Automatic detection of Spanish and Japanese modal markers and presence in spoken corpora

Carlos Herrero Zorita
Computational Linguistics Laboratory
Autonomous University of Madrid

Background

- BA East Asian Studies (Japanese itinerary) (2010)
- BA English Studies (2012)
- MA Applied Linguistics (2013)
- PhD Computational Linguistics Laboratory (Prof. Antonio Moreno Sandoval) (2017)

Structure

1) Definition of modality, classification, encoding

2) Modal markers in spoken corpora

3) Description of automatic detection of modality

Defining Modality

Defining Modality

- Universal, human-exclusive feature
- Same level as tense, aspect
- Very frequent in spoken discourse
- Well studied but difficult to define and classify

Defining Modality

	WEST	JAPAN
a.C.	Greek philosophers	
13th-17th	Modistae, logicians	Fujiwara
18th-19th	Kant, psycholinguists	Chinjutsu
19th-20th	Linguists. Lyons, Bally, Fillmore	Masuoka y Nitta
21st		

Defining Modality

Modality is everything that modifies the proposition, including negation, tense, case particles, discourse markers, etc. Present in every sentence (Fillmore, 1972; Masuoka, 1991; Wasa, 2005; Nuyts, 2006; Imithani, 2009)

Modality is the expression of the attitude or subjectivity of the speaker, also his or her emotions and opinions (Lyons, 1977; Palmer, 2001; Bybee et al., 1994; Nitta, 1991; Halliday, 1970 [2009])

Modality relates language with reality: expression of necessity/possibility, factuality, realis/irrealis in either the morphological mood, modal auxiliaries or both: (Givón, 1995; Palmer, 2001; Narrog, 2009a; Nomura, 2003;
Harada, 1999; Johnson, 1999)

Aims of the study

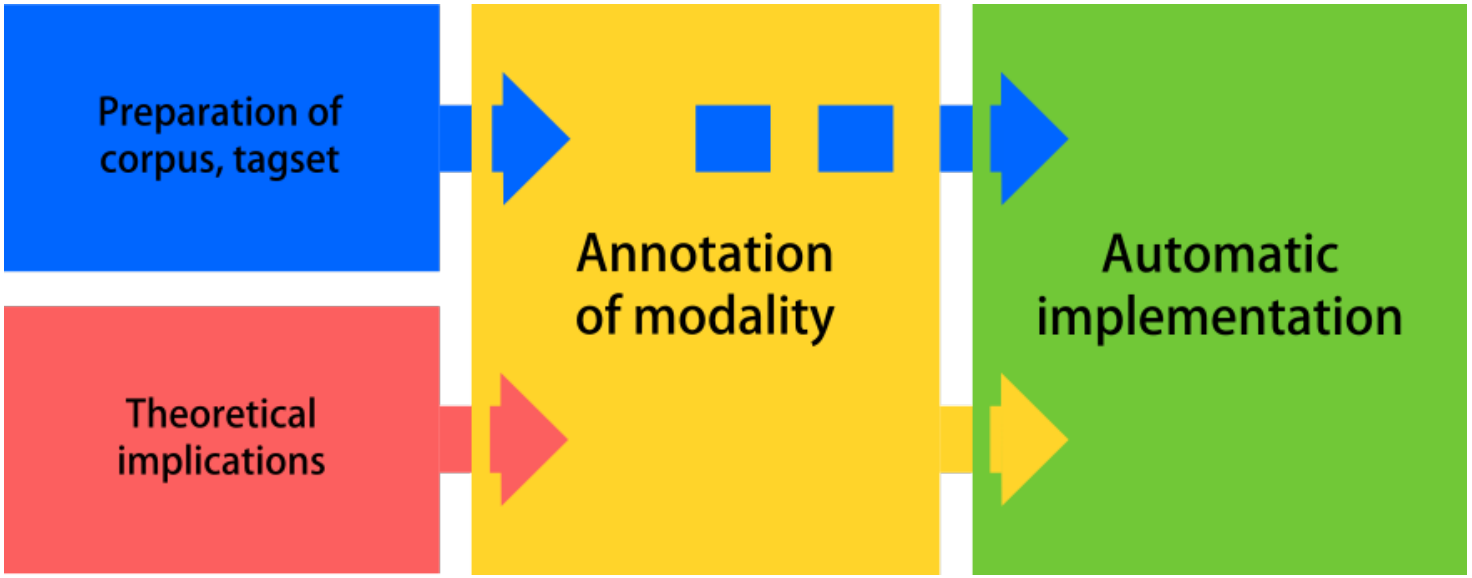
- Comparison of Spanish and Japanese modality from a computational perspective.

