```
////////////////////////////////////////////////
//                                            //
//  P4P, MSRP-A, and WRPA-authorship-A Corpora  //
//                                            //
////////////////////////////////////////////////
```


[Overview]

P4P, MSRP-A, and WRPA-authorship-A are paraphrase corpora manually annotated with the paraphrase phenomena they contain.

P4P stands for "Paraphrase for Plagiarism". It consists of a partition of the plagiarism cases in the PAN-PC-10 corpus [3], in concrete, it is composed of 856 source-plagiarism pairs in English. For further reading on the corpus, refer to [1] and [4].

MSRP-A stands for "Microsoft Research Paraphrase-Annotated". It consists of the positive cases in the MSRP corpus [2], in concrete, it is composed of 3,900 paraphrase pairs in English. For further reading on the corpus, refer to [4].

WRPA-authorship-A contains a subset of the authorship paraphrases in the "Wikipedia-based Relational Paraphrase Acquisition" corpus [5], in concrete, it is composed of 1,000 pairs in Spanish. For further reading on the corpus, refer to [4].

Corpora are freely available for research purposes at

        http://clic.ub.edu/corpus/en/paraphrases-download-en (package to download)
        http://clic.ub.edu/corpus/en/search (search interface)

Annotation guidelines are available at

        http://clic.ub.edu/corpus/en/paraphrases-en

[1] A. Barrón-Cedeño, M. Vila, M.A. Martí, and P. Rosso. 2013, to appear. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics 39(4), DOI: 10.1162/COLI_a_00153.

[2] W. B. Dolan and C. Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005), Jeju Island, pages 9-16.

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

[4] M. Vila, M. A. Martí, and H. Rodríguez. Corpus annotation with paraphrase types. A new annotation infrastructure (submitted).

[5] M. Vila, H. Rodríguez, and M. A. Martí. Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus (submitted).

[Folder Contents]

- The P4P corpus	P4P.xml
- The MSRP-A corpus	MSRP-A.xml
- The WRPA-authorship-A corpus	WRPA-authorship-A.xml
- This readme file	README.txt, README.pdf

[Format]

In order to set out the corpus format, we use the following example:

```
[1] <?xml version="1.0" encoding="UTF-8" standalone="no" ?>
[2] <corpus name="P4P" version="1.0" lng="en" xmlns="http://clic.ub.edu/mbertran/formats/paraphrase-corpus"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://clic.ub.edu/mbertran/
    formats/paraphrase-corpus http://clic.ub.edu/mbertran/formats/paraphrase-corpus.xsd" >
[3]     <snippets>
[4]             <snippet id="16488" source_description="type:plagiarism;plagiarism_reference:00061;
                offset:47727;length:182;source:P4P;wd_count:37">
[5]                     All art is imitation of nature. One does not need to recognize a tangible object
                        to be moved by its artistic representation. Here by virtue of humanity's vestures,
                        lies its appeal.
[6]             </snippet>
[...]
[7]     </snippets>
[8]     <paraphrase_candidates source_description="id:9249">
[9]             <snippet id="16488" />
[10]            <snippet id="16489" />
[11]            <annotation author="87" is_paraphrase="true" source_description="id:18689" >
[12]                    <phenomenon type="lex_same_polarity" projection="local" source_description="id:5528">
[13]                            <snippet id="16488" >
[14]                                    <scope offset="125" length="4"/>
[15]                            </snippet>
[16]                            <snippet id="16489" >
[17]                                    <scope offset="71" length="11"/>
[18]                            </snippet>
[19]                    </phenomenon>
[...]
[20]                    <phenomenon type="syn_diathesis"  source_description="id:5536">
[21]                            <snippet id="16488" >
[22]                                    <scope offset="32" length="92"/>
[23]                                    <key offset="32" length="3"/>
[24]                            </snippet>
[25]                            <snippet id="16489" >
[26]                                    <scope offset="0" length="70"/>
[27]                                    <key offset="21" length="2"/>
```

```
[28]                            </snippet>
[29]                    </phenomenon>
[...]
[30]            </annotation>
[31]    </paraphrase_candidates>
[...]
[32] </corpus>
```

The format of the corpus is xml.

The "corpus" tag (from lines [2] to [32]) contains the whole corpus and the associated information. In concrete, its attributes specify the name of the corpus, the version, the language ("lng"), and the location of the xsd schema file (http://clic.ub.edu/mbertran/formats/paraphrase-corpus.xsd). Moreover, it contains the "snippets" tag and a set of "paraphrase_candidates" tags.

The "snippets" tag (from [3] to [7]) contains the set of snippets that will be part of paraphrase candidates. Each "snippet" (from [4] to [6]) contains the actual text [5] and has two attributes: an id and a "source_description". The latter, which also appears in other tags, is used in the mapping to other corpora and databases.

Each "paraphrase_candidates" tag (from [8] to [31]) contains a pair of snippets ([9] and [10]), which constitute a potential paraphrase pair and are linked to the above mentioned snippets by the id. The "paraphrase_candidates" tag also contains the annotation results ("annotation" from [11] to [30]). The "annotation" tag has the following attributes: "author" (of the annotation) and "is_paraphrase", indicating whether the pair constitutes an actual paraphrase (with "true" or "false" values). The "annotation" tag, in turn, contains a set of "phenomenon" tags.

