XLike
FP7-ICT-2011.4.2
Contract no.: 288342
www.xlike.org

XLike

Deliverable D4.2.1

Early Semantic Triple Graphs Merging Prototype

Editor:	Xavier Carreras, Technical University of Catalunya
Author(s):	Lluís Padró, Technical University of Catalunya; Blaž Fortuna, Jožef Stefan Institute; Janez Starc, Jožef Stefan Institute
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M15
Actual Delivery Date:	5.4.2013
Suggested Readers:	All partners using the XLike Toolkit
Version:	1.0
Keywords:	Relation extraction, pattern learning, semantic indexing, grammatical relations, semantic roles, ontological relations

Disclaimer

This document contains material, which is the copyright of certain XLike consortium parties, and may not be reproduced or copied without permission.

All XLike consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the XLike consortium as a whole, nor a certain party of the XLike consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	XLike – Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP4
Document Title:	D4.2.1 - Early Semantic Triple Graphs Merging Prototype
Editor (Name, Affiliation)	Xavier Carreras, Technical University of Catalunya
Work package Leader (Name, affiliation)	Marko Grobelnik, Jožef Stefan Institute
Estimation of PM spent on the deliverable:	11

Executive Summary

This deliverable presents early prototypes to obtain cross-lingual representations of relational content of documents. These representations can be used to compare documents cross-lingually, or to aggregate the information of a collection of documents in a merged semantic graph that stores relations weighted by frequency.

We have followed two different approaches. The first takes a document analysed with linguistic tools of WP2, considers the grammatical relations appearing in the document, and generalizes such relations to obtain a cross-lingual representation of the relational content of the document. We have used WordNet as an interlingua representation lexical items, taking advantage that in WordNet words are already linked cross-lingually. To some extent, this is a simple approach that will work moderately well. One advantage is that WordNet has been built manually; hence the cross-lingual mappings are expected to be high quality. Another advantage is that there exist a number of similarity measures based on WordNet: hence, if two relations are not exactly identical, one can develop principled methods to compute a notion of similarity between the relations. Our next step is to develop alternative approaches to WordNet, where the cross-lingual representation of lexical items and named entities is given by unsupervised methods ---this is precisely the topic of WP3 in XLike [4].

We also present a second approach to semantic graph construction based on pattern rules that link textual patterns into logical patterns. In this case, the logical patterns are the building blocks of the semantic graphs representing documents. A tool has been developed to explore a large collection of documents (analysed linguistically) and manually construct patterns to map relational content into a cross-lingual semantic representation. Logical patterns based on the CyC ontology have been explored.

Finally, we present a set of experiments using a large collection of documents in English, Spanish, and Catalan, gathered during January 2013. Using this data, we illustrate the kind of representations that we obtain using the two different approaches.

Table of Contents

List of Figures

List of Tables

1 Introduction

The goal of task **T4.2 Semantic Graphs Construction** is to develop techniques and tools to obtain cross-lingual representations of relational content of documents. These representations are in the form of semantic graphs, that is, graphs where nodes represent entities mentioned in a document, and edges represent relations between entities also mentioned in the document. Hence, a semantic graph of a document is a relational summary of its content. In addition, the representation of relations in a semantic graph should be cross-lingual: equivalent relations expressed in different languages should map to the same form.

We explore three kinds of techniques in order to construct semantic graphs:

a) Merging predicate-argument relations extracted in WP2 into semantic graphs. The input consists of documents annotated with a set of grammatical relations, extracted in T2.2 and T2.4, and further linguistic analysis is used as the main guide for merging.

b) Linking annotations into coherent semantic graphs based on pattern analysis and ontological constraints. The input is a sequence of annotations, extracted by tools from WP2. The graphs will be constructed by combining statistical co-occurrence models (similar to Latent Semantic Analysis) and ontological constraints of annotations and their super-classes.

c) A combined approach.

This deliverable describes early approaches for (a) and (b). We first present each of the techniques, and then present experiments using real data where we illustrate the two approaches.

2 Merging Predicate-Argument Relations

In this approach we transform a document into a semantic graph based on the linguistic annotations obtained with the linguistic analysis tools in WP2 [2][3]. There are two main questions to solve to build semantic graphs:

1. How to build a graph of relations representing textual content?
2. How to make the graph cross lingual?

Next we describe how we solve these questions.

2.1 Extraction of Grammatical Relations

Following Task 2.2 [3], we employ grammatical relations in the form of triples to build a graph. There are two types of such linguistic triple relations:

1. Syntactic: subject-verb-object triples
2. Semantic: agent-predicate-theme triples

The following figure gives an example of syntactic dependencies (above the sentence) and semantic dependencies (below the sentence):

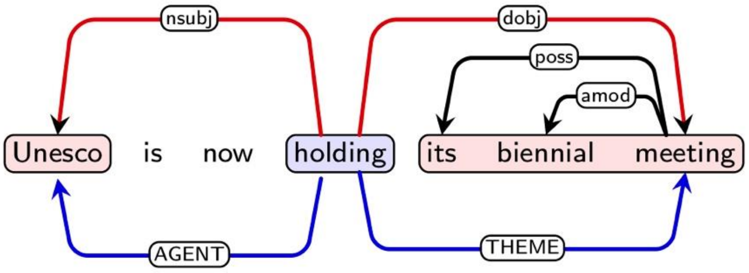


Figure 1. Syntactic dependencies (red) and semantic roles (blue).

Given a document, we will consider either syntactic or semantic triples, and build a semantic graph based on these. Essentially, each node in the graph corresponds to an entity mentioned in the document. Then, directed edges in the graph represent triple relations: the source node is the agent, the target node is the theme, and the label of the edge is the predicate.

For example, for the sentence in Figure 1 we would obtain a graph with two nodes, one for “Unesco” and another for “meeting”. A directed edge labelled “hold” would connect “Unesco” to “meeting”.

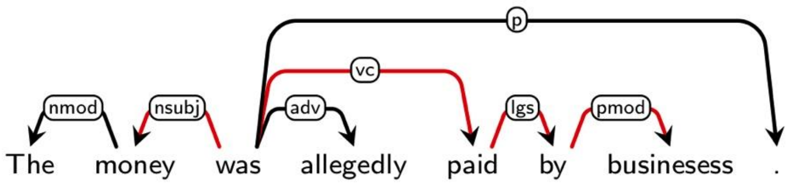
While in this example syntactic and semantic dependencies would result in the same graph, in general semantic dependencies should result in richer and more abstract relational representations. On the other hand, obtaining semantic dependencies requires running syntactic parsing first, and then run a semantic parser that is not always available for all languages. Sometimes, it can be more accurate and robust to rely on syntactic analysis only. In any case, it should be clear that our approach to construct semantic graphs does not really depend on the nature of grammatical relations obtained from the linguistic tools in WP2.

