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 analyze 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 that 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 analyzed 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 plagiarizing, and (iii) paraphrase mechanisms tend to shorten the plagiarized text. For the first time, the paraphrase mechanisms behind plagiarism have been analyzed, providing critical insights for the improvement of automatic plagiarism detection systems.

∗ 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.

doi:10.1162/COLI_a_00153

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). Although plagiarism may occur incidentally, it is often the outcome of a conscious process. Independently from the vocabulary or channel through which an idea is communicated, a person who 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 paraphrase 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 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, hereafter).[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.

1 http://www.webis.de/research/events/pan-10.

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 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 (Gülich 2003; Cheung 2009), 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 (Mel'čuk 1992; Milićević 2007). 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 (Chomsky 1957; Harris 1957; Levin 1993). They do not cover paraphrasing as a whole, however, 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 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,

CLASS	SUBCLASS	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 four classes, four subclasses, and 20 types.

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. 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 Figure 1. It consists of a three-level typology of 20 paraphrase types grouped in four classes and four subclasses. Paraphrase types stand for the linguistic mechanism triggering the paraphrase phenomenon. They are

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.

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 Example (1), where the nominal form *failure* is exchanged for the verb *failed*, has obvious syntactic implications; the paraphrase phenomenon, however, is triggered by the morpholexical change.[5] A structure-based change (diathesis) like the one in Example (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, as in Example (3). Miscellaneous changes comprise types not directly related to one single class. Finally, the subclasses 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"
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 Example (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 between cases such as Example (4) and, for example, Example (1) should be noted: In Example (1), the derivational change implies the syntactic one, so only one single paraphrase phenomenon is considered; in Example (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 Example (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*

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.

Modal verb changes are changes of modality using modal verbs, like *might* and *could* in Example (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 Example (7), the verbal form *differing* is changed to the adjective *different*, with the consequent structural reorganization.

(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 Example (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 Example (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 Example (10b), a (lexically emptied) specifier (*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 Example (11), *failed* is substituted for its antonym *succeed* and a negation is added. Second, there is the case

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).

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 Example (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

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 Example (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 (Example (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 Example (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

Ellipsis includes linguistic ellipsis (i.e, those cases in which the elided fragments can be recovered through linguistic mechanisms). In Example (16b), the subject *he* appears in both clauses; in Example (16a), 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

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.

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 Example (17a) are coordinated with the conjunction *and* in (17b).

(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 Example (18a) (*which limits the percentage of Jewish pupils in any school*) is part of the main clause in Example (18b).

(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 Example (19a), the list appears numbered and, in Example (19b), 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 Example (20a) and the indirect in Example (20b).

(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 Example (21a), interrogative sentences can be observed; they are changed to an affirmative sentence in Example (21b).

(21) 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 gather a wide variety of syntax/discourse reorganizations not covered by the types in the syntax and discourse subclasses above. An example can be seen in Example (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 Example (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*

Addition/deletion This type consists of all additions/deletions of lexical and functional units. In Example (25b), *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 bad 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, 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 Corpus 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

8 This type is based on the ideas of Talmy (1985).
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 at this Web site.
10 http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_corpus.html.
11 http://wicopaco.limsi.fr/.

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) built the 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 the following).

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 cases of plagiarism; 40% of them are exact copies, and 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 a 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

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`.

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. 2010a, b).

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 shortcut.

16 In contrast to *simulated plagiarism*, **paraphrase plagiarism** is a more general term referring to plagiarism based on paraphrase mechanisms.

each source/plagiarism pair with our tagset (each pair contains multiple paraphrase 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 Example (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 Example (28a): *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
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 Example (29), an isolated same-polarity substitution takes place, so the scope *older/aging* is annotated and the attribute LOCAL is used. In Example (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 tagged. In Example (31), the coordination change affects two juxtaposed sentences in Example (31a) and two coordinated clauses in Example (31b), 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 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

17 http://www.lsi.upc.edu/∼textmess/.

