

Paraphrase Typology Annotation Guidelines

Working paper 5: TEXT-MESS 2.0 (TEXT-Knowledge 2.0)

Marta Vila and M. Antònia Martí

July 2012



FPU AP2008-02185



TIN2009-13391-C04-04

Table of Contents

1. PRESENTATION	4
2. THE TASK	6
2.1 Is This a Paraphrase Pair?	6
2.2 The Tagset	7
2.3 The Scope	8
2.3.1 Kind of Units to Be Annotated	8
2.3.2 Scope Annotation Criteria	9
2.3.3 Should Punctuation Marks Be Included?	11
3. TAGSET DEFINITION	12
3.1 Morphology Based Change Tags	12
3.1.1 Inflectional	12
3.1.2 Modal Verb	12
3.1.3 Derivational	13
3.2 Lexicon Based Change Tags	14
3.2.1 Spelling	14
3.2.2 Same-polarity	15
3.2.3 Synthetic/Analytic	16
3.2.4 Opposite-polarity	18
3.2.5 Converse	18
3.3 Syntax Based Change Tags	19
3.3.1 Diathesis	19
3.3.3 Negation	19
3.3.4 Ellipsis	20
3.3.5 Coordination	20
3.3.6 Subordination & Nesting	21
3.4 Discourse Based Change Tags	21
3.4.1 Punctuation	21
3.4.2 Direct/Indirect	23
3.4.3 Sentence Modality	23
3.4.4 Syntax/Discourse Structure	23
3.5 Semantics Based Changes	24
3.6 Miscellaneous Changes	25
3.6.1 Format	25
3.6.2 Order	25
3.6.3 Addition/Deletion	26
3.7 Paraphrase Extremes	26
3.7.1 Identical	26
3.7.2 Entailment	27
3.7.3 Non-paraphrase	27

4. THE P4P, MSRP AND WRPA-AUTHORSHIP CORPUS	29
REFERENCES.....	30
ANNEX: LIST OF CONSULTED TYPOLOGIES.....	31

1. Presentation

This document sets out the guidelines for the paraphrase typology annotation task, which consists in annotating candidate paraphrase pairs with the paraphrase types they contain. Up to now, these guidelines have been used to annotate subsets of the following corpora, giving rise to their annotated versions:

<i>Corpus of origin</i>	<i>Annotated version</i>
PAN-PC-10 (Potthast et al., 2010), ¹ in English	P4P (Paraphrase for Plagiarism)
MSRP (Dolan and Brockett, 2005), ² in English	MSRP-A (Microsoft Research Paraphrase Corpus, Annotated)
WRPA-authorship (Vila et al., submitted; Vila et al., 2010), ³ in Spanish	WRPA-authorship-A (Wikipedia-based Relational Paraphrase Acquisition, Annotated)

CoCo (España et al., 2009)⁴ was the interface used for annotation.

P4P, MSRP-A and WRPA-authorship-A corpora are available at <http://clic.ub.edu/corpus/en/paraphrases-en>

Publications (to be completed, we are currently working on related publications):

Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*. To appear in issue 39:4. DOI: 10.1162/COLI_a_00153.

Barrón-Cedeño, Alberto, Marta Vila, and Paolo Rosso. 2012. Detección automática de plagio: De la copia exacta a la paráfrasis. In Elena Garayzábal, Miriam Jiménez, and Mercedes Reigosa (eds.). *Lingüística Forense: La Lingüística en el Ámbito Legal y Policial*, Euphonía Ediciones, pages 71-101.

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

Recasens, Marta and Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4):639-647.

For comments and other issues, refer to marta.vila@ub.edu

¹ <http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-10.html>


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

³ <http://clic.ub.edu/corpus/en/paraphrases-en>

⁴ <http://www.lsi.upc.edu/~textmess/>

The document is divided in three blocks: general considerations about the task (Section 2), the tagset definition (Section 3) and specificities in the annotation of the above mentioned corpora (Section 4). Examples are extracted/adapted from state of the art paraphrase typologies (See Annex) and the annotated corpora, or are our own. State of the art paraphrase typologies have sometimes also been inspiring for the creation of ours.

Marks and symbols used in this document:

- Fragments in the examples that should be annotated are underlined. When no fragment is underlined, it means that it is the whole example that should to be tagged.
- The so called key elements are **shadowed** in the examples.
-  stands for contrasts between different tags.
- **!!!** stands for warnings.

2. The Task

Paraphrasing stands for sameness of meaning between different wordings. For example, the pair of sentences in (a) are different in form but have the same meaning. Our **paraphrase typology** classifies paraphrases according to the linguistic nature of this difference in wording.

- a) John said "I like candies"/John said that he liked sweets

The task described in these guidelines consists in annotating **paraphrase pairs** with our paraphrase typology. These pairs may vary in the length, which can go from a few words to a full paragraph or bigger unit. Also, sometimes they correspond to full linguistic units (e.g., phrases or sentences); in other occasions, they consist of strings not corresponding to a complete linguistic unit. In (a), a paraphrase pair consisting of two sentences can be seen.

These paraphrase pairs are generally complex in the sense that they contain multiple atomic paraphrases. We call these atomic paraphrases **paraphrase phenomena** and they are what should be annotated with the typology. The paraphrase pair in (a) contains two paraphrase phenomena: the direct/indirect style alternation and a synonymy substitution.

In the annotation process, three main decisions should be made: determining whether a candidate paraphrase pair is effectively a **paraphrase** (Section 2.1), choosing the **tag** that best describes each phenomenon in the paraphrase pair (Section 2.2) and determining the **scope** of the fragment to be annotated with this tag (Section 2.3).

2.1 Is This a Paraphrase Pair?

The first step in the annotation process is determining whether a candidate paraphrase pair is actually a paraphrase. We consider **paraphrase pairs** those containing, at least, one paraphrase unit (a). We consider **paraphrase units** those having the same or an equivalent propositional content. Pairs without any paraphrase unit will be considered to be non-paraphrases (b). Only the pairs with a positive result will be subsequently annotated with the paraphrase typology.

- a) - Every Saturday I go to the cinema and then I have dinner with my friends
- I normally go to the swimming-pool in the morning and to restaurants with my colleagues at night
- b) - Every Saturday I go to the cinema and then I have dinner with my friends
- I have to buy a new computer for my sister

As can be seen, paraphrase pairs are understood as pairs containing at least a fragment that is a paraphrase, regardless of the content of the rest of the sentence. This decision is taken for not disregarding paraphrase fragments within sentences that are not full paraphrases. The subsequent annotation with paraphrase types will allow for distinguishing between paraphrase and non-paraphrase fragments within these sentences.

2.2 The Tagset

Our paraphrase typology consists of a three-level typology of 24 paraphrase **types** (lowercase non-bold cases) grouped into 5 **classes** (uppercase) and 4 **sub-classes** (bold) as follows. The tagset derived from the typology appears in small capitals on the right. There is one tag for each type.

