Paraphrasing

Potential Applications for Plagiarism Detection

Marta Vila

CLiC – Universitat de Barcelona

WIQE 2010 Valencia

September 13, 2010
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
Paraphrasing stands for (approximate) **sameness of meaning** between **different wordings**.

- He announced that the company wouldn’t participate
  He stated that the company wouldn’t participate
- My father built the house
  The house was built by my father
Why paraphrasing?

Paraphrasing stands for (approximate) **sameness of meaning** between **different wordings**.

- He announced that the company wouldn’t participate
  He stated that the company wouldn’t participate
- My father built the house
  The house was built by my father

**Why is paraphrasing relevant for plagiarism?**
Why paraphrasing?

Paraphrasing stands for (approximate) **sameness of meaning** between **different wordings**.

- He announced that the company wouldn’t participate
  He stated that the company wouldn’t participate
- My father built the house
  The house was built by my father

Why is paraphrasing relevant for plagiarism?

Paraphrasing is a linguistic ability used in plagiarism.
To plagiarize involves, on many occasions, paraphrasing.

- Hamlet’s pretense of madness
  Hamlet adopts a pretense of madness

1 http://www.princeton.edu/pr/pub/integrity/08/plagiarism/
What is relevant in paraphrasing?

- **Paraphrasing typologies**
  - A paraphrasing typology is also a plagiarism phenomena typology
What is relevant in paraphrasing?

- Paraphrasing typologies
  - A paraphrasing typology is also a plagiarism phenomena typology
- Paraphrasing corpora
- NLP approaches to paraphrasing
  - A number of techniques and corpora can also be useful in plagiarism detection
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
Paraphrasing Complexity

He announced that the company wouldn’t participate
He stated that the company wouldn’t participate

My father built the house
The house was built by my father
Paraphrasing Complexity

- Mary buttered her toast with expensive butter
- Emma cried
- Patrick Ewing scored a personal season high of 41 points
- He announced that the company wouldn’t participate
- My father built the house
- I want some fresh air
- She used to only eat hot dishes
- Steven made an attempt to stop playing Hearts
- "They got married last year"
- "They got married in 2004"
- "He stated that the company wouldn’t participate"
- "The pilot was having breakfast"
- "The house was built by my father"
- "Could you open the window?"
- "Did you like French movies?"
- "Are you interested in French cinema?"
- "Steven attempted to stop playing Hearts"
Let’s put it in order!
State of the Art

Paraphrasing typologies:

- **Dras (1999):** syntactic paraphrases
  
  \[
  LV + NP + \text{inf-VP} \iff V + \text{inf-VP}
  \]

  - (2) a. Steven *made an attempt* to stop playing Hearts.
  - b. Steven *attempted* to stop playing Hearts.

- **Fuijta (2005):** lexical and structural paraphrases

  Paraphrasing of common nouns to their synonyms (Fujita and Inui, 2001; Yamamoto, 2002b; Okamoto et al., 2003)

  s. *kyuryo-ni kinenkan-ga kansei-shi-ta.*  
  hill-LOC a memorial hall-NOM to build up-PAST  
  A memorial hall was completed on the hill.

  t. *takadai-ni kinenkan-ga kansei-shi-ta.*  
  hill-LOC a memorial hall-NOM to build up-PAST
Paraphrasing typologies:

- Bhagat (2009): lexical and structural paraphrases

\[ \text{Semantic implication: Replacing a word or a phrase denoting an action, event etc. by a word or phrase denoting its possible future effect, in the appropriate context, results in a paraphrase of the original sentence of phrase. This may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. This often generates a quasi-paraphrase.} \]

\[ \text{Accompanying structural changes: Substitution, Addition/Deletion, Permutation.} \]

\[ \text{Example: Google is in talks to buy YouTube.} \leftrightarrow \text{Google bought YouTube.} \]
\[ \text{The Marines are fighting the terrorists.} \leftrightarrow \text{The Marines are eliminating the terrorists.} \]


\[ a \text{ republic} / a \text{ republican state:} \]

\[ C_0 \leftrightarrow \text{Gener}(C_0) \xrightarrow{\text{ATTR}} A_0/\text{Adv}_1(C_0) \]
State of the Art

Problems:

- They do not cover the paraphrasing phenomenon as a whole
- Tied to a specific linguistic theory and formalism
  - Meaning-Text Theory in Žolkovskij and Mel’čuk, and Milićević
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
A Paraphrasing Typology

- **Substitution**
  → He *announced* that the company wouldn’t participate
    He *stated* that the company wouldn’t participate

- **Deletion**
  → Mary opened the *bottle of* wine
    Mary opened the wine

- **Transformation**
  → My father built the house
    The house was built by my father

- **Splitting**
  → Patrick Ewing scored a personal season high of 41 points
    Patrick Ewing scored 41 points. It was a personal season high

- **Change of order**
  → She used to *only* eat hot dishes
    She used to eat *only* hot dishes

---

A Paraphrasing Typology

- **Substitution**
  - He *announced* that the company wouldn’t participate
  - He *stated* that the company wouldn’t participate

- **Deletion (Addition)**
  - Mary opened the *bottle of* wine
  - Mary opened the wine

- **Transformation**
  - My father built the house
  - The house was built by my father

