Open Journal of Modern Linguistics, 2014, 4, 205-218
Published Online March 2014 in SciRes. http://www.scirp.org/journal/ojml
http://dx.doi.org/10.4236/ojml.2014.41016

Scientific Research

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

Marta Vila[1], M. Antònia Martí[1], Horacio Rodríguez[2]

[1]CLiC, Departament de Lingüística General, Universitat de Barcelona, Barcelona, Spain
[2]TALP, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain
Email: marta.vila@ub.edu, amarti@ub.edu, horacio@lsi.upc.edu

Received 10 November 2013; revised 17 December 2013; accepted 28 December 2013

Abstract

A precise and commonly accepted definition of paraphrasing does not exist. This is one of the reasons that have 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 unresolved issue. One of the main reasons is found in the multifaceted and

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

How to cite this paper: Vila, M., Martí, M.A. and Rodríguez, H. (2014) Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology. *Open Journal of Modern Linguistics, 4,* 205-218. http://dx.doi.org/10.4236/ojml.2014.41016

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, Sekine, & Sudo, 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 adhoc.

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, and 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 5700 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 & Brockett, 2005), and the Wikipedia-based Relational Paraphrase Acquisition corpus-WRPA (Vila, Rodríguez, & Martí, 2013). 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)[2].

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

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 is shown in the varied literature on the topic.

In what follows, we focus on how paraphrasing has been understood in linguistics (Section 2.1) and computational linguistics (Section 2.2)[4].

2.1. In Linguistics

In the field of linguistics, paraphrasing is at the core of two theories that set out language models focusing 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 identify recurrent patterns of semantic-syntactic correspondence, are a fundamental part of the ECD.

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

[3]The contents of this article were part of the PhD thesis by Vila (2013), built as an article compendium. It is available at http://hdl.handle.net/10803/117850.

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

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 representations 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 & Mel'čuk (1965), who built a paraphrasing system comprising lexical and syntactic paraphrasing rules[5]; paraphrasing at the semantic level was more recently described (Milićević, 2007a; Milićević, 2007b). 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 that maps equivalent structures. The main interest of this work is the effort to formalize paraphrasing; nevertheless, it only deals with those paraphrases that 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 structure is purely semantic, generative semantics appears to be a suitable means for describing paraphrasing[6]. 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 corresponds 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 focusing 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, focusing on paraphrasing by connotative variation, double-negation or double-inversion paraphrasing, and paraphrasing by synonymy substitution.

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 that underlie NLP paraphrase techniques. In Section 2.2.2, we overview paraphrase typologies built in this field.

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

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

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[7]. 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 & 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 & 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 & Pantel, 2001)[8], "named entities are preserved across paraphrases" (Shinyama, Sekine, & Sudo, 2002), "the meaning of the text around the source and target entities [in a concrete relation] will be similar throughout their different occurrences" (Vila, Rodríguez, & Martí, 2013), etc.

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 & 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 ngram overlapping (Dolan & Brockett, 2005); at syntax level, applying tree edit distance (Kouylekov & 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 & Lenci, 2010)[9]. 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[10].

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 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

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

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

[9]See Androutsopoulos & Malakasiotis (2010) for further reading on this topic.

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

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 focuses 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, McKeown, & Elhadad, 1999; Dutrey et al., 2011; Dolan, Quirk, & Brockett, 2004), lists of the paraphrases they address (Dorr et al., 2004; Kozlowski, McCoy, & Shanker, 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). These 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 & Witte, 1981).

There also exist exhaustive paraphrase typologies focusing 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 accompanying them (substitution, addition/deletion, and/or permutation). Finally, Fujita (2005) presents a general classification of lexical and structural paraphrases[11] 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 Section 2.2, we have shown how computational linguistics can "shed[s] new light on phenomena that traditional approaches fail to account for [and] bring refreshing insights and new points of view to all 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 focusing 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.

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-

[11]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 .

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 degrees 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 boundary cases are due to some kind of content loss. Content loss may be due to deletion [*my favorite* 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[12]. The ADDITION/DELETION type (types in our typology appear in **Tables 1** and **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

[12]Dras (1999: pp. 79-86) addresses the loss of meaning in paraphrasing regarding the paraphrase classes in his typology.

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). Spelling, punctuation, format, and paraphrase extremes are extracted from the MSRP-A corpus.

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

that remains, as the latter is a hyponym of the former. As shown in Vila et al. (Submitted), 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 US-led *invasion* of Iraq
b. The US-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)[13]. 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)[14]. Fujita (2005) talks

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

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

Table 2. *Paraphrase typology* (2).