- MORPHO-LEXICON BASED CHANGES
 - **Morphology based changes**
 - Inflectional changes INFLECTIONAL
 - Modal verb changes MODAL VERB
 - Derivational changes DERIVATIONAL
 - **Lexicon based changes**
 - Spelling changes SPELLING
 - Same-polarity substitutions SAME-POLARITY
 - Synthetic/analytic substitutions SYNTHETIC/ANALYTIC
 - Opposite-polarity substitutions OPPOSITE-POLARITY
 - Converse substitutions CONVERSE
- STRUCTURE BASED CHANGES
 - **Syntax based changes**
 - Diathesis alternations DIATHESIS
 - Negation switching NEGATION
 - Ellipsis ELLIPSIS
 - Coordination changes COORDINATION
 - Subordination and nesting changes SUBORDINATION & NESTING
 - **Discourse based changes**
 - Punctuation changes PUNCTUATION
 - Direct/indirect style alternations DIRECT/INDIRECT
 - Sentence modality changes SENTENCE MODALITY
 - Syntax/discourse structure changes SYNTAX/DISCOURSE STRUCTURE
- SEMANTICS BASED CHANGES
 - Semantics based changes SEMANTICS BASED CHANGES
- MISCELLANEOUS CHANGES
 - Format FORMAT
 - Change of order ORDER
 - Addition/deletion ADDITION/DELETION
- PARAPHRASE EXTREMES
 - Identical IDENTICAL
 - Entailment ENTAILMENT
 - Non-paraphrase NON-PARAPHRASE

The subclasses (morphology, lexicon, syntax and discourse based changes) follow the classical organisation in formal linguistic levels from morphology to discourse. Our paraphrase types are grouped in classes according to the nature of the underlying linguistic mechanism: (i) those types where the paraphrase arises at the morpho-lexicon level, (ii) those that are the result of a different structural organization and (iii) those types arising at the semantics level. Although the class stands for the trigger change, paraphrase phenomena in each class can entail changes in other parts of the sentence.

For instance, a morpho-lexicon based change (derivational) like the one in (a), where the verb *failed* is exchanged for its nominal form *failure*, has obvious

syntactic implications; however, the paraphrase is triggered by the morphological change. A structure based change (diathesis) like the one in (b) entails an inflectional change in *hear/was heard* among others. Finally, paraphrases in semantics are based on a different distribution of semantic content across the lexical units with, on many occasions, a complete change in the form (c).

- a) how the headmaster failed/the failure of the headmaster
- b) We were able to hear the report of a gun on shore intermittently/the report of a gun on shore was still heard at intervals
- c) I'm guessing we won't be done for some time/I've got a hunch that we're not through with that game yet

Miscellaneous changes comprise types not directly related to one single class. Finally, in paraphrase extremes, two special cases of paraphrase phenomena should be considered: they consist of the extremes of the paraphrase continuum, which goes from the highest level of paraphrasability (identity) to the lowest limits of the paraphrase phenomenon (entailment). Non-paraphrase fragments within paraphrase pairs are also part of the class paraphrase extremes.

As some of the names of our types explicitly reflect (e.g. ADDITION/DELETION), they are **bidirectional**: in a paraphrase pair, they can be applied from the first member of the pair to the second and vice versa.

2.3 The Scope

The scope refers to the selection of the tokens to be annotated within each tag. In what follows, we first define the type of units we are willing to annotate (Section 2.3.1), the criteria followed in the scope selection (Section 2.3.2) and when the punctuation marks should be included (Section 2.3.3).

2.3.1 Kind of Units to Be Annotated

We annotate **linguistic units**, not strings that do not correspond to a full linguistic unit. These linguistic units can go from the word to the (multiple-)sentence level.

In the paraphrase pair in (a), although a change takes place between the snippets *here by* and *it is there in*, two paraphrase mappings have to be established between *here* and *there* (1), and *by virtue of* and *in virtue of* (2), two different pairs of linguistic units.

- a) Here₁ by virtue of₂ humanity's vestures/It is there₁ in virtue of₂ the vesture of humanity in which it is clothed

However, selecting full linguistic units is not always possible or adequate from the paraphrase annotation point of view. In the following, we set out some exceptions to the above rule:

1. Cases in which only one member of the paraphrase pair corresponds to a linguistic unit. In (b), a SEMANTICS BASED CHANGE occurs between the underlined fragments. In the first sentence, it consists in a full linguistic unit, namely a causal clause; in the second sentence, the semantic content in the first appears divided into a nominal phrase and part of a verbal phrase, i.e., the verb *has impressed*. This nominal phrase plus the verb, although they do not constitute a full linguistic unit, are the scope of the phenomenon in the second sentence.

- b) - There is a pattern of regularity and order in the entire cosmos, due to some hints that science provides us
 - A presiding mind has impressed the stamp of order and regularity upon the whole cosmos

2. Cases in which non of the members of the paraphrase pair correspond to a linguistic unit. The prototypical example of this situation are contractions, within the SPELLING tag. In (c), *I* constitutes a nominal phrase and *will* is part of a verbal phrase. As the contraction is produced between these two pieces, they and only they constitute the scope of the phenomenon.

- c) I will go to the cinema/I'll go to the cinema

3. Cases of identical (see Section 2.3.2)

2.3.2 Scope Annotation Criteria

The way the scope should be annotated depends on the class of the tag. Three criteria should be followed:

1. Morpho-lexicon based changes, semantics based changes and miscellaneous changes: only the linguistic units affected by the trigger change are tagged (green rectangle in Figure 1). Moreover, as some of these changes may entail other modifications, mainly inflectional or structural, in the sentence (red E), two different tag attributes are provided: **local**, which stands for those cases in which the trigger change does not entail any other change in the sentence; and **global**, which stands for those cases in which the trigger change does entail other changes in the sentence (case of Figure 1). For the entailed changes pointed by the *global* attribute, neither the type of change nor the fragment suffering the change are specified in the annotation (absence of rectangle on the red E).

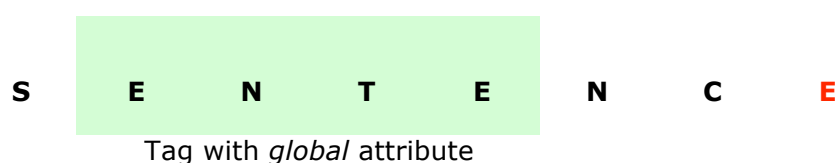


Figure 1. Criteria 1 in scope annotation

In (a), an isolated SAME-POLARITY (*motorists/drivers*) takes place, so the *local* attribute is used. In (b), the SAME-POLARITY (*rarely/has little to do with*) entails inflectional changes in other lexical units (*makes/making*). In this case, only *rarely/has little to do with* are tagged and the *global* attribute used.