2.2 Cross-lingual Representations

Generating cross-lingual representations of linguistic triples is essential in order to map content in different languages to a language-independent representation.

Our first approach, described here, is based in WordNet [5]. That is, we disambiguate nouns and verbs into WordNet concepts (synsets) using state-of-the-art disambiguation techniques. With this, nodes in the semantic graph are WordNet synsets, rather than the words themselves. Since WordNet synsets are cross-lingually linked, we obtain cross-lingual triple relations. Thus, WordNet naturally provides a bridge for lexical items from many languages to a common syntactic space.

The following figure gives an example of a sentence syntactically analysed. The red syntactic dependencies form a grammatical relation “business-pay-money”. The word “money” is disambiguated into synset “00582388-n” (described in WordNet as the most common medium of exchange), which is linked to the Spanish word “dinero” or the Catalan word “diners”. This example also illustrates the importance of having an appropriate syntactic representation: since the sentence is in passive form, we can adequately reorder the arguments such that they indicate who-does-what.



```
business/00582388-n pay/02251743-v money/13384557-n
```

Figure 2. An example of a syntactic relation and its representation with WordNet synsets.

We have used WordNet disambiguation tools based on the UKB state-of-the-art technique [6], which is available in FreeLing. At this point it is integrated for English, Spanish and Catalan. There exist WordNet versions for German [7] and Slovene [8] that we can integrate. Unfortunately, there does not exist a WordNet version of Chinese that is interconnected with the rest of languages. This illustrates a limitation of this approach: using it requires the availability of WordNet resources, which are expensive to produce.

A second limitation is that the accuracy of semantic disambiguation systems into WordNet concepts is only moderately accurate (performances of state-of-the-art systems are between 60% and 70% of accuracy).

3 Pattern Analysis and Ontological Constraints

3.1 Introduction

The goal of subtask (b) of task T4.2 is linking annotations into coherent semantic graphs based on pattern analysis and ontological constraints. The main effort has been in developing a system, which provides assistance for building pattern rules and applies the rules on a dataset annotated with linguistic information in order to produce semantic graphs of the documents.

3.1.1 Pattern rules

A pattern rule consists of a textual pattern, a logical pattern, and argument mappings. The textual patterns match with many different fractions of text. The arguments of patterns are non-basic tokens, which are described in the next section. These arguments (usually named entities) are connected with arguments of the logical pattern via argument mappings. When the rule is applied the arguments of the logical pattern are filled with arguments from the textual pattern, and the logical pattern becomes a relation that can become a potential part of the semantic graph. The static parts of the logical patterns, i.e., predicates, constants, should already be defined in the semantic graph to integrate well with it.

3.1.2 Dataset

Our system operates on the textual data annotated with linguistic tools from WP2. The structure is presented on Figure 3. There is one corpus for each language. The documents are split into sentences. Sentences are split into tokens. There are several layers for each tokenized sentence: lexical tokens, lemmas, part-of-speech tags, named entities. During the processing additional layer is generated – generalized tokens. In the process of generalization each generalized token is assigned value from another layer (Figure 4). One possibility of generalization is presented in Figure 4. If a token is a named-entity, then the generalized token will be the type of *[Named entity]*. If part-of-speech tag of the token is *CD*, then the generalized token will be assigned *[Number]*. Otherwise, the generalized token is the same as lexical token. These tokens are called *basic tokens* for the rest of this section. There are other possible generalizations. For instance, if no other rules apply the generalization token is the lemma of the token.

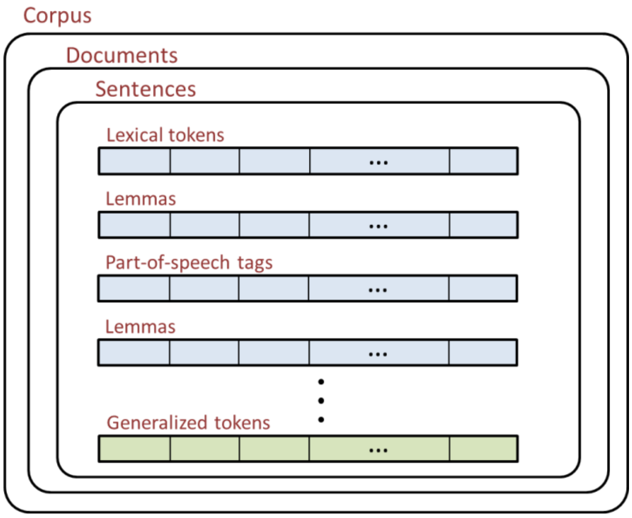
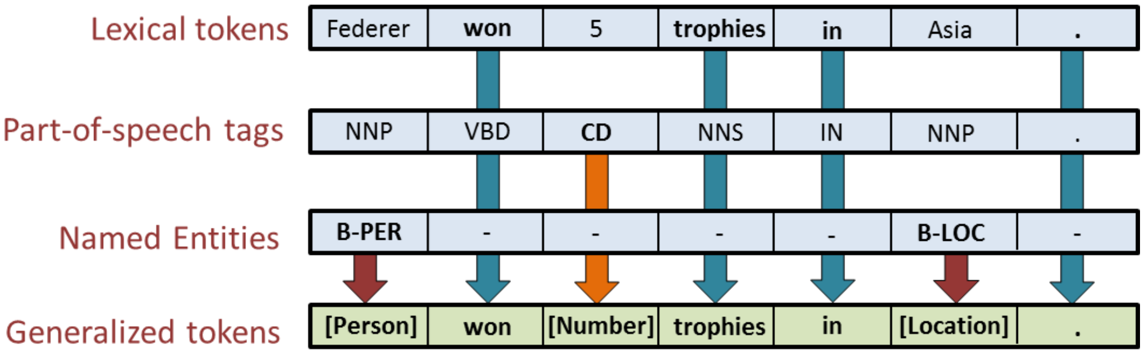


Figure 3. Backend data structure

After generalization, pattern frequencies are computed. All possible n-grams (substrings) of the generalized sentence are considered as patterns. Long patterns (more than 8 tokens) are not counted, because counting them would require a lot of computational resources. Also, these patterns rarely express a simple fact.

