annotation and plagiarism detection are performed at different levels of granularity: the scope of the paraphrase phenomenon goes from word to (multiple-)sentence level (cf. Section 3) and plagiarism detectors aim at detecting entire, in general, multiple-sentence fragments. We should bear in mind that plagiarism detectors do not try to detect a paraphrase instance, but a plagiarized fragment and its source, which may include multiple paraphrases. The detection of a paraphrase does not necessarily mean that the detector actually succeeded in identifying it, but that it probably uncovered a broader text fragment, a case of plagiarism. As a result, directly comparing paraphrase annotation and detectors' outcomes is not possible, and organizing the data in a way that makes them comparable is required. Thus, we grouped together cases of plagiarism with similar concentrations of paraphrases or in which a

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

kind or paraphrase clearly stands out from the rest in order to observe how the detectors performed on different profiles of plagiarism.[23] As we only take into account the type and number of paraphrase phenomena in a pair, the scope does not have an impact on the results and the difference in granularity becomes irrelevant.

In order to perform this process, we used k-means (MacQueen 1967), a popular clustering method. In brief, k-means performs as follows: (*i*) k, the number of clusters, is set up at the beginning, (*ii*) k points are selected as initial centroids of the corresponding clusters, for instance, by randomly selecting k samples, and (*iii*) the position of the centers and the members of each cluster are iteratively redefined to maximize the similarity among the members of a cluster (intra-cluster) and minimize the similarity among elements of different clusters (extra-cluster).

We first composed a vector of 22 features to represent each source–plagiarism pair in the P4P. Each feature corresponds to one paraphrase tag in our annotation, and its weight is the relative frequency of the type in the pair. However, as same-polarity substitutions occur so often in many different plagiarism cases (this type represents more than 45% of the paraphrase operations in the P4P corpus and 96% of the plagiarism cases include them), they do not represent a good discriminating factor. This was confirmed by a preliminary experiment carried out considering different values for k. Therefore, k-means was applied by considering 21 features only.

We carried out 100 clustering procedures with different random initializations and considering $k = [2, 3, \ldots 20]$. Our aim was twofold: (*i*) to obtain the best possible clusters for every value of k and (*ii*) to determine the number of clusters to better organize the cases. In order to determine a convenient value for k, we applied the elbow method

23 An analysis considering paraphrase fragments as the retrieval units was also carried out. However, the obtained results were practically random since, in the framework of plagiarism detection, detecting a paraphrase as plagiarized in general depends on its context.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

(cf. Ketchen and Shook 1996), which calculates the clusters' distortion evolution (also known as cost function) for different values for k. The inflection point, that is, "the elbow", was in $k = 6$.

On the basis of our findings, we analyze the characteristics of the resulting clusters. A summary is included in Fig. 4. Although same-polarity substitutions are not taken into account in the clustering, they obviously remain in the source–plagiarism pairs and their numbers are displayed. They are similarly distributed among all the obtained clusters and are the most frequent in all of them. Next, we describe the obtained results in the clusters that show the most interesting insights from the perspective of the paraphrase cases of plagiarism.

In terms of linguistic complexity, identical and semantics-based changes can be considered as the extremes of the paraphrase continuum: absolute identicality and a deep change in the form, respectively. In c_5 and c_2, identical and semantic types are the most frequent (after same-polarity substitutions), respectively, and more frequent than in the other clusters.[24] Moreover, the most common type in c_3 is spelling and format. We observed that 39.36% of the cases in spelling and format involve only case changes which can be easily mapped to the identical types by a case-folding process. In the other clusters, no relevant features are observed. In terms of quantitative complexity, we consider the amount of paraphrase phenomena occurring in the source–plagiarism pairs. It follows that c_5 contains the cases with the least phenomena on average. The remaining clusters have a similar number of phenomena. For illustration purposes, Table 4 includes instances of source–plagiarism pairs from clusters c_2 and c_5.

24 Identical and semantic fragments are also longer in the respective clusters than in the others.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Table 4
Instances of source–plagiarism (src–plg) pairs in clusters c_2 and c_5 of the P4P corpus. Semantic (identical) cases are highlighted in cluster c_2 (c_5). Subscripts link the corresponding source and plagiarized fragments.

	c_2; case id: 9623
src	*["What a darling!"]*$_\alpha$ she said; "I must give her *[something very nice]*$_\beta$." She hovered a moment over the child's head, "She shall marry the man of her choice," she said, "and live happily ever after." *[There was a little stir among the fairies.]*$_\gamma$
plg	*["Oh isn't she sweet!"]*$_\alpha$ she said, thinking that she should present with *[some kind of special gift]*$_\beta$. Floating just above the little one's head she declared that the child will marry whoever she chooses and live happily ever after. *[All of the other fairies found this quite astonishing.]*$_\gamma$
	c_5; case id: 9727
src	*[On the contrary, by plunging the red-hot shells in the saline solution the greatest uniformity is attained.]*$_\alpha$ *[Instead of using clam shells as the base of my improved composition, I may use other forms of sea shells– such as oyster shells, etc.]*$_\beta$ *[I claim as new:]*$_\gamma$ 1.
plg	*[On the contrary, by plunging the red-hot shells in the saline solution the greatest uniformity is attained.]*$_\alpha$ *[Instead of using clam shells as the base of my improved composition, I may use other forms of sea shells– such as oyster shells, etc.]*$_\beta$ *[I claim as new:]*$_\gamma$

5.2 Results and Discussion

Our in-depth analysis uses F-measure, precision, and recall as evaluation measures (cf. Section 4). Due to our interest in investigating the number of paraphrase plagiarism cases that state-of-the-art systems for plagiarism detection succeed in detecting, we pay special attention to recall.

As a starting point, Figure 5 (*a*) shows the evaluations computed by considering the entire PAN-PC-10 corpus (Stein et al. 2011). The best recall values are around 0.70, with very good values of precision, some of them above 0.90. The results, when considering only the simulated cases, that is, those generated by manual paraphrasing, are presented in Fig. 5 (*b*). In most of the cases, the quality of the detections decreases dramatically compared to the results on the entire corpus, which also contains translated, verbatim

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

and automatically modified plagiarism. Manually created cases seem to be much harder to detect than the other, artificially generated, cases.[25] The difficulty to detect simulated cases of plagiarism in the PAN-PC-10 corpus was stressed by Stein et al. (2011). This does not necessarily imply that automatically generated cases were easy to detect. When the simulated cases in the PAN-PC-10 corpus were generated, volunteers had specific instructions to create rewritings with a high obfuscation degree. Figure 5 (*c*) shows the evaluation results when considering only the cases included in the P4P corpus. Note that the shorter a plagiarized case is, the harder it seems to be to detect (cf. Potthast et al. 2010, Table 6), and the P4P corpus is composed precisely of the shortest cases of simulated plagiarism in the PAN-PC-10; that is, cases no longer than 50 words.

Figures 6 and 7 show the evaluations computed by considering the 6 clusters of the P4P corpus. We focus on the comparison between the results obtained in the extreme cases: c_5 versus c_2. Cluster c_5, which comprises the lowest linguistic (relevance of identical cases) and quantitative (less paraphrase phenomena) complexity, is the one containing plagiarism cases that are easiest to detect. Cluster c_2, which comprises the highest linguistic complexity (relevance of the semantics-based changes), is the one containing the most difficult plagiarism cases to detect. The results obtained over cluster c_3 are the nearest to those of c_5, as the high presence of spelling and format changes (most of which are similar to identical cases), causes a plagiarism detector to have relatively more success on detecting them. These results are clearly observed through the values of recall obtained by the different detectors. Moreover, a relation between recall and precision exists: in general terms, high values of recall come with higher values of precision. To sum up, there exists a correlation between linguistic and

25 This can be appreciated when looking at the difference of capabilities of the system applied at the 2009 and 2010 competitions by Grozea, Gehl, and Popescu (2009) and Grozea and Popescu (2010a), practically the same implementation. At the first competition, which corpus included artificial cases only, its recall was of 0.66 while in the second one, with simulated (i.e., paraphrastic) cases, it decreased to 0.48.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

quantitative complexity and performance of the plagiarism detection systems: more complexity implies worse performance of the systems.

Interestingly, the best performing plagiarism detection systems on the P4P corpus are not the ones that performed the best at the PAN-10 competition. By still considering recall only, the best approaches on the P4P corpus, those of Costa-jussà et al. (2010) and Nawab, Stevenson, and Clough (2010) (Figure 5 (*c*)), are far from the top detectors in the competition (Figure 5 (*a*)). On the one hand, Nawab, Stevenson, and Clough (2010) apply greedy string tiling, which aims at detecting as long as possible identical fragments. As a result, this approach clearly outperforms the rest of detectors when dealing with cases with a high density of identical fragments (c_5 in Figure 7). On the other hand, the approach of Costa-jussà et al. (2010) outperform the others when dealing with the cases in the remaining clusters. The reasons are twofold: (*i*) their pre-processing strategy (which includes case-folding, stopword removal, and stemming) looks at minimizing the differences in the form caused by some paraphrase operations; (*ii*) their technique based on dot–plot (which considers isolated words) is flexible enough to identify fragments that share some identical words only. Cluster c_3 is again somewhere in between c_5 and c_2. The results by Nawab, Stevenson, and Clough (2010) and Costa-jussà et al. (2010) are very similar in this case. The former shows a slightly better performance because the system is good at detecting identical cases and they have a high presence in spelling and format changes.

The best overall performance system (Grozea and Popescu 2010a) and the best system when dealing with paraphrase plagiarism (Costa-jussà et al. 2010) are both based on the dot–plot technique. Whereas Grozea and Popescu (2010a) employ character 16-grams without any pre-processing, Costa-jussà et al. (2010) apply case-folding, stopword removal, and stemming pre-processing and use word 1-grams. This latter

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

approach is much more flexible than the former one in terms of paraphrase plagiarism detection.

6. Conclusions and Future Insights

The starting point of this article is that paraphrasing is the linguistic mechanism many plagiarism cases rely on. Our aim was to investigate why paraphrase plagiarism is so difficult to detect by state-of-the-art plagiarism detectors, and, especially, to understand which types of paraphrases underlie plagiarism acts, which are the most challenging, and how to proceed to improve plagiarism detection systems.

In order to analyst the break-down of the detection systems when aiming at detecting paraphrase plagiarism, we annotated a subset of the manually simulated plagiarism cases in the PAN-PC-10 corpus with a paraphrase typology, spawning the P4P corpus. P4P is the only available collection of plagiarism cases manually annotated with paraphrase types, constituting a new resource for the Computational Linguistics communities interested in paraphrasing and plagiarism.

On the basis of this annotation, we grouped together plagiarism cases with a similar distribution of paraphrase mechanisms. In the light of these groupings, the performance of the systems in the Second International Competition on Plagiarism Detection was analyzed. The resulting insights are the following: (*a*) there exists a correlation between the linguistic (i.e., kind of paraphrases) and the quantitative (i.e., amount of paraphrases) complexity and performance of the plagiarism detection systems: more complexity results in a worse performance of the systems; (*b*) same polarity substitutions and addition/deletions are the mechanisms used the most when plagiarizing; and (*c*) plagiarized fragments tend to be shorter than their source. Interestingly, the latter two insights hold when analyzing real cases of paraphrase plagiarism and text re-use.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

These results can be used to guide future efforts in automatic plagiarism detection. On the basis of the idea that solving the most frequent paraphrase mechanisms means solving most paraphrase plagiarism cases and given that same-polarity substitutions and addition/deletion are the most used paraphrase mechanisms by far, we have identified the following promising lines for future research: (*i*) an appropriate use of already existing lexical knowledge resources, such as WordNet[26] and Yago[27]; (*ii*) the development and exploitation of new empirically built resources, such as a lexicon of paraphrase expressions that could be easily obtained from the P4P and other corpora annotated at the paraphrase level; and (*iii*) the application of measures for estimating the expected length of a plagiarized fragment given its source.

Acknowledgments
We would like to thank the people that participated in the annotation of the P4P corpus, Horacio Rodríguez for his helpful advice as experienced researcher, and the reviewers of this contribution for their valuable comments to improve this article.
This research work was partially carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results received funding from the EU FP7 Programme 2007-2013 (grant n. 246016), the MICINN projects TEXT-ENTERPRISE 2.0 and TEXT-KNOWLEDGE 2.0 (TIN2009-13391), the EC WIQ-EI IRSES project (grant n. 269180), and the FP7 Marie Curie People Programme. The research work of A. Barrón-Cedeño and M. Vila was financed by the CONACyT-Mexico 192021 grant and the MECD-Spain FPU AP2008-02185 grant, respectively. The research work of A. Barrón-Cedeño was partially done in the framework of his PhD at the Universitat Politècnica de València.

26 `http://wordnet.princeton.edu`
27 `http://www.mpi-inf.mpg.de/yago-naga/yago/`

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

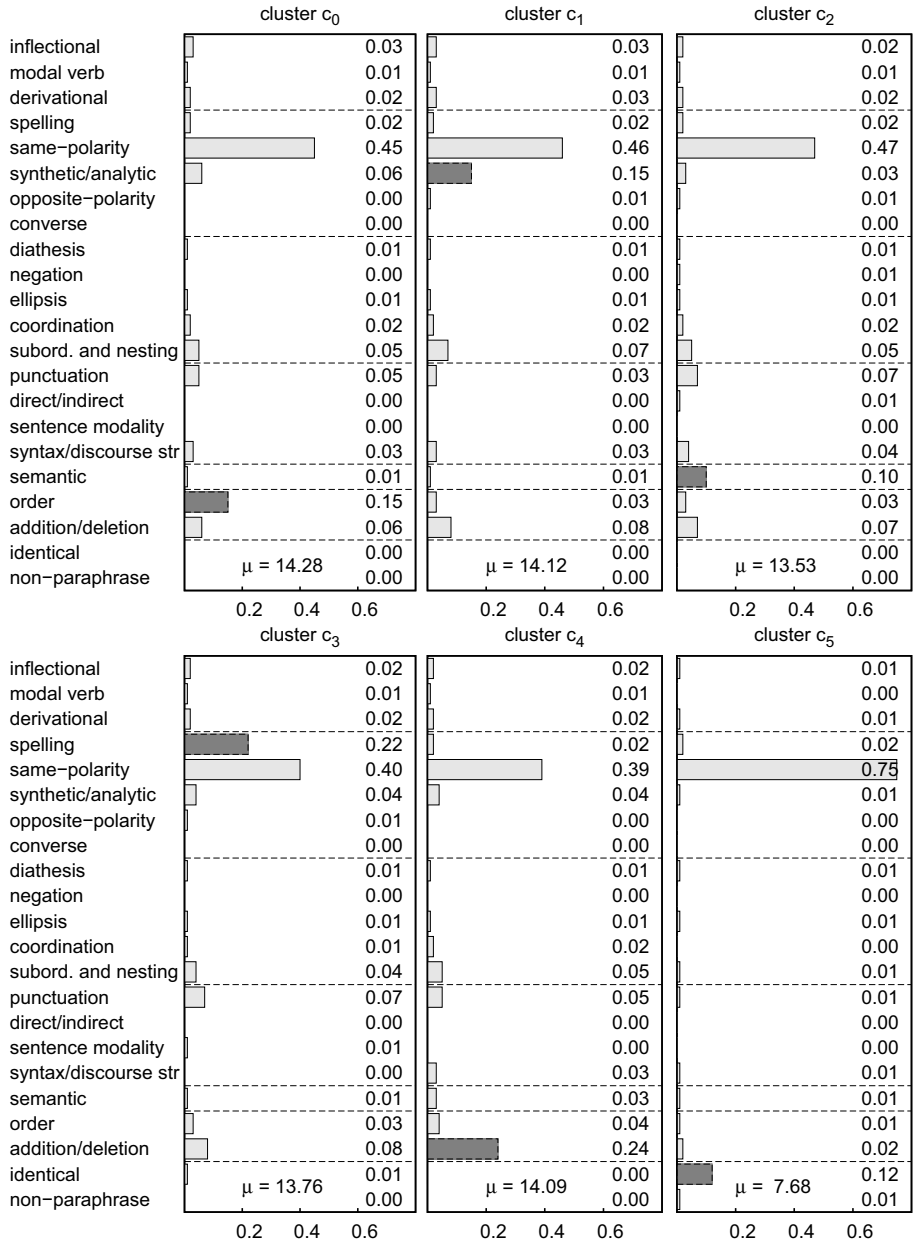

Figure 4
Average relative frequency of the different paraphrase phenomena in the source–plagiarism pairs of each cluster. The feature that stands out in the cluster and also respect to the rest of clusters, is represented darker (setting aside same-polarity substitutions). The value of μ refers to the average absolute number of phenomena per pair in each cluster.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