Each "phenomenon" tag (from [12] to [19], and from [20] to [29]) contains an annotated paraphrase phenomenon. It has the "type" attribute (see Tagset below) and contains the pair of snippets involved (e.g., "snippet" from [13] to [15] and [16] to [18])]. Each snippet has a set of scopes (e.g. "scope" in [14]), one for each continuous chunk covered. "Scope" has two attributes indicating its offset and length at character level. "Addition_deletion" (see Tagset) only contains a "scope" tag corresponding to one of the snippets.

Tags corresponding to morphology-, lexicon-, and semantics-based changes, as well as miscellaneous changes (see Tagset), such as the examples in [12-19], also include the "projection" attribute (required). It stands for the impact of the phenomenon in the rest of the snippet. "Projection" values may be "global", meaning there is an impact, or "local", meaning there is not. Syntax- and discourse-based change tags, such as the example in [20-29], generally include "key" tags (non-required), which contain the most relevant chunks in "scope". There is one "key" for each continuous chunk involved. In paraphrase-extreme tags, neither the "projection" attribute nor "key" tags are provided. In the WRPA-authorship-A corpus, "projection" and "key" are not included for any tag.

When "is_paraphrase" has the "false" value, a "phenomenon" tag of the type "non-paraphrases" and a "scope" covering the snippets as a whole appears. In the case of WRPA-authorship-A, the extremes of the snippet, fulfilled by variables, are not included within this tag.

[Tagset]

The tagset is presented in the following table. Tags are grouped into classes and its extended name or meaning is also provided. When a tag is only pertinent to certain corpora, these corpora are specified between parenthesis.

Class	Tag	Meaning
Morphology-based changes	mor_inflectional	inflectional changes
	mor_modal_verb	modal-verb changes
	mor_derivational	derivational changes
Lexicon-based changes	lex_spelling_and_format (P4P)	spelling-and-format changes
	lex_spelling (MSRP-A, WRPA-A)	spelling changes
	lex_same_polarity	same-polarity substitutions
	lex_synt_ana	synthetic/analytic substitutions
	lex_opposite_polarity	opposite-polarity substitutions
	lex_inverse	converse substitutions
Syntax-based changes	syn_diathesis	diathesis alternations
	syn_negation	negation switching
	syn_ellipsis	ellipsis
	syn_coordination	coordination changes
	syn_subord_nesting	subordination-and-nesting changes
Discourse-based changes	dis_punct_format (P4P)	punctuation-and-format changes
	dis_punctuation (MSRP-A, WRPA-A)	punctuation changes
	dis_direct_indirect	direct/indirect-style alternations
	dis_sent_modality	sentence-modality changes
	syn_dis_structure	syntax/discourse-structure changes
Semantics-based changes	semantic	semantics-based changes
Miscellaneous changes	format (MSRP-A, WRPA-A)	change of format
	order	change of order
	addition_deletion	addition/deletion
Paraphrase extremes	identical	identical
	entailment (MSRP-A, WRPA-A)	entailment
	non_paraphrases	non-paraphrase

[Referencing]

Please cite the following paper when using the P4P corpus:

```
@article{Barron+etal:13,
 AUTHOR = "Barr{\'o}n-Cede{\~n}o, Alberto and
           Vila, Marta and
           Mart{\'i}, {M. Ant{\`o}nia} and
           Rosso, Paolo",
Title = "Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection",
Journal = "Computational Linguistics",
Volume = "39",
Number = "4",
Note    = "{DOI}: 10.1162/COLI\_a\_00153",
Year = "2013, to appear"
}
```

or

```
@ARTICLE{
```

PENDING. [4] ABOVE.

}

Please cite the following paper when using the MSRP-A or WRPA-authorship-A corpora:

@ARTICLE{
PENDING. [4] ABOVE.

}

[Acknowledgements]

P4P, MSRP-A, and WRPA-authorship-A corpora:

We are grateful to the people that participated in the annotation of the corpora: Rita Zaragoza, Montse Nofre, and Oriol Borrega. This work was supported by the TEXT-KNOWLEDGE 2.0 (TIN2009-13391-C04-04) and KNOW2 (TIN2009-14715-C04-04) MICINN projects, as well as a MECD FPU grant (AP2008-02185).

P4P corpus:

This work was partially carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme. It was supported by the EU FP7 Programme 2007-2013 (grant n. 246016), the TEXT-ENTERPRISE 2.0 MICINN project (TIN2009-13391-C04-03), the ECWIQ-EI IRSES project (grant n. 269180), the FP7Marie Curie People Programme, and the CONACyT-Mexico 192021 grant.

[Contact]

We would like to know what these corpora are useful for.
For comments and other issues, refer to:

http://clic.ub.edu/en/users/marta-vila-rigat (P4P, MSRP-A, or WRPA-authorship-A corpora)

http://www.lsi.upc.edu/~albarron (P4P corpus)

[Last Revision]

February 2013