- a) I dislike rash motorists/I dislike rash drivers
 b) He rarely makes us smile/He has little to do with making us smile

!!! Initial letter case changes derived, basically, from ORDER or ADDITION/DELETION do not imply the use of the *global* attribute. In (c), the change of order of *first* makes the capital letter opening the sentence changing from *we* to *first*. This changes does not imply the use of the *global* attribute, and the *local* attribute is used instead.

- c) We got to some rather biggish palm trees first./First we got to some rather biggish palm trees

2. Structure based changes: the whole linguistic unit suffering the syntactic or discourse reorganization is tagged (light green rectangle in Figure 2). If the reorganization takes place within a phrase, the phrase is tagged. If the reorganization takes place within a clause, the clause is tagged. If the reorganization takes place within a sentence, the sentence is tagged. If the reorganization takes place between different phrases/clauses/sentences (mainly coordination and subordination phenomena), all and only the phrases/clauses/sentences affected are tagged. In the case of clause changes, if the reorganizations takes place within the subordinate clause, only this one is annotated (not the main clause) and vice versa.

Moreover, all structure based changes (except from diathesis alternations) have a **key element** that gives rise to the change and/or distinguishes it from others (dark green square in Figure 2). This key element is also annotated. First, the whole linguistic unit (including the key element) is tagged, and then the key element is annotated independently.

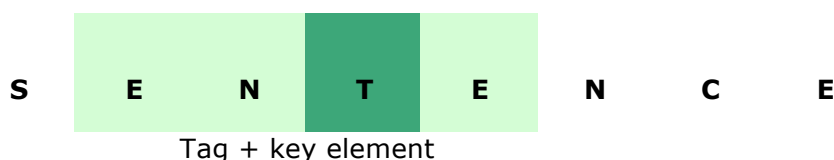


Figure 2. Criteria 2 in scope annotation

In (d), an active/passive alternation takes place (DIATHESIS tag). As the change takes place within the subordinate clause, only this clause is tagged. In (e), a change in the subordination form takes place (SUBORDINATION & NESTIG tag). As the change affects the way the two clauses (the main and the subordinate) are connected, the whole sentence is tagged. The connective mechanisms (the conjunction and the gerund clause) are annotated as key elements.

- d) When she sings that song, everything seems possible/When that song is sang, everything seems possible
- e) When we hear that song, everything seems possible/Hearing that song, everything seems possible

3. Entailment and non-paraphrase tags: the affected linguistic unit is tagged. The example in (f) is a case of ENTAILMENT; the example in (g) is a NON-PARAPHRASE.

- f) Google was in talks to buy Youtube/Google bought Youtube
- g) Mary and Wendy went to the cinema/Mary and Wendy like each other

4. Identical tag:

- Once all other phenomena are annotated, snippets which are identical in both sentences may remain. We should annotate as IDENTICAL these **snippet (not linguistic unit)** residues (h). In this case, we do not follow the linguistic unit criteria (Section 2.3.1).
- **Only one (discontinuous) identical tag** will be used in each pair of sentences.

- **Punctuation marks** will also be annotated as IDENTICAL if they effectively are.
- h) - The two argued that only a new board would have had the credibility to restore El Paso to health.
- The two believed that only a new board would have had the credibility to restore El Paso to health.

Finally, it should be noted that tags overlap on many occasions. In (i), a SAME-POLARITY tag overlaps an ORDER one.

- i) shaking his head wisely/sagely shaking his head

2.3.3 Should Punctuation Marks Be Included?

When a whole phrase/cause/sentence is annotated, **the closing (and opening) punctuation mark** (if any) is(are) included. Some examples are (a) and (b), which are cases of DIATHESIS and ADDITION/ELETION, respectively. In contrast, in (c) and (d), the commas are not included as they are not the opening and closing punctuation marks of the paraphrase phenomenon tagged (SAME-POLARITY), but of a bigger unit.

- a) This song (John sang it last year in the festival) will be a great success./
This song (it was sung by John last year in the festival) will be a great success.
- b) His judgement have kept equal pace in that conclusion/His judgment and interest may however have kept equal pace in that conclusion
- c) Before leaving and before saying goodbye, I looked around/Before leaving and before the bye bye moment, I looked around.
- d) My sisters lovely girls, live in Melbourne./ My sisters nice girls, live in Melbourne.

3. Tagset Definition

In the following, the annotation specificities are presented. For each tag, (1) the definition and (2) the scope of the fragment to be annotated are set out.

3.1 Morphology Based Change Tags

(1) Definition

Morphology based changes stand for those paraphrases that take place at the morphology level of language. Some changes in this class arise at the morphology level, but entail significant structural implications in the sentence.

(2) Scope

Only the linguistic unit affected by the trigger morphology change is annotated. The *global* attribute is used when changes in other parts of the sentence occur.

3.1.1 Inflectional

(1) Definition

Inflectional changes consist in changing inflectional affixes of words (a). In the case of verbs, this type includes all changes within the verbal paradigm (b).

(2) Scope

Inflectional changes generally affect isolated units of texts, so the linguistic unit in question is tagged using the *local* attribute (a). The *global* attribute is only used when other inflectional changes take place for agreement reasons: *occur/occurs* (c).

Auxiliaries and infinitive marks are also part of the scope (b).

- a) Unique reaction for man/Unique reactions for man
- b) Observing outside a phenomenon/To observe outside a phenomenon
- c) The rust that occurs/the rusts that occur

3.1.2 Modal Verb

(1) Definition

The MODAL VERB tag stands for changes of modality using modal verbs (a).

(2) Scope

Model verb changes generally affect isolated units of texts, so the linguistic unit in question is tagged using the *local* attribute (a).

Auxiliaries and infinitive marks are obviously part of the scope (a).

- a) You may be fine/You might be fine

3.1.3 Derivational

(1) Definition

The DERIVATIONAL tag stands for changes of category by adding derivational affixes to words. These changes comprise a syntactic reorganization in the sentence where they occur (a).

- a) The tenants wanted a reduction in the charge for electricity/The tenants wanted the charge for electricity to be reduced

☞ Although *drivers* and *driving* (b) are linked by a derivational process, this type is classified as SAME-POLARITY, and not as a DERIVATIONAL, because there is not an actual change of category, both are acting as nouns. In contrast, cases in which there is a change of category but the form remains the same are classified as DERIVATIONAL (c).

- b) I dislike rash drivers/I dislike rash driving
c) He was warned by the repeated flashing of a light/He was warned by a light flashing repeatedly

(2) Scope

Only the linguistic unit suffering the change of category is tagged, using the *global* attribute standing for the structural changes occurring in the sentence (a).