- **Splitting (Combining)**
  - Patrick Ewing scored a personal season high of 41 points
  - Patrick Ewing scored 41 points. It was a personal season high

- **Change of order**
  - She used to *only* eat hot dishes
  - She used to eat *only* hot dishes
A Paraphrasing Typology

- **Substitution**
  - → He **announced** that the company wouldn’t participate
    - He **stated** that the company wouldn’t participate

- **Deletion**
  - → Mary opened the **bottle of** wine
    - Mary opened the wine

- **Transformation**
  - → My father built the house
    - The house was built by my father

- **Splitting**
  - → Patrick Ewing scored a personal season high of 41 points
    - Patrick Ewing scored 41 points. It was a personal season high

- **Change of order**
  - → She used to **only** eat hot dishes
    - She used to eat **only** hot dishes

**Usually combined!**
A Paraphrasing Typology

Substitution

→ He **announced** that the company wouldn’t participate
   He **stated** that the company wouldn’t participate

- Deletion
  → Mary opened the **bottle of** wine
     Mary opened the wine

- Transformation
  → My father built the house
     The house was built by my father

- Splitting
  → Patrick Ewing scored a personal season high of 41 points
     Patrick Ewing scored 41 points. It was a personal season high

- Change of order
  → She used to **only** eat hot dishes
     She used to eat **only** hot dishes
Substitution

- Synonymy
  - He announced that the company wouldn’t participate
  - He stated that the company wouldn’t participate
  - WordNet, dictionary of synonyms, thesaurus
Substitution

- **Synonymy**
  - He *announced* that the company wouldn’t participate
  - He *stated* that the company wouldn’t participate
    - ✔️ WordNet, dictionary of synonyms, thesaurus
  - *Since* it was sunny yesterday, the laundry dried well
  - *As* it was sunny yesterday, the laundry dried well
    - ✔️ Dictionary of synonyms, thesaurus
Substitution

- **Synonymy**
  - He announced that the company wouldn’t participate
  - He stated that the company wouldn’t participate
    - ✔ WordNet, dictionary of synonyms, thesaurus
  - Since it was sunny yesterday, the laundry dried well
  - As it was sunny yesterday, the laundry dried well
    - ✔ Dictionary of synonyms, thesaurus

- **Generalization**
  - I have been studying the reproduction of cats for ten years
  - I have been studying the reproduction of felines for ten years
    - ✔ WordNet, taxonomies
Substitution

- **Synonymy**
  - He *announced* that the company wouldn’t participate
  - He *stated* that the company wouldn’t participate
  - ✓WordNet, dictionary of synonyms, thesaurus
  - *Since* it was sunny yesterday, the laundry dried well
  - *As* it was sunny yesterday, the laundry dried well
  - ✓Dictionary of synonyms, thesaurus

- **Generalization**
  - I have been studying the reproduction of *cats* for ten years
  - I have been studying the reproduction of *felines* for ten years
  - ✓WordNet, taxonomies
  - Mary *has worn* soft contact lenses since college
  - Mary *has used* soft contact lenses since college
  - ? WordNet, taxonomies
Substitution

- **Antonymy**
  - The neighboring town is **poorer** in forest resources than our town
  - Our town is **richer** in forest resources than the neighboring town
  - WordNet, dictionary of antonyms, thesaurus
  - Matching
Antonymy

→ The neighboring town is **poorer** in forest resources than our town.
Our town is **richer** in forest resources than the neighboring town.

✔ WordNet, dictionary of antonyms, thesaurus
✔ Matching

→ I **lost interest** in the endeavor.
I **developed disinterest** in the endeavor.

✔ WordNet, dictionary of antonyms, thesaurus
Substitution

- **Actant-Action Substitution**
  - I dislike rash *drivers*
  - I dislike rash *driving*

✓ Lemmatizer
Substitution

- Actant-Action Substitution
  - I dislike rash **drivers**
    - I dislike rash **driving**
    - ✓ Lemmatizer
  - Mary is John’s **student**
    - John **teaches** Mary
    - ✓ Argument structure

Use of the GL lexical entries (Pustejovsky, 2005) for paraphrasing
Substitution

- Word-Definition Substitution
  → Temporary space for rubble and **scrap** wood
  Temporary space for rubble and wood **that became unnecessary**

✓ MRD, WordNet
A Paraphrasing Typology

**Substitution**
- He *announced* that the company wouldn’t participate
  He *stated* that the company wouldn’t participate

**Deletion**
- Mary opened the *bottle of* wine
  Mary opened the wine

**Transformation**
- My father built the house
  The house was built by my father

**Splitting**
- Patrick Ewing scored a personal season high of 41 points
  Patrick Ewing scored 41 points. It was a personal season high

**Change of order**
- She used to *only* eat hot dishes
  She used to eat *only* hot dishes
Deletion

- Non-propositional content deletion
  - Steven *made an attempt* to stop playing Hearts
  - Steven *attempted* to stop paying Hearts
  - ✓ List of light verbs
Deletion

- **Non-propositional content deletion**
  - → Steven *made an attempt* to stop playing Hearts
  - Steven *attempted* to stop paying Hearts
  - ✔ List of light verbs

- **Argument deletion**
  - → *My father* built the house
  - The house was built
  - ✔ Argument structure, diathesis alternations
Deletion