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:

$$K_B = \frac{\sum_{b \in B} \min\left(1, \sum_{c \in C} overlapping(b, c)\right)}{|B|} \tag{2}$$

The *overlapping* measure is defined as:

$$overlapping(b, c) = \alpha \cdot \left(\frac{|b_s \cap c_s|}{|b_s|} + \frac{|b_p \cap c_p|}{|b_p|} \right) \tag{3}$$

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 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 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 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

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 the 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

to be shorter than their source (cf. top of Table 2). This means that, while reformulating (plagiarizing), people tend to use shorter expressions for the 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 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,

Table 2
Character-level lengths of the annotated paraphrases in the P4P corpus. At the top are 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

and Piao 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 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

18 `http://nlp.shef.ac.uk/meter/`.

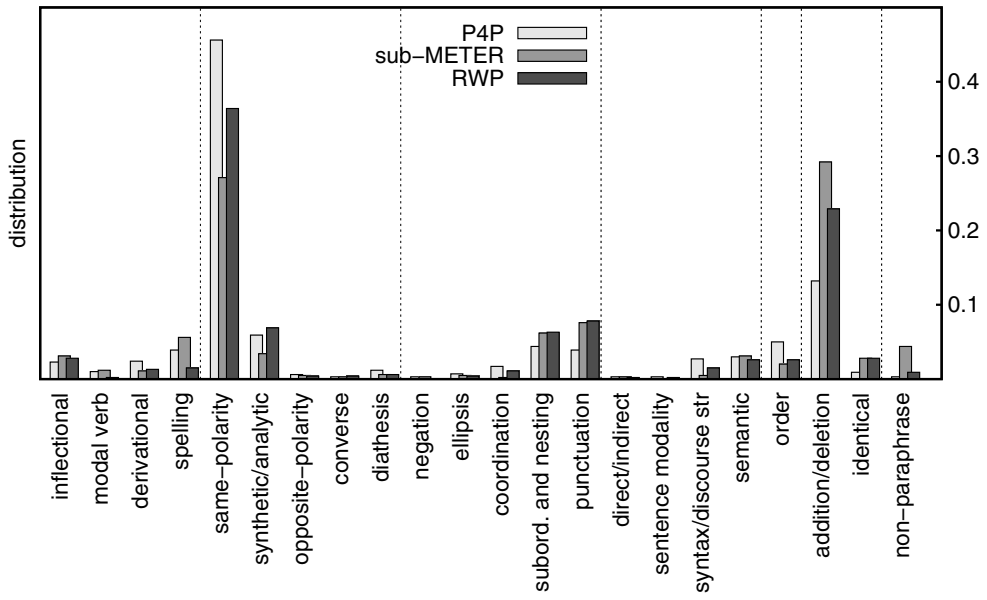

Figure 2
Overview of the paraphrase distribution in the P4P corpus with respect to the samples from the sub-METER and RWP corpora.

et al. 2011), representing a promising factor to consider in the detection of paraphrase plagiarism.

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 with the P4P data set.

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 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

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.

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 analysis**				**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						■

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

20 Refer to Clough (2000, 2003) and Maurer, Kappe, and Zaka (2006) for a general overview on approaches to plagiarism detection.

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).

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 (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, Wei 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).

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 $s_q \in d_q$, a 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

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.

22 In the dot–plot technique, documents are represented in an x, y plane: d is located in x, and 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."

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 Figure 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):

$$\begin{aligned} prec_{PDA}(S,R) &= \frac{1}{|R|} \cdot \left(\frac{|r_1 \sqcap s_1|}{|r_1|} + \frac{|r_2 \sqcap s_1|}{|r_2|} + \frac{|r_3 \sqcap s_1|}{|r_3|} + \frac{\overset{0}{\overbrace{|\emptyset|}}}{|r_4|} + \frac{|r_5 \sqcap s_2|}{|r_5|} \right) \\ &= \frac{1}{5} \cdot \left(\frac{2}{4} + \frac{1}{1} + \frac{2}{2} + \frac{3}{7} \right) = 0.5857 \quad \text{and} \end{aligned}$$

$$\begin{aligned} rec_{PDA}(S,R) &= \frac{1}{|S|} \cdot \left(\frac{|(s_1 \sqcap r_1) \bigcup (s_1 \sqcap r_2) \bigcup (s_1 \sqcap r_3)|}{|s_1|} + \frac{|s_2 \sqcap r_5|}{|s_2|} + \frac{\overset{0}{\overbrace{\emptyset}}}{|s_3|} \right) \\ &= \frac{1}{3} \cdot \left(\frac{5}{7} + \frac{3}{3} \right) = 0.5714 \end{aligned}$$

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.