!!! On occasions, a derivational change entails other derivational changes in the sentence. In these cases, the derivational changes occupying the most nuclear position in the sentence are chosen. The *global* attribute stands for the other derivational changes in the sentence. In (d), there are two dependent derivational changes: *essentially/essential* and *identical/identity*. *Identical/identity* occupy a more nuclear position and *essentially/essential* are their respective complements, so *identical/identity* are tagged using the *global* attribute.

- d) we can see that they are essentially identical/we shall find an essential identity

On other occasions, the most nuclear position is interchanged between the two units in dispute. In these cases, the unit occupying the most nuclear position in the first sentence is chosen. In (e), we have two dependent derivational changes: *equator/equatorial* and *encircling/circle*. In the first sentence, *equator* occupies a more nuclear position, and *circle* is more nuclear in the second. *Equator/equatorial* are then tagged in this case.

- e) The trajectory of the equator encircling it/The equatorial circle described around it

Auxiliaries and infinitive marks are not part of the scope in this case.

3.2 Lexicon Based Change Tags

(1) Definition

Lexicon based change tags stand for those paraphrases that arise at the lexical level. This type gathers phenomena that, all having a lexical basis, are different in nature: they go from simple lexical substitutions, to lexical substitutions entailing significant structural implications in the sentence.

(2) Scope

Always the smallest possible lexical unit has to be annotated. In (a), we should not consider one single paraphrase phenomenon because it can be divided into two lexical units pairs: *often-debated/much-disputed* (1) and *issue/question* (2). These SAME-POLARITY substitutions are independent paraphrase phenomena, as we could substitute *often-debated* by *much-disputed*, leaving *issue* unchanged (*much-disputed issue*). Thus, two different SAME-POLARITY tags with the attribute *local* should be used. In contrast, in (b), *lies* and *is revealed* should not be tagged on their own as SAME-POLARITY substitutions, as they are semantically embedded in the wider lexical units *lies its appeal* and *its appeal is revealed*, respectively. The tag used in this case is, again, SAME-POLARITY with the attribute *local*.

- a) often-debated₁ issue₂/much-disputed₁ question₂
- b) Here by virtue of humanity's vestures, lies its appeal/Here by virtue of humanity's vestures, its appeal is revealed

Auxiliaries and infinitive marks are not tagged within the lexical unit in question. Only the verb *to be*, when it is part of a passive voice, should be included in the scope (c).

- c) The viewpoint of these lands had been altered/The whole aspect of the land had changed

3.2.1 Spelling

(1) Definition

This type comprises spelling changes and changes in the lexical form in general, the following among them:

1. **Spelling**

- a) color/colour

2. **Acronyms**

- b) North Atlantic Treaty Organization/NATO

3. **Abbreviations**

- c) Mister/Mr.

4. **Contractions**

- d) you have/you've

5. **Hyphenation**

- e) flow-accretive/flow accretive

(2) Scope

Only the affected lexical unit is tagged and the attribute *local* will be generally used.

3.2.2 Same-polarity

(1) Definition

The SAME-POLARITY tag is used when a lexical unit is changed for another one with approximately the same meaning. Both lexical (a) and functional (b) units are considered within this type. Sameness of category is not a requisite to belong to this type (c).

- a) The pilot took off despite the stormy weather/The plane took off despite the stormy weather
- b) Despite the stormy weather/In spite of the stormy weather
- c) He rarely makes us smile/He has little to do with making us smile

!!! When prepositions are part of a larger lexical unit, changes or deletions of these prepositions are tagged as SAME-POLARITY and annotated together with the lexical unit where they are embedded (d).

- d) do away/do away with

SAME-POLARITY may be used to tag several linguistic mechanisms, the following among them:

1. Synonymy

- e) I like your house/I like your place

2. General/specific

- f) I dislike rash motorists/I dislike rash drivers

3. Exact/approximate

- g) They were 9/They were around 10

4. Metaphor

- h) I was staring at her shining teeth/I was staring at her shining pearls

5. Metonymy

- i) I read a book written by Shakespeare/I read a Shakespeare

6. Expansion/compression: expressing the same content with multiple pieces and/or in a more detailed way.

- j) Ended up causing a calm aura/Caused a rather sober and subdued air

7. Word/definition

- k) Heart attacks have experienced an increase in the last decades/Sudden coronary thromboses have experiences an increase in the last decades

8. Translation

- l) Jean-François Revel, in History of the Western Philosophy/Jean-François Revel, in Histoire de la philosophie occidentale

9. Idiomatic expressions

m) It is raining cats and dogs/It is raining a lot

10. Part/whole

n) Yesterday I cut my finger/ Yesterday I cut my hand

(2) Scope

Only the changing lexical unit is tagged. The *local* attribute will be generally used (a). In other occasions, we will use the *global* one: in (o), a structural change is entailed; in (c), repeated in (p), an inflectional change occurs (*makes/making*) and, in (q), a punctuation change takes place (parenthetical commas).

o) Give him a message/Communicate a message to him

p) He rarely makes us smile/He has little to do with making us smile

q) This can never be touched with the foot/This must, on no account, be touched with the foot

3.2.3 Synthetic/Analytic

(1) Definition

SYNTHETIC/ANALYTIC stands for those changes of synthetic structures to analytic structures and vice versa. It should be noted, however, that sometimes "syntheticity" or "analyticity" is a matter of degree. Consider examples (a) and (b). In (a), we would probably consider as analytic the genitive structure. In (b), in contrast, the genitive structure would probably be the synthetic one. Genitive structures are not synthetic or analytic by definition, but more or less synthetic/analytic compared to other structures. Thus, we could redefine this group as a change in the degree of syntheticity/analyticity.

a) the Met show/the Met's show

b) Tina's birthday/The birthday of Tina

SYNTHETIC/ANALYTIC comprises phenomena such as:

1. **Compounding/decomposition**

A compound is decomposed through the use of a prepositional phrase (a). The alternation adjectival/prepositional phrase (b) and single word/adjective+noun alternations (c) are also considered here.

a) The gamekeeper preferred to make wildlife television documentaries/The gamekeeper preferred to make television documentaries about wildlife

b) Chemical life-cycles of the sexes/Life-cycles for chemistry for genders

c) One of his works holding the title "Liber Cosmographicus De Natura Locorum" belongs to a category of physiography/One of his works bearing the title of "Liber Cosmographicus De Natura Locorum" is a species of physical geography

2. **Alternations affecting genitives and possessives**

Alternations between genitive/prepositional phrases (d), possessive/prepositional phrases (e), genitive/nominal phrases (f), genitive/adjectival phrases (g), etc.

d) Tina's birthday/The birthday of Tina

e) His reflection/The reflection of his own features

- f) the Met show/the Met's show
- g) Russia's Foreign Ministry/the Russian Foreign Ministry

☞ A distinction has to be established between this type and DERIVATIONAL. Some DERIVATIONAL cases also contain genitive alternations (h), but these alternations are part of a wider derivational change. In the cases of genitive alternations classified as SYNTHETIC/ANALYTIC, the alternation is an isolated and independent phenomenon.

