Lang Resources & Evaluation (2015) 49:77–105
DOI 10.1007/s10579-014-9272-5

ORIGINAL PAPER

Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures

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

Received: 18 February 2013 / Accepted: 13 June 2014 / Published online: 2 July 2014

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

Keywords Paraphrasing · Paraphrase typology · Corpus annotation · Inter-annotator agreement

1 Introduction

Paraphrasing, which stands for different wordings expressing (approximately) the same meaning, is omnipresent in the ordinary use of natural languages. This pervasiveness makes paraphrase knowledge indispensable in many natural language

M. Vila (✉) · M. Bertran · M. A. Martí
CLiC, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain
e-mail: marta.vila@ub.edu

M. Bertran
e-mail: manu.bertran@ub.edu

M. A. Martí
e-mail: amarti@ub.edu

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

processing (NLP) systems. In this sense, paraphrase corpora are essential as they allow for a better understanding of the linguistic nature of paraphrasing, as well as the development of paraphrase tools based on real data and their subsequent evaluation.

Paraphrasing is a complex phenomenon and, although it has been an object of study in NLP over the last few decades, it sometimes gives the sense of a still unexplored field. This is quite evident at the sphere of paraphrase corpora, where there is a lack of standard or reference corpora, due to the difficulties in compiling large, general, and accurate datasets.[1] The lack of standard datasets has complicated and sometimes impeded progress in the field; also, without these resources, researchers have resorted to developing their own small, ad hoc datasets (Chen and Dolan 2011).

A special type of paraphrase corpora are those containing information about the linguistic operations underlying paraphrases, in other words, corpora with the annotation of paraphrase types. Paraphrasing presents multiple and diverse linguistic manifestations; thus, this type of corpora show a great potential in order to go a step further in solving the puzzle of paraphrasing in NLP. Nevertheless, if paraphrase corpora are few, those with type annotation are, to the best of our knowledge, almost non-existent. Building annotation schemes is inherently difficult and the task is even more complicated for phenomena that are still not well understood (Zaenen 2006), as is the case of paraphrasing. A great variety of linguistic operations give rise to paraphrases, a single paraphrase may include multiple combined paraphrase phenomena, and determining the scope of each phenomenon is not a straightforward task. This scenario makes the creation of such corpora a complex, costly, time-consuming challenge. This will be shown in the following sections through Fig. 3, which shows two pairs of paraphrases annotated by two annotators. By way of illustration, Annotator B detected, in the first pair, six different paraphrase phenomena, which sometimes overlap.

The development of these annotated paraphrase corpora involves building a powerful infrastructure backed by solid linguistic bases. In this article, we present such an infrastructure, as well as three corpora annotated by applying it. Firstly, we set out a new annotation scheme based on our paraphrase typology (Vila et al. 2014). This scheme comprises a set of 24 paraphrase-type tags, as well as instructions to detect and annotate the scope of each of these tags within the paraphrases. Secondly, we set out new measures for inter-annotator agreement in order to guarantee the quality of these annotations. We finally present three corpora annotated with our infrastructure: the Paraphrase for Plagiarism corpus (P4P), the Microsoft Research Paraphrase corpus-Annotated (MSRP-A), and the Wikipedia-based Relational Paraphrase Acquisition corpus-Annotated (WRPA-A). The latter is in Spanish; the other two, in English. The annotation of such diverse corpora is proof of the adequacy and robustness of our proposal.

Section 2 sets out the state of the art on corpora or small datasets with some kind of paraphrase-related annotation. Sections 3 and 4 describe the two components of our annotation infrastructure: the annotation scheme and the inter-annotator

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

agreement measures, respectively. Section 5 sets out the figures for the three annotated corpora, as well as a discussion and error analysis. Finally, conclusions and future work appear in Sect. 6.

2 State of the Art on Paraphrase-related Annotations

One of the corpora of reference in the field of paraphrasing is the Microsoft Research Paraphrase corpus—MSRP (Dolan and Brockett 2005).[2] It contains 5,801 English sentence pairs from news articles hand-labelled with a binary judgement indicating whether human raters considered them to be paraphrases (67 %) or not (33 %). Cohn et al. (2008),[3] in turn, present a corpus of 900 paraphrase sentence pairs manually aligned at the word or phrase level. The pairs in this corpus were compiled from three different sources: (1) equivalent sentence pairs from the MSRP corpus, (2) the Multiple-Translation Chinese corpus (MTC)[4] and (3) the monolingual parallel corpus used by Barzilay and McKeown (2001).

Besides corpora containing paraphrase pairs with yes/no annotations or alignments at word or phrase level, there exist some works that have gone further in paraphrase or paraphrase-related annotations. In this section, we focus on such works.[5]

Bhagat (2009) presents a paraphrase typology of 25 lexical changes (e.g., actor/action substitution or noun/adjective conversion) and 3 structural modifications that can accompany them (substitution, addition/deletion, and permutation). He empirically quantifies the distributions of the types annotating a small dataset: 30 sentences from the MSRP corpus and 30 sentences from MTC. Regarding the lexical changes, he took advantage of the alignments by Cohn et al. (2008) and broke the sentences into 145 and 210 phrases, respectively. These phrases were the units used for annotation. Regarding the structural changes, he annotated the entire sentences allowing more than one phenomena per sentence. The main limitations of this proposal from the perspective of paraphrase-type annotations [it should be noted that the annotation is not the main objective of Bhagat (2009)] are (1) the fact that the typology presents all paraphrase phenomena as primarily lexical (with structural changes accompanying them), which does not fit paraphrase phenomena with an important/unique structural component; (2) the coarse-grained nature and the coarse-grained annotation methodology for structural changes; and (3) the small size of the annotated set and the fact that almost half of the lexical changes do not appear in it.

[2] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/. The readme of the corpus contains a discussion on when a pair of sentences should be considered a paraphrase and when it should not, according to their approach.

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

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

[5] See Vila et al. (2013) for a more general state of the art on paraphrase corpora. See Vila et al. (2014) for a state of the art on paraphrase typologies: "paraphrase typology" does not equal "paraphrase-type annotation scheme", but typologies are the linguistic knowledge in which annotation schemes may be based. In this section, and in this article in general, we focus on the latter.

Romano et al. (2006) annotated 575 snippets from the domain of protein interaction with, among other information, the syntactic paraphrase phenomena they contain from a list of eight possibilities (e.g., passive form and coordination). The main drawbacks presenting this approach are the fact that it limits to only eight syntactic phenomena and, again, the small set they annotated.

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

Dutrey et al. (2011) define a typology of local modifications which are present in Wikipedia Correction and Paraphrase Corpus (WiCoPaCo), a corpus of rewritings extracted from the revision history in the French Wikipedia (Max and Wisniewski 2010).[6] Although it is not a paraphrase typology, it accounts for the degree of semantic variation of the types and includes rephrasings, which roughly correspond to paraphrases. The authors present the results of the manual annotation of 200 pairs of modification segments from WiCoPaCo. The annotation scheme consisted of four main classes based on the typology: surface corrections, rephrasings, strong semantic variations, and misalignments. Each annotation had to cover the entire segment and it was possible to assign several labels to the same segment. After the annotation, they observe that rephrasings have the largest number of occurrences, followed by strong semantic variations. Although this work does not cover a paraphrase-type annotation strictly speaking, from the perspective of type annotations, the tagset is again too general and the dataset annotated small.

Liu et al. (2010) and the Semantic Textual Similarity task in Semeval 2012 (Agirre et al. 2012) address paraphrase-related annotations that, although are not about paraphrase types, are worth to be mentioned here. Liu et al. (2010) present Paraphrase Evaluation Metric (PEM), which evaluates the quality of paraphrases and that of paraphrase generation systems. This metric is based on three criteria: adequacy (semantic similarity), fluency, and lexical dissimilarity. For validation purposes, they manually annotated 1,200 paraphrases, some of them created by humans, some of them automatically built. The MTC corpus was used as a source of paraphrases. The annotation for each paraphrase pair consisted of four scores, each given in a five-point scale: the above three criteria plus an overall score.

The dataset of the Semantic Textual Similarity task in Semeval 2012 contains information about paraphrasability.[7] It consists of 5,250 sentence pairs coming from

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

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

different sources, the MSRP corpus among them, annotated from 0 to 5 according to their degree of semantic similarity.

The works presented in this section are very different in nature. Although all of them are small steps in the field of paraphrase-related annotations, they do not necessarily consist of annotations with paraphrase types and, if they do, their objective is not the creation of an annotation infrastructure but simply testing a list of types, which are sometimes very coarse-grained. Moreover, these tests have been performed on very small datasets. Therefore, the development of an annotation infrastructure for paraphrase-type annotation is necessary in order to move forward in this domain.

3 The Annotation Scheme

Our annotation infrastructure consists of an annotation scheme and inter-annotator agreement measures. In this section, we focus on the former. It comprises a set of 24 paraphrase-type tags and instructions to annotate the scope of each of these tags within the paraphrase pairs. This annotation scheme was specified in the annotation guidelines[8] and is based on our paraphrase typology (Vila et al. 2014).