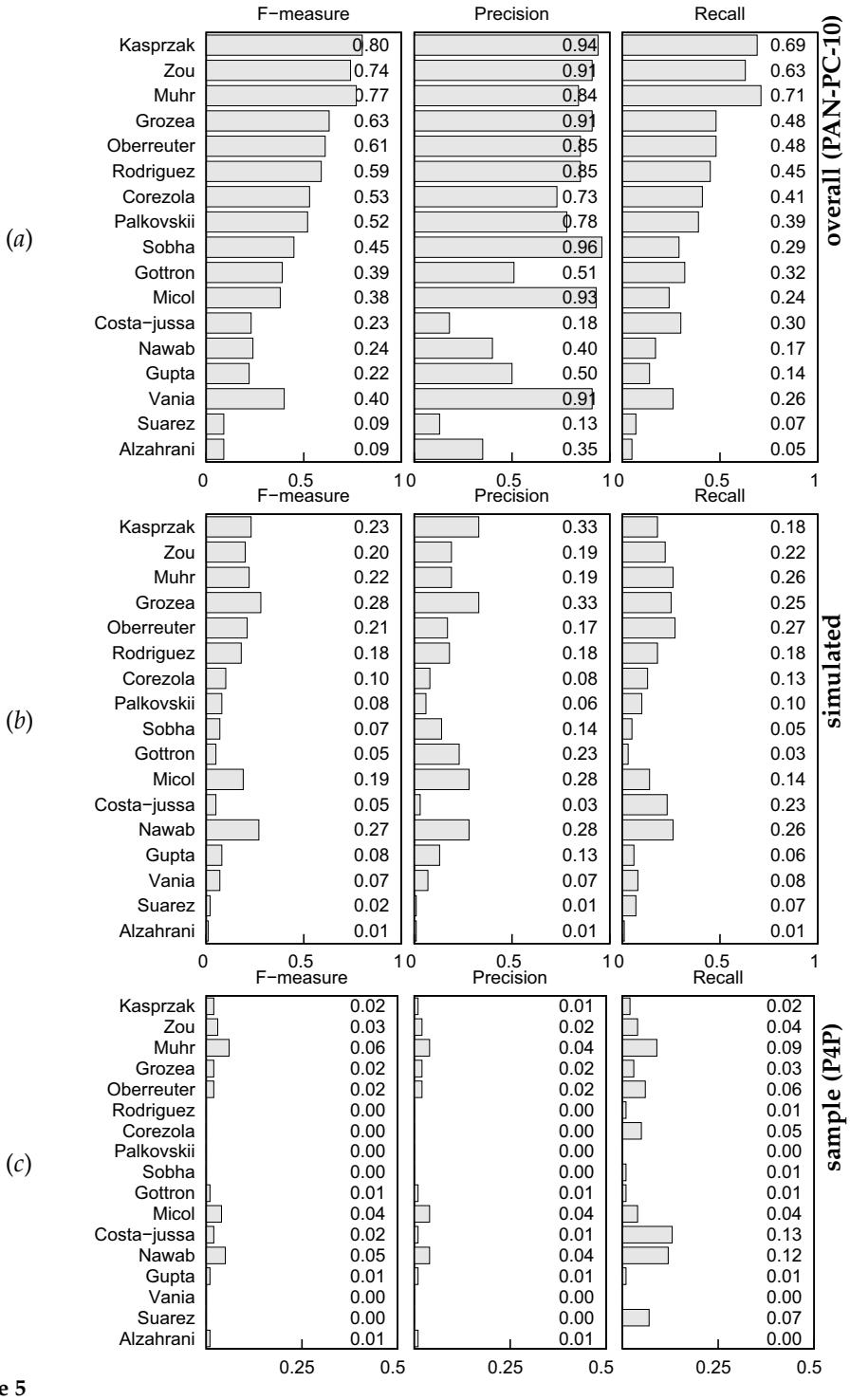

Figure 5
Evaluation of the PAN-10 competition participants' plagiarism detectors. Figures show evaluations over: (*a*) entire PAN-PC-10 corpus (including artificial, translated, and simulated cases); (*b*) simulated cases only; (*c*) sample of simulated cases annotated on the basis of the paraphrases typology: the P4P corpus. Note the change of scale in (*c*) .

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

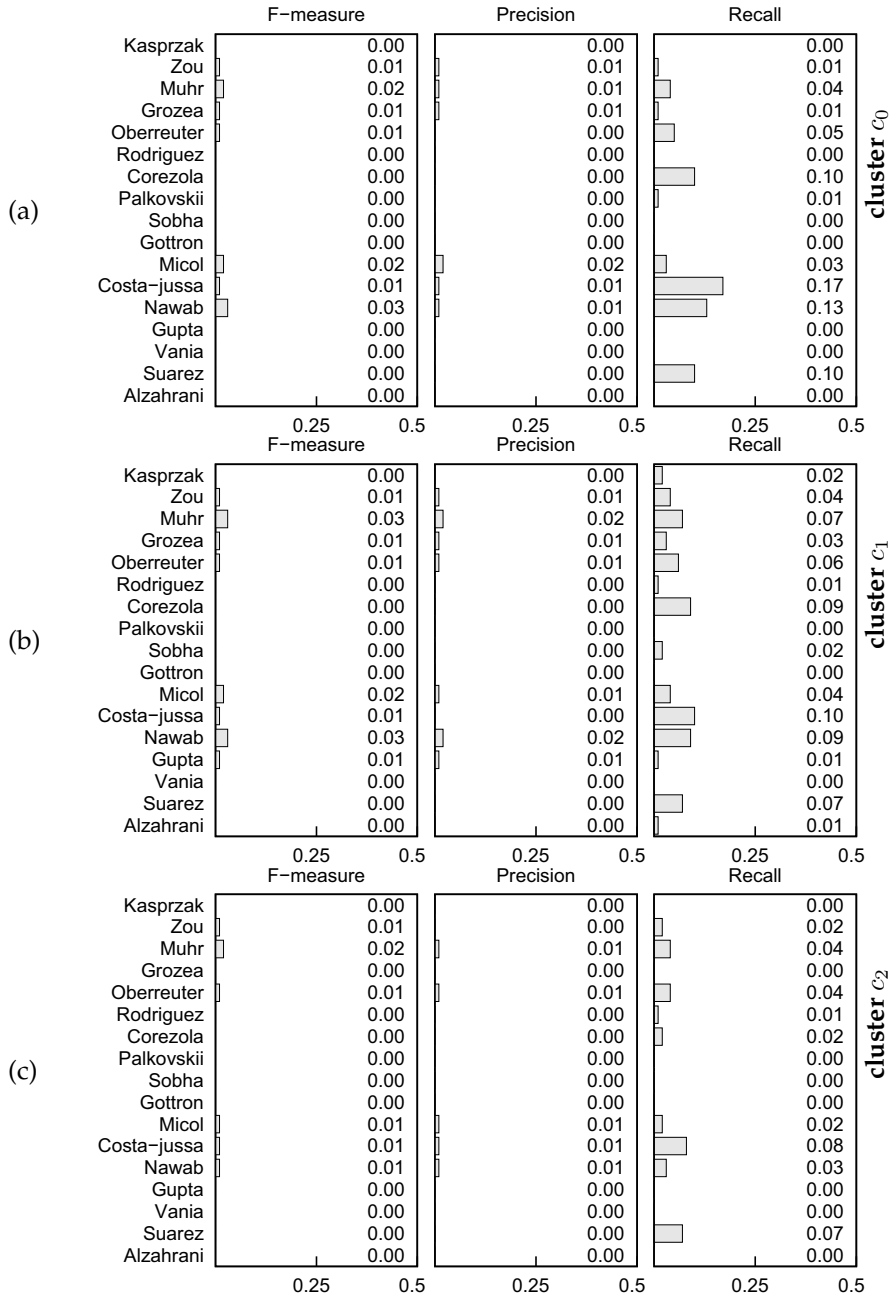

Figure 6
Evaluation of the PAN-10 competition participants' plagiarism detectors for (a) c_0; (b) c_1; (c) c_2.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

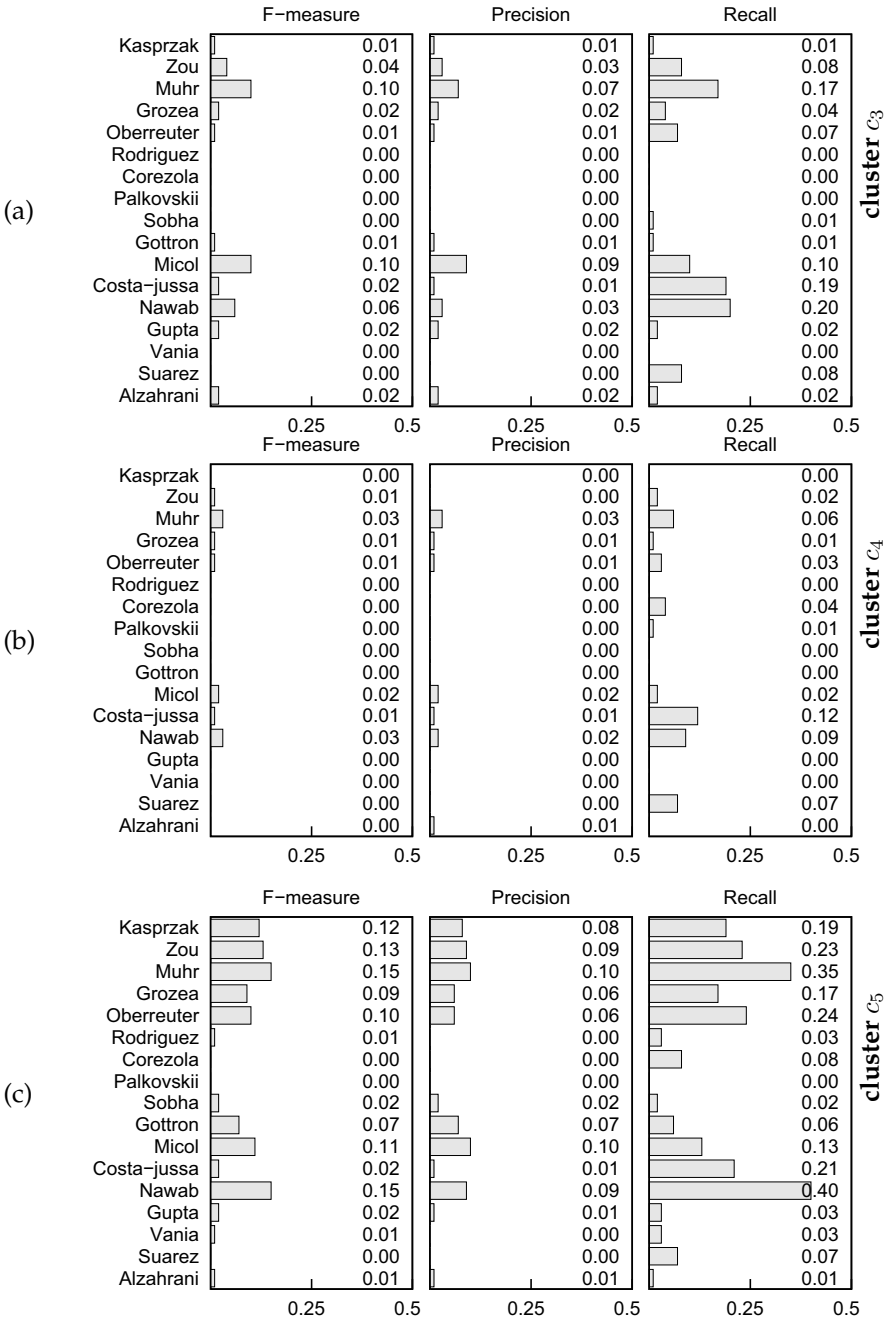

Figure 7
Evaluation of the PAN-10 competition participants' plagiarism detectors for (a) c_3; (b) c_4; (c) c_5.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

References

Alzahrani, Salha and Naomie Salim. 2010. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. In Braschler and Harman (Braschler and Harman 2010).

Association of Teachers and Lecturers. 2008. School work plagued by plagiarism - ATL survey. Technical report, Association of Teachers and Lecturers, London, UK. Press release.

Barrón-Cedeño, Alberto, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In Huang and Jurafsky (Huang and Jurafsky 2010), pages 37–45.

Barzilay, Regina. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York.

Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2003)*, pages 16–23, Edmonton.

Barzilay, Regina and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 50–57, Toulouse.

Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 550–557, Maryland.

Bhagat, Rahul. 2009. *Learning Paraphrases from Text*. Ph.D. thesis, University of Southern California, Los Angeles.

Braschler, Martin and Donna Harman, editors. 2010. *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy, September.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Burrows, Steven, Martin Potthast, and Benno Stein. 2012 (to appear). Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology*.

Cheung, Mei Ling Lisa. 2009. *Merging Corpus Linguistics and Collaborative Knowledge Construction*. Ph.D. thesis, University of Birmingham, Birmingham.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton & Co., The Hague/Paris.

Clough, Paul. 2000. Plagiarism in natural and programming languages: an overview of current tools and technologies. Technical Report CS-00-05, Department of Computer Science. University of Sheffield, Sheffield, UK.

Clough, Paul. 2003. Old and new challenges in automatic plagiarism detection. Technical report, National UK Plagiarism Advisory Service, UK.

Clough, Paul, Robert Gaizauskas, and Scott Piao. 2002. Building and annotating a corpus for the study of journalistic text reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1678–1691, Las Palmas.

Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–614.

Comas, Rubén, Jaume Sureda, Candy Nava, and Laura Serrano. 2010. Academic cyberplagiarism: A descriptive and comparative analysis of the prevalence amongst the undergraduate students at Tecmilenio University (Mexico) and Balearic Islands University (Spain). In *Proceedings of the International Conference on Education and New Learning Technologies (EDULEARN'10)*, Barcelona.

Corezola Pereira, Rafael, Viviane P. Moreira, and Renata Galante. 2010. UFRGS@PAN2010: Detecting external plagiarism lab report for PAN at CLEF 2010. In Braschler and Harman (Braschler and Harman 2010).

Costa-jussà, Marta R., Rafael E. Banchs, Jens Grivolla, and Joan Codina. 2010. Plagiarism detection using information retrieval and similarity measures based on image processing techniques. In Braschler and Harman (Braschler and Harman 2010).

Dolan, William B. and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, pages 9–16, Jeju Island.

Dorr, Bonnie J., Rebecca Green, Lori Levin, Owen Rambow, David Farwell, Nizar Habash, Stephen Helmreich, Eduard Hovy, Keith J. Miller, Teruko Mitamura, Florence Reeder, and Advaith Siddharthan. 2004. Semantic annotation and lexico-syntactic paraphrase. In *Proceedings of the LREC Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 47–52, Lisbon.

Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Sydney.

Dutrey, Camille, Delphine Bernhard, Houda Bouamor, and Aurélien Max. 2011. Local modifications and paraphrases in Wikipedia's revision history. *Procesamiento del Lenguaje Natural*, 46:51–58.

España-Bonet, Cristina, Marta Vila, Horacio Rodríguez, and M. Antònia Martí. 2009. CoCo, a Web interface for corpora compilation. *Procesamiento del Lenguaje Natural*, 43:367–368.

Faigley, Lester and Stephen Witte. 1981. Analyzing revision. *College Composition and Communication*, 32(4):400–414.

Fujita, Atsushi. 2005. *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Ph.D. thesis, Nara Institute of Science and Technology, Nara.

Gottron, Thomas. 2010. External plagiarism detection based on standard IR. Technology and fast recognition of common subsequences. In Braschler and Harman (Braschler and Harman 2010).

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Grozea, Cristian, Christian Gehl, and Marius Popescu. 2009. ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. In Stein et al. (Stein et al. 2009), pages 10–18.

Grozea, Cristian and Marius Popescu. 2010a. ENCOPLOT - performance in the Second International Plagiarism Detection Challenge lab report for PAN at CLEF 2010. In Braschler and Harman (Braschler and Harman 2010).

Grozea, Cristian and Marius Popescu. 2010b. Who's the thief? Automatic detection of the direction of plagiarism. *Computational Linguistics and Intelligent Text Processing, 10th International Conference*, LNCS (6008):700–710.

Gülich, Elisabeth. 2003. Conversational techniques used in transferring knowledge between medical experts and non-experts. *Discourse Studies*, 5(2):235–263.

Gupta, Parth, Rao Sameer, and Prasenjit Majumdar. 2010. External plagiarism detection: N-gram approach using named entity recognizer. Lab report for PAN at CLEF 2010. In Braschler and Harman (Braschler and Harman 2010).

Harris, Zellig. 1957. Co-occurence and transformation in linguistic structure. *Language*, 3(33):283–340.

Huang, Chu-Ren and Dan Jurafsky, editors. 2010. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, August. COLING 2010 Organizing Committee.

IEEE. 2008. A Plagiarism FAQ. [`http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html`]. Published: 2008; Last accessed 25/Nov/2012.

Kasprzak, Jan and Michal Brandejs. 2010. Improving the reliability of the plagiarism detection system. Lab report for PAN at CLEF 2010. In Braschler and Harman (Braschler and Harman

2010).

Ketchen, David J. and Christopher L. Shook. 1996. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17(6):441–458.

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

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley. University of California Press.

Martin, Brian. 2004. Plagiarism: policy against cheating or policy for learning? *Nexus (Newsletter of the Australian Sociological Association)*, 16(2):15–16.

Maurer, Hermann, Frank Kappe, and Bilal Zaka. 2006. Plagiarism - a survey. *Journal of Universal Computer Science*, 12(8):1050–1084.

Max, Aurélien and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3143–3148, Valletta.

McCarthy, Diana and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43:139–159.

Mel'čuk, Igor A. 1992. Paraphrase et lexique: la théorie Sens-Texte et le Dictionnaire Explicatif et Combinatoire. In Igor A. Mel'čuk, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, and Suzanne Mantha, editors, *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches Lexico-sémantiques III*. Les Presses de l'Université de Montréal, Montréal, pages 9–58.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Milićević, Jasmina. 2007. *La Paraphrase. Modélisation de la Paraphrase Langagière*. Peter Lang, Bern.

Muhr, Markus, Roman Kern, Mario Zechner, and Michael Granitzer. 2010. External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. In Braschler and Harman (Braschler and Harman 2010).