- h) Mary teaches John/Mary is John's teacher

☞ Cases of 1 (compounding/decomposition) and 2 (alternation involving genitives and possessives) in which the alternation takes place with a clause (with a verb) are not considered here but in SUBORDINATION & NESTING (i)

- i) Volcanoes which are now extinct/extinct volcanoes

3. Synthetic/analytic superlative

- j) He's smarter than everybody else/He's the smartest

4. Light/generic element addition: Changing a synthetic form *A* for an analytic form *BA* by adding a more generic element (*B* is more generic than *A*). *A* has to have the same lemma/stem in both member of the pair as in (k). Moreover, although the category of the phrase *A* and the phrase *BA* may differ, the change does not have structural consequences outside *A* or *BA*. In (l), although the adverbial phrase *cheerfully* is changed to the prepositional phrase *in a cheerful way*, the rest of the sentence remains unchanged. Finally, the order of the *A* and *B* units can be *BA* (k) or *AB* (l).

- k) John boasted about his work/John spoke boastfully about his work
- l) Marilyn carried on with her life cheerfully/Marilyn carried on with her life in a cheerful way

☞ When *B* is the verb *to be* and there is a change of category of *A* through a derivational process, the phenomenon is tagged as DERIVATIONAL (m)

- m) Sister Mary was helpful to Darrell/Sister Mary helped Darrell

5. Specifier addition: This type is parallel to the previous one, but the added element *B* is not more generic, but focuses on one of the components or characteristics of *A* (n), emphasises *A* (o) or determines *A* (p).

- n) I had to drive through fog to get there/I had to drive through a wall of fog to get there
- o) We are meeting at 5/We are meeting at 5 o'clock
- p) Translation is what they need/The translation is what they need

☞ Contrary to SAME-POLARITY or SEMANTICS BASED CHANGES, where words vary from one member of the paraphrase pair to the other, in synthetic/analytic substitutions

- although a change of category may take place, lexical word stems are the same (1 and 2) or
- a support element is added, but other lexical word stems are the same (4 and 5).

(2) Scope

The whole phrase affected by the change should be tagged: in (l), repeated in (q), *in a cheerful way* should be tagged, because it is this whole prepositional phrase that substitutes the adverb *cheerfully*. The attribute *local* will be generally used.

- q) Marilyn carried on with her life cheerfully/in a cheerful way

3.2.4 Opposite-polarity

(1) Definition

OPPOSITE-POLARITY stands for changes of one lexical unit for another one with opposite polarity. In order to maintain the same meaning, other changes have to occur. Two phenomena are considered within this type:

1. Double change of polarity

A lexical unit is changed for its antonym or complementary. In order to maintain the same meaning, a double change of polarity has to occur within the same sentence: another antonym (a) or complementary substitution (b), or a negation (c).

- a) John lost interest in the endeavour/John developed disinterest in the endeavour
b) Only 20% of the students were late/Most of the students were on time
c) He did not succeed in either case/He failed in both enterprises

2. Change of polarity and argument inversion

An adjective is changed for its antonym in comparative structures. In order to maintain the same meaning, an argument inversion has to occur (d).

- d) The neighbouring town is poorer in forest resources than our town/Our town is richer in forest resources than the neighbouring town

(2) Scope

In the case of double change of polarity, the two changes of polarity have to be tagged as a single (and possibly discontinuous, like in b) phenomenon and using a single tag, with the *local* or *global* attribute.

In the case of change of polarity and argument inversion, only the antonym adjectives are tagged using the *global* attribute standing for the argument inversion (d).

3.2.5 Converse

(1) Definition

A lexical unit is changed for its converse. In order to maintain the same meaning, an argument inversion has to occur (a).

(2) Scope

Only the converses are tagged using the *global* attribute standing for the argument inversion (a).

- a) Amy enjoyed the interaction/The interaction pleased Amy

3.3 Syntax Based Change Tags

(1) Definition

Syntax based change tags stand for those changes that involve a syntactic reorganization in the sentence. This type basically comprises changes within a single sentence; and changes in the way sentences, clauses or phrases are connected.

(2) Scope

The phrase/clause/sentence(s) suffering the modification is(are) tagged. All syntax tags but DIATHESIS have key elements that should be annotated as well.

3.3.1 Diathesis

(1) Definition

DIATHESIS gathers the diathesis alternations in which verbs can participate (a).

- a) Mike boiled the water/The water boiled

(2) Scope

The whole linguistic unit suffering the syntactic reorganization is tagged. In this case, there is no key element to be tagged.

3.3.3 Negation

(1) Definition

Changing the position of the negation within a sentence, like in (a), (b) or (c).

- a) One does not need to recognize a tangible object to be moved by its artistic representation/In order to move us, it needs no reference to any recognised original
b) No children came/Children didn't come
c) He refused to recognize his faults/He recognized no fault

(2) Scope

The whole linguistic unit suffering the modification is tagged (not only the negation scope). Negation marks are tagged as key elements.

!!! Remember the notation used in these guidelines to mark scope and key elements in the examples (Section 1).

3.3.4 Ellipsis

(1) Definition

This tag includes linguistic ellipsis, i.e., those cases in which the elided snippets can be recovered through linguistic mechanisms. In (a), in the first member of the pair the idea of “being able to change to” is expressed twice; in the second member of the pair it is only expressed once due to elision.

- a) - Thus, chemical force **can become** electrical current and that current **can change back into** chemical being.
- So we **can change** chemical force into the electric current, or the current into chemical force.

☞ When the elided snippets cannot be recovered solely through linguistic mechanisms, they must be considered DELETIONS.

(2) Scope

The whole linguistic unit suffering the modification is tagged (not only the elided snippets). All appearances of the elided snippet in both sentences are tagged as key elements: the idea of “being able to change to” in (a).

3.3.5 Coordination

(1) Definition

Changes in which one of the members of the pair contains coordinated snippets. This coordination is not present (in (a) it changes to a juxtaposition) or changes its position and/or form (b) in the other member of the pair.

- a) I like **pears and apples**/I like pears. I like apples
b) **Older plans and contemporary ones**/Old **and** contemporary plans

☞ When the alternation takes place between, on the one hand, coordinated or juxtaposed units and, on the other hand, subordinated or nested units, the phenomenon is tagged as SUBORDINATION & NESTING.

(2) Scope

Only the coordinated or juxtaposed linguistic units are tagged. In the first member of the pair in (a), the coordinated units are only the two noun phrases, so only they are tagged. In the second member of the pair, two full sentences are juxtaposed, so they constitute the scope of the annotation. Only the coordination (not juxtaposition) marks are tagged as key elements.

3.3.6 Subordination & Nesting

(1) Definition

Changes in which one of the members of the pair contains a subordination (a) or a nesting (b). This subordination or nesting is not present (in (a) and (b) it changes to a juxtaposition) or changes the position and/or form (c) in the other member of the pair. Nesting is understood as a general term meaning that something is embedded in a bigger unit.