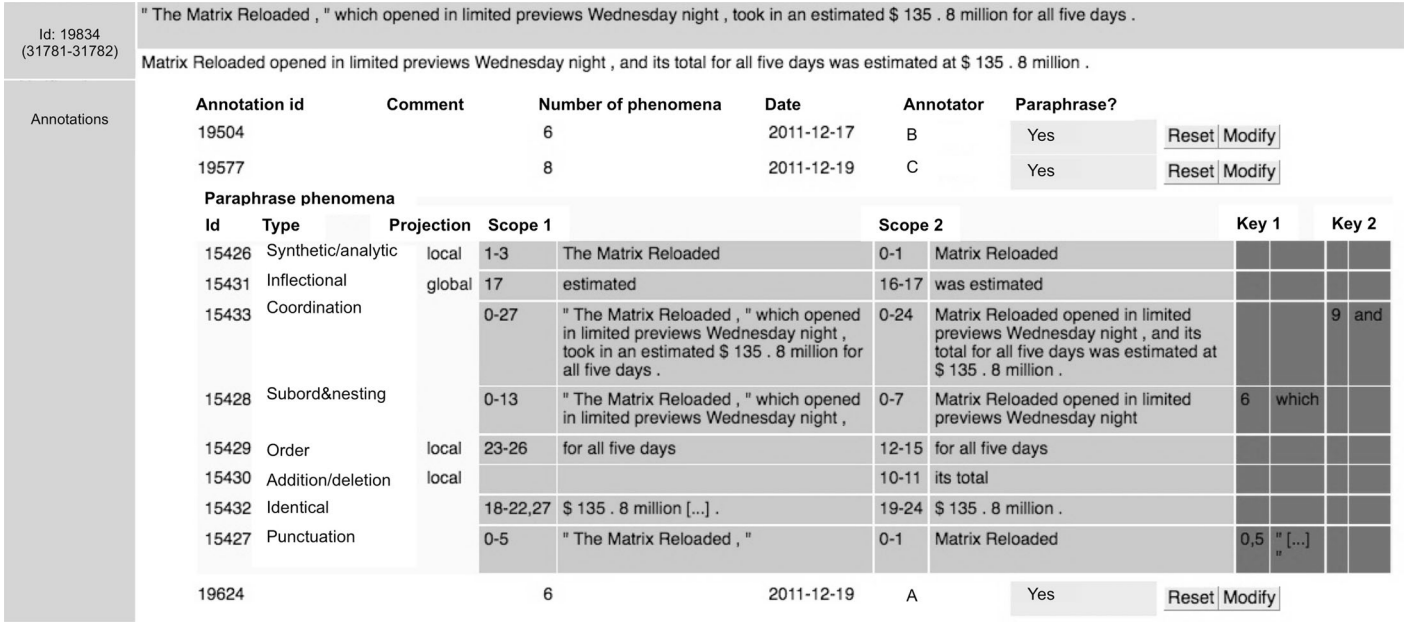
The development of paraphrase corpora annotated with types does not only involve building a consistent annotation scheme, it also has to be backed by solid linguistic bases: "annotations are not substitute for the understanding of a phenomenon. They are an encoding of that understanding" (Zaenen 2006). In this sense, our work relies on our thoughts and proposals on the paraphrase phenomenon presented in Recasens and Vila (2010) and Vila et al. (2014).

The annotation task comprises two steps: (1) the classification of pairs as paraphrases and non-paraphrases and (2) the annotation of paraphrase types within those pairs. Only pairs considered paraphrases in the first step will be subsequently annotated in the second. An example of an annotated paraphrase pair from the MSRP-A corpus can be seen in Fig. 1. This is used to illustrate the explanation below.[9]

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

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

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



Id: 19834 (31781-31782)

" The Matrix Reloaded , " which opened in limited previews Wednesday night , took in an estimated $ 135 . 8 million for all five days .

Matrix Reloaded opened in limited previews Wednesday night , and its total for all five days was estimated at $ 135 . 8 million .

Annotations

Annotation id	Comment	Number of phenomena	Date	Annotator	Paraphrase?	
19504		6	2011-12-17	B	Yes	Reset Modify
19577		8	2011-12-19	C	Yes	Reset Modify

Paraphrase phenomena

Id	Type	Projection	Scope 1		Scope 2		Key 1		Key 2	
15426	Synthetic/analytic	local	1-3	The Matrix Reloaded	0-1	Matrix Reloaded				
15431	Inflectional	global	17	estimated	16-17	was estimated				
15433	Coordination		0-27	" The Matrix Reloaded , " which opened in limited previews Wednesday night , took in an estimated $ 135 . 8 million for all five days .	0-24	Matrix Reloaded opened in limited previews Wednesday night , and its total for all five days was estimated at $ 135 . 8 million .			9	and
15428	Subord&nesting		0-13	" The Matrix Reloaded , " which opened in limited previews Wednesday night ,	0-7	Matrix Reloaded opened in limited previews Wednesday night	6	which		
15429	Order	local	23-26	for all five days	12-15	for all five days				
15430	Addition/deletion	local			10-11	its total				
15432	Identical		18-22,27	$ 135 . 8 million [...] .	19-24	$ 135 . 8 million .				
15427	Punctuation		0-5	" The Matrix Reloaded , "	0-1	Matrix Reloaded	0,5	" [...] "		

19624	6	2011-12-19	A	Yes	Reset	Modify

Fig. 1 Annotation example from the MSRP-A corpus. It consists of a screenshot of the CoCo interface used for annotation (see Sect. 5) with some terminology adaptations for consistency in this article. The paraphrase pair appears on the top. It has been annotated by three annotators ("Annotator"): A, B, and C, although only the annotation by C is displayed. For annotations by A, B, and C, the number of phenomena and a paraphrase judgement (yes/no) can be seen ("Number of phenomena" and "Paraphrase?", respectively). In the case of C, annotated phenomena are also shown, with information about the type ("Type") and the projection ("Projection"), as well as the scope ("Scope 1" corresponds to the scope in the first member of the paraphrase pair and "Scope 2", to the scope in the second; both the numeric range in tokens and the corresponding snippet of text are shown) and key elements ("Key 1" and "Key 2" are presented in the same format as scope)

using the NON-PARAPHRASE tag.[10] In the example in Fig. 1, the three annotators annotated the pair as a paraphrase.

Regarding (2) type annotation, the units we annotate are atomic paraphrase phenomena within possibly complex paraphrase pairs. Each paraphrase phenomenon is assigned a tag (the type) and a scope (the corresponding fragment in one member of the pair and the corresponding fragment in the other). In Fig. 1, eight paraphrase phenomena within the paraphrase pair are displayed for the annotator C.

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

In most of the cases, paraphrase phenomenon scopes correspond to standard linguistic units (e.g., phrase or clause), such as the nominal phrase in SYNTHETIC/ANALYTIC in Fig. 1 [SYNTHETIC/ANALYTIC is used to tag those paraphrase pairs showing differences in the degree of syntheticity; in the corresponding example in Fig. 1, *Matrix Reloaded* (without the article) is more synthetic].[12] Scopes can be discontinuous, such as the case of IDENTICAL in the same figure (indicated by [...]). Also, scopes corresponding to different paraphrase phenomena can overlap: in our example, SYNTHETIC/ANALYTIC overlaps with PUNCTUATION (this tag is used for the changes in the punctuation marks). Finally, the scope affects the annotation task differently depending on the class. In what follows, we present the three ways to annotate the scope that we defined, which can be seen in Fig. 1:

Morpholexicon-based changes, semantics-based changes, and miscellaneous changes only the affected linguistic unit(s) is(are) tagged. As some of these changes entail other changes (mainly inflectional or structural), the annotators can choose between two different facets for each phenomenon: LOCAL, which stands for those cases not entailing other changes; and GLOBAL, which stands for those cases entailing them. For these entailed changes, neither the type of change nor the fragment undergoing the change are specified in the annotation. We call this distinction between LOCAL/GLOBAL *projection* and it is compulsory for the tags in this group. By way of illustration, in Fig. 1, a change of order (ORDER tag) without any other implication takes place, so the LOCAL attribute is used.

Structure-based changes the whole linguistic unit undergoing the syntactic or discourse reorganisation is tagged. Moreover, most structure-based changes have a *key element* that gives rise to the change and/or distinguishes it from others. As the scope of structure-based changes is generally long, key elements allow for the identification of the structural change the annotator is referring to. Contrary to the

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

[11] See Vila et al. (2014) for a more detailed presentation of our paraphrase typology and Barrón-Cedeño et al. (2013) for a more detailed description of the types. In this article, we set out short definitions of the types for clarification purposes when required.