- **Non-propositional content deletion**
  - → Steven *made an attempt* to stop playing Hearts
  - Steven *attempted* to stop playing Hearts
  - ✓ List of light verbs

- **Argument deletion**
  - → *My father* built the house
  - The house was built
  - ✓ Argument structure, diathesis alternations

- **Adjunct deletion**
  - → John ran home *at noon*
  - John ran home
  - ? Argument structure, diathesis alternations
Deletion recoverable by coercion

→ John began **reading** a book
John began a book
✓ Qualia
A Paraphrasing Typology

- **Substitution**
  - He *announced* that the company wouldn’t participate
  - He *stated* that the company wouldn’t participate

- **Deletion**
  - Mary opened the *bottle of wine*
    - Mary opened the wine

- **Transformation**
  - My father built the house
    - The house was built by my father

- **Splitting**
  - Patrick Ewing scored a personal season high of 41 points
    - Patrick Ewing scored 41 points. It was a personal season high

- **Change of order**
  - She used to *only* eat hot dishes
    - She used to eat *only* hot dishes
Transformation

- **Syntactic transformation**
  - → My father built the house
    The house was built by my father
  - → The laundry sways in the breeze
    The breeze makes the laundry sway

✔ Diathesis alternations, Dras (1999)
A Paraphrasing Typology

- **Substitution**
  - He *announced* that the company wouldn’t participate
  - He *stated* that the company wouldn’t participate

- **Deletion**
  - Mary opened the *bottle of* wine
    - Mary opened the wine

- **Transformation**
  - My father built the house
    - The house was built by my father

**Splitting**

- Patrick Ewing scored a personal season high of 41 points
  - Patrick Ewing scored 41 points. It was a personal season high

- **Change of order**
  - She used to *only* eat hot dishes
    - She used to eat *only* hot dishes
Splitting

→ Patrick Ewing scored a personal season high of 41 points.
Patrick Ewing scored 41 points. It was a personal season high.
✓ Matching
A Paraphrasing Typology

- **Substitution**
  - He *announced* that the company wouldn’t participate
  - He *stated* that the company wouldn’t participate

- **Deletion**
  - Mary opened the *bottle of* wine
  - Mary opened the wine

- **Transformation**
  - My father built the house
  - The house was built by my father

- **Splitting**
  - Patrick Ewing scored a personal season high of 41 points
  - Patrick Ewing scored 41 points. It was a personal season high

**Change of order**

- She used to *only* eat hot dishes
  - She used to eat *only* hot dishes
Change of order

- She used to **only** eat hot dishes
  - She used to eat **only** hot dishes
- The student copied the critical diagrams **before returning the book**
  - Before returning the book, the student copied the critical diagrams
  - ☑️ Matching
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
Corpora used for paraphrasing treatment:

- **Parallel corpora (multiple parallel translations, source)**
  Barzilay and McKeown (2001)

- **Comparable corpora (newspaper articles, same event)**
  Barzilay, McKeown and Elhadad (1999)

- **Bilingual parallel corpora (source – translation)**
  Zhao et al. (2009)

- **Monolingual corpus**
  Bhagat and Ravichandran (2008)

- **Wikipedia**
  Vila, Rodríguez and Martí (2010)

- **Web**
  Dolan, Quirk, Brockett (2004)
Paraphrasing corpora:

- Microsoft Research Paraphrase Corpus
  Dolan, Brockett and Quirk (2005)

- Barzilay and Lee (2003)
  http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html

- CoCo interface
  España et al. (2009)
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
**COCO** The *Text-Mess* Corpora Compilation

Home | Goals | Statistics | Tasks

Current Task: Paraphrasing

Evaluate a pair  Generate a pair  Complete a pair  Generate templates  > Search <  > Modify <

---

**Evaluate an existing pair**

Choose Sentence 1 from **Microsoft corpus** and Sentence 2 from **Users generated (EN)**

Select sentences:  
- Randomly
- Sequentially
- Filtering

START

---

Sentence 1: Other, more traditional tests are also available.

Sentence 2: You can find more traditional tests as well.

Are they paraphrases?  
- **YES**
- **NO**

Why?  

SUBMIT

* FINISH *

Marta Vila  Paraphrasing  38/89
Authorship Generation

Fill in all or some of the gaps:
(Scroll down to see already available sentences)

Proposal: Frida Kahlo Las dos Fridas

All the available examples of this pair:
- Frida Kahlo es autora de Las dos Fridas
- La mexicana Frida Kahlo pintó Las dos Fridas
- Frida Kahlo, autora de Las dos Fridas
- La mexicana Frida Kahlo pintó Las dos Fridas en 1939
- La mexicana Frida Kahlo es la autora de Las dos Fridas
- Frida Kahlo pintó Las dos Fridas en 1939
- La pintora Frida Kahlo es la autora de Las dos Fridas
- Frida Kahlo es la autora de Las dos Fridas
- La mexicana Frida Kahlo es la autora de Las dos Fridas (1939)
- Frida Kahlo pintó Las dos Fridas (1939)
- Frida Kahlo es la autora, entre otras obras, de Las dos Fridas
- La mexicana Frida Kahlo, autora de Las dos Fridas, nació en 1907
- Frida Kahlo, autora de Las dos Fridas, nació en 1907
- Frida Kahlo pintó la famosa obra Las dos Fridas
- La artista Frida Kahlo pintó Las dos Fridas
- La autora mexicana Frida Kahlo es la autora de Las dos Fridas
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
State of the Art

Distributional Hypothesis (Harris, 1954)
Words that occur in the same contexts tend to have similar meanings.