Nawab, Rao Muhammad Adeel, Mark Stevenson, and Paul Clough. 2010. University of Sheffield lab report for PAN at CLEF 2010. In Braschler and Harman (Braschler and Harman 2010).

Potthast, Martin, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2010. Overview of the 2nd International Competition on Plagiarism Detection. In Braschler and Harman (Braschler and Harman 2010).

Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, 45(1):1–18.

Potthast, Martin, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In Huang and Jurafsky (Huang and Jurafsky 2010), pages 997–1005.

Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview of the 1st international competition on plagiarism detection. In Stein et al. (Stein et al. 2009), pages 1–9.

Recasens, Marta and Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.

Rodríguez Torrejón, Diego Antonio and José Manuel Martín Ramos. 2010. CoReMo system (Contextual Reference Monotony). In Braschler and Harman (Braschler and Harman 2010).

Shimohata, Mitsuo. 2004. *Acquiring Paraphrases from Corpora and Its Application to Machine Translation*. Ph.D. thesis, Nara Institute of Science and Technology, Nara.

Computational Linguistics Just Accepted MS.
doi: 10.1162/COLI_a_00153

Stamatatos, Efstathios. 2009. Intrinsic plagiarism detection using character n-gram profiles. In Stein et al. (Stein et al. 2009), pages 38–46.

Stein, Benno, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, 45:63–82.

Stein, Benno, Martin Potthast, Paolo Rosso, Alberto Barrón-Cedeño, Efstathios Stamatatos, and Moshe Koppel. 2011. Fourth International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. *ACM SIGIR Forum*, 45:45–48.

Stein, Benno, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors. 2009. *Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009)*, San Sebastian. Volume 502 of CEUR-WS.org.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen, editor, *Language Typology and Syntactic Description. Grammatical Categories and the Lexicon*, volume III. Cambridge University Press, Cambridge, chapter II, pages 57–149.

Vila, Marta, M. Antònia Martí, and Horacio Rodríguez. 2011. Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Vila, Marta, Horacio Rodríguez, and M. Antònia Martí. Submitted. Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus.

Zou, Du, Wei jiang Long, and Zhang Ling. 2010. A cluster-based plagiarism detection method. In Braschler and Harman (Braschler and Harman 2010).

Part III

Conclusions, Contributions, and Future Directions

Chapter 5

Conclusions, Contributions, and Future Directions

The broadness and multifaceted nature of paraphrasing has prevented the creation of a precise and commonly accepted paraphrase characterization, as well as making the construction of comprehensive and well-founded paraphrase corpora a challenge. This thesis has addressed these two weak points and provides new insights on paraphrase boundaries and typology, as well as a new paraphrase corpus and corpora annotated with paraphrase types. This new knowledge and resources have proved to be of interest for the construction of new-generation systems involving paraphrase knowledge, such as automatic plagiarism detection systems. This thesis also sheds light on potential lines for future research in paraphrasing, an area that sometimes appears to be a virgin field.

In this chapter, I set out what has been accomplished in this thesis (Section 5.1) and the future lines of research it opens (Section 5.2).

5.1 Conclusions and Contributions

In what follows, the main contributions of the present thesis are set out. They include both theoretical aspects, and the methods and resources created. Then, other important considerations regarding paraphrasing arising from this thesis are presented.

Main contributions:

1. Based on the idea that paraphrasing is located on a continuum of semantic similarity, I have defined a **border area between paraphrases and non-paraphrases** where those cases involving content

loss, pragmatic knowledge, and changes in some grammatical features are located (Section 2.1).

2. **Paraphrase and coreference** may overlap, which has sometimes led to confusion in computational linguistics. In collaboration with Marta Recasens, I have clarified the difference between these two phenomena from different perspectives, the most relevant involving the fact that paraphrasing concerns approximate sameness of meaning, whereas coreference is about discourse-referent correspondence. I have also shown how paraphrase and coreference tasks in NLP can mutually benefit (Section 2.2).

3. I have developed a new **paraphrase typology** that goes a step forward with respect to the state-of-the-art ones. It includes 24 types comprising the linguistic mechanisms that give rise to paraphrases. They are grouped into 5 classes according to the linguistic nature of the mechanism (Section 2.1).

4. Derived from the typology, I have defined a **paraphrase-type annotation scheme**. It describes the guidelines for assigning tags and scopes to paraphrase phenomena. Tags correspond to the types in the typology; the scope-annotation criterion, in turn, is based on the distinction between classes in the typology (Section 3.2).

5. I have created the **Inter-annotator Agreement for Paraphrase-Type Annotation measures (IAPTA)**. They compute agreement at different levels of granularity. The most fine-grained measure considers both the coincidence in the tag and the level of overlapping in the scope (Section 3.2).

6. I have developed the **Wikipedia-based Relational Paraphrase Acquisition method (WRPA)**, which automatically acquires paraphrases expressing a concrete relation between two entities from Wikipedia. Applying distant learning and based on the distributional hypothesis, WRPA extracts candidate paraphrases from Wikipedia articles. They are then generalized through generalization and ngram processes. Finally, two classifiers that take the candidate paraphrases as input are built: a relation and a paraphrase classifier (Section 3.1).

7. Using the WRPA method, I have created the **WRPA corpus**, which currently covers 16 relations and two languages (English and Spanish).

It addresses one of the most challenging facets of paraphrase in computational linguistics, namely those paraphrases that do not necessarily show a formal mapping between them (Section 3.1).

8. Using the paraphrase-type annotation scheme, I have annotated three corpora that are different in nature and in two different languages (English and Spanish), giving rise to **the Paraphrase for Plagiarism (P4P), the Microsoft Research Paraphrase-Annotated (MSRP-A), and the WRPA-A corpora**. The results of the IAPTA measures when used on these corpora demonstrate the adequacy of our annotation methodology. These corpora constitute a powerful resource for machine learning and a source for deriving new tools, such as paraphrase lexicons, and for theoretical research (Section 3.2).

9. Based on the idea that paraphrases are the linguistic phenomenon underlying many plagiarism acts, in collaboration with Alberto Barrón-Cedeño, I have demonstrated that there exists a correlation between the linguistic (i.e., type of paraphrases) and the quantitative (i.e., amount of paraphrases) complexity and the performance of the plagiarism detection systems: more complexity results in a worse performance of the systems. Also, I have shown that the most frequent paraphrase types underling plagiarism acts are same-polarity substitutions and addition/deletion. These issues provide **insights for future research on automatic plagiarism detection**. (Section 4.1).

The WRPA, WRPA-A, P4P, and MSRP-A corpora are available as a package to download and, in the case of the annotated corpora, also as a search interface at `http://clic.ub.edu/corpus/en/paraphrases-en`. The annotation guidelines are also available at the same website.

Besides the specific contributions mentioned above, the present research has unveiled some problematic and sometimes illuminating issues involved in paraphrasing, which are discussed below.

Other considerations:

10. The broad and multifaceted nature of paraphrasing has complicated the creation of a precise and generally assumed paraphrase characterization. Instead, it has been addressed from different perspectives and pursuing different objectives, giving rise to partial and ad-hoc analyses (Section 2.1).

11. Computational linguistics has not always found in linguistics adequate paraphrase knowledge to base its methods and systems on. As a conse-

quence, it has developed its own approaches to treating paraphrasing, which shed new light on the understanding of the phenomenon (Section 2.1).

12. Paraphrase boundaries are flexible and they depend on the working field, task, and objectives: some linguistic phenomena that are considered to be paraphrases in one approach, are not in another (Section 2.1).

13. Two main paraphrase genres exist: reformulative and non-reformulative paraphrases. The former are those paraphrases created in the framework of a reformulation, that is, the paraphrase is created by applying modifications to a source text snippet. The latter are those paraphrase pairs not born in reformulation processes. Reformulative paraphrases tend to be nearer in form than non-reformulative paraphrases. For computational linguistics, dealing with these two types of paraphrases requires different techniques (Sections 3.1 and 3.2).

14. In reformulative paraphrasing, the most frequent paraphrase types are same-polarity substitutions and addition/deletion. In non-reformulative paraphrasing, the most frequent types are same-polarity substitutions and semantics-based changes. This shows where to put the focus in improving paraphrase systems (Section 3.2).

15. The intrinsic variety of paraphrasing demands a highly expressive representation formalism. Nevertheless, high expressive capacity generally entails low computational efficiency, as, in general, there is a trade-off between the two. Finding an adequate balance is needed (Section 2.3).

16. The multifaceted nature of paraphrasing prevents the creation of corpora covering paraphrasing as a whole. Instead, state-of-the-art corpora cover specific facets of paraphrasing (Section 3.1).

5.2 Future Directions

Despite the efforts to apprehend paraphrasing from linguistics and computational linguistics, there is still an issue as to where further work needs to be undertaken. This thesis opens the path to a number of lines of research. Each article in the compendium is closed with concrete future direction on the topic it addresses. In this section, I will give a general overview of the most significant ones.

Further work should be done on paraphrase characterization. In concrete, I have detected three borderline-paraphrase areas; nevertheless, others may exist. On the other hand, two types in our typology, which are very general, require further analysis to see whether they accept a more fine-grained subdivision: semantics-based and syntax-discourse structure changes.

This thesis has made a small primary step in paraphrase representation. Further work needs to be done to find a representation approach capable of dealing with the complexity of paraphrase at a reasonable computational cost.

This is the first time an annotation infrastructure including annotation guidelines and inter-annotator agreement measures for paraphrase-type annotation has been created. It constitutes a pioneering work in an almost unexplored field, which opens the path to new proposals and improvements. In concrete, further work could be done on solving the issue of false negatives (cases considered to be disagreements when they should not) and false positives (cases erroneously considered as agreements) in the IAPTA measures. The corpora annotated using this infrastructure constitute a rich source of information and a powerful resource for deriving new tools. They can also be used in machine learning to learn to automatically annotate paraphrase phenomena.

Regarding the WRPA method and corpus, potential lines for future work are the application of our method to other types of patterns, to other relations, and to other languages. Extending WRPA to non-structured corpora to extract patterns from the whole web applying bootstrapping constitutes another promising line of research.

Bibliography

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38(1):135–187.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 597–604, Ann Arbor (MI).

Barrón-Cedeño, A. (2012). *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism*. PhD thesis, Universitat Politècnica de València, Valencia.

Barrón-Cedeño, A., Vila, M., Martí, M., and Rosso, P. (2013, to appear). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*. DOI: 10.1162/COLI_a_00153.

Barrón-Cedeño, A., Vila, M., and Rosso, P. (2012). Detección automática de plagio: De la copia exacta a la paráfrasis. In Garayzábal, E., Jiménez, M., and Reigosa, M., editors, *Lingüística Forense: La Lingüística en el Ámbito Legal y Policial*, pages 71–101. Euphonía Ediciones, Madrid.

Barzilay, R. and McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 50–57, Toulouse.

Bhagat, R. (2009). *Learning Paraphrases from Text*. PhD thesis, University of Southern California, Los Angeles.

Bhagat, R. and Ravichandran, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2008)*, pages 674–682, Columbus (OH).

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011)*, volume 1, pages 190–200, Portland (OR).

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge.

Cohn, T., Callison-Burch, C., and Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–614.

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

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, pages 9–16, Jeju Island.

Dras, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD thesis, Macquarie University, Sydney.

Dutrey, C., Bernhard, D., Bouamor, H., and Max, A. (2011). Local modifications and paraphrases in Wikipedia's revision history. *Procesamiento del Lenguaje Natural*, 46:51–58.

España-Bonet, C., Vila, M., Rodríguez, H., and Martí, M. A. (2009). CoCo, a web interface for corpora compilation. *Procesamiento del Lenguaje Natural*, 43:367–368.

Fuchs, C. (1988). Paraphrases prédicatives et contraintes énonciatives. In Bès, G. G. and Fuchs, C., editors, *Lexique et Paraphrase*, number 6 in Lexique, pages 157–171. Presses Universitaires de Lille, Villeneuve d'Ascq.

Fuchs, C. (1994). *Paraphrase et énonciation*. Ophrys, Paris.

Fujita, A. (2005). *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. PhD thesis, Nara Institute of Science and Technology, Nara.

Gonzàlez, E., Rodríguez, H., Turmo, J., Comas, P. R., Naderi, A., Ageno, A., Sapena, E., Vila, M., and Martí, M. A. (2013, to appear). The TALP participation at TAC-KBP 2012. In *Proceedings of the 5th Text Analysis Conference (TAC 2012)*, Gaithersburg (MD).

Harris, Z. (1954). Distributional structure. *Word*, 10(2–3):146–162.

Harris, Z. (1957). Co-occurence and transformation in linguistic structure. *Language*, 33(3):283–340.

Herrera, J., Peñas, A., and Verdejo, F. (2007). Paraphrase extraction from validated question answering corpora in Spanish. *Procesamiento del Lenguaje Natural*, 39:37–44.

Kouylekov, M. and Negri, M. (2010). An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations (ACLDemos 2010)*, pages 42–47, Uppsala.

Lin, D. and Pantel, P. (2001). DIRT-Discovery of Inference Rules from Text. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pages 323–328, San Francisco (CA).

Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Martin, R. (1976). *Inférence, antonymie et paraphrase*. Librarie C. Klincksieck, Paris.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 997–1005, Beijing.

Recasens, M. (2010). *Coreference: Theory, Annotation, Resolution, and Evaluation*. PhD thesis, Universitat de Barcelona, Barcelona.

Recasens, M. and Vila, M. (2010). On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.

Vila, M., Bertran, M., Martí, M. A., and Rodríguez, H. (submitted-a). Corpus annotation with paraphrase types. New annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation*.

Vila, M. and Dras, M. (2012). Tree edit distance as a baseline approach for paraphrase representation. *Procesamiento del Lenguaje Natural*, 48:89–95.

Vila, M., González, S., Martí, M. A., Llisterri, J., and Machuca, M. J. (2010a). ClInt: a biligual Spanish-Catalan spoken corpus of clinical interviews. *Procesamiento del Lenguaje Natural*, 45:105–111.

Vila, M., Martí, M. A., and Rodríguez, H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Vila, M., Martí, M. A., and Rodríguez, H. (submitted-b). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Lingua*.

Vila, M., Rodríguez, H., and Martí, M. A. (2010b). WRPA: A system for relational paraphrase acquisition from Wikipedia. *Procesamiento del Lenguaje Natural*, 45:11–19.

Vila, M., Rodríguez, H., and Martí, M. A. (submitted-c). Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus. *Natural Language Engineering*.

Žolkovskij, A. and Mel'čuk, I. (1965). O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija*, 5:23–28.

Wintner, S. (2009). What science underlies natural language engineering? *Computational Linguistics*, 35(4):641–644.

Appendix A

Previous-Version Publications

Marta Vila (Universitat de Barcelona)
M. Antònia Martí (Universitat de Barcelona)
Horacio Rodríguez (Universitat Politècnica de Catalunya)

(2011)

Paraphrase concept and typology. A linguistically based and computationally oriented approach

Procesamiento del Lenguaje Natural 46:83–90.

Journal URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln`

Abstract In this paper, we present a critical analysis of the state of the art in the definition and typologies of paraphrasing. This analysis shows that there exists no characterization of paraphrasing that is comprehensive, linguistically based and computationally tractable at the same time. The following sets out to define and delimit the concept on the basis of the propositional content. We present a general, inclusive and computationally oriented typology of the linguistic mechanisms that give rise to form variations between paraphrase pairs.

Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 83-90 recibido 17-01-11 revisado 16-02-11 aceptado 24-02-11

Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach*

Concepto y tipología de paráfrasis. Una aproximación lingüística orientada al tratamiento computacional

Marta Vila, M. Antònia Martí
CLiC-UB
Gran Via 585, 08007 Barcelona
marta.vila@ub.edu, amarti@ub.edu

Horacio Rodríguez
TALP-UPC
Jordi Girona 1-3, 08034 Barcelona
horacio@lsi.upc.es

Resumen: En este artículo, se presenta un análisis crítico de la bibliografía sobre la definición de paráfrasis y su tipología. Dicho análisis pone de manifiesto que no existe una caracterización de la paráfrasis completa y lingüísticamente fundamentada que, al mismo tiempo, sea tratable computacionalmente. Se propone una definición y delimitación del concepto fundada sobre el contenido proposicional. Sobre esta base, se ha elaborado una tipología general, inclusiva y orientada al tratamiento computacional de los mecanismos lingüísticos que dan lugar a la variación en la forma de los pares parafrásticos.
Palabras clave: Paráfrasis, límites de la paráfrasis, tipología de paráfrasis.

Abstract: In this paper, we present a critical analysis of the state of the art in the definition and typologies of paraphrasing. This analysis shows that there exists no characterization of paraphrasing that is comprehensive, linguistically based and computationally tractable at the same time. The following sets out to define and delimit the concept on the basis of the propositional content. We present a general, inclusive and computationally oriented typology of the linguistic mechanisms that give rise to form variations between paraphrase pairs.
Keywords: Paraphrasing, paraphrase boundaries, paraphrase typology.

1 Introduction

Paraphrasing stands for sameness of meaning between different wordings. Prototypical paraphrase examples can be seen in (1) and (2), where the semantic content remains the same despite the differences in the form: *significant* is substituted for its synonym *considerable* in (1-b), and (2) illustrates an active/passive diathesis alternation.

(1) a. This task requires *significant* knowledge to be successful.
b. This task requires *considerable* knowledge to be successful.

(2) a. The Romans constructed that bridge.
b. That bridge was constructed by the Romans.