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

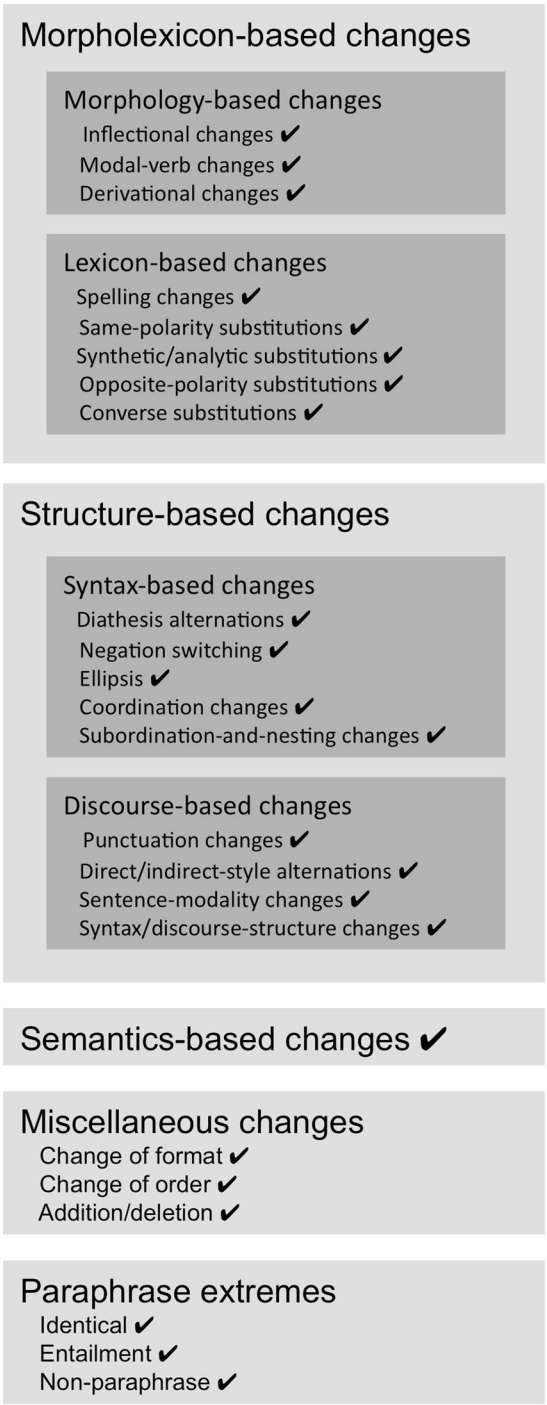


Fig. 2 Three-level paraphrase typology. Types are indicated with a tick

projection in the previous group, key elements are not compulsory for the tags in this group. In Fig. 1, the two full sentences in the pair constitute the scope of the coordination change (COORDINATION) and the conjunction *and* stands for the key element. The COORDINATION tag is used for changes in which one of the members of the pair contains coordinated snippets, and this coordination is not present or changes its position and/or form in the other member of the pair.

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

4 Inter-annotator agreement

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

Inter-annotator agreement in natural language annotation tasks is mostly calculated through observed agreement (Fleiss 1981) or the kappa measure (Cohen 1960). However, when the task is complex and the global score is the result of the combination of heterogeneous partial scores computed over smaller units, getting global agreement is rare and so these measures are close to 0. Moreover, the use of these smaller units increases the difficulty of computing agreement: they have to be carefully selected according to the task and the way to combine partial scores has to be set. These smaller units have been used in several NLP tasks, such as automatic summarisation evaluation: ROUGE (Lin and Hovy 2003), Basic Elements (Hovy et al. 2006), and the Pyramid method (Nenkova and Passonneau 2004) have been widely used in DUC and TAC contests.[13] More recently, ORANGE (Lin and Och 2004) and QARLA (Amigó et al. 2006) have been proposed as a way of combining heterogeneous measures and raters. Although these measures were created for evaluation, their application to inter-annotator agreement is straightforward. As will be shown in Sect. 4.2, a difference between evaluation and inter-annotator agreement measures is that evaluation measures distinguish between gold-standard and system and, in inter-annotator agreement, this distinction is not pertinent and annotations by different annotators have the same status. To apply evaluation measures to inter-annotator agreement, we first apply the measures taking one annotator as gold standard, then apply again the measures taking the other annotator as gold standard, and finally compute the average between the two results.

Another illustrative example, where the task of computing inter-annotator agreement is not a trivial one, is the case of annotations consisting of selecting a subset of tags from a set of interdependent ones. Kupper and Hafner (1989) proposed an agreement metric for these cases, derived from kappa. Cohn et al. (2008) argue for the usefulness of using this metric for the case of paraphrasing.

Comparing paraphrase annotations involving multiple pieces with variable type, scope, projection, and key elements is a challenge, and there are no established

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

approaches to do it. To fill this gap, we created the Inter-annotator Agreement for Paraphrase Type Annotation measures (IAPTA). These are ranged in $[0, 1]$ and classified in three groups of increasing granularity level:

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

Although the measures relevant to our work are DO-measures, because they are the most precise, we present more coarse-grained measures, namely N- and TPO-measures, because they may be useful for other approaches to paraphrase-type annotation, which are less precise in terms of scope and less costly in terms of human effort. N-measures and TPO-measures can be considered to be upper bounds for DO-measures, i.e., for a given pair, the easiest approach to compute agreement is considering only the counts of annotated tokens or phenomena (filtered by type or not), therefore, N-measures have the highest scores (Sect. 5.2); if we add the distinction between total and partial overlapping of the scope, agreement is more difficult and, therefore, TPO-measures have lower scores; if we finally consider the degree of overlapping of the scope, as well as projection and key elements, agreement is even more difficult and, thus, DO-measure scores are the lowest.

In what follows, we present each of these measures and illustrate them through two paraphrase pairs from the MSRP-A corpus annotated by B and C. Figure 3 shows these annotated pairs and Table 1 sets out the corresponding IAPTA-measure scores. Pair 1 in Fig. 3 is the same pair as in Fig. 1; also, the annotation displayed in Fig. 1 is the one corresponding to Annotator C in Fig. 3. Figure 3 shows one more annotation for this pair (the annotation by B) and also shows another paraphrase pair (Pair 2). Figure 3 shows more annotations and pairs in order to be able to illustrate inter-annotator agreement (Table 1). Nevertheless, for the sake of readability, the amount of information regarding annotations has been simplified with respect to Fig. 1 (basically, projection and key elements). As illustrative examples of Fig. 3, in Pair 1—Annotator B, we observe a PUNCTUATION change between *"The Matrix Reloaded"* and *Matrix Reloaded* (number 8); this paraphrase phenomenon overlaps with a SYNTHETIC/ANALYTIC tag (number 5); and the total number of annotated phenomena in the pair is six.

4.1 N-measures

N-measures are the most coarse-grained IAPTA measures. They only take into account the total number of tokens covered by the scope of paraphrase phenomena or the total number of phenomena annotated, sometimes filtered by type. They are ranged in $[0, 1]$. Projection or key elements are not considered.

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

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

Each agr_n measure can also be computed for each paraphrase type independently (agr_n^t) and combined into a global score by averaging ($agr_n^{\bar{t}}$). Moreover, it can be computed for each paraphrase pair (agr_n^p) and then all the pairs averaged ($agr_n^{\bar{p}}$). All the possibles combinations of (non-)typewise and (non-)pairwise approaches result in 8 measures. In order to calculate $agr_n^{\bar{p},\bar{t}}$, we first compute $agr_n^{\bar{p},t}$ for each type independently and then the average of all types ($agr_n^{\bar{p},\bar{t}}$).

This formula follows the widely used min/max (or intersection/union) way of computing similarities in disciplines such as information retrieval or text clustering, where the numerator (min/intersection) measures the similarity of the texts and the denominator (max/union) normalises it. Well known methods, such as Dice or Jaccard, are examples of this approach (Baeza-Yates and Ribeiro-Neto 1999).

In what follows, we will illustrate N-measures through agr_{ph} and agr_w^t on Pair 1 in Fig. 3. The scores corresponding to all the annotations and pairs in Fig. 3 are set out in Table 1.

The calculation of agr_{ph} is displayed in Eq. 2. As shown in Fig. 3, the number of phenomena annotated by B in Pair 1 is 6 and the number of phenomena annotated by C is 8. Figure 1 also shows these figures under "number of phenomena".

$$agr_{ph} = \frac{\min(6, 8)}{\max(6, 8)} = \frac{6}{8} = 0.75 \tag{2}$$

The calculation of agr_w^t for the tag SYNTHETIC/ANALYTIC is displayed in Eq. 3. As shown in Fig. 3, the number of tokens annotated by B under this tag is 8 and the number of tokens annotated by C under this tag is 5 (considering both members of the pair). In Fig. 1, for the case of C, we can obtain these numbers adding the ranges in "Scope 1" and "Scope 2". Then, agr_w^t would be applied to all the types in the pair and the results would be averaged in order to obtain $agr_w^{\bar{t}}$ for the pair.