The students solved the problem
The students found a solution to the problem

Bhagat and Ravichandran (2008)

Extended Distributional Hypothesis (Lin and Pantel, 2001)
Paths that link the same sets of words tend to have similar meanings.

The students ← N:subj:V ← solved → V:obj:N → the problem
The students ← N:subj:V ← find → V:obj:N → solution → N:to:N → the problem
State of the Art

Matching

Strings that share a high number of units tend to have similar meanings.

- Bag of words (NEs)
- N-grams

**DATE**: NUM1 are killed and around NUM2 injured when suicide bomber blows up his explosive-packed belt at X1 in X2.

palestinian suicide bomber blew himself up at X1 in X2 DATE, killing NUM1 and wounding NUM2 police said.

Barzilay and Lee (2003)
Edit distance (Levenshtein Distance)

Strings separated by a small edit distance tend to have similar meanings.

Dolan, Quirk and Brockett (2004)
Strings resulting from the translation of another string in another language are paraphrases.

Language 1

Language 2

String 1

String 2

String 3

= Paraphrases

Barzilay and McKeown (2001)

Zhao et al. (2009)
Rule application

To check if strings satisfy a set of (manually) created paraphrasing rules.

**Head omission**: group of students/students

**Ordering of sentence components**: Tuesday they met.../They met ... Tuesday

Barzilay, McKeown and Elhadad (1999)
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work

Marta Vila
Paraphrasing 47/89
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
WRPA
A System for Relational Paraphrase Acquisition from Wikipedia

Vila, Rodríguez and Martí (2010)
Relational Paraphrases

Paraphrases expressing a relation between two entities \(<Source, Target>\).

- Cervantes wrote El Quijote/El Quijote by Cervantes
- Joan Ponç was born in Barcelona/Barcelona, Joan Ponç’s home town
Hypothesis

Distributional Hypothesis (Harris, 1954)

Words that occur in the same contexts tend to have similar meanings.

The students solved the problem
The students found a solution to the problem

Our hypothesis:
same Sources and Targets ↔ paraphrase candidates

The designer of WORK was written by AUTHOR
WORK was created by AUTHOR
WORK by AUTHOR
WRPA—Relational Paraphrase Aquisition from Wikipedia

**WRPA**
- Complex relations
- Other sections
- Paraphrasing

**Others**
- Simple relations
- Infoboxes
- Information Extraction

Wu y Weld (2007, 2010)
Wu, Hoffmann y Weld (2008)
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
WRPA–Relational Paraphrase Acquisition from Wikipedia

Structured information
↓
{Source} ↓ {Target}
↓
NON-structured information
↓
{Source paraphrase Target}
WRPA–Relational Paraphrase Acquisition from Wikipedia

Structured information
↓
{Source}  {Target}
↓
NON-structured information
↓
{Source paraphrase Target}
↓
anchor anchor
↓
PATTERN
WRPA—Relational Paraphrase Acquisition from Wikipedia

Structured information

\{Source\} \downarrow \{Target\} \downarrow CFG

NON-structured information

\{Source \textbf{paraphrase} Target\} \downarrow anchor \downarrow anchor

PATTERN
WRPA–Relational Paraphrase Acquisition from Wikipedia

\{text\} \quad [X] \quad \{text\} \quad Y \quad \{text\} \quad [Z] \quad \{text\}
WRPA–Relational Paraphrase Acquisition from Wikipedia

\[
\begin{align*}
\{\text{text}\} & \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow \\
[X] & \downarrow \downarrow \downarrow \downarrow \downarrow \\
\{\text{text}\} & \downarrow \downarrow \downarrow \downarrow \downarrow \\
Y & \downarrow \downarrow \downarrow \downarrow \downarrow \\
\{\text{text}\} & \downarrow \downarrow \downarrow \downarrow \\
[Z] & \downarrow \downarrow \downarrow \\
\{\text{text}\} & \\
\text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \\
\text{Source} & \text{Target} & \text{Compl. info} & \\
\end{align*}
\]
WRPA–Relational Paraphrase Acquisition from Wikipedia

\[
\begin{array}{cccccc}
\{ \text{text} \} & [X] & \{ \text{text} \} & Y & \{ \text{text} \} & [Z] & \{ \text{text} \} \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \text{paraphr.} \\
\text{Source} & \text{Target} & \text{Compl. info} \\
\text{Person} & \text{Date of birth} & \text{Place of birth} & \text{Date of death} & -
\end{array}
\]
WRPA–Relational Paraphrase Acquisition from Wikipedia

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Compl. info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Date of birth</td>
<td>–</td>
</tr>
<tr>
<td>Place of birth</td>
<td>Date of death</td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td>Work</td>
<td>Date</td>
</tr>
</tbody>
</table>