The omnipresence of paraphrasing in natural language gives rise to the need to apprehend the mechanisms that govern this phenomenon from a linguistic perspective. Natural Language Processing (NLP) components dealing with paraphrasing, in turn, appear to have great potential for the improvement of systems for understanding and generation, such as question answering, summarization or machine translation. Despite its potential, a linguistically backed and, at the same time, computationally efficient account of the whole paraphrase phenomenon has not yet been developed.

In this work, a proposal for the characterization of paraphrasing is presented. We follow two different perspectives: an intensional perspective setting out the properties a linguistic expression needs to be considered a paraphrase (the concept), and an extensional perspective specifying the objects that fall under paraphrasing (typology). It consists in a comprehensive and linguistically founded

* This work is supported by the FPU grant AP2008-02185 from the Spanish Ministry of Education, and the Text-Knowledge 2.0 (TIN2009-13391-C04-04) and KNOW2 (TIN2009-14715-C04-04) projects from the Spanish Ministry of Science and Innovation.

ISSN 1135-5948

Marta Vila (Universitat de Barcelona)
Horacio Rodríguez (Universitat Politècnica de Catalunya)
M. Antònia Martí (Universitat de Barcelona)

(2010)

WRPA: A system for relational paraphrase acquisition from Wikipedia

Procesamiento del Lenguaje Natural 45:11–19.

Journal URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln`

Abstract In this paper we present WRPA, a system for Relational Paraphrase Acquisition from Wikipedia. WRPA extracts paraphrasing patterns that hold a particular relation between two entities taking advantage of Wikipedia structure. What is new in this system is that Wikipedia's exploitation goes beyond infoboxes, reaching itemized information embedded in Wikipedia pages. WRPA is language independent, assuming that there exists Wikipedia and shallow linguistic tools for that particular language, and also independent of the relation addressed.

Procesamiento del Lenguaje Natural, Revista nº45, septiembre 2010, pp 11-19 recibido 30-04-10 revisado 17-05-10 aceptado 20-05-10

WRPA: A System for Relational Paraphrase Acquisition from Wikipedia*

WRPA: Un sistema para la adquisición de paráfrasis de relaciones de la Wikipedia

Marta Vila
CLiC-UB
Gran Via 585, Barcelona
marta.vila@ub.edu

Horacio Rodríguez
TALP-UPC
Jordi Girona 1-3, Barcelona
horacio@lsi.upc.es

M. Antònia Martí
CLiC-UB
Gran Via 585, Barcelona
amarti@ub.edu

Resumen: En este artículo se presenta WRPA, un sistema para la Adquisición de Paráfrasis de Relaciones de la Wikipedia. Aprovechando la estructura de la Wikipedia, WRPA extrae patrones de paráfrasis que expresan una determinada relación entre dos entidades. La novedad de este sistema reside en que se explota dicha enciclopedia más allá de las fichas (o infoboxes), aprovechando información itemizada que contienen algunas de sus páginas. WRPA es independiente de la lengua, asumiendo la existencia, para la lengua en cuestión, de Wikipedia y de herramientas para el tratamiento superficial del lenguaje, así como independiente de la relación tratada.
Palabras clave: Paráfrasis, Extracción de Información, Extracción de Relaciones, Wikipedia.

Abstract: In this paper we present WRPA, a system for Relational Paraphrase Acquisition from Wikipedia. WRPA extracts paraphrasing patterns that hold a particular relation between two entities taking advantage of Wikipedia structure. What is new in this system is that Wikipedia's exploitation goes beyond infoboxes, reaching itemized information embedded in Wikipedia pages. WRPA is language independent, assuming that there exists Wikipedia and shallow linguistic tools for that particular language, and also independent of the relation addressed.
Keywords: Paraphrasing, Information Extraction, Relation Extraction, Wikipedia.

1 Introduction

Paraphrasing stands for (approximate) sameness or equivalence of meaning between different wordings. This definition puts into words a vague and complex phenomenon with a broad range of manifestations that can involve lexical, syntactic, semantic and pragmatic knowledge. NLP components dealing with paraphrasing appear to have great potential for the improvement of understanding and generation systems such as question-answering, summarization or machine translation. As a result, it has been the focus of a large amount of work in the last couple of decades.

In this paper we present WRPA, a system for Relational Paraphrase Acquisition from Wikipedia. Due to the vagueness and complexity of the paraphrasing phenomenon, we restrict ourselves to relational paraphrases, i.e., those expressing a relation between two entities, because they constitute a well delimited but in turn comprehensive set.

Our approach to paraphrasing has a close relationship with Information Extraction systems, as they are frequently used for extracting semantic relations. However, while IE systems are geared towards obtaining the semantic relation held by pairs of entities—named the source and the target—in a corpus (Figure 1), paraphrasing focusses on the wording used to express those relations (patterns and instances in Figure 1). A lot of techniques can be used in IE, e.g., machine learning and rule- or pattern-based techniques. WRPA is only related to the latter.

Our approach is based on Harris (1954)'s Distributional Hypothesis which states that

* This work is supported by the FPU Grant AP2008-02185 from the Spanish Ministry of Education, and the Text-Knowledge 2.0 (TIN2009-13391-C04-04) and KNOW2 (TIN2009-14715-C04-04) projects.

Alberto Barrón-Cedeño (Universitat Politècnica de València)
Marta Vila (Universitat de Barcelona)
Paolo Rosso (Universitat Politècnica de València)

(2012)

Detección automática de plagio: de la copia exacta a la paráfrasis (Automatic plagiarism detection: from exact copy to paraphrasing)

In Elena Garayzábal, Miriam Jiménez and Mercedes Reigosa, eds.*Lingüística Forense: La Lingüística en el Ámbito Legal y Policial*, Euphonía Ediciones, Madrid, pp. 71–101.

Abstract Plagiarism, unauthorized and non-referenced text reuse, is a phenomenon that has gained great interest because of the amount of bibliographic resources and information available online. Due to the magnitude of the problem, the manual revision of texts looking for plagiarism is virtually impossible. Plagiarism automatic detectors arise as a precautionary and corrective measure to assist humans to detect plagiarism in texts, which is a forensic-linguistics task. Automatic-detection tools only seek to assist humans in the detection process, providing evidence of potential cases of plagiarism. The final decision and subsequent actions must be taken by the expert. This chapter briefly introduces plagiarism and presents its relationship to paraphrasing. This linguistic phenomenon, although it is on the basis of plagiarism acts, has not received sufficient attention from the experts. In this regard, we believe that existing work on paraphrasing in the fields of linguistics and natural language processing show a great potential for the automatic detection of plagiarism.

Detección automática de plagio: de la copia exacta a la paráfrasis *

Alberto Barrón-Cedeño[1], Marta Vila[2] y Paolo Rosso[1]

[1]Natural Language Engineering Lab. - ELiRF
Universidad Politécnica de Valencia
{lbarron, prosso}@dsic.upv.es

[2]CLiC, Departament de Lingüística
Universitat de Barcelona
marta.vila@ub.edu

Resumen

El plagio, el reuso no autorizado y sin referencia de texto, es un fenómeno que ha cobrado gran interés debido a la enorme cantidad de recursos bibliográficos e información al alcance de la mano en Internet. Debido a la magnitud del problema, la revisión manual de los textos en busca de plagio es prácticamente imposible. Los conocidos como detectores automáticos de plagio surgen como una medida precautoria y correctiva para asistir al humano en la detección de plagio en textos, una tarea de la lingüística forense.

Debe observarse que las herramientas de detección automática de plagio buscan solamente asistir al humano en la detección, proveyéndole de las mayores pruebas posibles de un potencial caso de plagio. La decisión final, así como las acciones pertinentes, debe ser tomada por el experto.

En este capítulo se introduce brevemente el plagio y se presenta su relación con la paráfrasis. Este fenómeno lingüístico, si bien se encuentra en la base del acto de plagiar, no ha recibido atención suficiente por parte de los expertos. En este sentido, consideramos que los trabajos existentes sobre paráfrasis en el ámbito de la lingüística y el procesamiento del lenguaje natural son valiosas para la detección automática de plagio.

Palabras clave: detección de plagio, detección de paráfrasis, lingüística forense

*Esta contribución está orientada a la descripción de los conceptos y métodos subyacentes a la detección automática de plagio y no al análisis de las herramientas comerciales disponibles. Si el lector está interesado en las herramientas, puede considerar los servicios otorgados por compañías como Turnitin (iParadigms, 2010) o DOC Cop (McCrohon, 2010). Adicionalmente, sugerimos consultar (Maurer et˜al., 2006); particularmente las secciones 4 y 5. Por otro lado, este análisis está enfocado al plagio de texto. El lector interesado en el plagio de otro tipo de recursos, como por ejemplo música, puede consultar los trabajos de Robine et˜al. (2007) y Müllensiefen and Pendzich (2009).

Appendix B

Collateral Publications

Cristina España-Bonet (Universitat Politècnica de Catalunya)
Marta Vila (Universitat de Barcelona)
Horacio Rodríguez (Universitat Politècnica de Catalunya)
M. Antònia Martí (Universitat de Barcelona)

(2009)

CoCo, a web interface for corpora compilation

Procesamiento del Lenguaje Natural 43:367–368.

Journal URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln`

Abstract CoCo is a collaborative web interface for the compilation of linguistic resources. In this demo we are presenting one of its possible applications: paraphrase acquisition.

Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 367-368 recibido 1-05-2009; aceptado 5-06-2009

CoCo, a web interface for corpora compilation*

CoCo, una interfaz web para la compilación de corpus lingüísticos

C. España-Bonet(1), M. Vila(2), H. Rodríguez(1), M.A. Martí(2)
(1) TALP Research Center
LSI Department
Universitat Politècnica de Catalunya
Jordi Girona 1–3, 08034 Barcelona
(2) CLiC
Linguistics Department
Universitat de Barcelona
Gran Via 585, 08007 Barcelona
cristinae@lsi.upc.edu, marta.vila@ub.edu, horacio@lsi.upc.es, amarti@ub.edu

Resumen: CoCo es una interfaz web colaborativa para la compilación de recursos lingüísticos. En esta demo se presenta una de sus posibles aplicaciones: la obtención de paráfrasis.
Palabras clave: Paráfrasis, Web Colaborativa, Interfaces

Abstract: CoCo is a collaborative web interface for the compilation of linguistic resources. In this demo we are presenting one of its possible applications: paraphrase acquisition.
Keywords: Paraphrasing, Collaborative Web, Interfaces

1. Introduction

CoCo[1] (Corpora Compilation) is a web interface designed for the compilation of linguistic corpora. Similar tools for a specific task or corpus can be found, such as the work by Chklovski (Chklovski, 2005b; Chklovski, 2005a) for collecting paraphrases or the Anawiki web page[2] by Poesio *et al.* (Poesio, Kruschwitz, and Jon, 2008) devoted to creating anaphorically annotated resources. As CoCo, these tools take advantage of web cooperation.

The system is open to any volunteer interested in contributing to the creation and widening of linguistic corpora, and it is currently being used by undergraduates at the University of Barcelona. CoCo will deal with different tasks, Paraphrasing, Coreference or Textual Entailment among them. The system is now prepared to gather data in four working languages: Catalan, Spanish, English and Arabic.

As stated previously, anyone can register and contribute as a user. Moreover, there is a subgroup of expert users which are allowed to control, modify and validate what is being incorporated into the database.

In the following section, we describe the task for which CoCo is currently being used: paraphrase acquisition.

2. Paraphrase Acquisition

Up to now, the operative part of the web is devoted to compiling a corpus of paraphrases. Paraphrases are understood as the different ways in which the same (or similar) content is expressed linguistically. There are two different approaches to the task. The first one, *General Paraphrasing*, aims to collect paraphrases of any kind. As a first input, the database has been filled with the paraphrases from the Microsoft corpus (Microsoft, 2005). The second, *Relational Paraphrasing*, is restricted to the paraphrases that express some kind of relationship between two entities. For now, it is devoted to the relationship of authorship.

2.1. General Paraphrasing

In the *Paraphrasing* section, the user is encouraged to widen the paraphrase corpus in three different ways.

- *Pair Generation.* A pair of paraphrasing sentences must be introduced.

* This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02), TEXT-MESS Lang2World (TIN2006-15265-C06/06), Ancora-Nom (FFI2008-02691-E/FILO) and the DOI/REFLEX-NBCHC050031 program as well as the FI Grant (2009FI_B 00690) from the Generalitat de Catalunya.

[1] http://www.lsi.upc.edu/∼textmess/

[2] http://www.anawiki.com/

- *Pair Completion.* Given a fixed original sentence, the user proposes a paraphrase. The original sentence is chosen from the existing corpora either randomly, sequentially or filtering by some criteria such as length or words contained.
- *Template Generation.* The same as in the previous task, but now the user is given part of the requested paraphrase. Some of the words are hidden so that the sentence only needs to be completed. The amount of hidden information can be modified by the user, who can hide or reveal words.

These three main tasks are accompanied by a section that allows users to search within the corpora or to modify their items.

Moreover, users subscribed as experts can evaluate already existing paraphrases. A pair is not accepted (or rejected) as a paraphrasing pair until it has been validated by at least three expert users.

2.2. Relational Paraphrasing

The second approach to Paraphrase Acquisition in the CoCo tool is that devoted to the collection of relational paraphrases. For now, the task focuses on *Authorship Paraphrasing*, that is, on those paraphrases that express some kind of relationship between an author and their work. We understand the relationship of authorship in a broad sense. It includes the relationship between painters and their paintings, between scientists and their theories, or between businessmen and their companies, to mention some examples.

As in the case of *General Paraphrasing* the user can choose different subtasks:

- *Authorship Generation.* A pair (author, work) is randomly shown and the user is asked to write a sentence containing the two items in an order which is randomly determined. A visual example of this task can be seen in Figure 1.
- *Web Evaluation.* Sentences automatically extracted from the web can be evaluated.

This section is already being used by students of Linguistics and Documentation at the University of Barcelona.

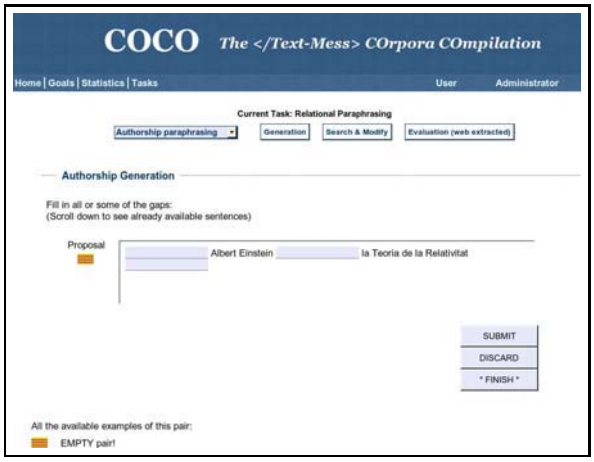

Figura 1: Screenshot of CoCo. This page allows the user to complete an authorship paraphrasing.

3. Conclusions

In this demo we are presenting a new tool for corpora compilation, currently being used for paraphrase acquisition. Up to now, the results obtained demonstrate CoCo's usefulness for the collection of corpora oriented to specific purposes. This is the case of the authorship paraphrase corpus that have been obtained for both Catalan and Spanish. The paraphrases obtained are being exploited in a current research on paraphrasing in the field of Linguistics.

References

Chklovski, Timothy. 2005a. 1001 paraphrases: Incenting responsible contributions in collecting paraphrases from volunteers. In *Proceedings of KCVC 2005*, pages 16–20.

Chklovski, Timothy. 2005b. Collecting paraphrase corpora from volunteer contributors. In *Proceedings of K-CAP 2005*, pages 115–120. ACM.

Microsoft. 2005. Microsoft research paraphrase corpus. http://research. microsoft.com/research/downloads.

Poesio, Massimo, Udo Kruschwitz, and Chamberlain Jon. 2008. Anawiki: Creating anaphorically annotated resources through web cooperation. In *Proceedings of LREC 2008*, pages 2352–2355.

Edgar Gonzàlez (Universitat Politècnica de Catalunya)
Horacio Rodríguez (Universitat Politècnica de Catalunya)
Jordi Turmo (Universitat Politècnica de Catalunya)
Pere R. Comas (Universitat Politècnica de Catalunya)
Ali Naderi (Universitat Politècnica de Catalunya)
Alicia Ageno (Universitat Politècnica de Catalunya)
Emili Sapena (Universitat Politècnica de Catalunya)
Marta Vila (Universitat de Barcelona)
M. Antònia Martí (Universitat de Barcelona)

The TALP participation at TAC-KBP 2012

To appear in *Proceedings of the 5th Text Analysis Conference (TAC 2012), Gaithersburg (MD).*

Task URL `http://www.nist.gov/tac/2012/KBP/`

Abstract This document describes the work performed by the Universitat Politècnica de Catalunya (UPC) in its first participation at TAC-KBP 2012 in both the Entity Linking and the Slot Filling tasks.

The TALP participation at TAC-KBP 2012

E. Gonzàlez[2], H. Rodríguez[1], J. Turmo[1], P.R. Comas[1], A. Naderi[1], A. Ageno[1], E. Sapena[1], M. Vila[3] and M.A. Martí[3]
[1] TALP Research Center, UPC, Spain.
[2] TALP Research Center, UPC, Spain. Now at Google.
[3] CLiC, Universitat de Barcelona, Spain.
{egonzalez,horacio,turmo}@lsi.upc.edu
{pcomas,anaderi,ageno,esapena}@lsi.upc.edu
{marta.vila,amarti}@ub.edu

Abstract

This document describes the work performed by the Universitat Politècnica de Catalunya (UPC) in its first participation at TAC-KBP 2012 in both the Entity Linking and the Slot Filling tasks.