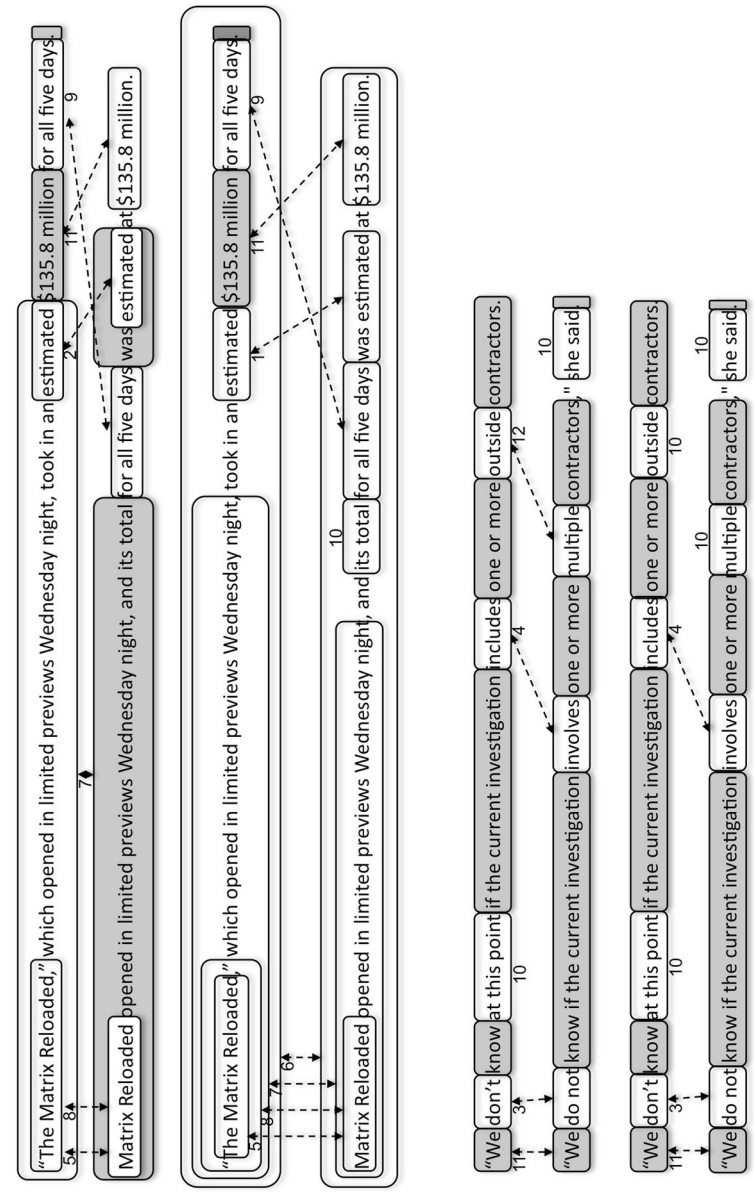
$$agr_w^{\text{SYNTHETIC/ANALYTIC}} = \frac{\min(5, 8)}{\max(5, 8)} = \frac{5}{8} = 0.625 \tag{3}$$

4.2 TPO-measures

TPO-measures are based on the evaluation measures in Dale and Narroway (2011), which were used in the pilot round of the Helping Our Own (HOO) shared task.[15] HOO aims to promote the development of automated tools and techniques that can assist authors in the writing task. Systems participating in the task had to detect

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

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


Pair 1
Annotator B
"The Matrix Reloaded," which opened in limited previews Wednesday night, took in an estimated $135.8 million for all five days.
Matrix Reloaded opened in limited previews Wednesday night, and its total for all five days was estimated at $135.8 million.
Annotator C
"The Matrix Reloaded," which opened in limited previews Wednesday night, took in an estimated $135.8 million for all five days.
Matrix Reloaded opened in limited previews Wednesday night, and its total for all five days was estimated at $135.8 million.
Pair 2
Annotator B
"We don't know at this point if the current investigation includes one or more outside contractors.
"We do not know if the current investigation involves one or more multiple contractors," she said.
Annotator C
"We don't know at this point if the current investigation includes one or more outside contractors
"We do not know if the current investigation involves one or more multiple contractors," she said.

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

Table 1 IAPTA-measure scores for annotation examples in Figure 3. For DO-measures, only F_1 is displayed

N-measures	
agr_w	0.98
$agr_w^{\bar{p}}$	0.98
$agr_w^{\bar{t}}$	0.56
$agr_w^{\bar{p},\bar{t}}$	0.57
agr_{ph}	0.81
$agr_{ph}^{\bar{p}}$	0.81
$agr_{ph}^{\bar{t}}$	0.62
$agr_{ph}^{\bar{p},\bar{t}}$	0.60
TPO-measures	
$agr_o^{partial}$	0.76
agr_o^{total}	0.55
DO-measures	
F_1	0.74

errors and infelicities in texts, indicate their extent and optionally a type, and correct them. For evaluation, a set of gold-standard edits were compared with the set of edits corresponding to the participating team's output. Three scoring measures were used: (1) detection, which indicates whether the system determined that an edit was required at some point in the text; (2) recognition, which indicates whether the system correctly determined the extent of the source text that requires editing; and (3) correction, which indicates whether the system offered a correction that is amongst the corrections provided in the gold standard.

We adapted the measures relevant to our work, that is, detection and recognition, which gave rise to $agr_o^{partial}$ and agr_o^{total}, respectively. The $agr_o^{partial}$ measure accounts for paraphrase phenomena of the same type that at least partially overlap (*partial* and *o* in the name of the measure) in scope ["lenient alignments" in Dale and Narroway (2011)]; the agr_o^{total} measure, in turn, accounts for phenomena that totally overlap ["strict alignments" in Dale and Narroway (2011)]. Positive cases in agr_o^{total} are also considered in $agr_o^{partial}$, but not vice versa. TPO-measures are ranged in $[0, 1]$. Projection and key elements are not considered.

Dale and Narroway (2011) compute precision, recall, and F_1 of the systems' edits compared to the gold-standard ones. Their approach is, therefore, directional. As inter-annotator agreement lacks directionality, we compute the precision, recall, and F_1 of one annotator taking the other as gold standard, and vice versa. Then we compute the average of the values of F_1.

In our calculation, B and C are the set of paraphrase phenomena annotated by annotators $\mathcal{B}$ and $\mathcal{C}$ (we consider independently all the phenomena occurring in all the pairs). Then, $agr_o^{partial}$ is computed as follows:

$$agr_o^{partial} = \frac{F_1^B + F_1^C}{2} \tag{4}$$

Precision, recall and F_1 are calculated as follows:

$$P^B = \frac{partialCount_B}{|B|} \tag{5}$$

$$R^B = \frac{partialCount_B}{|C|} \tag{6}$$

$$F_1^B = 2 \cdot \frac{P^B \cdot R^B}{P^B + R^B} \tag{7}$$

$PartialCount_B$ is obtained as follows:

$$partialCount_B = \sum_{b \in B} partialOverlap(b) \tag{8}$$

where $partialOverlap(b)$ is 1 if there is at least one phenomenon in C partially overlapping with b, and 0 otherwise. Finally, F_1^C is computed accordingly.

The formulae corresponding agr_o^{total} are the same as $agr_o^{partial}$ only changing *partialCount* and *partialOverlap* for *totalCount* and *totalOverlap*, i.e, moving from phenomena that, at least, partially overlap to phenomena that totally overlap.

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

- Their unit for comparison are "fragments"; in concrete, they have 19 fragments of approximately 1,000 words in length with gold standard edits and several systems' output. Our unit of comparison are pairs of snippets annotated by different annotators (see Table 3 for the figures corresponding to each corpus). To handle this, we concatenate the two members of the paraphrase pair into a single fragment.
- Their scores are calculated on a fragment-by-fragment basis and on a dataset as a whole (computing the average across the fragments). We calculate the scores in a non-pairwise way (pairwise and non-pairwise are explained in Sect. 4.1).
- In their case, participants were not required to indicate the type of error and this feature was not evaluated in that round. Type annotation was only used, when present, to obtain scores filtered for the individual types. As we are interested in taking types into account within our inter-annotator agreement calculation, we only consider the overlapping between paraphrase phenomena of the same type.

In cases where a paraphrase phenomenon only overlaps with phenomena of a different type, we consider there is no overlapping.
- Their distinction between optional and mandatory edits is not relevant to our work.
- They work at character level; we work at token level.

We will illustrate TPO-measures through $agr_o^{partial}$ in Pair 1 in Fig. 3. The scores corresponding to the whole Fig. 3 are again set out in Table 1.

In Pair 1 in Fig. 3, SUBORDINATION&NESTING, SYNTHETIC/ANALYTIC, PUNCTUATION, ORDER, and IDENTICAL show, at least, partial overlapping. From them, PUNCTUATION, ORDER, and IDENTICAL show total overlapping. Cases such as the DERIVATIONAL by annotator B and the INFLECTIONAL by annotator C, although they overlap, are not considered overlapping as their types are different. Equations 9–12 show the steps to calculate F_1^B.

$$partialCount_B = 5 \tag{9}$$

$$P^B = \frac{5}{6} = 0.833 \tag{10}$$

$$R^B = \frac{5}{8} = 0.625 \tag{11}$$

$$F_1^B = 2 \cdot \frac{0.833 \cdot 0.625}{0.833 + 0.625} = 0.714 \tag{12}$$

F_1^C would be calculated accordingly (obtaining the same result as F_1^B, that is, 0.714) and $agr_o^{partial}$ would consist of the average of F_1^B and F_1^C (again 0.714).