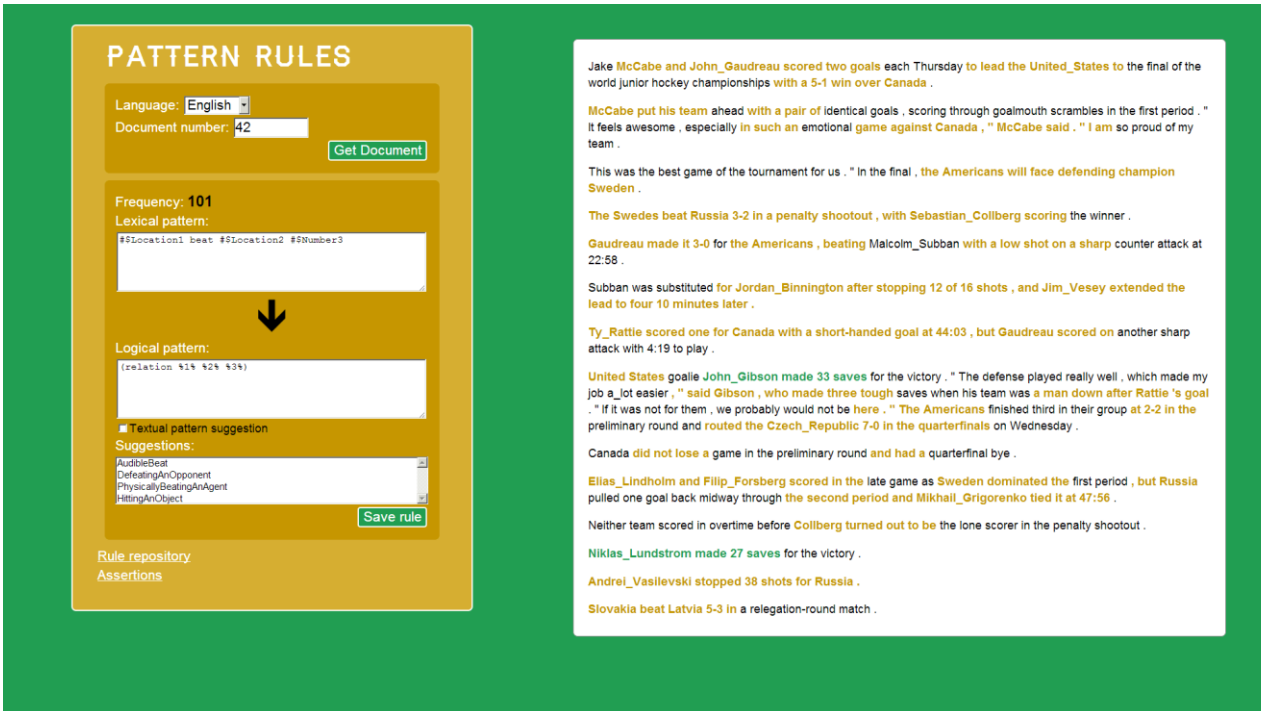


Lexical tokens	Federer	won	5	trophies	in	Asia	.
Part-of-speech tags	NNP	VBD	CD	NNS	IN	NNP	.
Named Entities	B-PER	-	-	-	-	B-LOC	-
Generalized tokens	[Person]	won	[Number]	trophies	in	[Location]	.

Figure 4. The process of generalization.

3.2 Description of the system

In this section, we will demonstrate how our system works together with a short user manual of the graphical user interface GUI. The GUI of our system is presented on Figure 5. It consists of several panels, which are presented in the following sections.



Language: English
Document number: 42

Frequency: 101
Lexical pattern:

```
#$Location1 beat #$Location2 #$Number3
```

Logical pattern:

```
(relation %1% %2% %3%)
```

□ Textual pattern suggestion
Suggestions:
AudibleBeat
DefeatingAnOpponent
PhysicallyBeatingAnAgent
HittingAnObject

Jake McCabe and John_Gaudreau scored two goals each Thursday to lead the United_States to the final of the world junior hockey championships with a 5-1 win over Canada .

McCabe put his team ahead with a pair of identical goals , scoring through goalmouth scrambles in the first period . " It feels awesome , especially in such an emotional game against Canada , " McCabe said . " I am so proud of my team .

This was the best game of the tournament for us . " In the final , the Americans will face defending champion Sweden .

The Swedes beat Russia 3-2 in a penalty shootout , with Sebastian_Collberg scoring the winner .

Gaudreau made it 3-0 for the Americans , beating Malcolm_Subban with a low shot on a sharp counter attack at 22:58 .

Subban was substituted for Jordan_Binnington after stopping 12 of 16 shots , and Jim_Vesey extended the lead to four 10 minutes later .

Ty_Rattie scored one for Canada with a short-handed goal at 44:03 , but Gaudreau scored on another sharp attack with 4:19 to play .

United States goalie John_Gibson made 33 saves for the victory . " The defense played really well , which made my job a_lot easier , " said Gibson , who made three tough saves when his team was a man down after Rattie 's goal . " If it was not for them , we probably would not be here . " The Americans finished third in their group at 2-2 in the preliminary round and routed the Czech_Republic 7-0 in the quarterfinals on Wednesday .

Canada did not lose a game in the preliminary round and had a quarterfinal bye .

Elias_Lindholm and Filip_Forsberg scored in the late game as Sweden dominated the first period , but Russia pulled one goal back midway through the second period and Mikhail_Grigorenko tied it at 47:56 .

Neither team scored in overtime before Collberg turned out to be the lone scorer in the penalty shootout .

Niklas_Lundstrom made 27 saves for the victory .

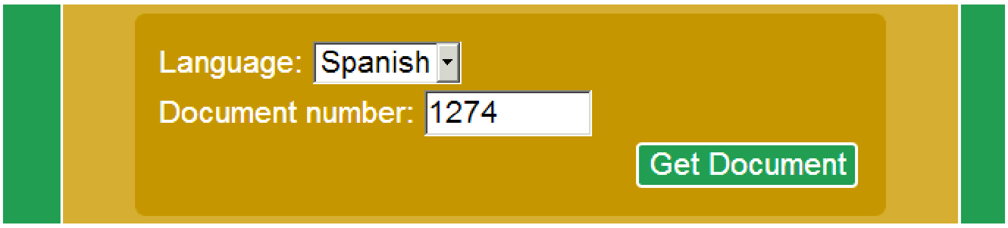
Andrei_Vasilevski stopped 38 shots for Russia .

Slovakia beat Latvia 5-3 in a relegation-round match .

Figure 5. The Graphical User Interface (GUI) of the system.

3.2.1 Document selection panel

In the beginning, the user selects a language and document number of his choice (Figure 6). The selected document will appear on the right hand side.