Diagram:

```
{text} [X] {text} Y {text} [Z] {text}
```

```
paraphr. paraphr. paraphr. paraphr.
```

```
↓ ↓ ↓ ↓ ↓ ↓ ↓
```

```
Source Target Compl. info
```

```
paraphr. paraphr. paraphr. paraphr.
```

```
↓ ↓ ↓ ↓ ↓ ↓ ↓
```

```
Date of birth Date of death
```
WRPA–Relational Paraphrase Acquisition from Wikipedia

\[
\begin{array}{cccccc}
\{\text{text}\} & [X] & \{\text{text}\} & Y & \{\text{text}\} & [Z] & \{\text{text}\} \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \text{paraphr.} & \\
\text{Source} & \text{Target} & \text{Compl. info} & & & & \\
\text{Person} & \text{Date of birth} & \text{Place of birth} & \text{–} & \text{Date} & \text{(English)} & \\
\text{Date of death} & & & & & & \\
\text{Author} & \text{Work} & \text{Date} & \text{(Spanish)} & & & \\
\end{array}
\]
Authorship

- Complex relation (painters, sculptors, architects, writers, composers, singer-songwriters, directors, philosophers, inventors and scientists).
- Rich casuistic of manifestations.
  → Appropriateness of this relation for research in paraphrasing.
Wikipedia components and sections relevant to our work:

- Categories
- Infoboxes
- Work section
- Work page
Paraphrasing and Plagiarism Detection
Typologies
Corpora
NLP Approaches
The WRPA System

Presentation
Methodology
Evaluation and Results
Conclusions and Future Work

Paraphrasing 61/89

Marta Vila
Paraphrasing 61/89
Wikipedia components and sections relevant to our work:

- Categories
- Infoboxes
- Work section
- Work page
System scheme
Corpus Set Up
Corpus Set Up

Date of birth
Place of birth → Persons
Date of death
Authorship → Authors
Corpus Set Up

Date of birth
Place of birth → Persons
Date of death
Authorship → Authors

*A category of authors does not exist in the Spanish Wikipedia → done manually*
Anchor Point Extraction
Anchor Point Extraction

**X Extraction**
- Titles
- Redirection pages

**Y and Z Extraction**

**Person:**
- 'Date of birth', 'place of birth' and 'date of death' attributes in infoboxes.

**Authorship:**
- 'Work' attribute in infoboxes.
- Work section
- Work page
Anchor Point Extraction

**X Extraction**
- Titles
- Redirection pages

**Y and Z Extraction**

**Person:**
- 'Date of birth', 'place of birth' and 'date of death' attributes in infoboxes.

**Authorship:**
- 'Work' attribute in infoboxes.
- Work section
- Work page

*Attributes are expressed in different ways → done manually/KBP track (TAC 2010)*
Anchor Point Extraction

X Extraction
- Titles
- Redirection pages

Y and Z Extraction

Person:
- 'Date of birth', 'place of birth' and 'date of death' attribute in infoboxes.

Authorship:
- 'Work' attribute in infoboxes.
- Work sections
- Work pages

- Attributes are univalued.
- The English Wikipedia is very extensive.
- There is a relatively high number of infoboxes in person pages (34%).
- The corresponding attributes generally appear in the infoboxes.

- Attributes are multivalued.
- The Spanish Wikipedia is smaller.
- Most author pages lack an infobox.
- Most infoboxes lack a work attribute.
- Infoboxes only contain the most important works (2 approx.).
Anchor Point Extraction

X Extraction

- Titles
- Redirection pages

Y and Z Extraction

**Person:**
- 'Date of birth', 'place of birth' and 'date of death' attribute in infoboxes.

**Authorship:**
- 'Work' attribute in infoboxes.
- Work sections
- Work pages

- Attributes are univalued.
- The English Wikipedia is very extensive.
- There is a relatively high number of infoboxes in person pages (34 %).
- The corresponding attributes generally appear in the infoboxes.

- Attributes are multivalued.
- The Spanish Wikipedia is smaller.
- Most author pages lack an infobox.
- Most infoboxes lack a work attribute.
- Infoboxes only contain the most important works (2 approx.).
Anchor Point Extraction

Paul Auster:

- Infobox: 0
- Work section: 33 (25 correct)
- Work page: 18 (all correct)
Anchor Point Extraction

<table>
<thead>
<tr>
<th>X</th>
<th>Y(Z)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Authorship</strong></td>
</tr>
<tr>
<td>Canaletto</td>
<td>La Riva degli Schiavoni (1730-31)</td>
</tr>
<tr>
<td>Edgar Allan Poe</td>
<td>[[Eureka]]</td>
</tr>
<tr>
<td>Luis Eduardo Aute</td>
<td>Templo de carne (1986)</td>
</tr>
<tr>
<td></td>
<td><strong>Date of birth</strong></td>
</tr>
<tr>
<td>David Kaye</td>
<td>[[October 14]], [[1964]]</td>
</tr>
<tr>
<td>Sara Rue</td>
<td>1979</td>
</tr>
<tr>
<td>Joan of Arc</td>
<td>c. [[1412]]</td>
</tr>
<tr>
<td></td>
<td><strong>Place of birth</strong></td>
</tr>
<tr>
<td>Giovanni Branca</td>
<td>[[San Angelo]] in [[Lizzola]], [[Pesaro]]</td>
</tr>
<tr>
<td>Grigore Preoteasa</td>
<td>[[Bucharest]], [[Romania]]</td>
</tr>
<tr>
<td>Tomas Plekanec</td>
<td>[[Kladno]], [[Czech Republic CZE]]</td>
</tr>
</tbody>
</table>
Pattern Extraction

Diagram:

- WP Articles
- Source Articles
- Infobox
  - Target Section
  - Target Page
- X Extraction
- Y Extraction
- Z Extraction
- X Expansion (PNG)
- {text} [X] [text] [Y] [text] [Z] [text]
- X sigue escribiendo la novela Y
  X was born in Y
- ? [I w?!]
- sigue escribiendo la novela, siguió escribiendo novelas, escribió la novela, escribía novelas...
Pattern Extraction

\[
\begin{align*}
\{\text{text}\} & \ [X] \ \{\text{text}\} \ Y \ \{\text{text}\} \ [Z] \ \{\text{text}\} \ .
\end{align*}
\]
En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYY " en (ZZZZZZZ).