- Two parts:
  - Corpus study
  - Development of a modal tagger

Questions

- What is the best definition and classification of modality for a cross-linguistic computational work?
- How is modality used in spoken Spanish and Japanese, and how are modal markers modified?
- How can we formalise this information into a program that can annotate modals automatically in new texts?

Methodology



Requirements for modality

- Cross-linguistic: Spanish and Japanese
- Easy to formalise
- Automatic tagging
- Objetive, context-independent
- Compatible with other elements such as negation

Modality in this study

- Based on the work of previous typologists.
- Modal logic.
- Modality signals the necessity or possibility of *P*.
- Encoded in grammatical mood in old languages, now needs additional elements.

Modality in this study

I **must** go home now

"The SOA of *going home* is necessary" ($\Box P$)
(True in all possible worlds)

Modality in this study

I **must** go home now

"The SOA of *going home* is necessary" ($\Box P$)
(True in all possible worlds)

A complete recovery is **possible**

"The SOA of *recovering completely* is possible" ($\Diamond$P)
(True in at least one possible world)

Modality in this study

Epistemic

“It may rain tomorrow”

Necessity / Possibility

Modality in this study

Epistemic

"It may rain tomorrow"

Necessity / Possibility

Deontic

"Come here!"

Modality in this study

Necessity / Possibility

Epistemic

"It may rain tomorrow"

Deontic

"Come here!"

Ambiguous

"John may enter the room"

Modal markers

- Same discrepancies as modality definition.
- Syntactic point of view.
- Fully grammaticalised/marked elements.
- Add modal meaning to the verb (i.e. mood).

Modal markers

- Auxiliaries

  Auxiliary + Verb

  Juan debe venir mañana

  Juan must come tomorrow

Modal markers

- Auxiliaries

Verb + Auxiliary

明日　　は、フアンが　来なきゃいけない

Tomorrow NOM Juan　NOM come-must

Juan must come tomorrow

Modal markers

- Adverbs

Mañana a lo mejor llueve

明日はおそらく雨が降るだろう

It'll probably rain tomorrow

Modal markers

- Adjetives

(Predicative position)

Es necesaria una transfusión de sangre

輸血が必要だ

A blood transfusion is necessary

Modal markers

- Mood: imperative and potential

¡Vete!

行け！

Leave!

Modal markers

	Spanish	Japanese
Auxiliaries	6	24 (60)
Adverbs	36	12
Adjectives	23	12
Mood	1	2

Presence in spoken corpora

Corpora

C-ORAL ROM

- 301,329 words
- 379 speakers
- Different contexts

C-ORAL JAPÓN

- 127,676 words
- 58 speakers
- Educational purpose

Tagset

Classification	NEC/POSS
Subclassification	EPIS/DEON/AMBG
Type	AUX/ADV/ADJ/MOOD
Negated	
Separation	ID/Ref
Ellipsis	
Value	0%/30%/50%/70%/100%

Annotation

C-ORAL ROM

```
<Turn>
<Name>SEV</Name>
<Utterance id="1882"
Type="enunciation">
pues
<w neg="Yes">no</w>
<m lang="ESP" modtype="NEC"
subtype="AMBG" neg="Yes"
class="mood_SUBJ"
value="0%">puedes</m>
trabajar ahí
</Utterance>
</Turn>
```

C-ORAL JAPÓN

```
<UNIT id="11550" speaker="MAS">
<m lang="JAP" modtype="NEC"
subtype="EPIS" neg="no" class="Adverb"
value="100%"> 絶対 </m>
スポーツ好きな人とか
</UNIT>
```