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

4.3 DO-measures

The DO-measures are the most fine-grained of the IAPTA measures. For each paraphrase phenomenon annotated, they calculate the degree of overlapping at token level with annotations of the other annotator of the same type. They do not only account for those annotations that totally or partially overlap, but determine to what extend they coincide. They also consider projection and key elements. DO-measures are ranged in $[0, 1]$.

In what follows, we define the computation of DO-measures: let B and C be the set of paraphrase phenomena annotated by annotators $\mathcal{B}$ and $\mathcal{C}$ (we consider

independently all the phenomena occurring in all the pairs). For a phenomenon $b \in B$, b_t refers to the type (t), b_{s_i} refers to the scope (s) in the i member of the pair ($i \in \{1, 2\}$), b_p refers to the projection (p), and b_{k_i} refers to the scope of the key element (k) in the i member of the pair.

The basic measure involved in DO-measures is the global overlapping of the annotations by one annotator and the annotations by the other: K_B is the global overlapping of B by C, that is, to what extent B cases are covered by C ones; and K_C is defined accordingly as the global overlapping of C by B. As the global overlapping is not symmetric ($K_B \neq K_C$), we average the two scores using the unweighted harmonic mean as depicted in Eq. 13.

$$F_1 = 2 \cdot \frac{K_B \cdot K_C}{K_B + K_C} \tag{13}$$

The global overlapping K_B is computed by combining the local overlapping of the different phenomena in B and C as shown in Eq. 14. K_C is computed accordingly.

$$K_B = \frac{\sum_{b \in B} \min\left(1, \sum_{c \in C} overlapping(b, c)\right)}{|B|} \tag{14}$$

The local overlapping (*overlapping*) measure of two phenomena x and y is defined in Eq. 15, ranging from 0 to 1, which goes together with Eqs. 16 and 17.

$$overlapping(x, y) = \left\{ \begin{array}{l} 0 \text{ if } x_t \neq y_t; \\ \text{otherwise,} \\ \alpha \cdot \pi \cdot \kappa \cdot (coverage(x_{s_1}, y_{s_1}, 0) + coverage(x_{s_2}, y_{s_2}, 0)) \end{array} \right\} \tag{15}$$

$$\kappa = \left\{ \begin{array}{l} 1 \text{ if all } b_{k_i} \text{ are empty;} \\ \text{otherwise,} \\ 0.75 + 0.125 \cdot coverage(b_{k_1}, c_{k_1}, 1) + 0.125 \cdot coverage(b_{k_2}, c_{k_2}, 1) \end{array} \right\} \tag{16}$$

$$coverage(x, y, \chi) = \left\{ \begin{array}{l} \chi \text{ if } |x| = 0; \\ \text{otherwise,} \\ \frac{|x \cap y|}{|x|} \end{array} \right\} \tag{17}$$

As shown in Eq. 15, if the two phenomena are of different type, the overlapping is 0; otherwise, the degree of overlapping is computed as follows. The *coverage* of the scopes of the annotations regarding the first (x_{s_1}, y_{s_1}) and second (x_{s_2}, y_{s_2}) components of the paraphrase pair are computed and added. The coverage of the two snippets x and y (Eq. 17) is computed by counting the number of tokens

occurring in both scopes and dividing by the length of x. In the case of Eq. 15, the third parameter of *coverage* is set to 0.

The nucleus of Eq. 15 is the sum of *coverages*. However, it is not sufficient as it does not consider the fact that ADDITION-DELETION only shows scope in one of the members of the paraphrase pair (x_{s_1} or x_{s_2}); moreover, *coverage* does not take projection and key elements into account. Therefore, we weighted this measure by three additional factors accounting for the ADDITION-DELETION issue (α), projection (π), and key elements (κ).[16]

Regarding the first factor, $\alpha = 1$ for ADDITION-DELETION phenomena and $\alpha = 0.5$ for others. All paraphrase phenomena but ADDITION-DELETION show scope in both members of the pair (x_{s_1} and x_{s_2}) and we need to average the two resulting *coverages* (adding them and multiplying the result by 0.5); in the case of ADDITION-DELETION, this averaging is not necessary as there is only scope and, therefore, actual *coverage* in one member of the pair (x_{s_1} or x_{s_2}).

Regarding projection, $\pi = 1$ if $b_p = c_p$, that is, if projection values coincide; otherwise, $\pi = 0.75$. In other words, if there is agreement in projection, the sum of *coverages*, which corresponds to the agreement in scope, is maintained; if there is disagreement in projection, the sum of *coverages* is reduced by 25 %. This means that we assign to the agreement in scope a relevance of 75 % and to the agreement in projection a relevance of 25 %. In our approach, the coincidence of scopes is more relevant than the coincidence in projection, which consists in secondary information, and the 75 versus 25 % partition was chosen as a general approach to reflect this.

Regarding key elements, Eq. 16 is used. Here, we give again a 75 % of relevance to the coincidence in scope, which correspond to the summand 0.75, and a 25 % of relevance to the coincidence in key elements, which corresponds to the two 0.125: as key elements may show scope in both members of the pair, we divide this 25 % into two 12.5 %. Finally, in Eq. 16, the value of χ is 1. This parameter was introduced in order to weight the spurious annotations, i.e., those annotations without counterpart. Assigning a different value to χ in Eqs. 15 and 16 comes from the fact that, in the case of key elements, we consider a disagreement to be more harmful than their simple omission.

K_B and K_C can be computed typewise and pairwise analogously to agr_n, and the notation used is the same. Therefore, we can obtain F_1 (non-typewise and non-pairwise), $F_1^{\bar{t}}$ (for each type and then averaged), $F_1^{\bar{p}}$ (for each pair and then averaged), and $F_1^{\bar{p},\bar{t}}$ (the combination of typewise and pairwise options). However, it should be taken into account that typewise measures ($F_1^{\bar{t}}$ and $F_1^{\bar{p},\bar{t}}$) are not relevant here, as *overlapping* (Eq. 15) is only computed over same-type phenomena. If we are interested in a global measure for all the types, we already have the non-typewise ones (F_1 and $F_1^{\bar{p}}$). The difference between pairwise ($F_1^{\bar{p}}$ and $F_1^{\bar{p},\bar{t}}$) and non-pairwise measures (F_1 and $F_1^{\bar{t}}$), in turn, is that pairwise variants give the same importance to each pair independently of the number of annotated phenomena; in the non-pairwise variants, all phenomena contribute equally to the final score

[16] The π and κ factors can be omitted from the calculus (i.e., they can be set to 1) if they are not relevant, as in Barrón-Cedeño et al. (2013).

independently of which pair they belong to. We consider that in approaches like ours the focus should be better on the phenomena than on the pair. Therefore, in our work, we decided to use the non-pairwise and non-typewise measure.

In what follows, we will illustrate DO-measures through the first pair in Fig. 3; in concrete, we will calculate K_B. The scores corresponding to the whole Fig. 3 are set out in Table 1.

Both $\mathcal{B}$ and $\mathcal{C}$ annotated a SYNTHETIC/ANALYTIC case (*The Matrix Reloaded* vs. *Matrix Reloaded*) and they overlap. Equations 15 and 17 are applied to this case as follows:

$$coverage(b_{s_1}, c_{s_1}, 0) = \frac{3}{6} = 0.5 \tag{18}$$

$$coverage(b_{s_2}, c_{s_2}, 0) = \frac{2}{2} = 1 \tag{19}$$

$$overlapping(x, y) = 0.5 \cdot 1 \cdot 1 \cdot (0.5 + 1) = 0.75 \tag{20}$$

From the 6 tokens annotated by $\mathcal{B}$ in the first member of the pair, $\mathcal{C}$ covers 3. In the second member of the pair, both annotators cover the same tokens. The *coverages* are therefore 0.5 and 1, as shown in Eqs. 18 and 19. Moving to the *overlapping* formula (Eq. 20), as the type addressed is not ADDITION/DELETION, α is 0.5; as both annotators determined that the projection is LOCAL, the value of π is 1; finally, κ is 1 because there are no key elements annotated (Eq. 16). Multiplying these three factors and the sum of *coverages*, we obtain the final value for *overlapping*, that is, 0.75. After repeating the procedure with the remaining tags, we can calculate K_B as follows:

$$K_B = \frac{0.75 + 1 + 0.67 + 0 + 1 + 1}{6} = 0.74 \tag{21}$$

From the 6 annotations by $\mathcal{B}$, the first value corresponds to the SYNTHETIC/ANALYTIC case, the three cases with value 1 correspond to PUNCTUATION, IDENTICAL, and ORDER, which show total overlapping; 0.67 corresponds to SUBORDINATION&NESTING; and 0 to DERIVATIONAL, as it does not have a correspondence in the other annotator.