1 Introduction

Both Entity Linking (EL) and Slot Filling (SF) tasks aim at extracting useful information in order to enrich a knowledge base. This document describes the work carried out by the TALP research group of the Universitat Politècnica de Catalunya in its first participation at TAC-KBP 2012 in both the Entity Linking and the Slot Filling tasks for English. The purpose of this first participation has been mainly exploratory, aiming at performing a preliminary assessment of our approaches (one for EL, two different ones for SF) and drawing conclusions on how to improve them.

EL is the task of referring a Named Entity mention to the unique entry within a reference knowledge base (KB). TAC-KBP track defines the task of EL as follows: having a set of queries, each one consisting of a target name string along with a background document in which the target name string can be found and a source document collection from which systems can learn, the EL system is required to select the appropriate KB entry. Queries generally consist of the same name string from different docids. The system is expected to distinguish the ambiguous names (e.g., Barcelona could refer to the sport team, the university, city, state, or person). In TAC-KBP 2012, we have sent one run and evaluated our EL system just for Mono-lingual Entity Linking. The run did not access the web and also did not use query offsets during the evaluation.

In the SF task, the given set of queries is a set of entity KB nodes that must be augmented by extracting all the new learnable slot values for the entity as found in a large corpus of documents. SF involves mining information from the documents and therefore applies Information Extraction (IE) techniques. We have only participated in the English Mono-lingual Slot Filling task, submitting two runs. Both runs differ in the IE approach employed to detect possible query slot fillers in the candidate documents. The first approach is supervised (based on distant learning), whereas the second one is completely unsupervised (based on minority clustering).

The rest of the document is structured as follows. Section 2 describes the query preprocessing step, shared by all the systems. Section 3 is devoted to our Entity Linking approach. In section 4 we describe our Slot Filling approaches, including the shared document preprocessing step and the two different IE approaches applied. Finally, section 5 presents and analyses the results obtained in KBP 2012 by our approaches in both tasks.

2 Query preprocessing

Query preprocessing consists of the following tasks:

- For both SF and EL, a crucial point is generating the set of alternate names, A, for the query entity. For generating A we have used 4 sources of information: The query name (either

a word or a multiword) and its type, the available structured information and textual (non structured) information from documents supporting the query.

- For EL, classifying the query entity into the appropriate query type (PER, ORG or GPE) using the Stanford NERC[1]), over the reference document attached to the query. For SF this process is not needed because the type of the entity is known.
- For SF, obtaining, when existing, the corresponding node in KB. The facts associated to this node are retrieved.
- For both SF and EL, we look at Wikipedia (WP) for the possible existence of the corresponding page. Disambiguation pages are discarded. If some infobox is found their slots and values are retrieved.
- For EL, if the type is GPE we look at geographic gazetteers (GNIS[2] and GEONAMES[3]) and select the corresponding entries.

The documents we use as knowledge sources are:

- The reference document attached to the query.
- For SF, when a KB node is included in the query, the attached description document, if existing, and the facts associated to this node when containing free text.
- For both SF and EL when a WP page exists. the textual content of the page is selected.

Using all these knowledge sources, our way of building set A is the following: The set is initialized with the query name scored with 1. Then a set of enrichment procedures are iteratively applied until no more alternate names are found. The new alternate names are scored decresingly. There are two types of procedures for generating altrenate names: generic and type-specific. Generic procedures are the following:

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] http://geonames.usgs.gov/geonames/stategaz

[3] http://www.geonames.org

Query	Name	Alternate names	#
SF558	Barbara Boxer	1.0 Barbara Levy Boxer 0.8 Barbara L. Boxer 0.64 B. L. Boxer 0.56 B. Boxer 0.49 Boxer ...	37
SF520	Hong Kong Disneyland	1.0 Hong Kong Disneyland 1.0 HKDL 0.8 H. Kong Disneyland 0.8 Hong K. Disneyland 0.7 Hong Disneyland ...	30

Table 1: Examples of alternate names

1. We select a set of pairs (*WP infobox, slot*) where slot refers to an alternate name (e.g. *formal name*, *alias*, *nickname*, *also_known_as*, etc.). If we have a WP page we extract these values and insert them into A also with the maximum score. We proceed in the same way with the KB nodes using in this case available facts.
2. We apply the SF corresponding to the generic slot *alternate_name* existing for both PER and ORG, as described in section 4.

Specific procedures, applied iteratively over all the current members of A:

1. For PER. We use a DCG grammar of English person names for extracting the structure of a complex name. For instance, from *Paul Auster* our aim is to detect that the first name is *Paul* and the family (main) name is *Auster*. We then generate valid variants of the original name always preserving the family name. These variants are scored accordingly with the generalization degree, in our example: (*P. Auster*, 0.8), (*Auster*, 0.6).
2. For ORG. We have developed a set of 12 acronym/expansion mapping functions owning credibility scores which can be applied in the two directions:
   - Starting from an acronym we look up in the textual description for the occurrence of valid expansions applying our mapping

functions. We score the valid variants with the credibility of the applied function.
- Starting from a complete form we perform acronym detection equally scored.
- New forms of ORG names can be found removing common company suffixes (e.g. *Inc*, *Company*, etc.).

3. For GPE we extract all the variants existing in the geographic gazetteers and score them with the edit distance between the original form and the variant.

Some examples of alternate names generated with this procedure are shown in Table 1.

3 Entity Linking

Our approach is inspired by recent works on EL using graph-based methods such as (Guo et al., 2011; Hachey et al., 2011; Han et al., 2011). It consists of three steps for each query. Briefly, given a query, we start by selecting those KB nodes which are candidates to be the correct entity for the query (candidate generation step). Then, we create a graph with the selected candidates and information related to them (graph generation step). Finally, we explore the graph relations for ranking the candidates in order to select the most appropriate one for the query (graph-based ranking step).

The rest of this section describes our methods for candidates generation and graph generation, as well as the graph-based ranking approach.

3.1 Candidate Generation

As KB usually contains a large number of entries, it is desirable to avoid brute force comparisions between a particular query and all KB entries and to reduce the search space of potential candidates. Our priority, however, is to generate a large candidate set instead of a smaller one in order to increase recall (McNamee et al., 2010; Lehmann et al., 2010).

In order to get the set of candidates for a particular query, q, our system performs two steps. First, the query is preprocessed using the procedure described in Section 2. So, q is classified as PER, ORG or GPE, and the set A of alternate names for the query name, m, is obtained. Then, the set of candidates, C, for q is retrieved from the KB being each $c_i \in C$ an entry corresponding to one of the alternate name.

3.2 Graph Generation

From C, we create a graph to represent knowledge related to the candidates, which will be useful for later selecting the most appropriate one for q. We can describe the graph we use as follows:

> *The directed graph* ***G****=(V, E), where the vertices set* ***V*** *contains nodes representing all the candidates in* C*, the query, and the property values for the candidates and the query; and the directed edges set* ***E*** *consists of all weighted labelled connections between the vertices.*

The graph is initialized to a set of disjoined nodes corresponding to the elements of C. To enrich the graph, we need to retrieve the informative parts of each candidate from the KB entry: the set of facts and the wikitext if it exists. In the case of facts, considering each one as a property with a particular value, the property is represented as the label of a directed edge in the graph, whilst the value is represented as a node connected by the edge from the candidate. In the case of wikitext, we extract all NE mentions of types PER, ORG, LOC and MISC from the first 30 tokens.[4] Here, we consider that the most relevant information related to the candidate in the wikitext is frequently described in the first part of the text. Each extracted NE is represented as a node connected with an unlabeled edge to the candidate.

Moreover, we also represent the query in the graph by including a new node, q. Then, we take all NEs occurring within the context of all sentences of the background document in which the query name occurs. These NEs are represented as new nodes in the graph connected to the query node by an unlabeled edge.

An example of a graph generated for the query related to "*Picasso*" is depicted in Figure 1. Candidates for this query are "*Pablo Picasso*" a Spanish painter, "*Paloma Picasso*" a fashion designer and the youngest daughter of "*Pablo Picasso*," and "*Francisco Picasso*" an Olympic and national-record holding swimmer from Uruguay. Some properties of the first candidate are *Place of Birth = Málaga, Spain* and *Children = Paloma Picasso*.

[4] The same NERC is used (Stanford NERC).

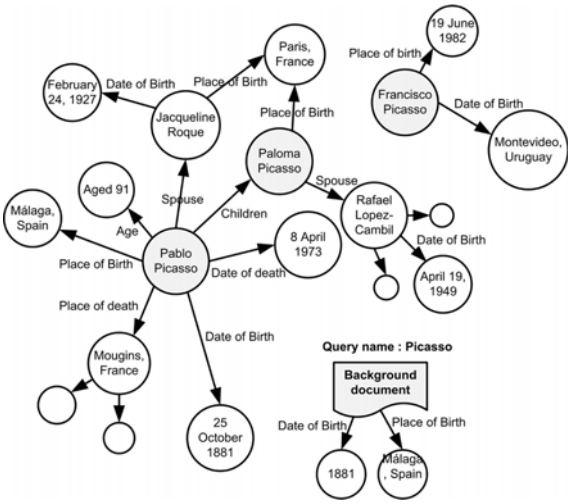

Figure 1: A graph for query name *Picasso*

Note that in the second case the relation is between two candidates.

All edges have a weight which represents the degree of dependency assigned to them. It is used to model and measure the connectivity between nodes. We have manually set these weights as follows. The weights of edges obtained from KB facts are set to 20, which is the highest weight, as we considered them as true information. The weights of those acquired from the candidate wikitext are set to 5. Moreover, the weights of edges related to the query are set to 1.

3.3 Graph-Based Ranking

Given the graph G, the system has to select the correct candidate as the KB reference of q. We score all the candidates by comparing their similarity/relatedness with the query node and select the one having the highest score.

Consider that $C = (c_1, c_2, \ldots, c_n)$ is the set of candidate nodes, q is the query node in the graph G, $P_{c_k} = (P^1_{c_k}, P^2_{c_k}, \ldots, P^m_{c_k})$ is the set of paths between q and c_k without considering direction of edges, where each $P^i_{c_k}$ is represented by the sequence of weights corresponding to the edges in the path, $P^i_{c_k} = \langle w_1, w_2, \ldots, w_r \rangle$, and s_{c_k} is the score of the candidate node c_k, then:

$$s_{c_k} = \begin{cases} \sum_{P^i_{c_k} \in P_{c_k}} \sum_{w_j \in P^i_{c_k}} w_j & \text{if } P_{c_k} \neq \emptyset \\ 0 & \text{if } P_{c_k} = \emptyset \end{cases} \quad (1)$$

Assuming m_q as the query name and $S = \{s_{c_k}\}$, the link between m_q and KB is obtained as follows:

$$link(m_q) = \begin{cases} c & \text{if } \exists c \in C : s_c = \max(S) \geq \beta \\ \text{NIL} & \textit{otherwise} \end{cases} \quad (2)$$

where, β is a threshold, different for each query type estimated using the first 100 queries of KBP 2011 manually tagged.

Figure 2 shows a sample graph structure. As shown in this figure, consider three candidates $C = (c_1, c_2, c_3)$ for a particular query q in the graph. Each candidate is connected to their corresponding properties by the directed edges. Each edge has an assigned weight, w. The initial score for the candidates, s_{c_1}, s_{c_2} and s_{c_3}, is 0 and the initial one for the query, s_q, is 1. Then, each candidate is scored by the products of s_q and the sum of weights for all paths from q to the candidate. In the example:

$$\begin{aligned} s_{c_1} &= s_q \cdot (w^2_q + w^1_{c_1}) \\ s_{c_2} &= s_q \cdot (w^2_q + w^1_{c_2}) + s_q \cdot (w^1_q + w^2_{c_2}) \\ s_{c_3} &= 0, \end{aligned} \quad (3)$$

where w^j_i stands for the weight of the j-th edge from node i.

We select the best scored candidate and return it as our solution if the score is over the threshold β. Otherwise the result is NIL.

In the case of NIL, sometimes, several EL queries refer to the same non-KB (NIL) entity. In these cases, these queries should be collected into one identifiable NIL cluster. For NIL clustering, those queries belonging to the same cluster take the same NIL id in the form of *NILxxxx* being *xxxx* a natural number. The method that we apply for NIL clustering is similar to the approach for ranking candidates.

If query q in the graph G results NIL, then we create a NIL graph (G_{NIL}) that represents several clusters, each one including previous queries related to the same non-KB entity. Each cluster is represented just with its first NIL query (i.e. the medoid) and its

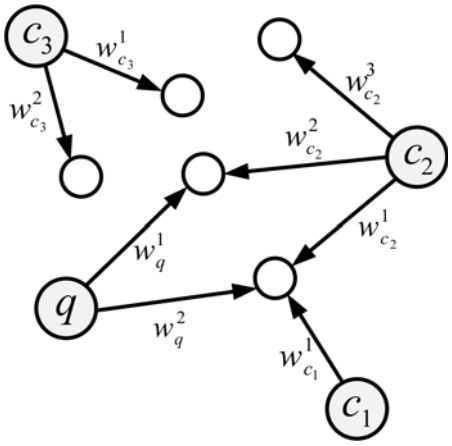

Figure 2: A sample view of our graph structure

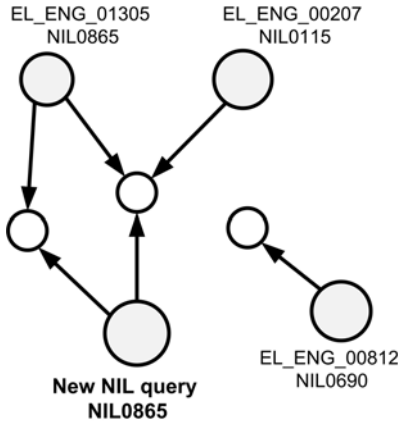

Figure 3: Sample NIL clustering graph

properties. The goal of our NIL clustering method is to select the cluster (i.e., the medoid) to which the query belongs and to assign the corresponding NIL id.

In Figure 3, we show a sample NIL clustering graph, (G_{NIL}). This graph contains three clusters. The medoid of each cluster is labeled by both the query id and the NIL id. As depicted in this figure, these medoids are "EL_ENG_01305," "EL_ENG_00207" and "EL_ENG_00812." These nodes are linked to their corresponding properties. Additionally, a NIL query is temporarily joined to G_{NIL} to infer the cluster to which NIL query belongs. The NIL query has two distinct paths to "EL_ENG_01305" and one path to "EL_ENG_00207."

In order to generate G_{NIL}, we use the same procedure as the one used for generating G when using the textual information. However, in this case, the weights of the resulting edges are set to 1 given that all the properties are extracted from the background documents.

Then, we proceed to find the most appropriate medoid for the NIL query node. This is performed using Equations 1 and 2 with medoids as candidates, C, and $\beta_\emptyset$ as threshold. If the NIL query node is linked to a medoid following Equation 2, then the id of the medoid is assigned to the NIL query and the NIL query node is deleted. Otherwise, a new id is assigned the NIL query and the NIL query node is joined to G_{NIL} as the medoid of a new NIL cluster.

In Figure 3, if the score of the node labelled "EL_ENG_01305" is greater than $\beta_\emptyset$, its NIL id ("NIL0865") is taken for the query result and NIL query is eliminated from G_{NIL}.

4 Slot Filling task

The UPC system for Slot Filling consists in three steps: 1) preprocessing the document collection in order to collect those documents relevant for each query, 2) applying Information Extraction (IE) patterns to the relevant documents to achieve possible fillers for the slots required for each query, and 3) integrating the resulting slot fillers into the KB knowledge base by normalising extracted fillers (i.e., selecting the most specific fillers under subsumtion for a particular slot, and normalising dates).

We have developed two different IE pattern learning approaches for our exploratory participation in KBP 2012: the first approach based on distant learning and the second one based on unsupervised learning. The rest of this section describes the preprocessing of the document collection as well as both learning approaches.

4.1 Document preprocessing

Prior to evaluation of KBP 2012, the document collection was indexed using Lucene[5] by all the words occurring in the documents.

At evaluation time, a preprocess has been performed in two steps for each query. The first step

[5] http://lucene.apache.org/

consists in retrieving the set D of documents containing at least one alternate name of the query expanded as described in Section 2 for the SF task. However, given the ambiguity of proper names, some of the retrieved documents could be related to a real entity different to the required one (e.g., retrieving documents related to Paul Watson -the environmental activist- can result with some documents related to other Paul Watson -the writer, the film maker, and so on-). This is why the second step of the preprocess consists in selecting the set $\hat{D} \subset D$ of documents really relevant for the query.

In order to obtain $\hat{D}$ from D, a particular relevance-feedback approach is performed. This approach is based on the assumption that lemmas frequently found in the close context to an occurrence of a NE can be useful to disambiguate it. The procedure starts preprocessing all the documents in D to get lemmas and POS tags of all words, as well as to detect NE occurrences. The initialization step consists of:

$L = \emptyset$
$\hat{D} = \{d_q\}$, the query reference document

Then, the following steps are iteratively performed:

1. Grow the set L of contextual lemmas[6] for all the alternate names of the query occurring as NEs in documents belonging to $\hat{D}$.
2. Select the subset $K \subseteq L$ of the most relevant contextual lemmas as described below.
3. Grow $\hat{D}$ with those documents from D in which at least one lemma belonging to K occurs within the context of an alternate name.
4. Repeat from step 1 until $\hat{D}$ does not change.