- a) - A building, **which** was devastated by the bomb, was completely destroyed.
- A building was devastated by the bomb. It was completely destroyed.
- b) - Patrick Ewing scored **a personal season high** of 41 points.
- Patrick Ewing scored 41 points. It was a **personal season high**
- c) The conference venue is in **the building whose roof is red**/The conference venue is in **the building with red roof**.

(2) Scope

Only the linguistic units involved in the subordination or nesting, as well as the coordinated and juxtaposed units, are tagged. In the first member of the pair in (a), the subordinated and main clauses are tagged. In the second member, the two juxtaposed sentences should be tagged.

In case a conjunction, a relative pronoun or a preposition are present, they are tagged as the key elements (a and c). In case they are not present, the whole subordinated or nested snippet is tagged (b). Juxtaposition or coordination elements are not tagged as key elements.

3.4 Discourse Based Change Tags

(1) Definition

These tags stand for those changes that take place at the discourse level of language. This type gathers phenomena that are very different in nature, though all having in common that consist in structural changes not affecting the argumental elements in the sentence.

(2) Scope

The phrase/clause/sentence(s) suffering the modification is(are) tagged. Moreover, a key element should be tagged in all discourse based tags.

3.4.1 Punctuation

(1) Definition

Changes in the punctuation (a). Cases consisting of linguistic mechanisms parallel to punctuation like (b) are also considered here.

- a) This, as I see it, is wrong/This—as I see it—is wrong.

- b) - You will purchase a return ticket to Streatham Common and a platform ticket at Victoria station
 - At Victoria Station you will purchase (1) a return ticket to Streatham Common and (2) a platform ticket

Sometimes occurs that several changes in the punctuation take place at the same time. These multiple changes are considered as a single phenomenon if they take place at the same level (between phrase, between clause or between sentence), like in (c). If they belong to different levels, they are tagged as separate phenomena: two changes in the punctuation take place in (d), repeated in (e), but they are annotated as independent paraphrase phenomena: one of them is tagged in (d) and the other in (e).

- c) I know she is coming. She will be fine. I know it/I know she is coming. she will be fine. I know it
- d) I need to buy a couple of things. Then, I will come/I need to buy a couple of things; then I will come
- e) I need to buy a couple of things. Then, I will come/I need to buy a couple of things; then I will come

☞ Deleting punctuation marks is considered here, not as a DELETION.

(2) Scope

The whole linguistic unit(s) suffering the modification is(are) tagged. Compare (d) and (e): in (d) the whole example should be tagged; in (e), only the underlined part. The changing punctuation signs are tagged as key elements.

☞ The limits between PUNCTUATION, COORDINATION and SUBORDINATION & NESTING are sometimes fuzzy. In the following, we set out a summary table of the three:

Tag	Units involved
PUNCTUATION	Juxtaposed units
COORDINATION	At least one of the members has to contain a coordination. The other can contain a coordination or a juxtaposition
SUBORDINATION & NESTING	At least one of the members has to contain a subordination or nesting. The other can contain a subordination/nesting, a coordination or a juxtaposition.

3.4.2 Direct/Indirect

(1) Definition

Changing direct style for indirect style, and vice versa.

(2) Scope

The whole linguistic unit suffering the modification is tagged (a). The conjunction in the indirect style is tagged as key element. If no conjunction is present, the whole subordinate clause is tagged.

- a) John said "I like football"/John said **that** he liked football

3.4.3 Sentence Modality

(1) Definition

Cases in which there is a change of modality (a). We are referring strictly to changes between affirmative, interrogative, exclamatory and imperative sentences.

- a) **Can** I make a reservation?/I'd **like to** make a reservation

☞ In MODAL VERB tags, in contrast, only modal verb alternations are involved.

(2) Scope

The whole unit suffering the modification is tagged. The elements that change are tagged as key elements (a).

3.4.4 Syntax/Discourse Structure⁵

(1) Definition

This tag is used to annotate other changes in the structure of the sentences not considered in the syntax and discourse based tags above: (a), (b) and (c).

(2) Scope

The linguistic unit(s) suffering the modification is(are) tagged. The elements that change are tagged as key elements.

- a) John wore his best suit to the dance last night/**It was** John **who** wore his best suit to the dance last night
b) He wanted to eat **nothing but** apples/**All** he wanted to eat **were** apples.
c) **You are very** courageous/**You have shown how** courageous **you are**

⁵ This tag is part of the class structure based changes (see Section 2.2). As it contains both changes at the syntax level and changes at the discourse level, it is not embedded in any of these subclasses. However, to simplify indexing, we include it in the discourse based tag section.

3.5 Semantics Based Changes

(1) Definition

SEMANTICS BASED CHANGES tag stands for changes that imply a different lexicalisation pattern of the same content units, like in (a), (b) or (c).

- a) It's a rare day that they manage to make their linguistic units happy/It is not always so fortunate as to make its supporters happy
- b) No one can stand against the effect of Giacomo's words/Nothing could equal the effect produced by Giacomo's words.
- c) One of his English acquaintances is here/An Englishman he had known is here

☞ The boundaries between SEMANTICS BASED CHANGES and SAME-POLARITY are not always clear. The following table sets out the criteria to distinguish them:

CASE	EXAMPLE	REQUIRED TAGS
One lexical unit is a paraphrase of another lexical unit.	my <u>mother</u> my <u>mum</u>	1 same-polarity
Two independent lexical unit substitutions.	<u>linked₁ closely₂ to₁</u> <u>intimately₂ connected with₁</u>	2 same-polarity
One single lexical unit is expanded to more than one in the other member of the pair.	calm rather sober and subdued	1 same-polarity
Change affecting more than one lexical unit and a clear cut of these units in the paraphrase mapping is not possible. In the example, the content units of TROPICAL-LIKE ASPECT and INCREASE OF THIS ASPECT are present in both snippets, but there is not a clear-cut mapping between the two.	which added to the tropical appearance the scenery was altogether more tropical	1 semantics
Sometimes a clear-cut mapping would be possible, but a different lexicalization exists.	Georges, a Mount Avery native Georges grew up in Mount Avery	1 semantics

(2) Scope

The affected linguistic unit is tagged. The *local* attribute is used when no other changes are entailed (c). The *global* attribute is used when there are other implications in the sentence: in (a), only the second member of the pair introduces an infinitive clause.

3.6 Miscellaneous Changes

This class gathers those changes that are related to more than one of the classes and subclasses in our typology, as they can take place in any of them.

3.6.1 Format

(1) Definition

This tag stands for changes in the format. In the following, some examples are set out:

1. **Digits/in letters**

a) 12/twelve

2. **Case changes**

b) Chapter 3/CHAPTER 3

3. **Format changes**

c) 03/08/1984 / Aug 3 1984

(2) Scope

Only the affected unit is tagged. The *local* attribute will be generally used.