5 The annotated corpora

The annotation scheme and inter-annotator agreement measures presented above were used to annotate three corpora. These corpora were compiled from different sources and by applying diverging techniques, which make them different in nature. Also, two of them are in English and one in Spanish. Paraphrase examples from the corpora appear in Table 2. In concrete, we annotated:[17]

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

- 847 paraphrases in the "simulated" cases of plagiarism in the PAN-PC-10 corpus (Potthast et al. 2010).[18] This gave rise to the P4P corpus, first presented in Barrón-Cedeño et al. (2013). The PAN-PC-10 corpus was created in the plagiarism domain. The "simulated" subset contains paraphrases manually created by reformulation: people were asked to simulate cases of plagiarism by rewording a given text snippet. This corpus is in English.
- The 3,900 paraphrases in the MSRP corpus (see Sect. 2). This gave rise to the MSRP-A corpus. The MSPR corpus was built using, among other complementary techniques, simple string edit distance and a heuristic strategy that pairs initial (presumably summary) sentences from different news stories in the same cluster; sentences were collected from the web over a 2-year period. This corpus is also in English.
- 1,000 paraphrases in the authorship cases in the WRPA corpus (Vila et al. 2013).[19] This gave rise to the WRPA-authorship-A corpus (simplified as WRPA-A). WRPA was built extracting relational paraphrases from Wikipedia applying the distributional hypothesis (two units of text are paraphrases if they share the same context). The subset used for annotation contains paraphrases expressing the authorship relation, that is, the relationship between an author and his work. This corpus is in Spanish.

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

The distinction between reformulative and non-reformulative paraphrases has been mentioned by other authors in the paraphrase literature. Milićević (2007), in the framework of the Meaning-Text Theory, distinguishes between "virtual" and "reformulative paraphrases": the former consists of those paraphrases sharing the same semantic representation; the latter consists of those paraphrases created by

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

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

Table 2 Paraphrase examples from the annotated corpora

P4P

(a) Bonaparte retreated to Lausanne to prepare to go to Mount St. Bernard. The veteran Austrian general did not sufficiently prepare to fight Bonaparte's arrival, as he did not think such an expedition likely

(b) Bonaparte repaired to Lausanne to prepare the expedition of Mount St. Bernard; the old Austrian general could not believe in the possibility of so bold an enterprise, and in consequence made inadequate preparations to oppose it

MSRP-A

(a) Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence

(b) Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence

WRPA-A

(a) AUTHOR es autor de WORK ('is author of')

(b) AUTHOR lanzó su primer álbum: WORK ('released his first album')

reformulating a given sentence. Bès and Fuchs (1988) also explain that "paraphrasing can be addressed at the level of virtual relation between sentences (language plane) or at the level of actual documented reformulations (discourse plane)".[20] Chen and Dolan (2011), in turn, created a paraphrase corpus by asking different annotators to describe the same video segments. They also experimented with the task of presenting a sentence to an annotator and explicitly asking for a reformulation. They observed that paraphrases in the second task diverged less, "since annotators were inevitably biased by lexical choices and word order in the original sentences;" paraphrases in the first task, in contrast, were not "based on linguistic closeness, but rather on visual similarity." Paraphrases created by reformulation showed a PINC (Paraphrase In N-gram Changes) score of 70.08, while parallel video descriptions had a score of 78.75. PINC is an evaluation metric created by the same authors and also presented in Chen and Dolan (2011) that relies on simple BLEU-like n-gram comparisons to measure the degree of novelty of automatically generated paraphrases.

The diverse levels of semantic similarity and formal correspondence, the latter being intimately linked to the distinction between reformulative and non-reformulative paraphrasing, put on the table paraphrase multifaceted nature, and the creation of annotation schemes should bear this in mind. In this sense, our annotation scheme can adapt to the multifaceted nature of paraphrasing. In concrete, WRPA-A annotation required criterion simplification: instead of using NON-PARAPHRASE or ADDITION/DELETION to tag the semantically diverging or non-parallel fragments in the pairs (which would result in an excessive number of less informative tags), only actual paraphrase phenomena were annotated, leaving the remaining fragments without any type of annotation. Therefore, the absence of tag here means non-paraphrase or that an element has been added or deleted. Also, overlapping between tags was not allowed and only the most representative tag was annotated on each

[20] The translation is ours.

fragment. Finally, projection and key elements were not used. All this constraints were incorporated to a file complementary to the guidelines.

Three annotators participated in the annotation of each corpus. They were all linguists, native Spanish speakers with an advanced level of English. The annotation of each corpus was performed in three phases: annotator training, inter-annotator agreement calculation, and final annotation. In the annotator training phase, one set of 50 cases was annotated by all annotators. Then, problems and disagreements were discussed, the guidelines were better specified regarding these issues, and the 50 annotations by one of the annotators was revised to be included in the corpus. In the inter-annotator agreement phase, one set of 100 cases was again annotated by all annotators, the inter-annotator agreement was computed obtaining an acceptable agreement (see Sect. 5.2) to proceed to the the final annotation phase. In this phase, the remaining cases in each corpus were annotated only once by one of the three annotators. The examples to be annotated in each phase (training, inter-annotator agreement, and final annotation) were randomly selected. CoCo (España-Bonet et al. 2009)[21] was the interface used for annotation.

The typology, the annotation scheme, and the inter-annotator agreement measures are independent up to a point: other tagsets can be used, some features in the annotation can be obviated, and modifications in the metrics are allowed. The annotation infrastructure can be applied to any corpora satisfying the following constraints: (1) units to be annotated are paraphrase pairs; (2) the pair is a complex paraphrase where a set of paraphrase phenomena are annotated; (3) each phenomenon is tagged with a paraphrase type from a closed tagset and eventually a scope consisting of a mapping between not necessarily contiguous spans of the two members of the pair.

In what follows, the results of the corpus annotation (Sect. 5.1) and the inter-annotator agreement scores (Sect. 5.2) are presented, together with a discussion and error analysis.

5.1 Annotation results and discussion

Table 3 shows the figures for the three annotated corpora. The lower figures in WRPA-A for average phenomena per pair and per word can be explained by the shorter length of the pair members and the adaptation of the annotation scheme mentioned above.

Table 4 shows the details for each paraphrase type in the three corpora; in concrete, their relative frequencies and average length are displayed. Empty cells are due to different reasons. In P4P, FORMAT (changes in the format) and ENTAILMENT (entailment relations) are empty because these tags did not exist in that annotation process. In MSRP-A, SENTENCE MODALITY is empty because no cases of change in the modality of the sentences were found there, as sentences in news articles are generally affirmative. In WRPA-A, many tags are empty due to the different nature of the corpus and the adaptation of the guidelines mentioned above.

[21] http://www.lsi.upc.edu/~textmess/.

Regarding relative frequencies, SAME-POLARITY and ADDITION/DELETION are the most prominent types both in P4P and MSRP-A. This is due to the accessibility of these types in reformulation processes, as they are mechanisms that are relatively simple to apply to a text by humans: changing one lexical unit for its synonym (understanding synonymy in a general sense) and deleting a text fragment, respectively. In P4P, SAME-POLARITY clearly surpasses ADDITION/DELETION, showing the high accessibility of this mechanism in conscious human reformulations. ADDITION/DELETION slightly surpasses SAME-POLARITY in MSRP-A, pointing to the recurrence in adding or deleting certain details depending on the newspaper.

The most frequent type in WRPA-A is again SAME-POLARITY; and, at a considerable distance, we find SEMANTIC, which involves a different lexicalization of the same content units. The nature of the corpus and the adaptation to the guidelines meant that the annotators tended to use the SAME-POLARITY tag when the fragment to map was a single lexical unit and the SEMANTIC tag when it was a more complex unit.

IDENTICAL is the third most frequent type in MSRP-A and WRPA-A, but among the least frequent in P4P. This is due to a change in the way the scope of this tag was marked: in P4P, IDENTICAL was only used when the identical fragment appeared between strong punctuation marks.[22] In the other corpora, all identical fragments in the pair were tagged as a single discontinuous tag. Almost all the sentences have some identical words; therefore, this is a frequent type in MSRP-A and WRPA-A.

Finally, distributions are clearly biased towards two or three types. This can correspond to either a real distribution of paraphrase phenomena or some inertia in the way of annotating. We are confident of the first interpretation because of the relatively high correlation of relative frequencies, with 0.74, 0.63, and 0.86 of Pearson's correlations. Pearson's correlations are calculated between corpus 1 and 2, corpus 2 and 3, and corpus 1 and 3. The relatively high results in all the cases show that there exists certain coherence between corpora.

Comparing our distributions with those of Bhagat (2009) (see Sect. 2), all types appear in our corpora to a greater o lesser extent, which is not the case in Bhagat (2009), where many types are not present in the corpus—in part explained by its small size. Also, in Bhagat (2009)'s resulting distributions, synonymy substitutions, function word variations, and external knowledge have the highest frequency. In structural changes, substitutions and additions/deletions are more frequent than permutations. As can be seen, there are points in common with our results.