Set K is obtained in two steps. First, L is sorted by score s_i^* as follows:

$$s_i^* = \frac{s_i - \min_j s_j}{\max_j s_j - \min_j s_j}$$

$$s_i = \log \frac{F(l_i)}{f(l_i)} \cdot \frac{f(l_i)}{\sum f(l_j)}$$

where $1 \leq i, j \leq |L|$, $F(l_i)$ is the frequency of lemma l_i in D and $f(l_i)$ is the frequency of lemma l_i when it occurs as contextual lemma in $\hat{D}$. Then, the minimun set of lemmas $\{l_i\} \subset L$ with greater score is automatically selected as K. Intuitivelly, this can be approached by selecting as threshold l_{th} that l_i supporting the maximum convexity of the curve defined by sorting set L by score s_i^*. This can be computed using the following equation:

$$l_{th} = argmin_i \sqrt{s_i^{*2} - (i/\max i)^2}$$

where $1 \leq i \leq |L|$.

4.2 Distant-Learning Approach

Our first run in the SF task of KBP 2012 follows the distant learning (DL) paradigm for Relation Extraction (RE). DL was initially proposed as a RE approach by (Mintz et al., 2009) and applied to the SF task in preceeding KBP contests by several groups such as (Agirre et al., 2009; Surdeanu et al., 2010; Garrido et al., 2011). DL uses supervised learning but the supervision is not provided by manual annotation but from the occurrence of positive training instances in a KS or reference corpus. In the first proposal, (Mintz et al., 2009) used Freebase, an online database of structured semantic data, as KS. In subsequent applications, Wikipedia (WP) infoboxes have been preferred due to its better precision, at a cost of a drop in recall. In our case we have chosen WP too. Our distant learning approach to the task consisted of the following steps:

1. From a local copy of the English WP,[7] we automatically locate the set of pages corresponding to PER and the corresponding to ORG. For doing so we used the links between WP pages and WP categories as well as the graph structure of WP categories. Let *PagesPER* and *PagesORG* be these sets.
2. We used the mapping between the generic slots and the specific slots occurring in WP infoboxes provided by the organization. Table 2 shows, as an example, the set of specific slots corresponding to the generic slot

[6] We use a centered window of 5 noun, verb or adjective lemmas to the left/right of each alternate name occurrence.

[7] `http://en.wikipedia.org/wiki/English_Wikipedia`. We use for this purpose the JWPK software by Iryna Gurevich: `http://www.ukp.tu-darmstadt.de/software/jwpl`

full_name	nickname	full name
othername	full name	name
burthname	pseudonym	nicknames
othername(s)	name	alias
native_name	playername	fullname
birth_name	birth name	stage/screen
aliases	subject_name	name
other_names	alias	other names
birthname	birth_name	realname
othernames	othername(s)	names
also known as	nickname	

Table 2: Specific slots for the generic slot *per:alternate_names*

Figure 4: Example of WP page

Occured Value	Extracted Value
[October 16] , [1952]	October 16, 1952
[March 7] [322 BC]	March 7 322 BC
[748]([Arabian Peninsula])	748
[1368] or [1377]	?
[406 AH] (1015 AD)	1015 AD
25 June , 1274	25 June , 1274
(still alive in 1974)	?
alive	?
'circa' 1126	1126 circa
[1663] (age 23)	1663

Table 3: Examples of values found for the generic slot *per:date_of_death*

per:alternate_names. As shown in Figure 4, WP pages can include both structured (infoboxes, itemized lists...) and unstructured material (text). We took advantage of page infoboxes and page textual content. For all the pages in either *PagesPER* or *PagesORG* we collected all the occurring infoboxes, slots and values resulting in a set of tuples: <*page name, generic slot, infobox name, specific slot, slot value*>. Let *PagesSlotsValuesPER* and *PagesSlotsValuesORG* be these sets. Extracting the values of an specific slot is in some cases easy (e.g. for single-valued slots with a precise type, as *per:date_of_birth*) but in many others it is difficult. In Table 3 some examples of values for the generic slot *per:date_of_death* are shown. Using the Alergia system, (Carrasco and Oncina, 1994), we have learned regular grammars of the slots' values for allowing their extraction. In fact, the number of learned grammars is smaller than the number of slots because some of the values are of the same type, for example the DATE grammar can be used for the slots *date_of_birth* and *date_of_death*.

3. For each of the tuples in *PagesSlotsValuesPER* and *PagesSlotsValuesORG* we extracted the patterns occurring in the text corresponding to the page. For doing so we obtained the possible alternate names of the page name using the same procedure described in section 2. A similar process is carried out for the slot values, for instance for the slot *per:date_of_birth* if the value is *27 April 1945*, also *27-04-1945*, *April 1945*, and *1945* are considered as valid variants (the same grammars used for extraction are used here for generation). As can be seen the process is far to be simple. Two sets of alternate names, *alternateNamesX* and *alternateNamesY* were obtained. We looked on the text for all the occurrences of *alternateNamesX*($X_0, \ldots X_n$) and *alternateNamesY*($Y_0, \ldots Y_m$). For each pair of occurrences (X_i, Y_j) we collected the sequence of words occurring between them and we grouped together all the patterns corresponding to each generic slot. We built in this way the multiset (a set with fre-

quency counts for all the members) *PatternsGenericSlot*. This process resulted in collecting 9,064 patterns for ORG (ranging from 70 for *org:city_of_headquarters*, up to 2,573 for *org:political_religious_affiliation*) and 6,982 patterns for PER (from 23 for *per:cause_of_death* to 588 for *per:title*) with very variable accuracy. In Table 7 some examples of the 57 patterns for the generic slot *per:date_of_birth* are shown.

Once the set of patterns for each generic slot was built (only the most frequent patterns are selected) the process of extraction can be performed as shown in the following steps.

1. For each query we expanded the name onto a set of alternate names containing name variants of the query name (the corresponding *alternateNamesX*).

2. We retrieve from Lucene the documents containing any of the variants in *alternateNamesX*. Some filtering processes were performed in the case of recovering a huge amount of documents (looking only for the more precise variants, e.g., for *John Smith* one of the variants is *Smith* which results on a extremely huge number of mostly irrelevant documents, constraining the search to the whole term *John Smith* could reduce this set to a manageable size, namely, a maximum of 1,000 documents per query.

3. For each query we tried to apply all the patterns corresponding to each generic slot to all the retrieved documents. So if $(X_0, \ldots X_n)$ are the variants of the query name and *PatternsgenericSlot* contains the patterns of a generic slot we look for the occurrences of an X_i followed by a pattern. The text following this pattern is thus a candidate to be the value of such slot. For locating the right limit of this text we used the same grammars used for extraction in step 2.

4.3 Unsupervised learning approach

Our second approach for learning IE patterns is completely unsupervised from the point of view of using annotated slot-filler examples. Our goal is to explore the approapriateness of using clustering techniques to discover patterns useful to detect relevant relations between pairs of named entity types occurring in text, and then, classifying the relevant relations into the set of possible slots in an unsupervised maner. Following, we describe both the relation detection pattern learning approach and the relation classification approach.

4.3.1 The relation detection approach

For each slot in a template of the KBP scenario of extraction, we can define the pair (t_1, t_2) as the pair of entity types associated to the template itself (t_1 can be ORG or PER) and to the slot (t_2 can be AGE, PER, ORG, CITY, COUNTRY, RELIGION, CHARGE, and so on). For each (t_1, t_2), the procedure starts by gathering the set X of entity pairs, $x_i = (e_1, e_2)$, being t_1 and t_2 the entity types of e_1 and e_2, respectively, and co-occurring in sentences of the document collection. Most of the pairs x_i will not be linked by any particular relation. In fact, a minority of them will be effectively related. In this context, minority clustering can be used to detect groups of related entity pairs as foreground clusters and discard non-related ones as background noise.

Based on these assumptions, our goal in KBP 2012 is to perform initial experiments using the Ensemble Weak minOrity Cluster Scoring (EWOCS) algorithm (Gonzàlez and Turmo, 2012). Concretely, we have used the default configuration to deal with the relation detection task (Gonzàlez and Turmo, 2009; Gonzàlez, 2012), RD-EWOCS up to now, which is briefly described below.

Figure 5 depicts the RD-EWOCS general algorithm. It requires to represent each example as a binary feature vector. The default features used to represent each entity pair $x_i \in X$ are described in Table 4. The algorithm consists in two main steps: the *scoring* of the set of entity pairs related to a particular (t_1, t_2) and the *filtering* of the relevant pairs.

Scoring. Briefly, using an individual weak clustering algorithm f, we randomly produce R clustering models, $\pi = \pi_1, \ldots, \pi_R$ where $R = 100$ by default, from X. The default f for RD-EWOCS is a *Random Bregman Clustering* algorithm, a partition clustering algorithm which consists of the following steps:

- For each clustering model $\pi_c = \{\pi_1^c, \ldots, \pi_k^c\}$ randomly select both the number of clusters

	Feature	Description
structural	rightly/lefty	the first NE type t_1 occurs to the right/left of t_2
	dist_X	distance in tokens between the pair is X
	ch_dist_X	distance in chunks between the pair is X
word based	left_X_Y/right_X_Y	token X positions before/after to the left/rightmost NE of the pair has POS Y
	lmid_X_Y/rmid_X_Y	token X positions after/before to the left/rightmost NE of the pair has POS Y
	l_left_X_Y/l_right_X_Y	token X positions before/after to the left/rightmost NE of the pair has lemma Y
	l_lmid_X_Y/l_rmid_X_Y	token X positions after/before to the left/rightmost NE of the pair has lemma Y
	n_left_X/n_right_X	token X positions before/after to the left/rightmost NE is a negative word
	n_lmid_X/n_rmid_X	token X positions after/before to the left/rightmost NE is a negative word
chunk based	ch_left_X_Y/ch_right_X_Y	chunk X positions before/after to that containing the left/rightmost NE of the pair has type Y
	ch_lmid_X_Y/ch_rmid_X_Y	chunk X positions after/before to that containing the left/rightmost NE of the pair has type Y
	chl_left_X_Y/chl_right_X_Y	chunk X positions before/after to that containing the left/rightmost NE of the pair has a head with lemma Y
	chl_lmid_X_Y/chl_rmid_X_Y	chunk X positions after/before to that containing the left/rightmost NE of the pair has a head with lemma Y
	cht_left_X_Y/cht_right_X_Y	chunk X positions before/after to that containing the left/rightmost NE of the pair has a head with POS Y
	cht_lmid_X_Y/cht_rmid_X_Y	chunk X positions after/before to that containing the left/rightmost NE of the pair has a head with POS Y

Table 4: Default feature set for RD-EWOCS

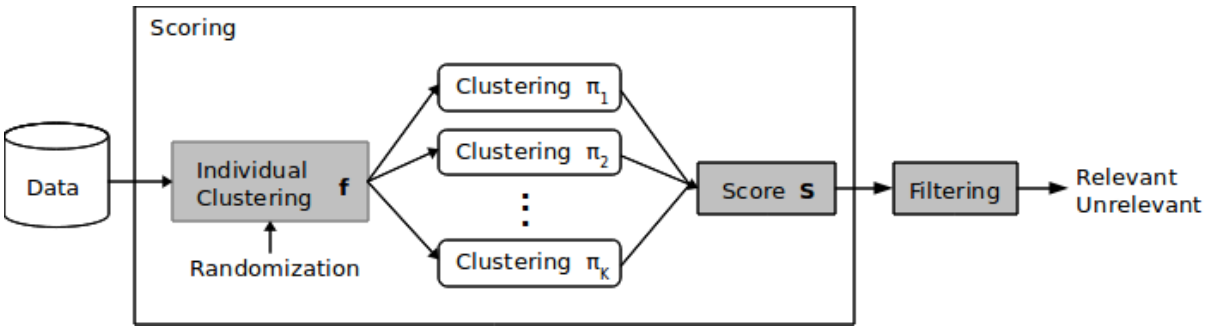

Figure 5: RD-EWOCS general algorithm

$k \in [2, k_{max}]$, where $k_{max} = 50$ by default, and the k seeds, $\{x_1^c, \ldots, x_k^c\}$.

- For each entity pair $x_i \in X$ and cluster $\pi_j^c \in \pi_c$ compute membership grades using a Gaussian-kernel distance as Bregman divergence as follows:

$$grade(x_i, \pi_j^c) = \frac{e^{-D(x_j^c, x_i)}}{\sum_{q=1}^{k} e^{-D(x_q^c, x_i)}}$$

$$D(x, y) = 2\alpha(1 - e^{-\gamma\|x-y\|^2})$$

where, parameters α and γ are automatically tuned in an unsupervised maner with the SOFTBBC-EM algorithm (Gupta and Ghosh, 2006).

- For each cluster $\pi_j^c \in \pi_c$ compute normalized sizes, $size^*$, as the product of the number of non-empty clusters[8], K_c, with the sum of membership grades of all pairs $x_i \in X$:

$$size^*(\pi_j^c) = K_c \cdot size(\pi_j^c)$$

$$size(\pi_j^c) = \sum_{x_i \in X} grade(x_i, \pi_j^c)$$

$$K_c = |\{\pi_j^c | size(\pi_j^c) \geq 1\}|$$

[8] A cluster is non-empty if its size is greater or equal than a threshold. By default, this threshold is 1.

Once π has been computed, each pair x_i is scored as the average of scores s_i^c achieved with each clustering model $\pi_c \in \pi$:

$$s_i^* = \frac{\sum_{\pi_c \in \pi} s_i^c}{R}$$
$$s_i^c = \sum_{\pi_j^c \in \pi_c} grade(x_i, \pi_j^c) \cdot size^*(\pi_j^c)$$

Filtering. Using the same idea as for filtering the most relevant documents in the preprocess (see Section 4.1), the set $\hat{X}$ of those pairs having greater or equal score than the one supporting the maximun convexity of the curve, x_{th} with score s_{th}, is considered as the set of relevant entity pairs:

$$\hat{X} = \{x_i \in X | s_i^* \geq s_{th}\}$$
$$s_{th} = min_i \sqrt{s_i^{*2} - (i/\max i)^2}$$

4.3.2 The relation classification approach

The unsupervised pattern-detection we have described so far, produces a set of entity pairs (e_1, e_2) that are related. But the exact nature and meaning of this relation remains unknown. Thus, we implement an unsupervised classification method that assigns each entity pair to the most likely template slot defined in the KBP evaluation.

Our method comprises two steps: first, the relations are separated according to the entity types (t_1, t_2). For each type pair we do an agglomerative clustering that groups the similar relations into some clusters. Second, we use an unsupervised similarity measure to map each cluster to one of the template slots available for this specific pair of types.

Clustering. To cluster the relation examples, we group them according to the entity types such as (*person*,*date*), (*person*,*location*), (*organization*,*date*), and then perform a clustering of each group. Each example is represented as a binary feature vector, as in the relation detection step. Here we use a subset of the features from Table 4, namely lemmas of tokens and lemmas of chunks in a window of size 5. The clustering algorithm we use is a simple agglomerative clustering with euclidean distance. The hierarchical clustering produces a dendrogram and we use the Calinski criterion (Calinski and Harabasz, 1974) to find an optimum cutting level. A separate clustering is performed for each pair of entity types.

The idea behind this process is to obtain groups of similar relations, ideally, one different relation per cluster.

Mapping. Assuming that each cluster obtained in the previous step corresponds to one different relation, in this step we try to map each cluster to one of the template's suitable slots for that pair (e.g. the pair (*person*,*organization*) can correspond to the slots: *employee_of* and *member_of*). Human experts have selected what t_2 types are the most suitable for each slot.

The mapping is set through an unsupervised process as follows: we take the *description* field (d_s) from the official slots definition document (Ellis, 2012) corresponding to the pair (t_1, t_2). This description is compared to the set of all relation examples in a cluster (each one is a sentence) concatenated in a single text S_s. We compare them using the textual semantic similarity measure of (Corley and Mihalcea, 2005).

This scoring scheme considers the similarity between pairs of words from two text segments, attempting to find for each word the most similar word in the other segment. The similarity between a pair of words is scored using the metric introduced by (Lin, 1998), which takes into account the information content (IC) of each word and their least common subsumer (LCS) in the WordNet taxonomy:

$$sim(v, w) = \frac{2 \cdot IC(LCS(v, w))}{IC(v) + IC(w)}$$

Finally, the word-to-word similarities are combined together into a text-to-text similarity using this function:

$$sim(d_s, S_s) = \frac{\sum_{pos} \sum_{w \in \{d_{s_{pos}}\}} maxSim(w) \cdot idf_w}{\sum_{w \in \{T_{i_{pos}}\}} idf_w}$$

which takes into account part-of-speech tags. This function is directional, we combine both directions by averaging them into a single symmetric similarity measure.

		All	PER	ORG	GPE
All Docs	Overall	0.421	0.599	0.382	0.194
	In-KB	0.311	0.603	0.138	0.192
	NIL	0.545	0.595	0.538	0.203
NW Docs	Overall	0.460	0.620	0.426	0.201
	In-KB	0.344	0.630	0.150	0.197
	NIL	0.582	0.611	0.587	0.232
Web Docs	Overall	0.344	0.533	0.322	0.181
	In-KB	0.253	0.535	0.126	0.183
	NIL	0.461	0.531	0.463	0.169

Table 5: TALP_UPC ML-EL results in TACKBP 2012 (B-cubed+ F-score)

5 Results and Analysis

5.1 Entity Linking

We sent one run for the TACKBP 2012 EL evaluation with following specifications: using wikitext, no access to the Web and without using offset. Table 5 shows our results. It shows B-cubed+ F-scores for both In-KB and NIL queries. Our overall result for all entities and both Newswire (NW) and Web documents is 0.421. We have better score for the PER entity type (0.599) in comparison to ORG (0.382) and GPE (0.194) types. For ORG, one reason is because of difficulty to expand correct forms from acronyms, for instance "ABC" can refer to "American Broadcasting Company" or "Australian Broadcasting Corporation." In the case of GPE, the problem occurs because there are many geopolitical entities with the same name, for instance "Hamilton" may refer to a region in the "New South Wales," "Queensland," "South Australia," "Tasmania" or "Victoria." The results are also better for NW documents in comparison to Web documents. We think that the reason can be the grammar irregularities found in the Web documents.