3.6.2 Order

(1) Definition

This tag includes any type of change of order from the word level to the sentence level: (a), (b) and (c).

a) She used to only eat hot dishes/She used to eat only hot dishes

b) "I want a beer", he said/"I want a beer", said he

c) They said : "We believe that the time has come for legislation to make public places smoke-free./ "The time has come to make public places smoke-free," they wrote in a letter to the Times newspaper.

(2) Scope

Only the linguistic unit changing its position is tagged. In case of doubt of which is the unit changing its position, this order should be followed:

- Non argumental elements (a)
- Internal arguments (c)
- The subject (b)

In the case of enumerations (A, B, C/B, A, C), the first element in the list in the first member of the paraphrase pair is selected for annotation (A). In the case of copulatives (A is B/B is A), the first element in the first member is also selected (A).

The *global* attribute should be used if other changes are implied, changes in the punctuation included (c).

3.6.3 Addition/Deletion

(1) Definition

Deletion of lexical (a) and functional units (b).

- a) I would like pears, apples and strawberries/ I would like pears and strawberries
- b) However, I don't want to be here/I don't want to be here

☞ The limits between ADDITION/DELETION and light or specifier addition in SYNTHETIC/ANALYTIC tags are sometimes not clearly cut. In (c), the SYNTHETIC/ANALYTIC tag should be used, as *group* does not add any lexical content to the phrase, but emphasizes its plural nature. In (d), in contrast, an ADDITION/DELETION tag should be used, as *history* does add lexical content.

- c) Students/group of students
- d) Students/History students

☞ Regarding the structure based change class and ADDITION/DELETION, we only use the ADDITION/DELETION tag when the deletion is independent of any structural reorganization (e). In (f), the conjunction has also been deleted, but this deletion is the result of a COORDINATION change. The same criteria applies to ORDER: it has to take place as an independent phenomenon to be tagged as so.

- e) Actually, you shouldn't be here/You shouldn't be here.
- f) I like the beach and I want to go there/I like the beach. I want to go there.

☞ With the ADDITION/DELETION tag, we will only annotate a fragment in one of the members of the pair. If there is a deletion in each member of the pair, they will be tagged as independent deletion phenomena (1 and 2 in g), not under the same tag.

- g) Actually₁ you shouldn't be here/You shouldn't be here with me₂.

(2) Scope

Only the linguistic unit deleted is tagged. When a functional unit is deleted together with a lexical unit, this functional unit is included in the scope (h). The *local* attribute is normally used (a), but cases of *global* are also possible: in (h), a duplication/merging of coreferent units takes place.

- h) John said that he is a lawyer/John is a lawyer

3.7 Paraphrase Extremes

The following types stand for the extremes of the paraphrase continuum: identity on the one hand, and entailment and non-paraphrase on the other.

3.7.1 Identical

(1) Definition

We annotate as IDENTICAL those linguistic units that are exactly the same in wording (a).

- a) - The two argued that only a new board would have had the credibility to restore El Paso to health.
- The two believed that only a new board would have had the credibility to restore El Paso to health.

(2) Scope

See Section 2.3.2.

3.7.2 Entailment

(1) Definition

Fragments having an entailment relation (a).

☞ It should be noted that entailment relations are present in many paraphrase types (e.g. general/specific in SAME-POLARITY or ADDITION/DELETION). We will only use the ENTAILMENT tag when no other tag suits the phenomenon in question.

(2) Scope

Only the entailed linguistic units are tagged (a).

- a) Google was in talks to buy Youtube/Google bought Youtube

3.7.3 Non-paraphrase

(1) Definition

Non-paraphrase includes fragments which do not have the same meaning (a), as well as cases in which we need extralinguistic information in order to establish a link between the members of the paraphrase pair: cases of same illocutive value but different meaning (b), cases of subjectivity (c), cases of potential coreference (d), (e) and (f), etc.

- a) - The two had argued that you shouldn't go there
- He and Zilkha believed that this is unfair
- b) I want some fresh air/Could you open the window?
- c) The U.S.-led invasion of Iraq/The U.S.-led liberation of Iraq
- d) They got married last year/They got married in 2004
- e) I live here/I live in Barcelona
- f) They will come later/They will come this afternoon

!!! Paraphrase and coreference overlap considerably. Those cases that may corefer, but at the same time are paraphrases, should be annotated as paraphrases.

In cases (d), (e) and (f), the linguistic information is not enough to link the two members of the pair, we need to know which point in the time or in the space are we taking as reference. Thus, they are annotated as non-paraphrases.

Cases in (g), (h) and (i) can be linked only through linguistic information (a year in the past, a 'city' type of entity, a masculine singular entity,

respectively). Thus, they are annotated as paraphrases.

- g) They got married last year/They got married a year ago
- h) I live in Barcelona/I live in a city
- i) I love John/I love him

!!! Although sometimes a non-paraphrase fragment may actually affect the meaning of the full sentence, only the fragment in question will be tagged as NON-PARAPHRASE (j) and the rest of the sentence will be annotated independently of this fact.

- j) Mike and Lucy decided to leave/Mark decided to leave

☞ When two linguistic units having a different meaning are not aligned formally nor informatively, they should be tagged as two different ADDITION/DELETION cases (1 and 2 in k), not as NON-PARAPHRASES.

- k) Yesterday₁ Google failed/Google failed because of the crisis₂.

(2) Scope

Only the non-paraphrase linguistic unit is tagged.

4. The P4P, MSRP and WRPA-authorship Corpus

As stated in Section 1, these guidelines have been used to annotate three paraphrase corpora, giving rise to P4P, MSRP-A and WRPA-authorship-A. The guidelines have evolved during these three annotations processes, done successively in the previous order. In what follows, we present some disparities from the final guidelines shown in the P4P annotation. They refer to three sections in the present document:

2.1 Is this a paraphrase pair?

In spite of establishing paraphrase judgement in a fragmented way (considering as paraphrase pairs those containing, at least, one paraphrase unit), we carry this out considering the text fragment as single unit (it should be considered a paraphrase as a whole).

2.2 The tagset

The tags SPELLING&FORMAT and PUNCTUATION&FORMAT covered, in P4P annotation, the phenomena annotated within SPELLING, PUNCTUATION and FORMAT in the other annotation processes. Moreover, the tag ENTAILMENT is not present in the P4P.

2.3.2 Scope annotation criteria

The scope corresponding to identical and non-paraphrase was performed differently. They were only annotated as so when they appeared between strong punctuation marks.