En los años siguientes XXXXXX escribió prolíficamente: "YYYYYY" apareció en **z4**, "****y5***" en **z5**, "****y6***" en **z6**.

En los años siguientes XXXXXX escribió prolíficamente: "****y4***" apareció en **z4**, "YYYYYY" en **z5**, "****y6***" en **z6**.

En los años siguientes XXXXXX escribió prolíficamente: "****y4***" apareció en **z4**, "****y5***" en **z5**, "YYYYYY" en **z6**.

La última obra de XXXXXX es un poema: "YYYYYY" (**z35**).

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYY", el drama.

En Mijáilovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "****y15***", el drama.

:*Ágora - "YYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" ("Trompetas para Isabel", **z3** y "****y4***" ("Las XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "****y3***" ("Trompetas para Isabel", **z3**) y "YYYYYY" ("Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".

Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZZ: "No soy socióloga, ni psicóloga ni he querido

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfrute.

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "****y5***" (**z2**), dos pastorales, que reflejan el disfrute.
En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYY " en (ZZZZZZZ).

En los años siguientes XXXXXX escribió prolíficamente: "YYYYYY" apareció en **z4**, "**y5**" en **z5**, "**y6**" en **z6**.

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "YYYYYY" en **z5**, "**y6**" en **z6**.

En los años siguientes XXXXXX escribió prolíficamente: "**y4**" apareció en **z4**, "**y5**" en **z5**, "YYYYYY" en **z6**.

La última obra de XXXXXX es un poema: "YYYYYY" (**z35**).

En Mijálovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYY", el dra.

En Mijálovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "**y15**", el dra.

:*Ágora - "YYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" («Trompetas para Isabel», **z3**) y "**y4**" («Las

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "**y3**" («Trompetas para Isabel», **z3**) y "YYYYYY" («Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".

Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZZZ: "No soy socióloga, ni psicóloga ni he querid

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfr

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "**y5**" (**z2**), dos pastorales, que reflejan el disfr.
En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYYYY" en (ZZZZZZZZ).

En los años siguientes XXXXXX escribió prolificamente: "YYYYY" apareció en **z4**, "**y5**" en **z5**, "**y6**

En los años siguientes XXXXXX escribió prolificamente: "***y4***" apareció en **z4**, "YYYYYY" en **z5**, "**y6**

En los años siguientes XXXXXX escribió prolificamente: "**y4**" apareció en **z4**, "**y5**" en **z5**, "YYYY" en **z6**

La última obra de XXXXXX es un poema: "YYYYY" (**z35**).

En Mijálovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYY", el drama

En Mijálovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "**y15**", el drama

:*Ágora - "YYYYY"
:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYY" («Trompetas para Isabel», **z3**), y "**y4**" («Las

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "**y3**" («Trompetas para Isabel», **z3**) y "YYYYY" («Las

En enero de 1845, publicó un poema que le haría célebre: "YYYYY".

Algunas de estas leyes inspirarían en su momento una de sus obras fundamentales: "YYYYY".

Sobre "YYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZ: "No soy socióloga, ni psicóloga ni he querido.

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYY" (**z2**), dos pastorales, que reflejan el disfru

Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "**y5**" (**z2**), dos pastorales, que reflejan el disfru
En contra de la guerra, con ocasión de su adhesión al "Consejo Mundial de la Paz" pintó el famoso "YYYYYY " en (ZZZZZZZ).
En los años siguientes XXXXXX escribió prolificamente: "YYYYYY" apareció en **z4**, "****y5***" en **z5**, "****y6***" en **z6**.
En los años siguientes XXXXXX escribió prolificamente: "****y4***" apareció en **z4**, "YYYYYY" en **z5**, "****y6***" en **z6**.
En los años siguientes XXXXXX escribió prolificamente: "****y4***" apareció en **z4**, "****y5***" en **z5**, "YYYYYY" en **z6**.
La última obra de XXXXXX es un poema: "YYYYYY" (**z35**).
En Mijálovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "YYYYYY", el dra...
En Mijálovskoye, tras la reprimenda paterna y acogido por su amada aya, XXXXXX compuso seis capítulos de "****y15***", el dra...

*:Ágora - "YYYYYY"

:*YYYYYY de l"Assut de l"Or

XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "YYYYYY" («Trompetas para Isabel», **z3**) y "****y4***" («Las
XXXXXX escribió dos libros sobre la reina Isabel I de Inglaterra, "****y3***" («Trompetas para Isabel», **z3**) y "YYYYYY" («Las
En enero de 1845, publicó un poema que le haría célebre: "YYYYYY".
Algunas de estas leyendas inspirarían en su momento una de sus obras fundamentales: "YYYYYY".