Analyzing the results shows that we should improve our system in several directions:

- In our run, the pair of offsets for start and end location of the query name in the background document was not used. Then, in the case that for a particular query (e.g., *Hamilton*) two or more different NEs in the background document (e.g., *David Hamilton* and *Daniel Hamilton*) are found, then the offsets are needed to solve the ambiguity.

- When classifying a query, our NERC could not properly identify the query types PER, ORG, or GPE for some queries. This problem caused the generation of irrelevant potential candidates.

- We need to develop an appropriate method to estimate the NIL threshold ($\beta_\emptyset$). In our participation the selection was done *adhoc*.

- We did not take into account the edges labels during the computation of the scores for candidate ranking. For this reason our ranking procedure is not able to discriminate between very similar relations or properties (e.g., *date_of_birth* and *date_of_death*). The lack of this analysis caused a big drop in the EL scores.

- We did not use any external resource such as: 1) The lists of name variation based on hyperlinks and redirects, 2) a particular KB derived from Wikipedia or external corpora to check the correctness of facts or aliases, or 3) a training data set derived from Wikipedia.

From our point of view, the reasons described above do not invalidate the graph-based approach, as most of the recent research devoted to EL explores similar approaches. In this sense, we think that there is room enough to improve our results.

Run	P	R	F1
Run1	0.224	0.043	0.072
PER	0.241	0.058	0.093
ORG	0.152	0.017	0.031
Run2	0.013	0.005	0.007
PER	0.015	0.002	0.003
ORG	0.012	0.011	0.012

Table 6: TALP_UPC SF results in TAC KBP 2012

5.2 Slot Filling

Regarding SF, we submited the distant-learning based approach and the unsupervised based one as Run1 and Run2, respectively. Table 6 shows the results achieved.

For Run1, the statistics of the official results were of 0.04 Recall, 0.22 Precision and 0.072 F1. These results are not bad in terms of Precision (0.11 median) but are very low in terms of Recall (0.08 median). As we do not use any confidence scoring for our answers, NIL is assigned to slots to which no valid assignement has been found. So, for analysing our errors we focus on not NIL answers. For Run2, the results for both types of queries, PER and ORG, are very poor. We achieved 0.005 and 0.016 for Recall and Precision, respectively.

First we present an analysis of the query and document pre-processes, common to both runs.

Regarding both query and document preprocess, a white box evaluation has been carried out taking as a reference V1 of the Assessment Results provided by the organisation. Therefore, we have computed the total recall according to the documents that have been successfully used to extract any correct slot value for any of the slots of the 80 queries (filler judgement column equal to 1).

The recall of the IR phase has been 0.96. A 20% of the documents not found in this step were due to the fact that we failed to include the query names followed by a saxon genitive in the list of alternate names, while 77% of them were due to problems in the generation of the alternate names (missing diminutives, such as "Cathie Black," too general names such as "Arsenal," etc.). Recall for PER queries was 0.98, whilst for ORG queries it was 0.95. Even though this difference is not very significant, we have seen that the generation of alternate names for ORG queries performed worse than for PER queries, due to the less robust methods applied to the task, specially for the case of acronym expansion/compression, as discussed in Section 5.1. The average number of alternate names per query was of 10.5 for PER and 2.9 for ORG.

As to the result of the subsequent process of selection of relevant documents, the recall was 0.83 (0.92 for PER queries, 0.78 for ORG ones). This is partly due to the fact that there were some queries for which, wrongly, just the reference document was found as relevant. The reason for such behaviour was our assumption that alternate names of queries occur as NEs in preprocessed documents. However, this fact strongly depends on the accuracy of the NERC system used. In particular, no alternate name has been recognised as NE for the reference document of some queries with the NERC system we used. As a consequence, the set of keywords useful to retrieve more relevant documents is empty for these queries. This makes our relevant feedback approach stop without providing more documents than the reference ones. Specifically, we discovered that for 13 queries no document other than the reference one was retrieved and for 8 other queries less than 4 document were retrieved. On the other hand, this filtering process turns out to be important as to the reduction in the number of documents: the average number of 1,866 documents found by the IR process is reduced to 611 documents, with an average reduction of 49.94% (56% for PER queries, 43% for ORG ones).

Now, focussing on the analysis of features specific this run, we proceed grouping the results in two axes: queries and slots.

From the queries axe we observe that the distribution of correct answers is extremelly query biased. In fact most of the queries have no answers at all (only 13 from the 40 PER queries and 4 from 40 ORG queries generated some results). This explain our low Recall figures. A second observation is the extremelly unbalanced performance of our system for PER and ORG: 66 correct answers were extracted in top position for PER (0.24 Precision) but only 12 for ORG (0.15 Precision).

Moving to the slot axe we discover that 12 out of the 16 ORG slots produce no results (only 7 for PER). We have manually analyzed a sample of 25 patterns from the pattern sets of all the slots. The results were significant: for PER, all but one (*per:age*) of the slots got an accuracy over 0.9, while for ORG only one slot (*org:alternate_name*) got an accuracy over 0.5.

The reasons why this happens are multiple:

- *PagesORG* are less accurate than *PagesPER* possibly due to the difficulty of obtaining the set of relevant categories for ORG. Getting the relevant WP categories for PER is straigthforward. This is not the case of ORG where categories are spread within the whole set of WP categories.

- Less infoboxes are filled for ORG.

Pattern
was born in
born
on <DATE> in
in
<DATE> in
born in
was born on
was born on <DATE> in
<DATE>
was born
was born and raised in

Table 7: Some of the best scored patterns for the generic slot *per:date_of_birth*

- ORG generic slots mappings are less reliable than PER ones, For many slots the grammars used are really precise (as DATE or PLACE) in the case of PER, but present a great variability in the case of ORG. Locating a PERSON, a DATE or a LOCATION within a value string is easier than locating an ORGANIZATION.

- The patterns extracted for PER are in many cases very short (as shown in Table 7) and frequent. This is not the case for ORG where many patterns are long and occur with very low frequency.

- Most of generic slots for PERSON are single-valued, in the case of ORG the situation is the contrary.

- While persons use to show a similar profile, organizations, present a great variability, for instance a political PARTY or a football TEAM have few points in common.

- Sometimes the mappings between generic and specific slots provided by KBP organizers were not accurate enough. For instance, for *per:age*, the slots contain a large number of varied wordings cointaining the age together with many other useless and noisy information. The grammar learned from this material is obviously extremelly unaccurate.

Focusing on Run1, the main reasons why we obtain poor results, besides those presented above for preprocessing steps, are the following:

- According to (Gonzàlez, 2012), EWOCS performance improves if the size of the ensemble of clusterings is selected taking into account the size of the data set, so that large data sets require large ensembles. In this sense, we think that our unsupervised approach requires much more than 100 clusterning models to achieve good results for detecting slot fillers in KBP corpus. This does not unduly penalize the efficiency of the system given that the computation of the clustering models can be paralelized.

- The process of clustering and mapping we have presented in Section 4.3.2 finds the KBP slot that best matches the semantic content of the examples present in each cluster. But due to the high degree of unsupervision, these procedure does neither guarantee that all clusters will be mapped to a different slot, nor that all slots will have an assigned cluster. Additionally, with this method it is not possible to decide that a cluster is not capturing any of the relations expressed by the slots and therefore these examples should be filtered out. This may be a serious drawback in some cases. For example, nothing prevents the system from learning a classifier that splits the relations involving the (*organization,person*) into three clusters and assigns all of them to the *org:shareholders* slot and none to the *org:founded_by* slot.

Taking into account all these points, we think that there is room enough for improvements in both approaches presented to deal with SF task.

Acknowledgments

This work has been produced with the support of the project KNOW2 (TIN2009-14715-C04-04).

References

E. Agirre, A. X. Chang, D. S. Jurafsky, C. D. Manning, V. I. Spitkovsky, and E. Yeh. 2009. Stanford-UBC at TAC-KBP. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*.

T. Calinski and J. Harabasz. 1974. A dendrite method for cluster analysis. In *Communications in Statistics-theory and Methods*.

R. C. Carrasco and J. Oncina. 1994. Learning stochastic regular grammars by means of a state merging method. In *Grammatical Inference and Applications*. Springer-Verlag.

C. Corley and R. Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05.

J. Ellis. 2012. *TAC KBP Slots, Version 2.4*. Linguistic Data Consortium.
http://www.nist.gov/tac/2012/KBP/task_guidelines/.

G. Garrido, B. Cabaleiro, A. Pe
nas, A. Rodrigo, and D. Spina. 2011. Distant supervised learning for the TAC-KBP Slot Filling and Temporal Slot Filling Tasks. In *Text Analysis Conference (TAC 2011)*.

E. Gonzàlez and J. Turmo. 2009. Unsupervised relation extraction by massive clustering. In *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM)*.

E. Gonzàlez and J. Turmo. 2012. Unsupervised ensemble minority clustering. In *Research report. Department of Llenguatges i Sistemes Informátics (LSI - UPC)*.

E. Gonzàlez. 2012. *Unsupervised Learning of Relation Detection Patterns*. Ph.D. thesis, UPC Programme in Artificial Intelligence.

Y. Guo, W. Che, T. Liu, and S. Li. 2011. A graph-based method for entity linking. In *5th International Joint Conference on Natural Language Processing*.

G. Gupta and J. Ghosh. 2006. Bregman bubble clustering: A robust , scalable framework for locating multiple, dense regions in data. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*.

B. Hachey, W. Radford, and J. R. Curran. 2011. Graph-based named entity linking with wikipedia. In *Lecture Notes in Computer Science, Volume 6997*.

X. Han, L. Sun, and J. Zhao. 2011. Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th International ACM SIGIRC-conference on Research and Development in Information Retrieval*.

J. Lehmann, S. Monahan, L. Nezda, A. Jung, and Y. Shi. 2010. LCC approaches to knowledge base population at tac 2010. In *Text Analysis Conference*.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of 15th International Conference on Machine Learning (ICML)*.

P. McNamee, H. T. Dang, H. Simpson, P. Schone, and S. M. Strassel. 2010. An evaluation of technologies for knowledge base population. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC)*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the ACL-AFNLP Joint Conference*. ACL, August.

M. Surdeanu, D. McClosky, J. Tibshirani, J. Bauer, A. X. Chang, V. I. Spitkovsky, and C. D. Manning. 2010. A simple distant supervision approach for the TAC-KBP slot filling task. In *Proceedings of the TAC-KBP 2010 Workshop*.

Marta Vila (Universitat de Barcelona)
Santiago González (Universitat de Barcelona)
M. Antònia Martí (Universitat de Barcelona)
Joaquim Llisterri (Universitat Autònoma de Barcelona)
María Jesús Machuca (Universitat Autònoma de Barcelona)

(2010)

ClInt: A bilingual Spanish-Catalan spoken corpus of clinical interviews

Procesamiento del Lenguaje Natural 45:105–111.

Journal URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln`

Abstract In this paper we present ClInt (Clinical Interview), a bilingual Spanish-Catalan spoken corpus that contains 15 hours of clinical interviews. It consists of audio files aligned with multiple-level transcriptions comprising orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding. This is a previously non-existent resource for these languages and it offers a wide-ranging exploitation potential in a broad variety of disciplines such as Linguistics, Natural Language Processing and related fields.

Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 105-111 recibido 26-04-10 revisado 18-05-10 aceptado 20-05-10

ClInt: a Bilingual Spanish-Catalan Spoken Corpus of Clinical Interviews

ClInt: un corpus oral bilingüe español-catalán de entrevistas clínicas

Marta Vila, Santiago González, M. Antònia Martí
CLiC – Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
08007 Barcelona
{marta.vila, santiago.gonzalez, amarti}@ub.edu

Joaquim Llisterri, María Jesús Machuca
Grup de Fonètica – Departament de Filologia Espanyola
Universitat Autònoma de Barcelona
Edifici B, 08193 Bellaterra, Barcelona
{joaquim.llisterri, mariaJesus.Machuca}@uab.cat

Resumen: En este artículo se presenta ClInt (*Clinical Interview*), un corpus oral bilingüe español-catalán que contiene un total de 15 horas de entrevistas clínicas. Está formado por archivos sonoros alineados con transcripciones a varios niveles que comprenden información ortográfica, fonética y morfológica, además de codificación lingüística y extralingüística. Se trata de un recurso hasta el momento inexistente para estas lenguas que ofrece múltiples posibilidades de explotación desde una amplia variedad de disciplinas, tanto las vinculadas a la Lingüística como las que se relacionan con el Procesamiento del Lenguaje Natural.
Palabras clave: Corpus oral, corpus bilingüe, entrevista clínica.

Abstract: In this paper we present ClInt (Clinical Interview), a bilingual Spanish-Catalan spoken corpus that contains 15 hours of clinical interviews. It consists of audio files aligned with multiple-level transcriptions comprising orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding. This is a previously non-existent resource for these languages and it offers a wide-ranging exploitation potential in a broad variety of disciplines such as Linguistics, Natural Language Processing and related fields.
Keywords: Spoken corpus, bilingual corpus, clinical interview.

1 Introduction

Corpus availability has become indispensable for performing studies in many scientific fields. Nowadays, these language resources are fundamental in disciplines such as Linguistics, Natural Language Processing (NLP) and related fields.

Spoken corpora are those most in demand, probably due to their shortage and the difficulty involved in obtaining them, not only in the transcription procedure, but also in the recording. In this sense, one of the most valuable types is the one that captures real —not artificially elicited— communicative situations. Spoken corpora in professional situations are especially difficult to obtain, because it is not easy to gain access to certain environments, such as trials, business meetings, or clinical interviews.

In this paper we present ClInt (Clinical Interview),[1] a bilingual Spanish-Catalan spoken corpus of clinical interviews, a hitherto non-existent resource for these languages. It consists of audio files aligned with multiple-level transcriptions containing orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding.

The remainder of this paper is structured as follows: in Section 2, we present the related work done in this area. In Section 3, we provide an overview of the corpus. Section 4 is devoted

[1] The corpus and source URLs mentioned in this paper appear in the appendix.

ISSN 1135-5948

to corpus development. In Section 5, future research possibilities are suggested. Finally, Section 6 sets out some final remarks about this project.

2 Related Work

To the best of our knowledge, ClInt is the first bilingual Spanish-Catalan spoken corpus of clinical interviews. Moreover, there are very few corpora of this type in other languages. The DiK-corpus is particularly relevant in this sense. It consists of the transcriptions of 25 hours of audio recordings of monolingual and interpreted doctor-patient communication in German, Turkish, Portuguese and Spanish.

Despite the shortage of clinical interview corpora, in more general terms, there do exist spoken conversational corpora, both in Spanish and Catalan. In Spanish, some examples are CORLEC (*Corpus Oral de Referencia de la Lengua Española Contemporánea*[2]) (Marcos, 1991), the *Corpus de conversaciones coloquiales*[3] (Briz, 2001) and the spoken section in CREA (*Corpus de Referencia del Español Actual*[4]) (RAE). Our major reference in Catalan is COC (*Corpus Oral de Conversa Coloquial*[5]) (Payrató and Alturo, 2002), contained in the CCCUB (*Corpus del Català Contemporani de la Universitat de Barcelona*[6]).

Moreover, there exist corpora including speech by sick and disabled people, and by people with language disorders (Peraita and Grasso, 2009; Navarro and San Martín, 2009). Also, recorded clinical interview simulations for doctor training can be found (Borrell, 2000).

Finally, there has been some work in the literature with regard to clinical therapist skills training in virtual environments. In this context, the patient is a virtual human and the doctor has to interact with this virtual human in order to improve his skills in the process (Kenny et al., 2007; 2008).

3 Corpus Overview

The corpus is comprised of a total of 15 hours of recordings divided into 40 clinical interviews of an average of 22 min each. These interviews correspond to four different residents (ten interviews for each resident).

The recordings were carried out in the pneumology clinic of a hospital in the Barcelona metropolitan area. Catalonia is a bilingual community where Catalan and Spanish coexist. As the recordings were made giving absolute freedom to participants with respect to their language usage, this bilingualism is reflected in the corpus. Furthermore, the corpus displays Spanish and Catalan dialectal variants.

The ClInt corpus consists of the audio files aligned with their orthographic transcriptions (with linguistic and extralinguistic encoding), their phonetic transcriptions, as well as their morphosyntactic analysis. All this information is stored in a database.

4 Corpus Development

The ClInt corpus (Figure 1) was recorded using a stereo digital recorder (SANYO, ICR-RS176NX) and a uni-directional condenser microphone (FoneStar, BM-704BL). The characteristics of this equipment ensure that the corpus is available for further phonetic studies. These recordings were manually transcribed using conventional spelling and encoded in XML format using the Transcriber (Barras et al., 2001), a tool for assisting in the manual transcription and encoding of speech signals that provides a user-friendly interface. This tool allows for the alignment between the audio and the transcription.