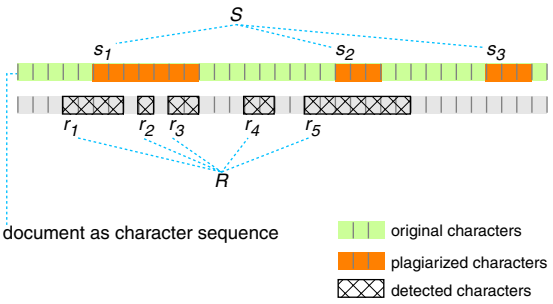

Figure 3
A document as character sequence, including plagiarized sections S and detections R returned by a plagiarism detection algorithm (used with permission of Potthast et al. [2010]).

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 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 with respect to cases of verbatim copy and algorithmically simulated paraphrasing. Then we analyze the detectors' performance when considering the previously 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 analyses 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 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).

23 An analysis considering paraphrase fragments as the retrieval units was also carried out. The obtained results were practically random, however, because in the framework of plagiarism detection, detecting a paraphrase as plagiarized in general depends on its context.

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. Because 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), however, 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 (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 Figure 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 that 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.

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

24 Identical and semantic fragments are also longer in the respective clusters than in the others.

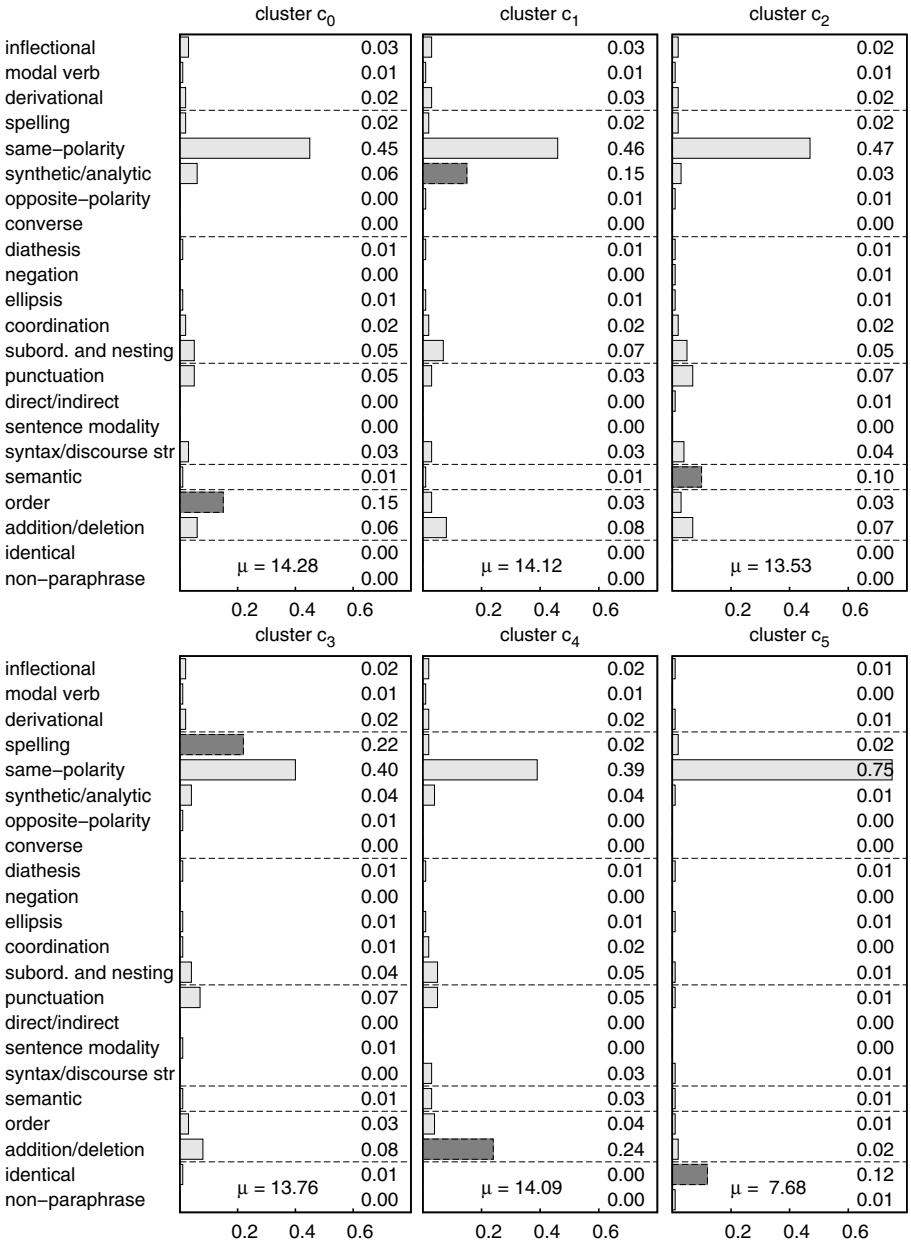

Figure 4
Average relative frequency of the different paraphrase phenomena in the source–plagiarism pairs of each cluster. The feature that stands out in the cluster and also with respect to the rest of the clusters is represented by a darker bar (setting aside same-polarity substitutions). The value of μ refers to the average absolute number of phenomena per pair in each cluster.

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 with the results on the entire corpus, which also contains translated, verbatim, and automatically modified plagiarism. Manually created cases seem to be

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:]*$_\gamma$

much harder to detect than the other, artificially generated, cases.[25] The difficulty of detecting 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 constitutes 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 constitutes 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 in detecting them. These results are clearly observed through the values of recall obtained by the different detectors. Moreover, a relation

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, whereas in the second one, with simulated (i.e., paraphrastic) cases, it decreased to 0.48.