Sobre "YYYYYY", XXXXXX dijo a la agencia española EFE el 25 de junio de ZZZZZZZ: "No soy socióloga, ni psicóloga ni he quier...
Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "YYYYYY" (**z2**), dos pastorales, que reflejan el disfr...
Las primeras obras poéticas compuestas por XXXXXX son "L"Allegro" e "****y5***" (**z2**), dos pastorales, que reflejan el disfr...
Pattern Extraction

. \{\text\} [X] \{\text\} Y \{\text\} [Z] \{\text\} .
\downarrow
. X \{\text\} Y .
Pattern Extraction

Authorship
1. X sigue escribiendo la novela Y
2. X comenzó a grabar su álbum debut, “Y”
3. X dirigió ‘Educando a Rita” (YEAR) y “Y”

Date of birth
4. X known as PERSON was born in Y
5. X was born in Y
6. X was born in PLACE on Y

Date of death
7. X DATE-Y
8. X DATE to Y
9. X PERSON DATE - Y

Place of death
10. X born DATE in Y
11. X DATE in Y
12. X DATE Y
Pattern Generalization

- Corpus Setup
- Source Articles
- WP Articles
- Infobox
  - Target Section
  - Target Page
- X Extraction
- Y Extraction
- Z Extraction
- X Expansion
  - PNG
- Pattern Extraction
  - (text) [X] (text) Y {text} [Z] (text)
  - X sigue escribiendo la novela Y
  - X was born in Y
- Generalization
  - sigue escribiendo la novela, siguió escribiendo novelas, escribió la novela, escribía novelas...
Pattern Generalization

To gather the variants of a generic pattern (only for authorship).

- To lemmatize and PoS tag the patterns (Freeling).
- To represent each pattern as a sequence of \(<word, \text{lemma, PoS}>\) tuples.

<table>
<thead>
<tr>
<th>Word</th>
<th>seguiente</th>
<th>escribiendo</th>
<th>la</th>
<th>novela</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemma</td>
<td>seguir</td>
<td>escribir</td>
<td>el</td>
<td>novela</td>
</tr>
<tr>
<td>PoS</td>
<td>v</td>
<td>v</td>
<td>da</td>
<td>nc</td>
</tr>
</tbody>
</table>

- To use an A* approach until achieving \(n\) matches with other patterns.
Pattern Generalization

- Each state is represented as a sequence of tuples consisting of a token (word, lemma or PoS) and a condition (obligatory, skippable or omitted):

<table>
<thead>
<tr>
<th>Initial state</th>
<th><a href="">sigue:w</a> <a href="">escribiendo:w</a> <a href="">la:w</a> <a href="">novela:w</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching pattern</td>
<td>sigue escribiendo la novela</td>
</tr>
<tr>
<td>Matching patterns</td>
<td>1) sigue escribiendo la novela</td>
</tr>
<tr>
<td></td>
<td>2) siguió escribiendo novelas</td>
</tr>
<tr>
<td></td>
<td>3) escribió la novela</td>
</tr>
<tr>
<td></td>
<td>4) escribía novelas</td>
</tr>
</tbody>
</table>
Evaluation

**Precision**

To apply each pattern to its original page and verify whether the output is a Y (or a variant) of the corresponding infobox (correct Ys).

\[
\frac{\text{Num. of correct Ys}}{\text{Num. of extracted Ys}}
\]

**Recall**

**Person:** To apply the pattern to its original page and verify whether the output is a Y (it can be applied to that page).

Conservative assumption: all the pages contain a Y.

\[
\frac{\text{Num. of extracted Ys}}{\text{Num. of pages of the corpus}}
\]