Regarding the length of the annotated fragments, the paraphrase types with the greatest average length are those in the class of structure-based changes (see Fig. 2). The reason is to be found in the above distinction between the two ways to annotate the scope: in structural reorganisations, we annotate the whole linguistic unit undergoing the change.

One of the difficulties we had to deal with during annotation was that, in some pairs, the rewording made it difficult to isolate the paraphrase phenomena. Example

[22] Strong punctuation marks are full stops, semi-colons, question marks, exclamations, and other punctuation marks that can divide autonomous text fragments (in general, sentences, or clauses), such as parentheses, hyphens, or colons.

Table 3 Global figures for the annotated corpora

	P4P	MSRP-A	WRPA-A
Words	83,745	186,616	20,544
Pairs	856	3,900	1,000
Paraphrase phenomena	11,420	22,105	1,332
Word average in pair members	48.92	23.93	10.27
Average phenomena per pair	13.34	5.67	1.33
Average phenomena per word	1.47	1.28	0.36

Table 4 Per-type figures for the annotated corpora

Type	P4P		MSRP-A		WRPA-A	
	RF	AL	RF	AL	RF	AL
Inflectional	2.22	1.30	2.78	1.45	3.75	1.07
Modal verb	1.02	2.47	0.83	2.37	0.38	2.00
Derivational	2.29	1.03	0.85	1.05	1.73	1.00
Spelling	3.83	1.58	2.91	2.06		
Same-polarity	**44.41**	1.53	**24.81**	1.75	**53.15**	1.76
Synthetic/analytic	5.86	3.33	4.42	3.53		
Opposite-polarity	0.57	2.65	0.09	2.03		
Converse	0.29	2.09	0.20	1.95		
Diathesis	1.14	13.28	0.73	11.52		
Negation	0.29	11.73	0.09	6.88		
Ellipsis	0.76	11.08	0.30	12.88		
Coordination	1.84	26.02	0.22	14.61		
Subordination&nesting	5.23	18.83	2.14	12.31		
Punctuation	4.71	23.16	3.77	18.09		
Direct/indirect	0.32	20.40	0.30	21.06		
Sentence modality	0.31	18.37				
Syntax/discourse structure	2.74	17.36	1.39	16.33		
Semantic	2.98	9.10	1.53	6.25	**16.22**	3.35
Format			1.10	1.69		
Order	5.04	5.39	3.89	5.98	0.08	2.00
Addition/deletion	**12.91**	1.67	**25.94**	1.41		
Identical	0.88	14.20	**17.54**	13.80	**14.57**	4.15
Entailment			0.37	6.82	4.05	2.33
Non-paraphrase	0.39	15.35	3.81	2.73	6.08	9.15

RF relative frequency (percentage), *AL* average length (tokens). Figures above 10 in the RF column are in bold

(1) from P4P illustrates this situation. In these cases, we tried to isolate as many paraphrase phenomena as possible, assuming that other changes could remain without annotation. In (1), a SAME-POLARITY between *thought of* and *conceive* can be

isolated, among others. When isolating some phenomena was not possible, the SEMANTIC tag was used, as in the sentence in square brackets in (1).

(1)

(a) No longer was the body *thought of* as just a vessel, it was treated with the most respect and reverence. [From that time on artists have shown the human body to be worth of royalty and utmost fidelity.]

(b) Men began to *conceive* that the human body is noble in itself and worthy of patient study. [The object of the artist then became to unite devotional feeling and respect for the sacred legend with the utmost beauty and the utmost fidelity of delineation.]

As mentioned at the beginning of Sect. 5, our annotation scheme can adapt to the multifaceted nature of paraphrasing; in this sense, WRPA-A annotation, a corpus of non-reformulative paraphrases, required criterion simplification. Many tags were not present in WRPA and the annotators tended to use basically SAME-POLARITY and SEMANTIC tags. This article put on the table that there is a clear difference between reformulative and non-reformulative paraphrases and further work should address this.

5.2 Inter-annotator agreement scores and discussion

Table 5 shows the IAPTA scores for the three annotated corpora. Each column corresponds to the agreement between two annotators. Only one column appears for P4P as only two annotators participated in the inter-annotator agreement phase of this corpus.

The score values are consistent between the three corpora and are in line with our expectations. In N-measures, the scores for agr_w, agr_{ph}, and their pairwised versions are the highest, almost all above 0.90. The scores decrease when analysed by type, generally not being below 0.50.

In TPO- and DO-measures, the scores are in general lower than in N-measures, because TPO- and DO-measures are more fine-grained. In TPO-measures, the scores for $agr_o^{partial}$ are higher (around 0.75) than agr_o^{total} (around 0.50), also in line with our expectations. DO-measure scores are below $agr_o^{partial}$ and above agr_o^{total} (nearer $agr_o^{partial}$). Both TPO- and DO-measures take into account the scope of the phenomena, but do it to different degrees: $agr_o^{partial}$ is the loosest, because, if there is overlapping at some point, whatever its degree, the example is considered positive; agr_o^{total} is the most strict measure, because it only accepts as positive a total overlapping; finally, DO-measures are not discrete but consider the degree of overlapping.

The scores for the P4P corpus tend to be lower than those of MSRP-A, as the former was more complex to annotate: the pair members were longer and there was a higher concentration of paraphrase phenomena, as shown in Table 3 (see also examples in Table 2). Despite the lower semantic similarity and formal correspondence of the paraphrase pairs in WRPA-A, which would have made the annotation

more complex, scores in WRPA-A are higher due to the simplification in the annotation scheme applied.

Regarding individual scores per type, some aspects should be pointed out.[23] Our types (1) are generic (e.g., NEGATION covers multiple and diverse phenomena), but, at the same time, (2) precisely define which linguistic phenomenon they refer to (e.g., it stands for those paraphrases where the negation has changed its position in the sentence). However, one type in our tagset, namely SYNTAX&DISCOURSE, does not fulfil property (2) : this type is a kind of "others" for the structure-based class. As a result, this is one of the types with the lowest inter-annotator agreement. The other tag with the lowest agreement is SEMANTICS, which, up to a point, is again a by-default tag standing for cases involving multiple and varied paraphrase changes. In future work, an analysis of the phenomena annotated under these tags to see whether they accept a more fine-grained classification could be performed. The types with the highest agreement are IDENTICAL and SAME-POLARITY.

As explained in Sect. 4, in our work, we consider the most precise and adequate measure to be F_1 in DO-measures. DO-measures are the most precise of IAPTA measures and F_1 is the most adequate of DO-measures because it is the non-typewise and non-pairwise one. We propose the score of this measure as the one that should be taken as definitive in our work. The scores obtained for F_1 (generally above 0.70) are satisfactory given the difficulty of the task: it requires thorough annotator training and even experienced annotators make errors due to the complexity in the annotation of some paraphrase pairs (long snippets, high paraphrase phenomena density) and the number of features they have to take into account (type, scope, projection, and key elements). Moreover, we cannot avoid some degree of subjectivity: on occasions, different ways to annotate the same phenomenon are acceptable depending on the perspective (see "false negatives" below). In a much simpler task, the binary decision of whether two sentences are paraphrases in the MSRP corpus, a similar agreement was obtained (Dolan and Brockett 2005).[24]

It should be pointed out that, at the end of the MSRP-A annotation process, we performed a new inter-annotator agreement calculation with a new set of 100 cases. The scores of F_1 in DO-measures are 0.79, 0.78, and 0.78, respectively. These results are slightly higher than those corresponding to the first inter-annotator agreement phase (Table 5), which shows that the annotation guidelines succeed in its cohesive function by reducing disagreements.

Finally, we performed a manual analysis of a sample of annotated pairs in the inter-annotator agreement set. We classified the infelicities found into two classes: false negatives and false positives, which stand for complementary situations.

False negatives are those cases considered to be disagreements in the inter-annotator agreement calculation when they should not be considered as such in

[23] For reasons of space, we do not include the per-type scores of inter-annotator agreement. Instead, we point out the most relevant issues in this respect.

[24] Dolan and Brockett (2005)'s agreement value and ours are not directly comparable, as they represent different measures in diverging tasks with different degrees of complexity. Nevertheless, we consider that obtaining a value in the line of that of Dolan and Brockett (2005)'s simpler task shows that ours can be considered a satisfactory result.