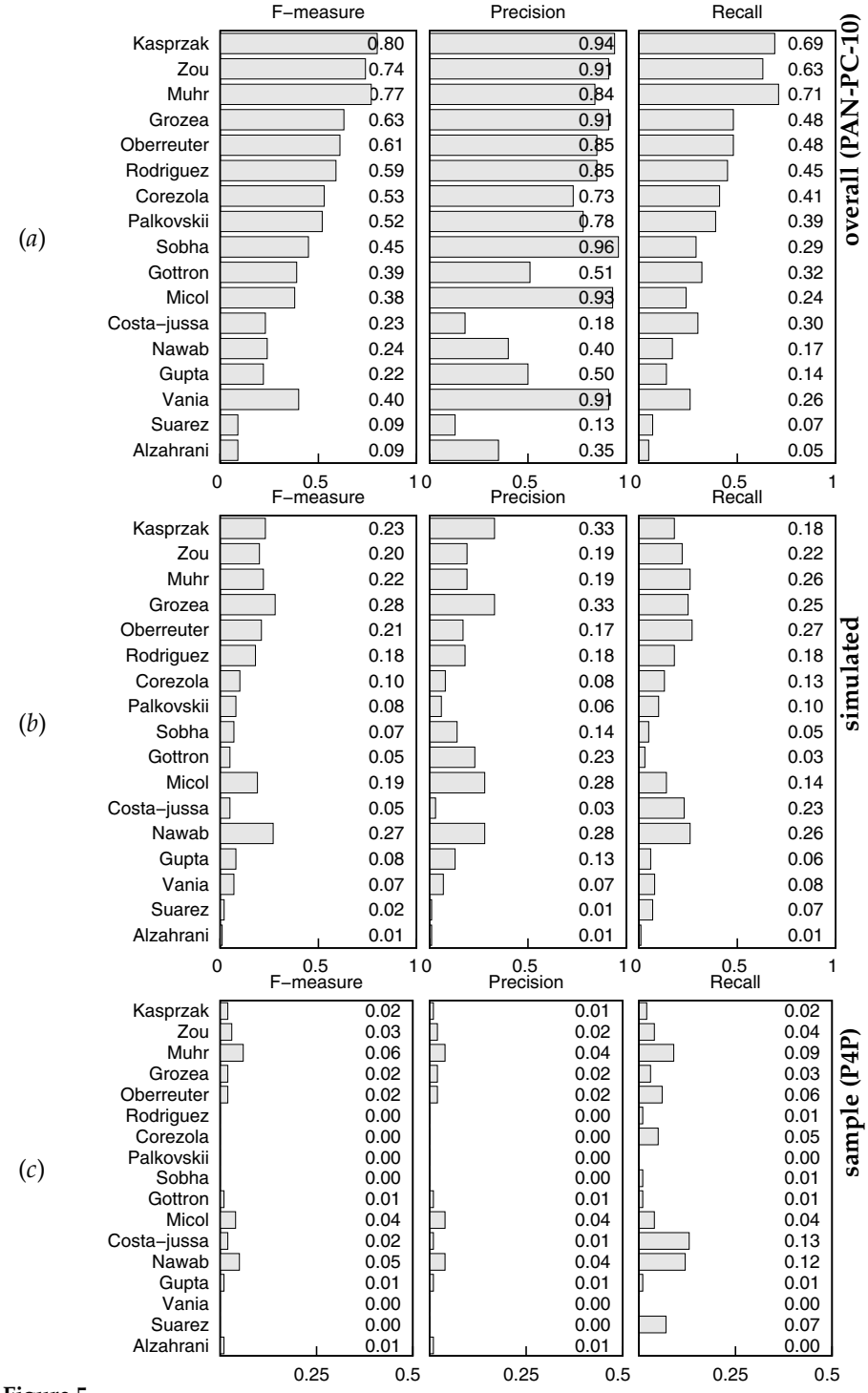

Figure 5
Evaluation of the Pan-10 competition participants' plagiarism detectors. Figures show evaluations over: (a) entire PAN-PC-10 corpus (including artificial, translated, and simulated cases); (b) simulated cases only; and (c) sample of simulated cases annotated on the basis of the paraphrases typology: the P4P corpus. Note the change of scale in (c).