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*

about "pragmatic paraphrases" (6) and "referential paraphrases" (9). Dorr et al. (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) points 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 quasilinguistic."

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 we move from the idea of

Table 3. Prototypes for SYNTHETIC/ANALYTIC SUBSTITUTIONS. These examples also appear in the annotation guidelines (see footnote 2) and, as all the examples there, are extracted/adapted from state of the art paraphrase typologies (see the annex of the guidelines) and the annotated corpora, or are our own.

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

"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. 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 both coreference and textual entailment, leading to recurrent confusions. In what follows, the main differences 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 & Vila, 2010). In example (17), a paraph-

rase 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 store, 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. 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. 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).

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[15].

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 focuses 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 column), two of them having two sub-classes each (second column)[16]. In what follows, an overview of our typology is set out. In concrete, we describe its scope, the type of units it classifies, its structure, 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

[15]See Section 2 in this article and the appendices in the annotation guidelines (footnote 2) for a complete list of the consulted typologies.
[16]The typology was first presented (with some slight differences) in Barrón-Cedeño et al. (2013). The present article focuses on the nature and structure of the typology; Barrón-Cedeño et al. (2013), in contrast, focuses on the definition of each type.

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 DERIVATIONAL 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).

The types[17]**.** 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 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[18]. 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 & 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.

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

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

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 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, Peñas, & Verdejo, 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 5700 paraphrase pairs from three corpora that are different in nature and in two languages (Vila et al., Submitted). Moreover, our typology proposal has already been tested in the automatic plagiarism detection field with promising results (Barrón-Cedeño et al., 2013).

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.

Acknowledgements

This work was supported by the MINECO projects DIANA (TIN2012-38603-C02-02) and SKATER (TIN2012-38584-C06-01), as well as a MECD FPU grant (AP2008-02185).

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. http://dx.doi.org/10.5539/elt.v4n3p223

Bannard, C., & Callison-Burch, C. (2005). Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* Ann Arbor (MI), 597-604.

Baroni, M., & Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics, 36(4),* 673-721. http://dx.doi.org/10.1162/coli_a_00016

Barreiro, A. (2008). *Make It Simple with Paraphrases. Automated Paraphrasing for Authoring Aids and Machine Translation*. Ph.D. Thesis, Porto: Universidade do Porto.

Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics, 39(4),* 917-947. http://dx.doi.org/10.1162/COLI_a_00153

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

Barzilay, R., & McKeown, K. (2001). Extracting Paraphrases from a Parallel Corpus. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001),* Toulouse, 50-57.

Barzilay, R., McKeown, K., & Elhadad, M. (1999). Information Fusion in the Context of Multi-Document Summarization.

Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL 1999), College Park (MD), 550-557.

Bhagat, R. (2009). *Learning Paraphrases from Text*. Ph.D. Thesis, Los Angeles: University of Southern California.

Boonthum, C. (2004). iSTART: Paraphase Recognition. *Proceedings of the ACL 2004 Student Research Workshop,* Barcelona, 31-36. http://dx.doi.org/10.3115/1219079.1219089

Boyer, M., & Lapalme, G. (1985). Generating Paraphrases from Meaning-Text Semantic Networks. *Computational Intelligence, 1(1),* 103-117. http://dx.doi.org/10.1111/j.1467-8640.1985.tb00063.x

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

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

Clough, P. (2003). *Measuring Text Reuse*. Ph.D. Thesis, Sheffield: University of Sheffield.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

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. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004),* Geneva, 350-356. http://dx.doi.org/10.3115/1220355.1220406

Dolan, B., & Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005),* Jeju Island, 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. *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora,* Lisbon, 47-52.

Dras, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. Thesis, Sydney: Macquarie University.

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. http://dx.doi.org/10.2307/356602

Fuchs, C. (1988). Paraphrases Prédicatives et Contraintes Énonciatives. In G. G. Bès, & C. Fuchs (Eds.), *Lexique et Paraphrase,* number 6 in *Lexique* (pp. 157-171). Villeneuve d'Ascq: Presses Universitaires de Lille.

Fuchs, C. (1994). *Paraphrase et Énonciation*. Paris: Ophrys.

Fujita, A. (2005). *Automatic Generation of Syntactically Well-Formed and Semantically Appropriate Paraphrases*. Ph.D. Thesis, Nara: Nara Institute of Science and Technology.