Table 5 IAPTA-measure scores for the three corpora

	P4P	MSRP-A			WRPA-A		
	A–B	A–B	A–C	B–C	A–B	A–C	B–C
N-measures							
agr_w	0.96	0.99	0.99	1.00	0.91	0.95	0.96
$agr_w^{\bar{p}}$	0.96	0.99	0.99	1.00	0.91	0.94	0.97
$agr_w^{\bar{l}}$	0.65	0.62	0.59	0.69	0.70	0.79	0.79
$agr_w^{\bar{p},\bar{l}}$	0.65	0.63	0.60	0.69	0.70	0.77	0.78
agr_{ph}	0.98	0.98	0.99	0.99	0.97	0.99	0.96
$agr_{ph}^{\bar{p}}$	0.85	0.89	0.88	0.88	0.92	0.92	0.91
$agr_{ph}^{\bar{l}}$	0.67	0.64	0.65	0.74	0.70	0.84	0.79
$agr_{ph}^{\bar{p},\bar{l}}$	0.36	0.49	0.42	0.48	0.56	0.52	0.60
TPO-measures							
$agr_o^{partial}$	0.67	0.80	0.77	0.77	0.78	0.75	0.77
agr_o^{total}	0.42	0.54	0.49	0.51	0.55	0.63	0.53
DO-measures							
F_1	0.62	0.74	0.73	0.73	0.73	0.75	0.74

Each column corresponds to the agreement between two annotators. For DO-measures, only F_1 is displayed

absolute terms. They are due to the assumption in the inter-annotator agreement formulae that, when different-type tags by two annotators overlap, they are not referring to the same phenomenon. However, this is not always true. In the example in (2) from the P4P corpus, *regular soldiers/soldiers* was annotated by B as a change from an analytic structure to a synthetic one (SYNTHETIC/ANALYTIC tag); in contrast, A used an ADDITION/DELETION tag for *regular*. In our calculation, it is considered that B lacks an ADDITION/DELETION tag and A, a SYNTHETIC/ANALYTIC one. Nevertheless, although annotators define it differently, both tags refer to the same phenomenon and they are not contradictory. Therefore, it would be better to consider these cases as partial agreement.

(2)

(a) [...] *Regular soldiers* and the militia maintained order and discipline [...]
(b) [...] *Soldiers* and militia kept everyone in line [...]

These cases cannot be solved straightforwardly, because different-type overlapping between annotators is also due to different phenomena that simply occur together, and these two types of overlapping are not easily automatically distinguishable. We are therefore forced to accept that there is some hidden agreement in our scores.

False positives are those cases erroneously considered as agreements. They are due to the assumption in the inter-annotator agreement formulae that, when same-type tags by two annotators overlap, they refer to the same phenomenon. This is not always true. In the example (3) from P4P, B annotates with the PUNCTUATION tag the absence of a comma before *and* in (3-a) versus its presence in (3-b) (the

corresponding scope appears in curly brackets); A annotates with the same tag the change from the full stop before *taxes* in (3-a) to the comma before *those* in (3-b) (scope in square brackets). The corresponding punctuation marks are the key elements of the annotations, which allow us to detect the annotators' intention. As there exists same-tag overlapping, these cases are considered positive in the calculation when they should not be, as they are referring to different phenomena.

(3)

(a) [...] [He remitted the excise duties on beer, {cider and leather}$_B$. Taxes on spirits were increased.]$_A$

(b) [...] [The excise duties on beer, {cider, and leather}$_B$ were now totally remitted, those on spirits being somewhat increased.]$_A$

A possible way to solve this problem would be to discard those cases with excluding key elements. However, once again, this is not a straightforward task, due to the relatively freedom in key element annotations and their variability. Moreover, only some paraphrase types have key elements. It should be pointed out that false positives are rare by their very nature.

These infelicities have slightly biased our results. Given the nature of each of them, we assume that false negatives are more frequent. Therefore, the bias affects our results negatively. Addressing these issues is left for future work.

6 Conclusions and future work

In this article, we have presented a new annotation infrastructure for paraphrase-type annotation consisting of an annotation scheme and inter-annotator agreement measures, as well as three corpora annotated accordingly. The main components in the annotation scheme are a tagset and instructions on how to annotate the scope of each paraphrase phenomena; the IAPTA measures, in turn, compute agreement at different levels of granularity. The annotation of such diverse corpora as P4P, MSRP-A, and WRPA-A, which are different in nature and in two languages, has demonstrated the comprehensiveness of the annotation scheme. IAPTA measures, in turn, have shown the quality of the annotations and the adequacy of the annotation scheme to annotate new paraphrase corpora.

Paraphrasing presents multiple and diverse linguistic manifestations; therefore, this type of resource shows a great potential in order to better understand the linguistic nature of paraphrasing and to go a step further towards solving the puzzle of paraphrasing in NLP. In concrete, these corpora constitute a powerful resource for machine learning and a source for deriving new tools, such as paraphrase lexicons. These annotated corpora show which are the most frequent paraphrase types and, consequently, where to put the focus in improving NLP systems. In this sense, the P4P corpus has already been used to determine the most frequent paraphrase types in plagiarism and which types are the most difficult to detect for plagiarism detection systems (Barrón-Cedeño et al. 2013).

This is the first time that this type of annotation infrastructure and corpora have been built, which makes our work experimental. They constitute a primary step in

an almost unexplored field and open the path to new proposals and improvements. In concrete, further work could be done in (1) seeing whether the most coarse-grained tags in our proposal (SYNTAX&DISCOURSE and SEMANTICS) accept a more fine-grained classification, (2) solving the issue of false positives and false negatives in the IAPTA measures, and (3) addressing the differences between reformulative and non-reformulative paraphrases.

Acknowledgments We are grateful to the people that participated in the annotation of the corpora: Rita Zaragoza, Montse Nofre, Patricia Fernández, and Oriol Borrega. We would also like to thank Alberto Barrón-Cedeño for his help in shaping inter-annotator agreement measure formulae. This work is supported by the Spanish government through the projects DIANA (TIN2012-38603-C02-02) and SKATER (TIN2012-38584-C06-01) from Ministerio de Ciencia e Innovación, as well as a FPU Grant (AP2008-02185) from Ministerio de Educación, Cultura y Deporte.

References

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 1st joint conference on lexical and computational semantics (*SEM 2012)* (pp. 385–393). Montréal.

Amigó, E., Giménez, J., Gonzalo, J., & Màrquez, L. (2006). MT evaluation: Human-like vs. human acceptable. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics (COLING/ACL 2006)* (pp. 17–24). Sydney.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Boston: Addison-Wesley Longman Publishing Co.

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.

Barzilay, R., & McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the association for computational linguistics (ACL 2001)* (pp. 50–57). Toulouse.

Bès, G. G., & Fuchs, C. (1988). Introduction. In *Lexique et paraphrase* (pp. 7–11). Presses Universitaires de Lille.

Bhagat, R. (2009). *Learning paraphrases from Text*, Ph.D. thesis. University of Southern California, Los Angeles.

Chen, D. L., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 2011)* (Vol 1, pp. 190–200). Portland.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

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

Dale, R., & Kilgarriff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European workshop on natural language generation (ENLG 2011)* (pp. 242–249). Nancy.

Dale, R., & Narroway, G. (2011). The HOO pilot data set: Notes on release 2.0. Resource document. http://clt.mq.edu.au/research/projects/hoo/hoo2011/files/HOOReleaseNotes20110621.pdf. Accessed 8 February 2013

Dale, R., & Narroway, G. (2012). A framework for evaluating text correction. In *Proceedings of the 8th international conference on language resources and evaluation (LREC 2012)* (pp. 3015–3018). Istanbul.

Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the 3rd international workshop on paraphrasing (IWP 2005)* (pp. 9–16). Jeju Island.

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

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

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

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

Hovy, E., Lin, C. Y., Zhou, L., & Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *Proceedings of the 5th international conference on language resources and evaluation (LREC 2006)* (pp. 899–902). Genoa.

Kupper, L. L., & Hafner, K. B. (1989). On assessing interrater agreement for multiple attribute responses. *Biometrics*, *45*(3), 957–967.

Lin, C. Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 4th annual meeting of the north american chapter of the association for computational linguistics: Human language technologies (NAACL/HLT 2003), Edmonton* (Vol. 1, pp. 71–78).

Lin, C. Y., & Och, F. J. (2004). ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on computational linguistics (COLING 2004)*, Geneva.

Liu, C., Dahlmeier, D., & Ng, H. T. (2010) PEM: A paraphrase evaluation metric exploiting parallel texts. In *Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP 2010), Cambridge* (pp. 923–932).

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

Max, A., & Wisniewski, G. (2010). Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010), Valletta* (pp. 3143–3148).

Milićević, J. (2007). *La paraphrase. Modélisation de la paraphrase langagière*. Bern: Peter Lang.

Nenkova, A., & Passonneau, R. (2004). Evaluating content selection in summarization: the pyramid method. In *Proceedings of the 5th annual meeting of the North American chapter of the association for computational linguistics: human language technologies (NAACL/HLT 2004), Boston* (pp 145–152).

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010), Beijing* (pp. 997–1005).

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

Romano, L., Kouylekov, M., Szpektor, I., Dagan, I., & Lavelli, A. (2006). Investigating a generic paraphrase-based approach for relations extraction. In *Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL 2006), Trento* (pp. 409–416).

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

Vila, M., Rodríguez, H., & Martí, M. A. (2013). *Relational paraphrase acquisition from Wikipedia*. The WRPA method and corpus: Natural language engineering. doi:10.1017/S1351324913000235.

Vila, M., Martí, M. A., & Rodríguez, H. (2014). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, *4*, 205–218.

Zaenen, A. (2006). Mark-up barking up the wrong tree. *Computational Linguistics*, *32*(4), 577–580.