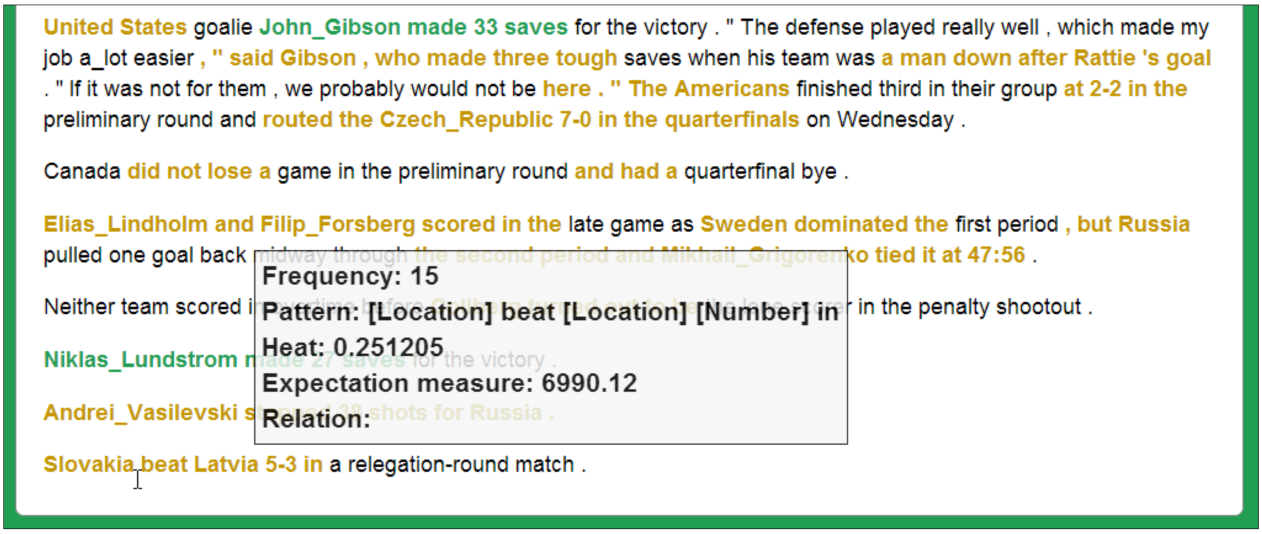
Language: Spanish

Document number: 1274

Figure 6. Document selection panel.

3.2.2 Document panel

The selected document is presented in this panel. The last part of document panel is presented on Figure 7. Each sentence is in its own paragraph. Plain text is in black colour. Parts, which are in orange or green represent special patterns. If the user moves the mouse over such pattern, the statistics of the pattern will appear in a hint box near the pattern (Figure 6, Figure 7).



United States goalie John_Gibson made 33 saves for the victory . " The defense played really well , which made my job a_lot easier , " said Gibson , who made three tough saves when his team was a man down after Rattie 's goal . " If it was not for them , we probably would not be here . " The Americans finished third in their group at 2-2 in the preliminary round and routed the Czech_Republic 7-0 in the quarterfinals on Wednesday .

Canada did not lose a game in the preliminary round and had a quarterfinal bye .

Elias_Lindholm and Filip_Forsberg scored in the late game as Sweden dominated the first period , but Russia pulled one goal back midway through the second period and Mikhail_Grigorenko tied it at 47:56 .

Neither team scored in overtime before Gallberg turned out to be the lone scorer in the penalty shootout .

Niklas_Lundstrom made 27 saves for the victory .

Andrei_Vasilevski stopped 38 shots for Russia .

Slovakia beat Latvia 5-3 in a relegation-round match .

Frequency: 15

Pattern: [Location] beat [Location] [Number] in

Heat: 0.251205

Expectation measure: 6990.12

Relation:

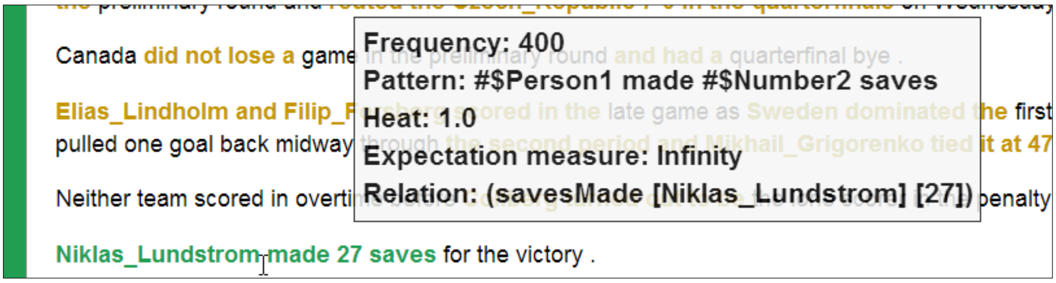
Figure 7. The bottom part of the document panel. The statistics box is displayed for text: *Slovakia beat Latvia 5-3.*

The parts of text that are highlighted in *orange*, match with patterns suggested by the system. These patterns occur in the corpus more times than a predefined threshold; they include at least one non-basic token; and have a higher expectation measure than other patterns which overlap with them. The expectation measure of the pattern $em(pat)$ is calculated in the following way:

$$em(pat) = \frac{p(pat)}{\prod_{i=1}^{n} p(pat_i)}$$

Where $p(pat)$ is the probability of the pattern - pat, $p(pat_i)$ is the probability of the i -th token in the pattern, and n the length of the pattern. This measure rewards the patterns whose tokens frequently co-occur together. However, sometimes patterns that have some redundant tokens, such as punctuations, are scored highly. In the future, we plan to combine this measure with dependency parse tree data to improve the suggestions. Additional research on identifying good patterns has been done in [1].

The parts of text that are highlighted in *green* have already pattern rules defined for them. Consequently, the corresponding relation is already constructed and presented in the hint box, like on Figure 8.



Canada did not lose a game ... round and had a quarterfinal bye .