Halliday, M. (1994). *An Introduction to Functional Grammar* (2nd ed.). New York: Edward Arnold.

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. http://dx.doi.org/10.2307/411155

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. http://dx.doi.org/10.1016/S0022-5371(71)80035-X

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. http://dx.doi.org/10.1017/S1351324910000124

Kouylekov, M., & Magnini, B. (2005). Recognizing Textual Entailment with Tree Edit Distance Algorithms. *Proceedings of the 1st PASCAL Recognising Textual Entailment Challenge* (*RTE I*), Southampton, 11-13 April 2005, 17-20.

Kozlowski, R., McCoy, K. F., & Shanker, V. K. (2003). Generation of Single-Sentence Paraphrases from Predicate/Argument Structure Using Lexico-Grammatical Resources. *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP 2003*), Sapporo, 1-8.

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

Lareau, F. (2002). *La Synthèse Automatique de Paraphrases Comme Outil de vérication des Dictionnaires et Grammaires de Type Sens-Texte*. Master's Thesis, Montreal: Université de Montréal.

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

Lin, D., & Pantel, P. (2001). DIRT-Discovery of Inference Rules from Text. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (*KDD 2001*), San Francisco (CA), 20-23 August 2001, 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. http://dx.doi.org/10.1162/coli_a_00002

Martin, R. (1976). *Inférence, Antonymie et Paraphrase*. Paris: Librarie C. Klincksieck.

McKeown, K. (1983). Paraphrasing Questions Using Given and New Information. *American Journal of Computational Linguistics, 9(1)*.

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

Milićević, J. (2007a). *La Paraphrase. Modélisation de la Paraphrase Langagière*. Bern: Peter Lang.

Milićević, J. (2007b). Semantic Equivalence Rules in Meaning-Text Paraphrasing. In L. Wanner (Ed.), *Selected Lexical and Grammatical Issues in the Meaning-Text Theory* (pp. 267-296). Amsterdam: John Benjamins.

Peñas, A., & Ovchinnikova, E. (2012). Unsupervised Acquisition of Axioms to Paraphrase Noun Compounds and Genitives. In A. F. Gelbukh (Ed.), *CICLing 2012, Part I, LNCS 7181* (pp. 388-401). Berlin: Springer-Verlag.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. *Proceedings of the 23rd International Conference on Computational Linguistics* (*COLING 2010*), Beijing, 23-27 August 2010, 997-1005.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.

Recasens, M., & Vila, M. (2010). On Paraphrase and Coreference. *Computational Linguistics, 36(4),* 639-647. http://dx.doi.org/10.1162/coli_a_00014

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., & Mollá, D. (2003). Exploiting Paraphrases in a Question Answering System. *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP 2003*), Sapporo, 11 July 2003, 25-32.

Romano, L., Kouylekov, M., Szpektor, I., Dagan, I., & Lavelli, A. (2006). Investigating a Generic Paraphrase-Based Approach for Relation Extraction. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL 2006*), Trento, 3-7 April 2006, 409-416.

Rus, V., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2009). Identification of Sentence-to-Sentence Relations Using a Textual Entailer. *Research on Language and Computation, 7(2-4),* 209-229. http://dx.doi.org/10.1007/s11168-009-9065-y

Shimohata, M. (2004). *Acquiring Paraphrases from Corpora and Its Application to Machine Translation*. Ph.D. Thesis, Nara: Nara Institute of Science and Technology.

Shinyama, Y., Sekine, S., & Sudo, K. (2002). Automatic Paraphrase Acquisition from News Articles. *Proceedings of the 2nd International Conference on Human Language Technology Research* (*HLT2002*), San Francisco (CA), 24-27 March 2002, 313-318.

Smaby, R. M. (1971). *Paraphrase Grammars,* volume 2 of *Formal Linguistics Series.* Dordrecht: D. Reidel Publishing Company.

Vila, M. (2013). *Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics*. Ph.D. Thesis, Barcelona: Universitat de Barcelona.

Vila, M., Bertran, M., Martí, M. A., & Rodríguez, H. (Submitted). Corpus Annotation with Paraphrase Types. New Annotation Scheme and Inter-Annotator Agreement Measures.

Vila, M., Rodríguez, H., & Martí, M. A. (2013). Relational Paraphrase Acquisition from Wikipedia. The WRPA Method and Corpus. *Natural Language Engineering.* http://dx.doi.org/10.1017/S1351324913000235

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. http://dx.doi.org/10.1162/coli.2009.35.4.35409