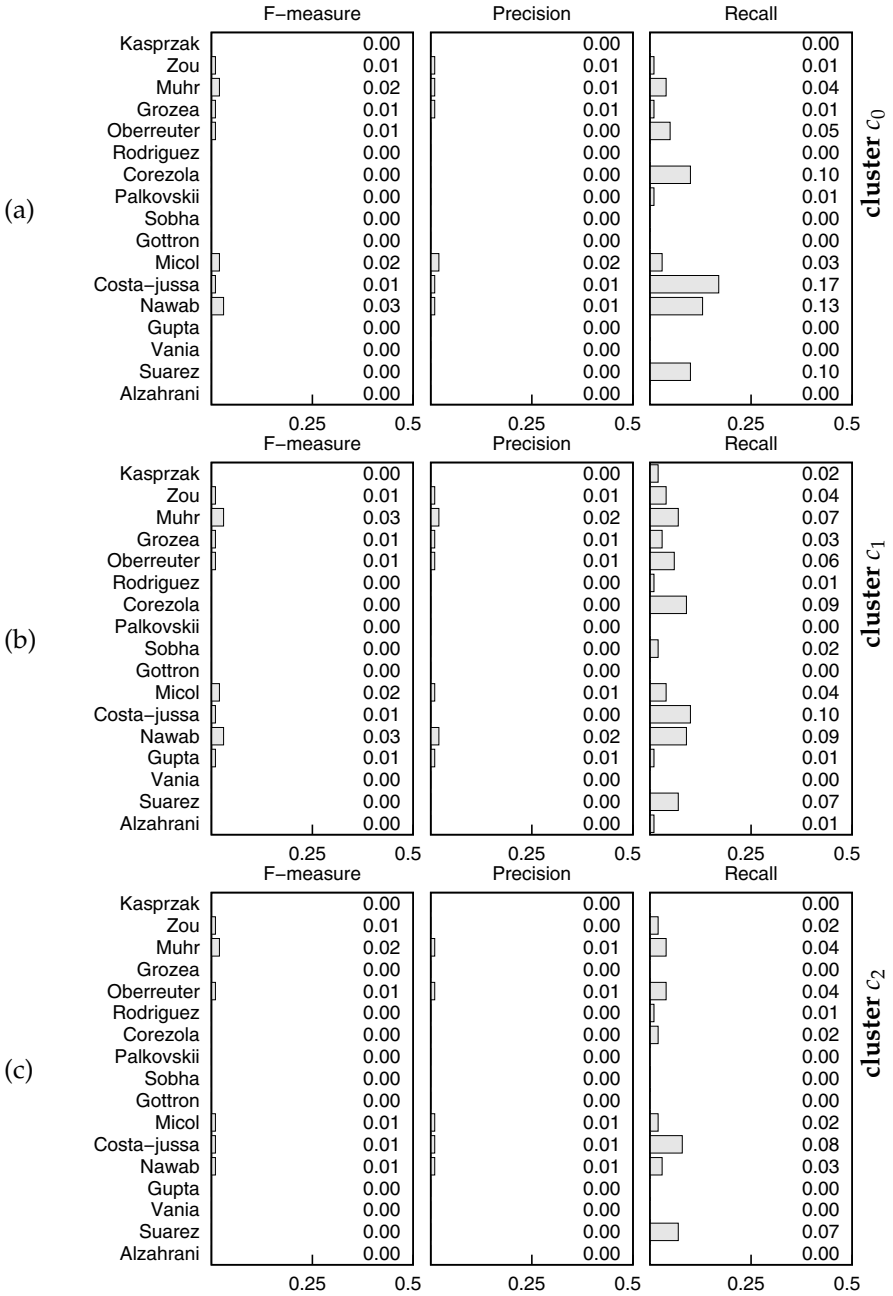

Figure 6
Evaluation of the Pan-10 competition participants' plagiarism detectors for (a) c_0; (b) c_1; and (c) c_2.

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 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