Elias_Lindholm and Filip_F... scored in the late game as Sweden dominated the first

pulled one goal back midway through the second period and Mikhail_Grigorenko tied it at 47

Neither team scored in overti... penalty

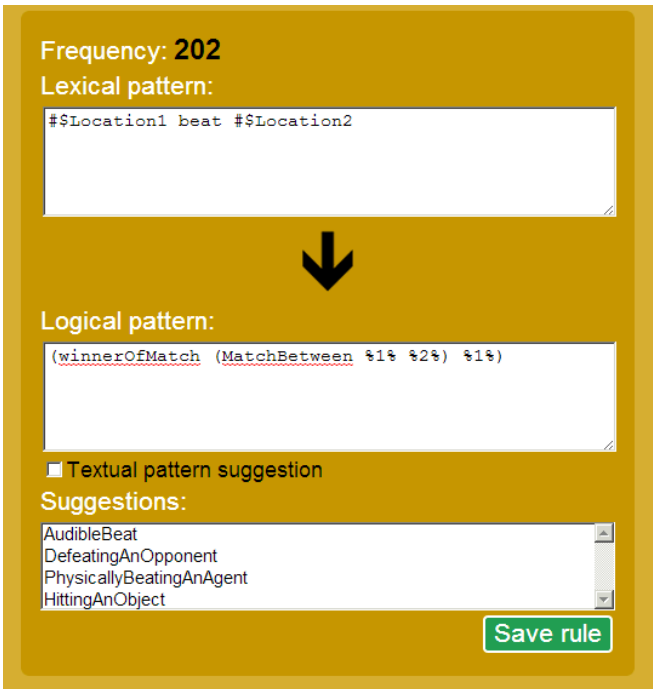
Niklas_Lundstrom made 27 saves for the victory .

Frequency: 400
Pattern: #$Person1 made #$Number2 saves
Heat: 1.0
Expectation measure: Infinity
Relation: (savesMade [Niklas_Lundstrom] [27])

Figure 8. Applied pattern rule on *Niklas_Lunstrum made 27 saves.*

3.2.3 Pattern rule panel

This is the panel, where users construct pattern rules (Figure 9). To construct a rule, a valid textual pattern needs to be in the *lexical pattern* box. This can be achieved either by clicking on the text highlighted in orange, or by selecting some text and dragging it the lexical pattern box. In both cases the system generalizes the text to become a pattern. The frequency of the pattern is calculated and displayed above. In the same moment, a generic pattern is displayed in the *logical pattern box*, for example *(relation %1% %2%)*. Each argument in the lexical pattern box starts with *#$* and ends with its serial number. To construct argument mappings, the user must define arguments in the logical pattern by stating their serial number encapsulated with percentage characters, for example *%1%*. The user constructs the logical pattern with concepts from the semantic graph. In the *suggestions* box, the system suggests few concepts from the semantic graph that are denoted by the basic words in the lexical pattern box. These concepts might be included in the logical pattern. When the rule is complete, the user clicks on the button *Save rule*. The system applies the rule on all the documents and reports the number of matches. The matches of the current document become highlighted in green.



Frequency: 202

Lexical pattern:

```
#$Location1 beat #$Location2
```

Logical pattern:

```
(winnerOfMatch (MatchBetween %1% %2%) %1%)
```

☐ Textual pattern suggestion

Suggestions:

AudibleBeat
DefeatingAnOpponent
PhysicallyBeatingAnAgent
HittingAnObject

Figure 9. Pattern rule panel

3.2.4 Displaying output

To see all constructed rules, the user must click on *Rule Repository* link bellow the pattern rule panel. To see all the assertions with corresponding sentences for each rule, the user must click on the *Assertions* link (Figure 5)

4 Experiments

We prepared a dataset from the NewsFeed of WP1 [1]. In particular, we took all articles from January 2013 that are in English, Catalan and Spanish, and analysed them with the linguistic tools in WP2. This results in a large corpus where we can extract relations and aggregate them over documents. Such large corpus is also useful in order to test the system for developing pattern rules.

For this experimentation in particular, we selected a subset of documents that are about corruption cases in Spain. Appendix A shows some example sentences in the three languages. We will first present experiments using the techniques to merge grammatical relations. Then we will present experiments using the pattern learning tool. In both cases we illustrate the type of patterns that can be extracted.

4.1 Using Grammatical Relations

We were interested to see cross-lingual matches of relational content. That is, assuming that the documents in different languages convey the same information, we should be able to extract similar relation triples from documents in different languaes. Since our documents are from the same time span (January 2013), and we selected a hot topic that has gotten international coverage (corruption), it is reasonable to assume that similar information is expressed in the document collections in three different languages.

Table 1. Examples of syntactic triples and their cross-lingual representation

Language	Subject	Predicate	Object
ca	diari/06267145-n	publicar/00967625-v	anotació/06763273-n
ca	paper/14974264-n	publicar/00967625-v	rotatiu/08288291-n
en	el pais	publish/00967625-v	detail/05817845-n
en	newspaper/06267145-n	publish/00967625-v	ledger/13404248-n
es	diario/06267145-n	publicar/00967625-v	fotografia/03931044-n
es	periódico/06267145-n	publicar/00967625-v	imagen/03931044-n
ca	pp	negar/02212825-v	document/06470073-n
ca	partit/08256968-n	negar/02212825-v	pagament/13278375-n
en	pp secretary general	deny/02212825-v	knowledge/00023271-n
en	spanish pm mariano rajoy	deny/02212825-v	allegation/07236077-n
es	pp	negar/02212825-v	financiacion
es	partido/08256968-n	negar/02212825-v	acusación/07234230-n
ca	pp	pagar/02251743-v	sobresou
ca	partit/08256968-n	pagar/02251743-v	vestit/03236735-n
en	business/00582388-n	pay/02251743-v	money/13384557-n
es	extesorero	pagar/02251743-v	sobresueldo
es	luis barcenas	pagar/02251743-v	sobresueldo
ca	tresorer/10727256-n	portar/02686471-v	registre/06507041-n
en	treasurer/10727256-n	write/01744611-v	ledger/13404248-n
es	tesorero/10727256-n	repartir/02294436-v	sobre/03291819-n

Table 1 presents some examples of triples, separated in blocks. In the first block the predicate corresponds to WordNet synset 00967625-v, which can be lexicalized as *publish* (in English) or *publicar* (in Spanish and Catalan). It is also interesting to observe that these triples share similar subjects and objects. For example, some subjects are labelled 06267145-n (*newspaper/periodico/diari*) while others refer to close words. One subject is *El Pais* which is the name of one newspaper: at this point we do not exploit cross-lingual representations of named entities (mainly because these are not covered in WordNet). If we look at objects, we observe less exact matches according to the synset; however, these words are semantically related. One idea for exploration is to exploit existing WordNet similarity measures in order to develop similarity measures of cross-lingual relations. This will allow to compare relational content of documents either inter-lingually or cross-lingually.