**Baseline**

**Person:** The most frequent pattern.
## Results

<table>
<thead>
<tr>
<th>Person</th>
<th>Date of birth</th>
<th>X born Y</th>
<th>0.95</th>
<th>0.57</th>
<th>0.71</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X Y</td>
<td>0.80</td>
<td>0.12</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 8 patterns</td>
<td>0.92</td>
<td>0.75</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>0.95</td>
<td>0.57</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Date of death</td>
<td>X DATE-Y</td>
<td>0.95</td>
<td>0.21</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>X PERSON DATE-Y</td>
<td>0.96</td>
<td>0.10</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 3 patterns</td>
<td>0.95</td>
<td>0.42</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>0.95</td>
<td>0.21</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Place of birth</td>
<td>X born DATE in Y</td>
<td>0.98</td>
<td>0.13</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>X DATE in Y</td>
<td>0.94</td>
<td>0.10</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 3 patterns</td>
<td>0.92</td>
<td>0.26</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>0.98</td>
<td>0.13</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>Authorship</td>
<td>&lt;pintó:w&gt;<a href="">su:w?</a><a href="">cuadro:w?</a></td>
<td>0.46</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;pintó:w&gt;<a href="">su:w?</a><a href="">primera:l?</a><a href="">obra:w?</a></td>
<td>0.49</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date of birth</td>
<td>Date of death</td>
<td>Place of birth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------</td>
<td>---------------------------</td>
<td>-------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Person</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date of birth</td>
<td>X born Y</td>
<td></td>
<td>X born DATE in Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 8 patterns</td>
<td></td>
<td>Top 3 patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td></td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td></td>
<td>Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td></td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td></td>
<td>0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td></td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date of death</td>
<td>X DATE-Y</td>
<td></td>
<td>X DATE in Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X PERSON DATE-Y</td>
<td></td>
<td>Top 3 patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.96</td>
<td></td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 3 patterns</td>
<td></td>
<td>Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td></td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td></td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.34</td>
<td></td>
<td>0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Place of birth</td>
<td>X born DATE in Y</td>
<td></td>
<td>X DATE in Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.98</td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X DATE in Y</td>
<td></td>
<td>Top 3 patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.94</td>
<td></td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 3 patterns</td>
<td></td>
<td>Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td></td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td></td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authorship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;pintó:w&gt;<a href="">su:w?</a><a href="">cuadro:w?</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;pintó:w&gt;<a href="">su:w?</a><a href="">primera:l?</a><a href="">obra:w?</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Person</th>
<th>Date of birth</th>
<th>Date of death</th>
<th>Place of birth</th>
<th>Authorship</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X born Y</td>
<td>X DATE-Y</td>
<td>X born DATE in Y</td>
<td><code>&lt;pintó:w&gt;&lt;su:w?&gt;&lt;cuadro:w?&gt;</code></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>X Y</td>
<td>X PERSON DATE-Y</td>
<td>X DATE in Y</td>
<td><code>&lt;pintó:w&gt;&lt;su:w?&gt;&lt;primera:l?&gt;&lt;obra:w?&gt;</code></td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.96</td>
<td>0.94</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Top 8 patterns</td>
<td>Top 3 patterns</td>
<td>Top 3 patterns</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.95</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>Baseline</td>
<td>Baseline</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>F1</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.71</td>
<td>0.21</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.83</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.57</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.71</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.34</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.41</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.23</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>

**Baseline**

<table>
<thead>
<tr>
<th>Person</th>
<th>Date of birth</th>
<th>Date of death</th>
<th>Place of birth</th>
<th>Authorship</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X born Y</td>
<td>X DATE-Y</td>
<td>X born DATE in Y</td>
<td><code>&lt;pintó:w&gt;&lt;su:w?&gt;&lt;cuadro:w?&gt;</code></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>X Y</td>
<td>X PERSON DATE-Y</td>
<td>X DATE in Y</td>
<td><code>&lt;pintó:w&gt;&lt;su:w?&gt;&lt;primera:l?&gt;&lt;obra:w?&gt;</code></td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.96</td>
<td>0.94</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Top 8 patterns</td>
<td>Top 3 patterns</td>
<td>Top 3 patterns</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.95</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>Baseline</td>
<td>Baseline</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>F1</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.71</td>
<td>0.21</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.83</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.57</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.71</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.34</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.41</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.23</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>

**Baseline**

<table>
<thead>
<tr>
<th>Person</th>
<th>Date of birth</th>
<th>Date of death</th>
<th>Place of birth</th>
<th>Authorship</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X born Y</td>
<td>X DATE-Y</td>
<td>X born DATE in Y</td>
<td><code>&lt;pintó:w&gt;&lt;su:w?&gt;&lt;cuadro:w?&gt;</code></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>X Y</td>
<td>X PERSON DATE-Y</td>
<td>X DATE in Y</td>
<td><code>&lt;pintó:w&gt;&lt;su:w?&gt;&lt;primera:l?&gt;&lt;obra:w?&gt;</code></td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.96</td>
<td>0.94</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Top 8 patterns</td>
<td>Top 3 patterns</td>
<td>Top 3 patterns</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.95</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>Baseline</td>
<td>Baseline</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>F1</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.71</td>
<td>0.21</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.83</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.57</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.71</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.34</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.41</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.23</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>

**Baseline**

- **Precision**: The percentage of correctly identified patterns.
- **Recall**: The percentage of correctly identified patterns.
- **F1**: The harmonic mean of Precision and Recall.
Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.
Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
- ...can deal with **complex relationships**.
Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
- ...can deal with **complex relationships**.
- ...is **language independent** and also independent of the **relation**.
Conclusions

Finding good and numerous anchor points is essential in systems based on the Distributional Hypothesis.

WRPA...

- ...guarantees the **quality** of these anchor points as they are directly extracted from structured and semantically labelled data.
- ...guarantees a large **quantity** of anchor points as they are not only extracted from infoboxes but also from section and work pages.
- ...can deal with **complex relationships**.
- ...is **language independent** and also independent of the **relation**.
- ...only relies on **Wikipedia** and **shallow linguistic processing**.
Future Work

- Recall and baseline for authorship.
Future Work

- Recall and baseline for authorship.
- Manual validation of paraphrase candidates: CoCo (España et al., 2009)
  - Do they express an authorship relation?
  - Establishment of paraphrasing relations between them in the framework of a typology.
Outline

1. Paraphrasing and Plagiarism Detection
2. Typologies
   - Paraphrasing Complexity
   - State of the Art
   - A Paraphrasing Typology
3. Corpora
   - State of the Art
   - CoCo Interface
4. NLP Approaches
   - State of the Art
5. The WRPA System
   - Presentation
   - Methodology
   - Evaluation and Results
   - Conclusions and Future Work
Thank you!

Questions?

marta.vila@ub.edu