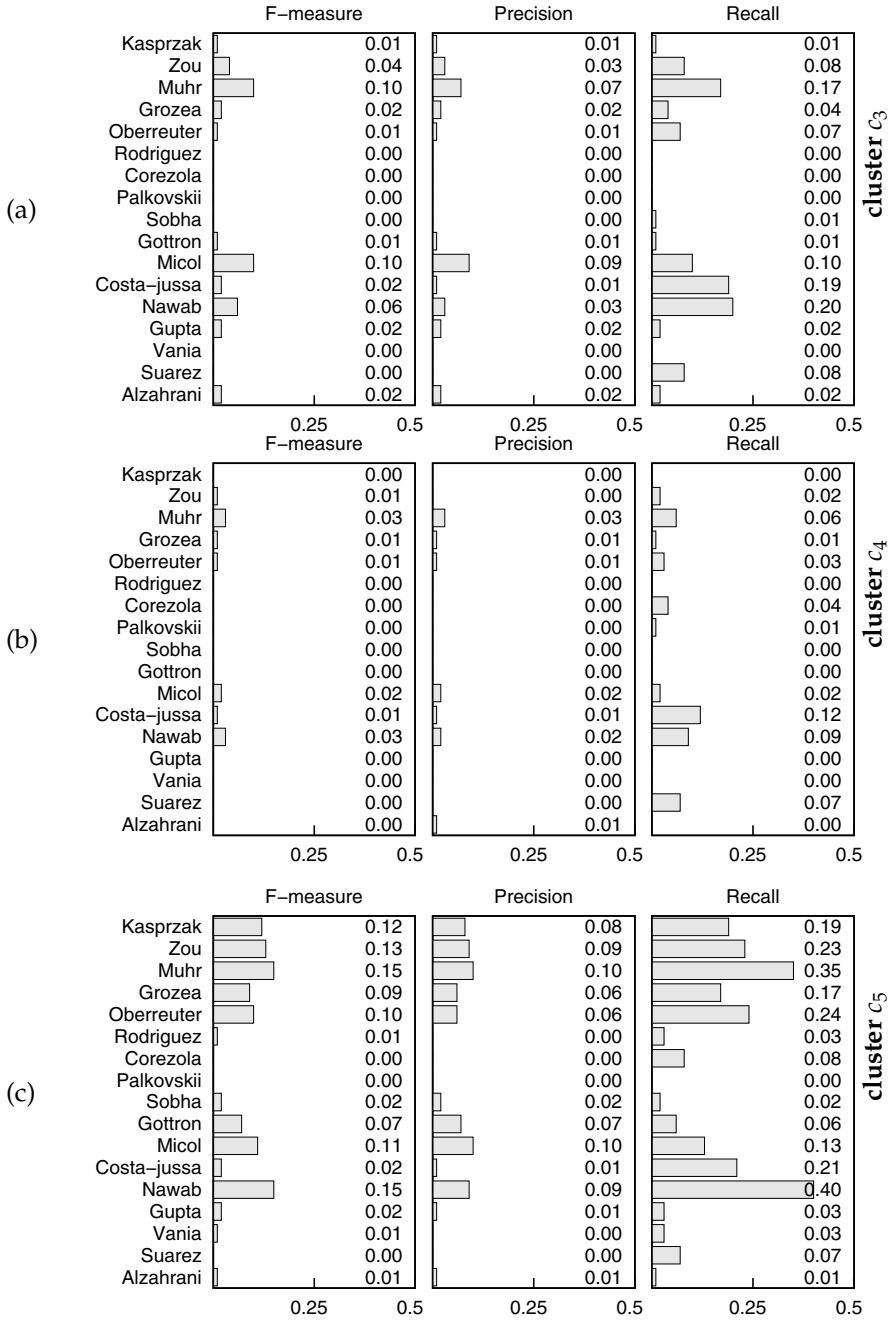

Figure 7
Evaluation of the Pan-10 competition participants' plagiarism detectors for (a) c_3; (b) c_4; and (c) c_5.

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) use 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 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 analyze 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/deletion 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.

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 who 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 no. 246016), the MICINN projects TEXT-ENTERPRISE 2.0 and TEXT-KNOWLEDGE 2.0 (TIN2009-13391), the EC WIQ-EI IRSES project (grant no. 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 Ph.D. at the Universitat Politècnica de València.

References

Alzahrani, Salha and Naomie Salim. 2010. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: `www.informatik.uni-trier.de/~ley/db/conf/clef/clef2010w.html`.

Association of Teachers and Lecturers. 2008. School work plagued by plagiarism—ATL survey. Technical report, Association of Teachers and Lecturers, London, UK. Available at: `www.atl.org.uk/Images/FrontlineSpring08.pdf`.

Barrón-Cedeño, Alberto, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, 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, College Park, MD.

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

Burrows, Steven, Martin Potthast, and Benno Stein. 2012. 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.

26 `http://wordnet.princeton.edu`.

27 `http://www.mpi-inf.mpg.de/yago-naga/yago/`.

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 1,678–1,691, 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)*, pages 3,450–3,455, Barcelona.
Corezola Pereira, Rafael, Viviane P. Moreira, and Renata Galante. 2010. UFRGS@PAN2010: Detecting external plagiarism lab report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: `www.informatik.uni-trier.de/~ley/db/conf/clef/clef2010w.html`.
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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: `www.informatik.uni-trier.de/~ley/db/conf/clef/clef2010w.html`.
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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: `www.informatik.uni-trier.de/~ley/db/conf/clef/clef2010w.html`.
Grozea, Cristian, Christian Gehl, and Marius Popescu. 2009. ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. In *Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009)*, San Sebastian, 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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: `www.informatik.uni-trier.de/~ley/db/conf/clef/clef2010w.html`.