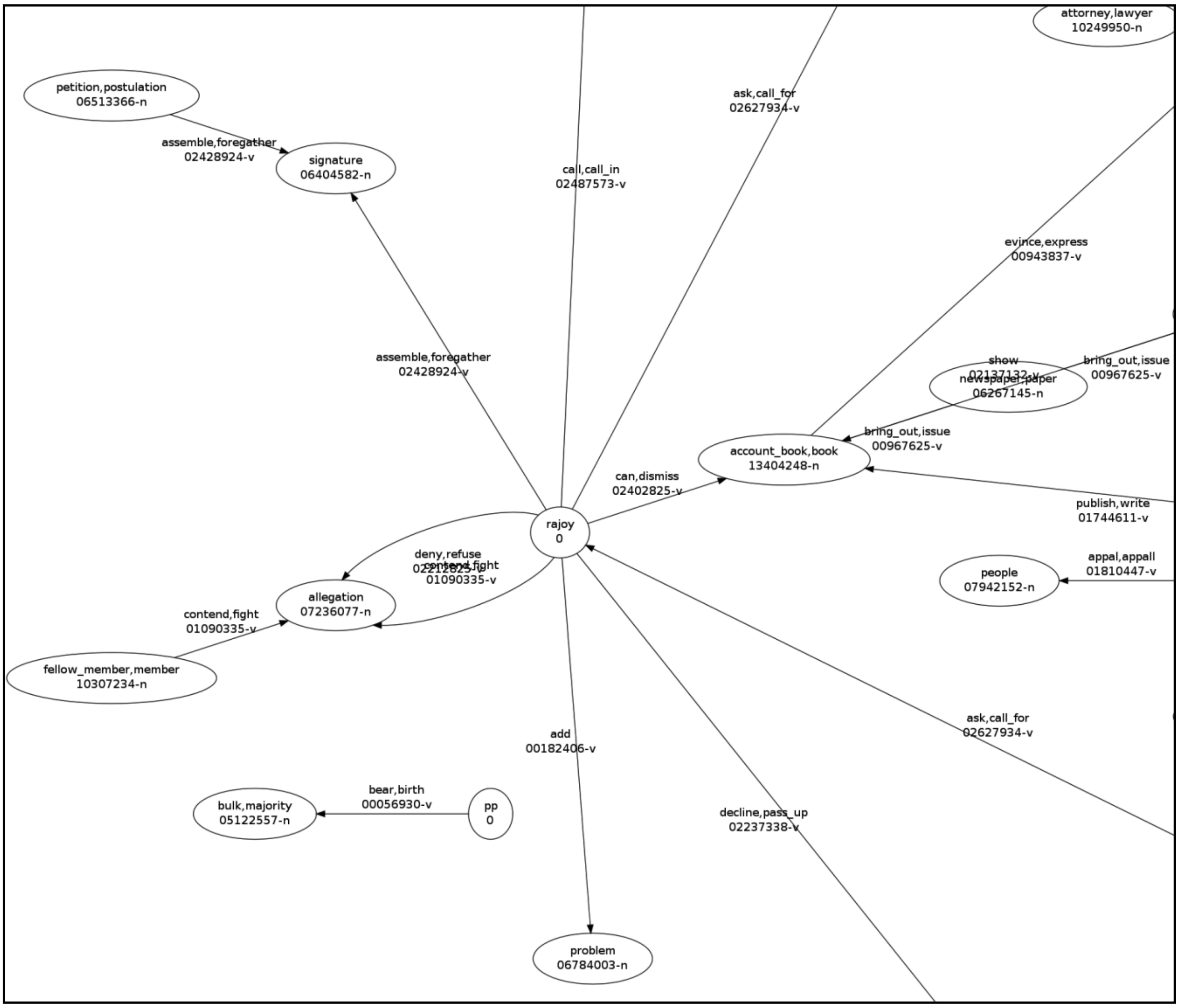


Figure 10. A graph based on syntactic triples.

Figure 10 presents a graph where we have merged cross-lingual grammatical relations from many documents in different languages. In this case the relations were extracted from the syntactic tree. We can see that merging works, in the sense that relations that have been mentioned with different lexical items have been aggregated on the same relation (for example, we can see that *Rajoy deny/refuse the allegation*). Figure 11 presents a similar graph where the relations have been extracted on the basis of semantic roles instead of the syntactic tree. The graphs are constructed using the same set of documents, thus it is clear that semantic-based extraction provides a much richer set of relations.

We leave as future work evaluating the precision and recall of the extraction and merging processes.