The basic unit of the text in the corpus is the 'breath group',[7] understood as a discourse stretch of speech between pauses (a pause is defined as a period of silence between 200 and 500 ms). Breath groups can be full (with speech uttered), empty (pauses above 500 ms) or with overlapping (when two people speak at the same time). A breath group generally corresponds to a register in the database and it is the unit of alignment, i.e, the audio files and the different transcription levels are synchronized at the level of breath groups.

From the manual transcription, called the Base Transcription (BT), an Orthographic Transcription (OT) and an Enriched Orthographic Transcription (EOT) were automatically obtained. The raw OT was used

[2] Reference Corpus of Contemporary Spoken Spanish

[3] Corpus of Colloquial Conversations

[4] Reference Corpus of Current Spanish

[5] Spoken Corpus of Colloquial Conversation

[6] Corpus of Contemporary Catalan of the University of Barcelona

[7] Also called 'phonic group' in the Spanish tradition.

in turn for the generation of the Phonetic Transcription (PhT) and as input for the Morphological Analysis (MA).

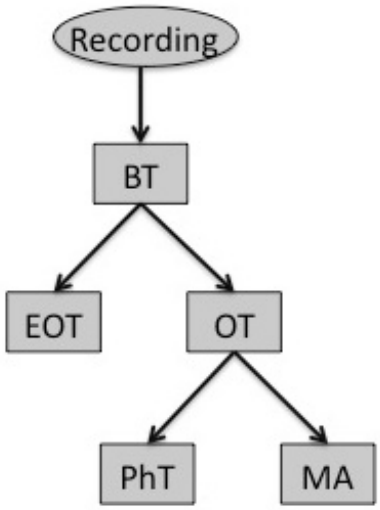

Figure 1: Corpus development scheme

4.1 Base Transcription

The BT (Figure 2) consists of a manual orthographic transcription and encoding of the audio files in XML format. For this purpose, we developed annotation guidelines and carried out a training process for all the annotators in order to avoid incoherencies in the transcription.

The orthographic transcription guidelines follow EAGLES - Expert Advisory Group on Language Engineering Standard (1996) recommendations. EAGLES general philosophy is always to use prescriptive forms and to document all the cases where this is not possible. Following these recommendations, the annotators used, whenever possible, the orthographic forms that appear in the Spanish and Catalan prescriptive dictionaries. However, with the aim of being faithful to the speakers' pronunciation, some non-prescriptive words (i.e. some onomatopoeias, interferences, unknown and mispronounced words, and abbreviated forms) were maintained and tagged. All of them are collected in a document accompanying the transcription. Numbers, acronyms and spelled words are represented as the speakers pronounce them, i.e., using the orthographically complete form. Prosodic tags are used instead of punctuation marks to ensure the correct interpretation of the text and, at the same time, to accurately reflect the spoken nature of the corpus.

The encoding is intended to be as general and scalable as possible in order to ensure the widest possible exploitation potential for ClInt. Below we list the tags corresponding to the information and phenomena that are encoded in the BT. For the sake of simplicity, we classify them into groups according to the type of information encoded.

Recording and transcription files (information about every recording and transcription file in the corpus): recording identification and date, person responsible for transcription, and transcription date.

Speakers (information about the speakers participating in the interaction): speakers' identification and sex, languages in which they are competent, and the language they (generally) use in the interview.

All the languages in which each speaker is (not) competent have a code (from 0 to 3) indicating the level of competence:

-The speaker does not understand the language.

-The speaker is able to understand the language, but is not able to speak it.

-The speaker is able to speak the language, but with certain limitations.

- The speaker is perfectly able to speak the language.

All the information related to languages is extracted from the recordings themselves. Information that is not specified or deductible from the recordings does not appear, since it is considered to be subjective.

Discourse interaction-related phenomena (information about turn-taking): turn-taking, overlaps, pauses above 500 ms.

Lexical and semi-lexical phenomena:

-Named entities: people, medicines and active principles.

-Acronyms: word formed from the initial letters of other words (e.g. *TAC* for *Tomografía Axial Computarizada*, 'computed tomography' in Spanish) [8].

-Spelled words: words uttered naming the letters that form them (e.g., *a-a-ese* for *AAS,* in this example, the patient is trying to spell the name of a medicine).

-Syllabification: words uttered separating the syllables that form them (e.g., *se-tan-ta-dos* for *setanta-dos*, 'seventy-two' in Catalan).

[8] For the sake of simplicity, we do not exemplify these phenomena using the XML tags.

-Onomatopoeias: words that reproduce the sound associated with what is named (e.g., *bumbum,* in this example, the patient is trying to reproduce the sound of fast walking).

-Interjections: words used for expressing the speaker's attitude (e.g., *ai*, in this example, the speaker is expressing pain) or for maintaining the communication between speakers (e.g., *ahà,* in this example, the speaker is communicating that he is following the conversation), among other uses.

-Abbreviated forms: words that have lost a sound or sounds at the end (e.g., *químio* for *quimioterapia*, 'chemotherapy' in Spanish).

-Mispronounced words: words that are uttered in the wrong way (e.g., *otroscopia* for *artroscopia*, 'arthroscopy' in Spanish).

-Truncated words: words that have been truncated in the interview for different reasons such as an interruption by another speaker (e.g., *magat* for *magatzem*, 'warehouse' in Catalan).

-Emphasis: words uttered prominently.

-Long sounds: lengthened sounds in a word.

-Non-understandable snippets: incomprehensible fragments.

-Unknown words: words that can be partially understood. The tag indicates that the interpretation is a guess.

-Voiced pauses: pauses in the speech in which the speaker produces a semi-lexical sound (e.g., *eee*).

Non-lexical phenomena: human and non-human noises (e.g., laughing, slams, typing).

Code-related phenomena: mixing and code switching.

Prosodic phenomena: terminal and truncated tones, following Payrató and Fitó (2008).

<Turn speaker="spk4" startTime="702.244" endTime="705.062">
<Sync time="702.244"/>
y cuando haces
<Event desc="voiced_pause" type="lexical" extent="begin"/>mmm<Event desc="voiced_pause" type="lexical" extent="end"/>
ejercicio
<Event desc="noise" type="noise" extent="begin"/>
<Event desc="long" type="pronounce" extent="begin"/>**s**<Event desc="long" type="pronounce" extent="end"/>**ientes**
<Event desc="noise" type="noise" extent="end"/>
que te falta un poco el aire<Pro desc="asc"/>
<Turn/>
<Turn speaker="spk2" startTime="705.062" endTime="706.059">
<Sync time="705.062"/>
sí el aire<Pro desc="desc"/>
<Turn/>

Figure 2: Example of Base Transcription[9]

4.2 Enriched Orthographic Transcription

The EOT (Figure 3) was automatically obtained from the BT just by changing the XML tags for more readable marks, e.g., <Turn speaker="spk4"> in Figure 2 has been changed to "Doctor" in Figure 3; or <Event desc="noise" type="noise" extent="end"/> in Figure 2 has been changed to [-noise] in Figure 3. This makes the transcription more readable.

Doctor y cuando haces <mmm> ejercicio [noise-] s:ientes [-noise] que te falta un poco el aire/
Patient sí el aire\

Figure 3: Example of Enriched Orthographic Transcription

4.3 Orthographic Transcription

The OT (Figure 4) was automatically obtained from the BT by eliminating all XML tags. Moreover, truncated words were reconstructed when they could be inferred from the context. When they could not, they were eliminated. Voiced pauses were not included either.

The OT has a neutral intermediate format suitable for automatically deriving the PhT and for carrying out the MA.

Doctor y cuando haces ejercicio sientes que te falta un poco el aire
Patient sí el aire

Figure 4: Example of Orthographic Transcription

4.4 Phonetic Transcription

The PhT is derived from the OT using SAGA (Moreno and Mariño, 1998), an automatic

[9] And when you do exercice, you feel you are breathless / Yes, I do.

Spanish phonetic transcriber, for the fragments in Spanish (Figure 5), and SEGRE (Pachès et al., 2000), an automatic Catalan phonetic transcriber, for the fragments in Catalan (Figure 6). The phonetic alphabet used in both cases is SAMPA. Although both SAGA and SEGRE take into account contextual phonetic phenomena (both inter and intra word), SEGRE considers resyllabification phenomena corresponding to spontaneous speech, e.g., the transcription [s 'i | **e** | **l** 'a j | r e] in Figure 6 considers resyllabification (in bold), while the transcription corresponding to SAGA [s 'i / **e l** / 'a j - r e] in Figure 5 does not.

Doctor i / k w a n - d o / 'a - T e s / e - x e r - T 'i - T j o / s j 'e n - t e s / k e / t e / f 'a l - t a / 'u m / p 'o - k o / e l / 'a j - r e
Patient s 'i / e l / 'a j - r e

Figure 5: Example of phonetic transcription using SAGA

Doctor i | k w a n | d o 'a | T e | s e | x e r | T 'i | T j o | s j 'e n | t e s | k e | t e | f 'a l | t a 'u m | p 'o | k o e | l'a j | r e
Patient s 'i | e | l 'a j | r e

Figure 6: Example of phonetic transcription using SEGRE[10]

4.5 Morphosyntactic Analysis

The MA (Table 1) is derived from the OT using FreeLing toolbox (Atserias et al., 2006). The MA is not strictly speaking a transcription, but a morphosyntactic analysis of all words in the corpus. A lemma and a category are assigned to every word in the corpus. Because of the spoken and sometimes non-prescriptive nature of the corpus, some questions were not held by the analyzer correctly. Thus, a manual revision of the morphosyntactic analysis had to be carried out.

Lemma	Word	Code
y	y	cc
cuando	cuando	cs
hacer	haces	vmip2s0
ejercicio	ejercicio	ncms000
sentir	sientes	vmip2s0
que	que	cs
tú	te	pp2cs000
faltar	falta	vmip3s0
el	el	da0ms0
aire	aire	ncms000
sí	sí	rg
el	el	da0ms0
aire	aire	ncms000

Table 1: Example of Morphosyntactic Analysis

5 Research Exploitation and Future Work

This corpus opens up a wide variety of possibilities in research. We want to emphasize the relevance of this corpus to disciplines such as Linguistics and NLP. Three main lines of research are being carried out. Firstly, ClInt constitutes part of a wider project, Text-Knowledge 2.0, aimed at studying language use. For this project we are developing several Catalan and Spanish corpora representative of different communicative situations. Our hypothesis is that there are fundamental differences between how linguistic structure is postulated on the basis of imagined configurations, and how it is actually expressed in live conversational contexts. More specifically, we want to identify memory storage units, that is, the way in which language is broken down into chunks based on the frequency of items and strings of items (Bybee and Hopper, 2001; Bybee, 2010).

Secondly, this corpus is especially relevant for the study of paraphrasing occurring over different registers. On many occasions, during the clinical interview, the same information is uttered by the doctor and the patient. However, in general terms, whereas doctors talk objectively using a technical register conferred by their medical knowledge and experience, patients talk subjectively, expressing their personal experience of illness, due to their lack of medical knowledge.

Thirdly, from a phonetic point of view, this type of corpora corresponds to spontaneous speech providing physical evidence of how we actually speak. Disfluencies can be studied in order to analyze how speakers plan their speech and which planning problems there are when someone says something in real conversation (Clark and Wasow, 1998). Moreover, modeling variation in spontaneous speech is also important to improve speech recognition systems. According to Nakamura, Iwano and Furui (2007) recognition performance

[10] Although SEGRE only works for Catalan, we have used the same snippet in Spanish in order to facilitate the comparison.

drastically decreases for spontaneous speech, so a paradigm shift from speech recognition to understanding is required when underlying messages of the speaker are extracted.

Finally, clinical-interview corpora are indispensable in the medical communication field. Many experts point out that doctor-patient communication has been given little attention (Clèries, 2006). Nowadays, doctors are more encouraged to perform therapeutic procedures than to talk to the patient, although on many occasions the diagnosis may be obtained solely through communication. According to some experts, the problem is that young doctors are not sufficiently trained and cover up their lack of experience with technique. Hence, many point out the need for communicative skills training in Medical Schools. Clinical-interview corpora are indispensable for doing research in this area and as real material to work with in communicative-skills training courses.

6 Final Remarks

In this paper, we have presented ClInt, a corpus of 15 hours of clinical interviews. It consists of audio files aligned with multiple-level transcriptions containing orthographic, phonetic and morphological information, as well as linguistic and extralinguistic encoding. The encoding is intended to be as general and scalable as possible, as ClInt's exploitation potential is very wide-ranging. We have shown the linguistic richness of this resource, partly due to its bilingual nature. We have also described the interest of this corpus from the Linguistics and NLP perspectives, as well as from a medical point of view.

Acknowledgments

We are very grateful to Gustavo Tolchinsky, the doctor responsible for the recordings, as well as to the annotators: Alba Vindel and Esther Arias. The construction of the ClInt corpus would not have been possible without their collaboration. This work is supported by the FPU Grant AP2008-02185 from the Spanish Ministry of Education, and the Text-Knowledge 2.0 (TIN2009-13391-C04-04) and ClInt (FFI2009-06252-E/FILO) projects.

Bibliografía

Atserias, J., B. Casas, E. Comelles, M. González, Ll. Padró and M. Padró. 2006. FreeLing 1.3: Syntactic and Semantic Services in an Open-source NLP Library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*, pages 18-25, Genoa.

Barras, C., E. Geo, Z. Wu and M. Liberman. 2001. Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. *Speech Communication,* 33:5-22.

Borrell, F. 2000. *Entrevista clínica. Manual de estrategias prácticas*. SemFYC, Barcelona.

Briz, A. (Coord.) 2001. Corpus de conversaciones coloquiales. Appendix 1 in *Oralia*. ArcoLibros, Madrid.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Bybee, J. and P. Hopper. 2001. *Frequency and the Emergency of Linguistic Structure*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Clark, H. and T. Wasow. 1998. Repeating Words in Spontaneous Speech, *Cognitive Psychology*, 37:201-242.

Clèries, X. 2006. *La comunicación. Una competencia esencial para los profesionales de la salud*. Elsevier-Masson, Barcelona.

EAGLES. 1996. *Preliminary Recommendations on Spoken Texts*. EAGLES Document EAG-TCWG-STP/P, May 1996.

Kenny, P., Th. Parsons, J. Gratch, A. Leuski and A. Rizzo. 2007. Virtual Patients for Clinical Therapists Skills Training. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, *IVA,* pages 197-210, Paris.

Kenny, P., Th. Parsons, J. Gratch and A. Rizzo. 2008. Evaluation of Justina: A Virtual Patient with PTSD. In *Proceedings of 8th International Conference on Intelligent Virtual Agents, IVA,* pages 394-408, Tokio.

Marcos, F. 1991. Corpus lingüístico de referencia de la lengua española. *Boletín de la Academia Argentina de las Letras* 56: 129-155.

Moreno, A. and J. B. Mariño. 1998. Spanish Dialects: Phonetic Transcription. In *Proceedings of the 5th International*

Conference on Spoken Language Processing, ICSLP, pages 189-192, Sydney.

Nakamura, M., K. Iwano and S. Furui. 2008. Differences Between Acoustic Characteristics of Spontaneous and Read Speech and Their Effects on Speech Recognition Performance. *Computer Speech & Language*, 22(2):171-184.

Navarro, M. I. and C. San Martín. 2009. Estudio comparativo de las habilidades metalingüísticas de un niño con Trastorno Específico del Lenguaje basado en un corpus de niños con este trastorno y niños que siguen la pauta estándar de desarrollo procedentes de un corpus de niños normohablantes. In *Proceedings of the 1st International Conference on Corpus Linguistics, CILC,* pages 326-344, Murcia.

Pachès, P., C. Mota, M. Riera, M. P. Perea, A. Febrer, M. Estruch, J. M. Garrido, M. J. Machuca, A. Ríos, J. Llisterri, I. Esquerra, J. Hernando, J. Padrell, C. Nadeu. 2000. Segre: An Automatic Tool for Grapheme-to-allophone Transcription in Catalan. In *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and strategic priorities, 2nd International Conference on Language Resources and Evaluation, LREC,* pages 52-61, Athens.

Payrató, Ll. and N. Alturo (Eds.). 2002. *Corpus oral de conversa col·loquial. Materials de treball*, Publicacions de la Universitat de Barcelona, Barcelona.

Payrató, Ll. and J. Fitó. 2008. *Corpus audiovisual plurilingüe*. Publicacions de la Universitat de Barcelona, Barcelona.

Peraita, H. and L. Grasso. 2009. Corpus lingüístico de definiciones de categorías semánticas de sujetos ancianos sanos y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina. In *Proceedings of the 1st International Conference on Corpus Linguistics*, *CILC,* pages 78-88, Murcia.

A Appendix 1: Corpus and Source URLs

CCCUB corpus
<http://www.ub.edu/cccub/>

ClInt corpus
< http://clic.ub.edu/en/clint-en>

CORLEC corpus
<http://www.lllf.uam.es/ESP/Corlec.html>

Corpus de conversaciones coloquiales
<http://www.valesco.es/>

CREA corpus
<http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/D55F5BFB05D63980C1257164003F02E5?OpenDocument&i=2>

DiK corpus
<http://www1.uni-hamburg.de/exmaralda/files/k2-korpus/index.html>

EAGLES standard
<http://www.ilc.pi.cnr.it/EAGLES96/spokentx/spokentx.html>

FreeLing
<http://www.lsi.upc.es/~nlp/freeling/>

SAGA
<http://www.talp.cat/talp/index.php/ca/recursos/eines/saga>

SEGRE
<http://www.talp.cat/Joomla_1.5.7_nou/index.php/ca/recursos/eines/segre>

Transcriber
<http://trans.sourceforge.net/en/presentation.php>