Objectives

- Frequency distribution according to linguistic and non-linguistic factors

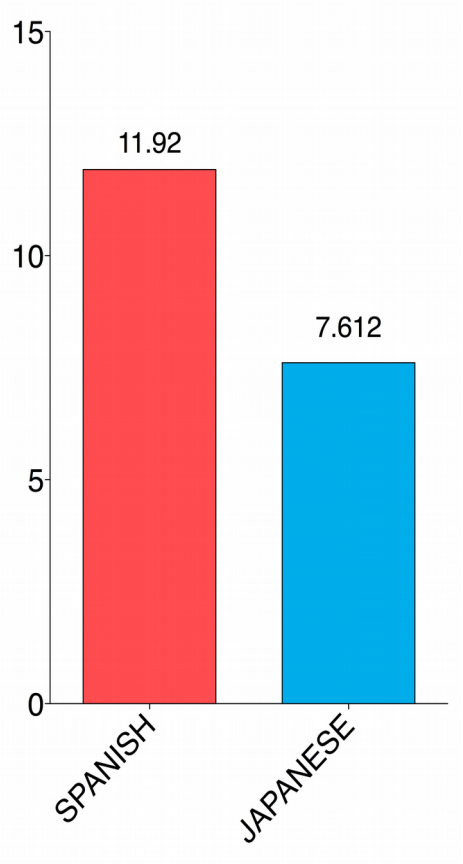
- Features that could modify the modal markers

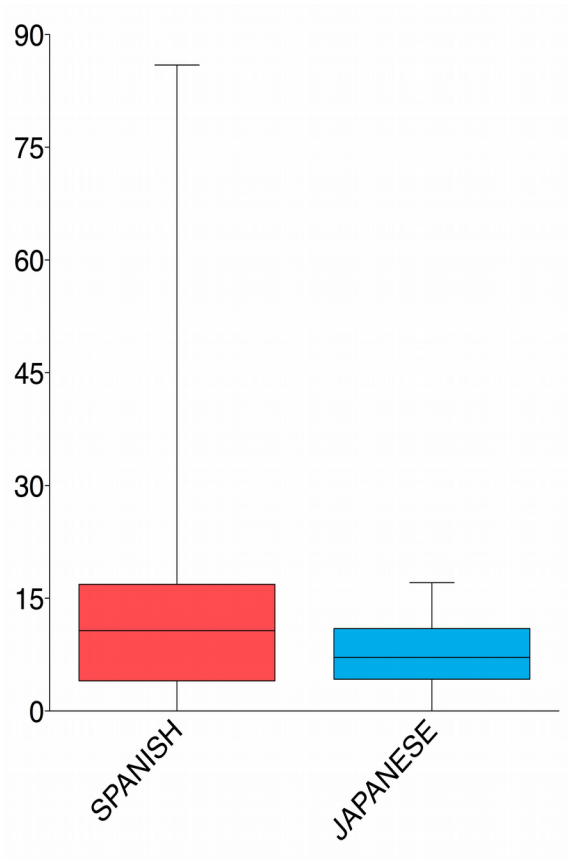
Objectives

- Is modality frequency significally different depending on the language, type of discourse, sex, age of the speakers?

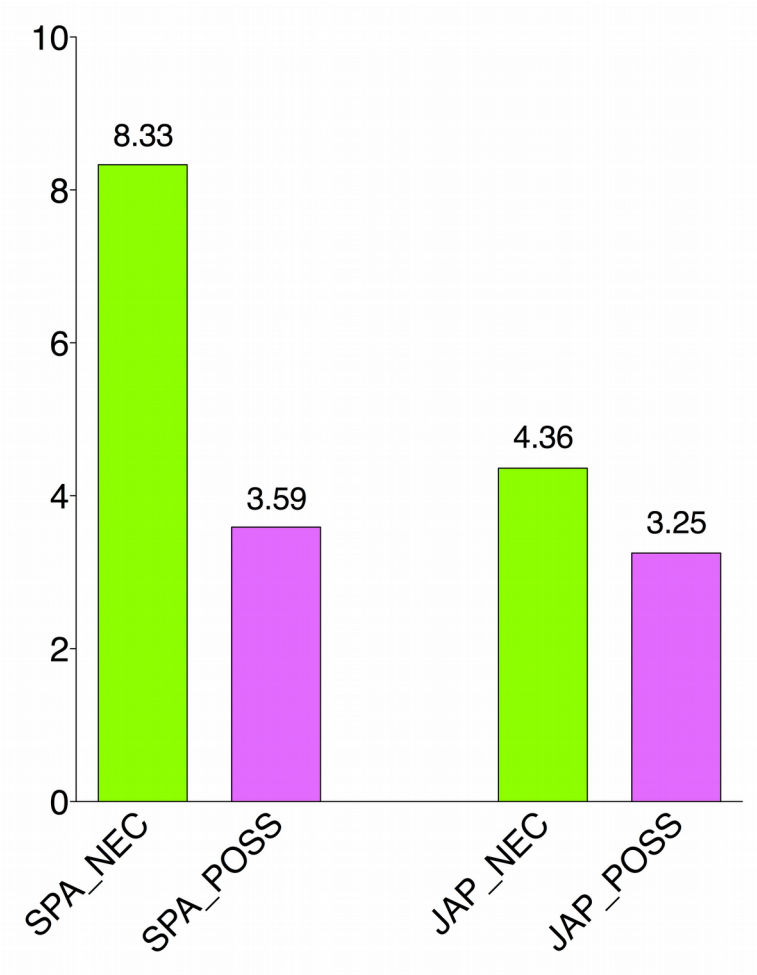
- Are external factors modifying the markers frequent enough to be taken into account by the tagger?