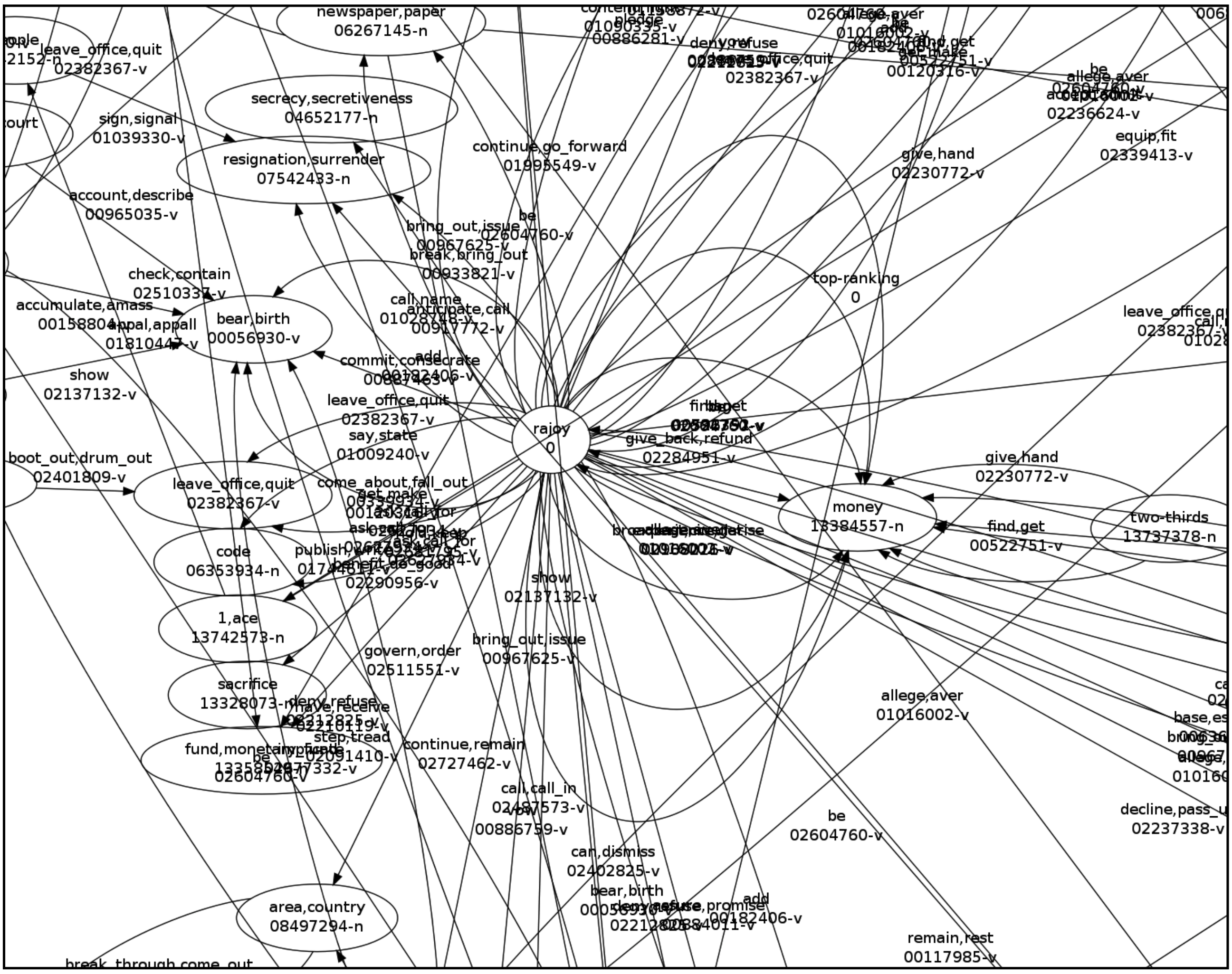


Figure 11. A graph based on semantic triples.

4.2 Using Patterns and Ontological Constraints

In this part we present an experiment using pattern analysis, where a few pattern rules are constructed and applied on the small set of documents.

We constructed several pattern rules on these articles with our system. In this experiment we took CyC as the semantic graph. All the constants that are not arguments were already defined in CyC. We will present several tables, where the first part of the table is the rule. In the second part of the table relations produced by the rule are stated. The first relation comes from the document, where the rule was created. We will first present five patterns for English language. The first two patterns connect entities with their types via “is a” predicate.

Table 2. Example patterns of “is a” predicates

Lexical pattern	Logical pattern	Times applied
Attorney #$Person1	(#$isa %1% #$Attorney)	435
Relations		
(#$isa [General_Eduardo_Torres-Dulce] #$Attorney)		
(#$isa [General_Jesus_Murillo] #$Attorney)		
(#$isa [Charles_Wycoff] #$Attorney)		
(#$isa [General_Mukti_Pradhan] #$Attorney)		

Lexical pattern	Logical pattern	Times applied
`newspaper #$Organization1`	`(#$isa %1% #$NewspaperOrganization)`	**394**
Relations		
`(#$isa [El_Pais] #$NewspaperOrganization)`		
`(#$isa [Bild] #$NewspaperOrganization)`		
`(#$isa [Wainstein] #$NewspaperOrganization)`		
`(#$isa [Al-Khaleej] #$NewspaperOrganization)`		

The pattern on table below does not produce relations, but concepts that represent the downtown of a particular city. These concepts can be used further in other relations.

Table 3. Pattern for the concept “downtown of a city”.

Lexical pattern	Logical pattern	Times applied
`downtown #$Organization1`	`(#$DowntownFn %1%)`	**1798**
Relations		
`(#$DowntownFn [Madrid])`		
`(#$DowntownFn [Singapore])`		
`(#$DowntownFn [Cleveland])`		
`(#$DowntownFn [Auckland])`		

The next pattern connects a person with his age.

Table 4. Pattern for “person-age” relation.

Lexical pattern	Logical pattern	Times applied
`#$Person1 , #$Number2 ,`	`(#$age %1% (#$YearsDuration %2%))`	**25046**
Relations		
`(#$age [Miguel_Gomez] (#$YearsDuration [30]))`		
`(#$age [Matias_Dellanno] (#$YearsDuration [37]))`		
`(#$age [Taylor_Nanz] (#$YearsDuration [18]))`		
`(#$age [Yvonne_Gomez] ($#YearsDuration [53]))`		

The next pattern is more complex. The relation that a person received some amount of money is split into three more basic relations. For example, there exists an event of money transfer (first assertion), where there was $34.000 transferred, and *Rajoy* is the person who received the money.

Table 5. Pattern for a “money transfer” predicate.

Lexical pattern	Logical pattern	Times applied

`#$Person1 received`
`about #$Location2`
`#$Number3`
	`(#$thereExists ?EVENT`
	` (#$and`
	`  (#$isa ?EVENT #$MoneyTransfer)`
	`  (#$moneyTransferred ?EVENT %3%)`
	`  (#$beneficary ?EVENT %1%))))`
		1

Relations		

`(#$thereExists ?EVENT`
` (#$and`
`  (#$isa ?EVENT #$MoneyTransfer)`
`  (#$moneyTransferred [$34000])`
`  (#$beneficary ?EVENT [Rajoy]))))`		

This pattern is very similar to the last one, except that it works on Catalan language. Unfortunately, money generalization was not set for Catalan language, therefore the rule only works on 9.900 euros.

Table 6. Another pattern for a “money transfer” predicate.

Lexical pattern	Logical pattern	Times applied

#$Person1 rebien 9.900_euros	(#$thereExists ?EVENT
	 (#$and
	  (#$isa ?EVENT #$MoneyTransfer)
	  (#$moneyTransferred ?EVENT (Euros 9900))
	  (#$beneficary ?EVENT %1%))))
		1

Relations		

(#$thereExists ?EVENT
 (#$and
  (#$isa ?EVENT #$MoneyTransfer)
  (#$moneyTransferred (Euros 9900))
  (#$beneficary ?EVENT [Álvarez_Cascos]))))		

The last rule presented is for Spanish language, stating that a person is a prime minister.

Table 7. Pattern for “prime minister” attribution.

Lexical pattern	Logical pattern	Times applied
primer ministro , #$Person1 ,	(isa %1% #$PrimeMinister-HeadOfGovernment)	253
Relations		
(isa [Mariano_Rajoy] #$PrimeMinister-HeadOfGovernment)		
(isa [Recep_Tayyip_Erdogan] #$PrimeMinister-HeadOfGovernment)		
(isa [David_Cameron] #$PrimeMinister-HeadOfGovernment)		
(isa [Shinzo_Abe] #$PrimeMinister-HeadOfGovernment)		

5 Conclusion

We have presented early prototypes to obtain cross-lingual representations of relational content of documents. These representations can be used to compare documents cross-lingually, or to aggregate the information of a collection of documents into a merged semantic graph that organizes relations weighted by frequency.