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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.

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

IEEE. 2008. A Plagiarism FAQ. [http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html]. Last accessed 25 November 2012.

Kasprzak, Jan and Michal Brandejs. 2010. Improving the reliability of the plagiarism detection system. Lab report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.

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, IL.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley.

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):1,050–1,084.

Max, Aurélien and Guillaume Wisniewski. 2010. Mining naturally occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3,143–3,148, 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.

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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.

Nawab, Rao Muhammad Adeel, Mark Stevenson, and Paul Clough. 2010. University of Sheffield lab report for PAN at CLEF 2010. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.

Potthast, Martin, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2010. Overview of the 2nd International Competition on Plagiarism Detection. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.

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. 2010b. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, pages 997–1,005.

Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview of the 1st international competition on plagiarism detection. In *Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009)*, San Sebastian, 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 *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at:

www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.

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

Stamatatos, Efstathios. 2009. Intrinsic plagiarism detection using character *n*-gram profiles. In *Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009)*, San Sebastian, 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.

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. To appear. Relational paraphrase acquisition from Wikipedia: The WRPA method and corpus, *National Language Engineering*.

Zou, Du, Wei jiang Long, and Zhang Ling. 2010. A cluster-based plagiarism detection method. In *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua. Available at: www.informatik.uni-trier.de/∼ley/db/conf/clef/clef2010w.html.