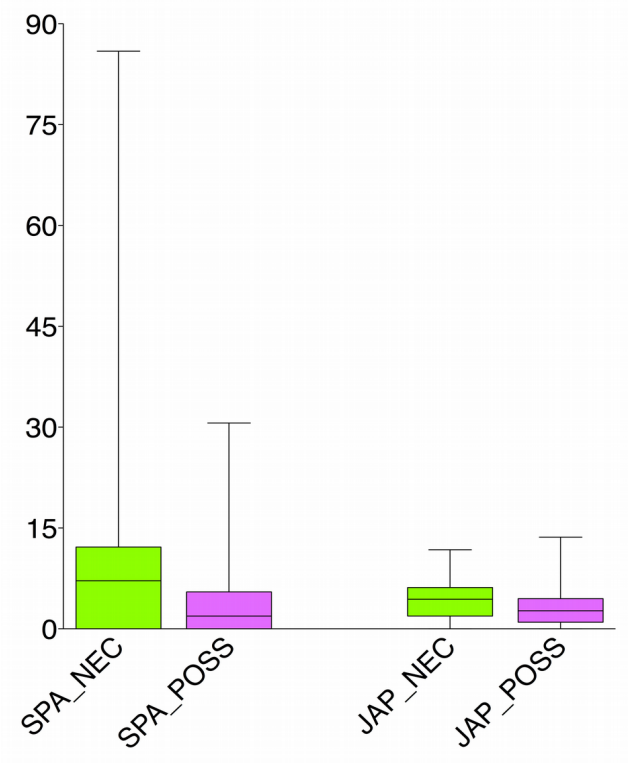
General numbers



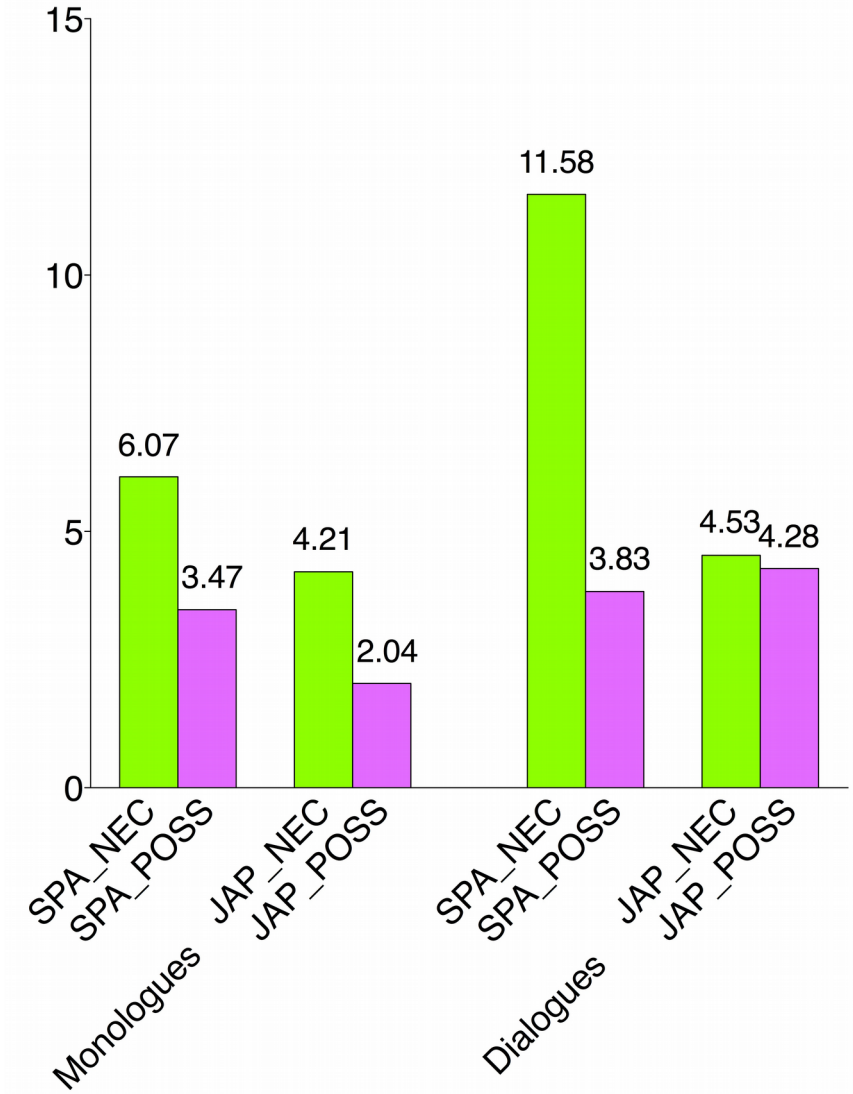


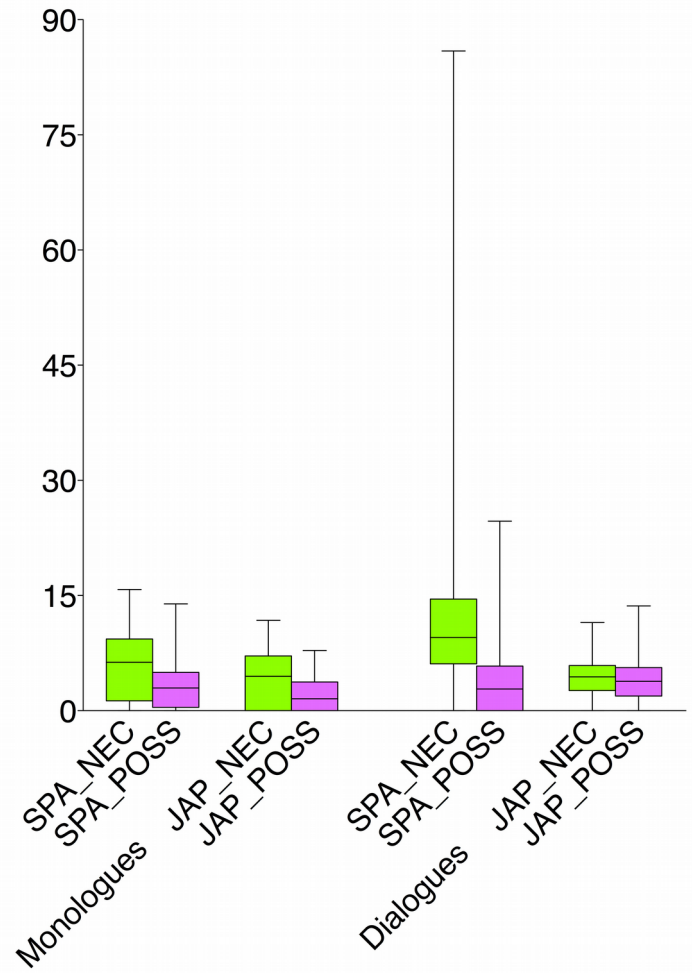
NEC vs POSS


10
8
6
4
2
0
8.33
3.59
4.36
3.25
SPA_NEC
SPA_POSS
JAP_NEC
JAP_POSS

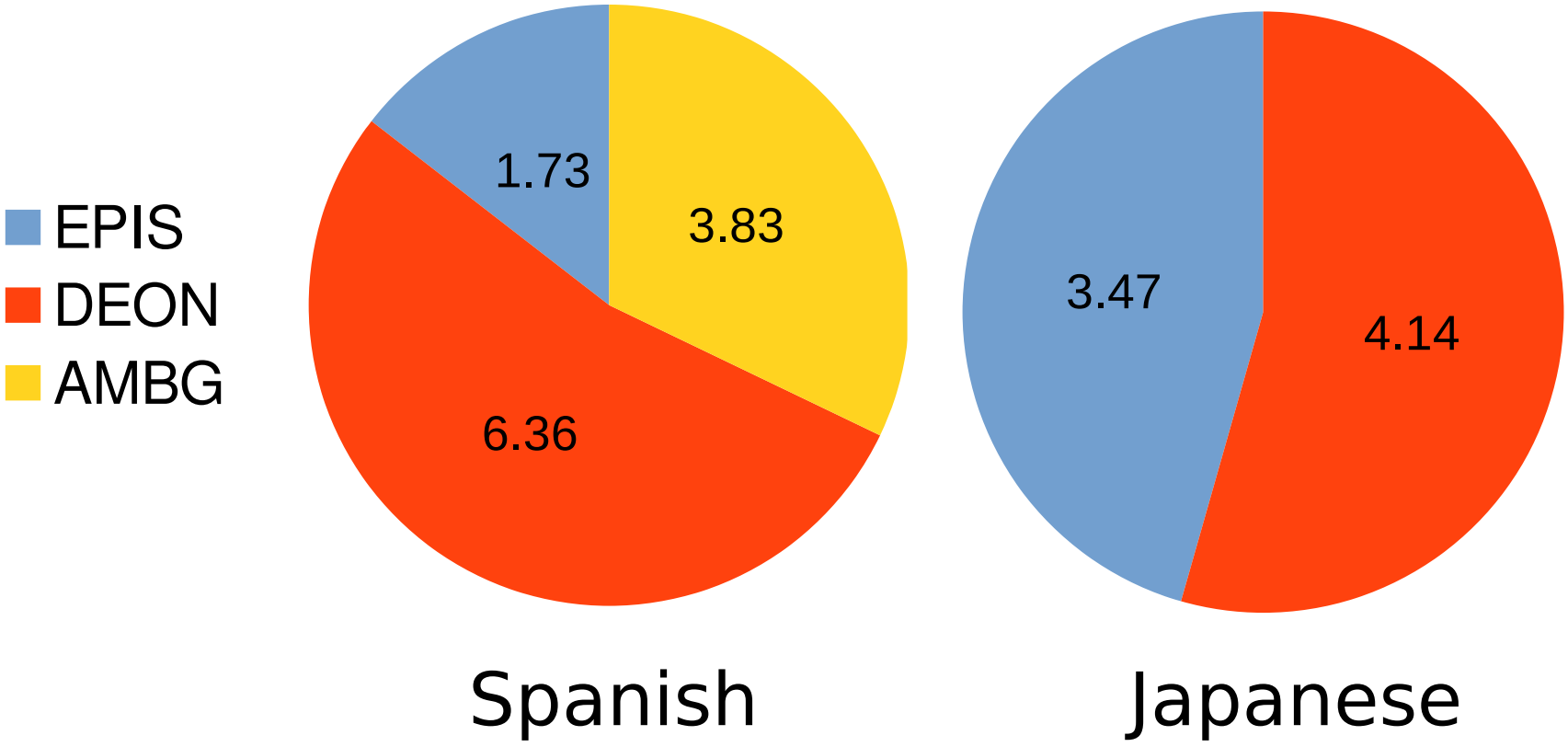

90
75
60
45
30
15
0
SPA_NEC
SPA_POSS
JAP_NEC
JAP_POSS

NEC vs POSS: Discourse

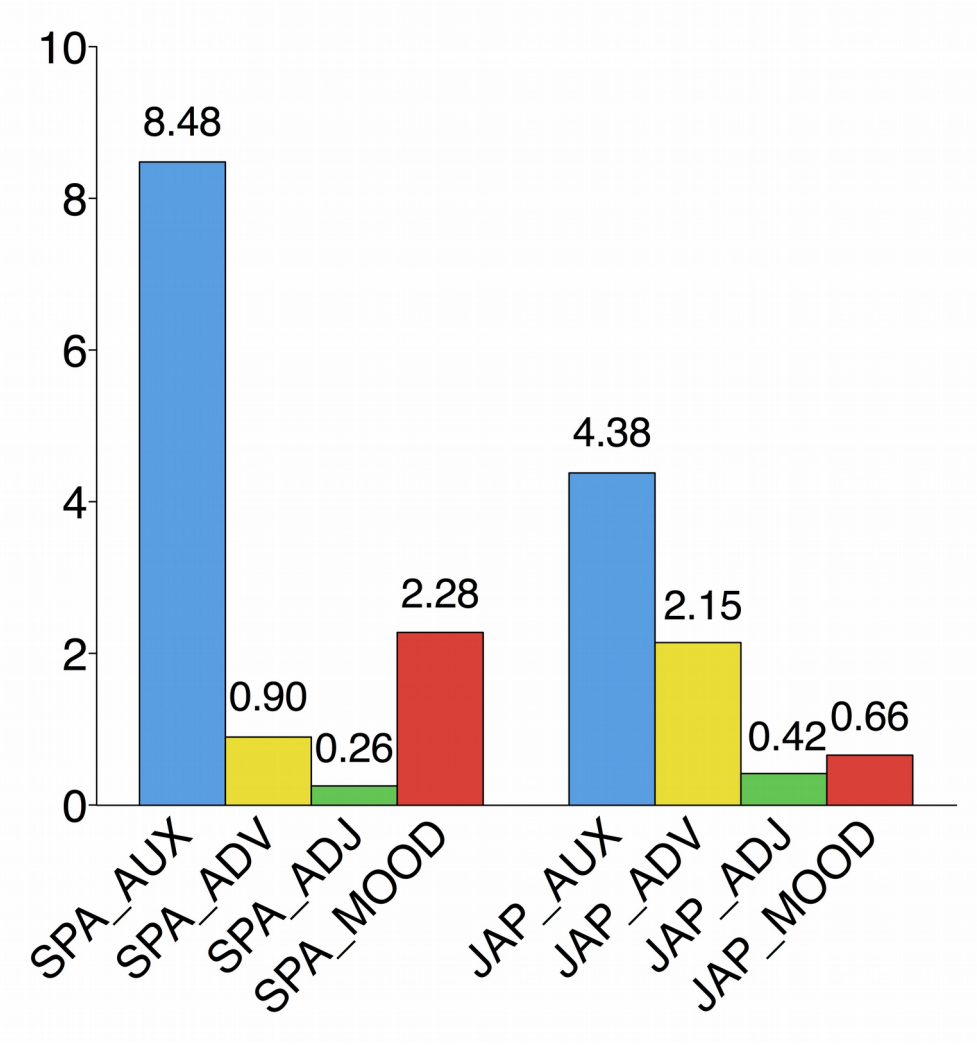




EPIS vs DEON



Type of marker



Modification of markers

Spanish

- Negation
- Syntactic separation
- Ellipsis