We have followed two different approaches. The first generalizes linguistic grammatical relations, either syntactic or semantic. Then we use WordNet as an interlingua representation of lexical items, taking advantage of the fact that in WordNet words are already linked cross-lingually. To some extent, this is a simple approach that will work moderately well. One advantage is that WordNet has been built manually; hence the cross-lingual mappings are expected to be high quality. Another advantage is that there exist a number of similarity measures based on WordNet: hence, if two relations are not exactly identical, one can develop principled methods to compute a notion of similarity between relations, using WordNet similarity measures as primitive functions to compute the the similarity between the nodes of a relation.

There are some limitations in using WordNet. First, our approach requires disambiguating the sense of each word into WordNet concepts, and this task is known to perform only at moderate accuracy. A second disadvantage is coverage. Since WordNet is produced manually, there only exist WordNets for a limited number of languages, and for some of them the coverage is not too good. Similarly, we cannot expect to find in WordNet specialized terms, named entities, or jargon that spontaneously arises in social media. To overcome these issues about coverage, our next step is to develop alternative approaches where the cross-lingual representation of lexical items and named entities is given by unsupervised methods ---this is precisely the topic of WP3 in XLike [4].

In the second part we have presented an alternative approach based on pattern rules that link textual patterns into logical patterns. In this case, the logical patterns are the building blocks of the semantic graphs representing documents. In the experiments we have illustrated the construction of different patterns using CyC as the set of logical patterns, which provides rich semantic representations of documents. We should investigate techniques that automatically learn to construct patterns. Also, as in the previous approach, we should integrate techniques to obtain cross-lingual representations of named entities and constants.

A final approach will combine the two methods. The aim here is to obtain rich semantic representations given by ontologies like CyC, and increase their coverage employing the linguistic representations given statistical methods for syntactic/semantic analysis.

References

[1] XLike Deliverable D1.3.2 --- Final prototype on data infrastructure.

[2] XLike Deliverable D2.1.1 --- Shallow linguistic processing prototype.

[3] XLike Deliverable D2.2.1 --- Early deep linguistic processing prototype.

[4] XLike Deliverable D3.1.1 --- Early text annotation prototype.

[5] Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670

[6] Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece.

[7] GermaNet: German WordNet, http://www.sfs.uni-tuebingen.de/lsd.

[8] SloWNet: Slovene WordNet, http://lojze.lugos.si/darja/slownet.html.

[9] J. Starc and B. Fortuna, “Identifying good patterns for relation extraction,” in Proceedings of the 15th International Multiconference Information Society - IS 2012, vol. A, pp. 205-208, 2012.

Annex A Example Sentences

These are some example sentences about corruption cases in Spain, in English, Spanish and Catalan, that appeared in media during January 2013.

English:

Riot police clashed with protesters in Madrid and impromptu demonstrations broke out in several other Spanish cities following the prime minister's televised denial that he had accepted under-the-table payments.

Prime Minister Mariano Rajoy promised to publicly disclose the amount of funds in all his personal bank accounts, denying recent media reports that allege he and members of his governing conservative Popular Party accepted or made under-the-table payments.

Speaking at a special executive committee meeting at his party's Madrid headquarters, Rajoy said "it is false" that he received or distributed undeclared money.

"Next week, my statements of income and assets will be made available to all citizens," he said, adding they would be published on the official website of the prime minister.

By late Saturday it was clear Rajoy's pledge had failed to defuse popular disquiet as riot police cordoned off several of Madrid's main avenues in a bid to stop protesters from gathering in large groups.

Spanish:

Un gigantesco escándalo estalló hoy cuando el diario El Mundo denunció que el ex tesorero del Partido Popular durante veinte años, Luis Bárcenas, repartió sobres con dinero en negro a dirigentes que procedía de sobornos de empresas constructoras. Cinco fuentes proporcionaron la información.

El episodio se ha convertido en una "bomba atómica", según la expresión de un comentarista político que recordó palabras del máximo acusado y está ligado a otros hechos de corrupción vinculados a Bárcenas como el notorio caso "Gurtel" una trama que afecta a decenas de personalidades de primer nivel vinculadas al PP, imputadas en un vasto proceso judicial.

Bárcenas, que está entre los acusados, figura en la documentación como "Luis el cabrón". Hace pocos días se conoció que el ex tesorero, tenía 22 millones de euros en cinco cuenta abiertas en Suiza y unos 4,5 millones en Estados Unidos.

Lo más sorprendente es que Bársenas fue suspendido de militancia y dimitió a su cargo de senador por Cantebria. El PP ha utilizado esta situación para insistir en que asi se demuestra que el los populares no tienen nada que ver con el extesorero. Antes fue defendido a capa y España por el partido y por su máximo dirigente, Mariano Rajoy, que incluso le pagado un abogado de oro para defenderlo.

Pero, también hoy, el diario El País publica con gran despliegue en su edición online, que Luis Barcenas , todavía conserva un despacho en la sede del Partido Popular, que frecuentaba constantemente. "Solo Rajoy tiene autoridad para permitir una situación asi", confió una de las fuentes.

Catalan:

El secretari general del PSOE, Alfredo Pérez Rubalcaba, va demanar ahir la dimissió de Mariano Rajoy per els presumptes casos de corrupció que esquitxen el Partit Popular.

En una declaració institucional, Rubalcaba va demanar que el també líder del PP abandoni la presidència del govern espanyol per tal de donar pas a una persona "que pugui restablir la fortalesa, credibilitat i estabilitat que necessita el nostre país", va assenyalar.

El dirigent socialista va explicar que el desencadenant de la petició ha estat la compareixença que Rajoy va fer dissabte, on no va donar als ciutadans "explicacions convincents" sobre el 'Cas Bárcenas' o la trama Gürtel.

En els seus arguments, el dirigent del PSOE va assenyalar que Mariano Rajoy "no pot dirigir el nostre país en un moment tant delicat com aquest". "La seva permanència al capdavant del govern no permetrà superar la crisi política, l'agreujarà dia a dia", va afegir. "S'ha plantejat si la seva presència és millor per a la imatge exterior d'Espanya?", va qüestionar. "La nostra obligació i deure amb els espanyols és dir que no", va sentenciar el líder de la oposició.

Tot i així, no va especificar si el PSOE impulsarà una moció de censura o demanarà la convocatòria d'eleccions anticipades. Va assenyalar que, en aquest moment, el problema de fons és que el president del PP està en una situació que l'impedeix dirigir el govern d'Espanya. "Creiem que el problema de fons del nostre país és que el president no pot fer front a la situació gravíssima per la que passa Espanya", afirmà. "No s'adona que la seva presencia és un llast per al nostre país?", va ressaltar. "Cal un govern fiable, fort, confiable, i el del PP ha deixat de ser-ho, començant pel propi president", va constatar.