References

- Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*. To appear in issue 39:4. DOI: 10.1162/COLI_a_00153.
- Barrón-Cedeño, Alberto, Marta Vila, and Paolo Rosso. 2012. Detección automática de plagio: De la copia exacta a la paráfrasis. In Elena Garayzábal, Miriam Jiménez, and Mercedes Reigosa (eds.). *Lingüística Forense: La Lingüística en el Ámbito Legal y Policial*, Euphonía Ediciones, pages 71-101.
- Dolan, William B. and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, pages 9-16.
- España-Bonet, Cristina, Marta Vila, Horacio Rodríguez, and M. Antònia Martí. 2009. CoCo, a web interface for corpora compilation. *Procesamiento del Lenguaje Natural* 43:367-368.
- Potthast, Martin, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Proceedings of COLING 2010: Posters*, pages 997-1005.
- Recasens, Marta and Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4):639-647.
- Vila, Marta, M. Antònia Martí, and Horacio Rodríguez. 2011. Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83-90.
- Vila, Marta, Horacio Rodríguez, and M. Antònia Martí. Relational paraphrase acquisition from Wikipedia. The WRPA method and corpus. Submitted.
- Vila, Marta, Horacio Rodríguez, and M. Antònia Martí. 2010. WRPA: A system for relational paraphrase acquisition from Wikipedia. *Procesamiento del Lenguaje Natural*, 45:11-19.

Annex: List of Consulted Typologies

This list contains all the typologies consulted. They have sometimes been inspiring for the creation of ours and some of the examples in these guidelines are extracted from them. Although all of them share the characteristic of setting paraphrase types, they come from different fields (discourse analysis, linguistics and computational linguistics), they are very different in nature (e.g., different levels of granularity or diverse presentation formats) and they pursue different objectives. Moreover, apart from typologies *sensu stricto*, this list contains other works related to paraphrasing and types of paraphrases in some way.

To read more about the state of the art on paraphrase typologies, refer to Barron-Cedeño et al. (2013) and Vila et al. (2011).

Apresjan, Jurij Derenikowicz. 1973. Synonymy and synonyms. In Ferenc Kiefer (ed.). *Trends in Soviet Theoretical Linguistics*, D. Reidel Publishing Company, pages 173-199.

Barzilay, Regina. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. PhD Thesis. Columbia University.

Barzilay, Regina and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the ACL 2001*, pages 50-57.

Barzilay, Regina, Kathleen McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the ACL 1999*, pages 550-557.

Bhagat, Rahul. 2009. *Learning Paraphrases from Text*. PhD thesis. University of Southern California.

Boonthum, Chutima. 2004. iSTART: Paraphrase recognition. In *Proceedings of the Fifth ACL Workshop on Student Research*, pages 55-60.

Cheung, Mei Ling Lisa. 2009. *Merging Corpus Linguistics and Collaborative Knowledge Construction*. PhD Thesis. University of Birmingham.

Chomsky, Noam. 1957. *Syntactic Structures*, Mouton & Co.

Culicover, Peter. 1968. Paraphrase generation and information retrieval from stored text. *Mechanical Translation and Computational Linguistics* 11(1,2):78-88.

Dolan, William B., Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the COLING 2004*, pages 350-356.

Dorr, Bonnie J., Rebecca Green, Lori Levin, Owen Rambow, David Farwell, Nizar Habash, Stephen Helmreich, Eduard Hovy, Keith J. Miller, Teruko Mitamura, Florence Reeder, and Advaith Siddharthan. 2004. Semantic annotation and lexico-syntactic paraphrase. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*.

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

Dutrey, Camille, Delphine Bernhard, Houda Bouamor, and Aurélien Max. 2011. Local modifications and paraphrases in Wikipedia's revision history. *Procesamiento del Lenguaje Natural* 46: 51-58.

Faigley, Lester and Stephen Witte. 1981. Analysing revision. *College Composition and Communication* 32(4): 400-414.

- Fernández, Patricia. 2006. Hacia una propuesta de clasificación de unidades de traducción. *RAEL: Revista Electrónica de Lingüística Aplicada* 5: 141-154.
- Fujita, Atsushi. 2005. *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. PhD thesis. Nara Institute of Science and Technology.
- Fujita, Atsushi. 2010. Typology of paraphrases and approaches to compute them. Invited talk at the workshop *Corpus-based Approaches to Paraphrasing and Nominalization* (CBA 2010). Slides available at <http://paraphrasing.org/~fujita/publications/fujita-CBA2010-slides.pdf>
- Fujita, Atsushi and Kentaro Inui. 2005. A class-oriented approach to building a paraphrase corpus. In *Proceedings of the 3rd International Workshop on Paraphrasing* (IWP 2003), pages 25-32.
- Gülich, Elisabeth. 2003. Conversational techniques used in transferring knowledge between medical experts and non-experts. *Discourse Studies* 5(2):235-263.
- Harris, Zellig. 1957. Co-occurrence and transformation in linguistic structure. *Language* 33(3): 283-340.
- Hirst, Graeme. 2003. Paraphrasing paraphrased. Keynote address for *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*. Slides available at <ftp://ftp.cs.toronto.edu/pub/gh/Hirst-IWP-talk.pdf>
- Honeck, Richard P. 1971. A study of paraphrases. *Journal of Verbal Learning and Verbal Behavior* 10: 367-381.
- Jasmina Milićević. 2007. *La Paraphrase*, Peter Lang.
- Kozłowski, Raymond, Kathleen F. McCoy, and Vijay K. Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexicogrammatical resources. In *Proceedings of IWP 2003*, pages 1-8.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
- Martin, Robert. 1976. *Inférence, Antonymie et Paraphrase*, Librairie C. Klincksieck.
- Max, Aurélien and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the LREC 2010*, pages 3143-3148.
- Mel'čuk, Igor A. 1992. Paraphrase et lexique: la théorie Sens-Texte et le Dictionnaire Explicatif et Combinatoire. In Igor A. Mel'čuk, Nadia Arbatchewsky-Jumarie, André Clas, Suzanne Mantha and Alain Polguère. *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches Lexico-sémantiques III*, Les Presses de l'Université de Montréal, pages 9-59.
- Ohtake, Kiyonori and Kazuhide Yamamoto. 2003. Applicability analysis of corpus-derived paraphrases toward example-based paraphrasing. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 380-391.
- Rinaldi, Fabio, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. 2003. Exploiting paraphrases in a question answering system. In *Proceedings of IWP 2003*, pages 25-32.
- Shimohata, Mitsuo. 2004. *Acquiring Paraphrases from Corpora and Its Application to Machine Translation*. PhD thesis. Nara Institute of Science and Technology.
- Turell, M. Teresa. 2011. La tasca del lingüista detectiu en casos de detecció de plagi i determinació d'autoria en textos escrits. *Llengua, Societat i Comunicació* 9:67-83.

Zangenfeind, Robert. 2009. Types of paraphrase rules in practice. German paraphrases of a Russian text. In Proceedings of the 4th International Conference on Meaning-Text Theory, pages 389-398.