- Errors

Japanese

- Negation
- Syntactic separation
- Ellipsis
- Writing variation
- Variation according to politeness

Modification of markers

- Negation of modality

Change in the classification:

A crash is possible ($\Diamond P$)

A crash is not possible ($\neg\Diamond P$) = ($\Box\neg P$)

Modification of markers

- Negation of modality

Change in the classification:

I have to go ($\Box$P)

I don’t have to go ($\neg\Box$P) = ($\Diamond P$)

Modification of markers

- Negation of modality:
  - Change:
    Neg. + can go (POSS) = NEC
    Neg. + have to go (NEC) = POS
  - No change:
    Neg. + must go (NEC) = NEC

Modification of markers

- Negation of modality:
  - Change:
    Neg. + can go (POSS) = NEC
    Neg. + have to go (NEC) = POS
  - No change:
    Neg. + must go (NEC) = NEC
  - Fairly frequent:
    12%-13% in Spanish and Japanese

Modification of markers

- Separation
  (1.48% in SPA, max 4 / 0.18% in JAP, max 2)

  Podrías, no sé, venir aquí
  You could, I don’t know, come here

- Ellipsis of AUX/Main Verb
  (1.08% in Spanish / 3.89% in Japanese)

  Sí, puedes.
  Yes, you can.

Modification of markers

◇ ***Errors*** made by Spanish native speakers (1.74% of the constructions)

- *Deber* (“must”, deontic) vs *deber de* (“must”, epistemic)

- Using the infinitive as imperative

Modification of markers

- Variation in the writing system

多分 vs たぶん

- Variation according to politeness

行かなければなりません

行かなければいけない
行かなきゃいけません
行かなきゃだめ
行かなきゃ

Automatic annotation

Objectives

- Automatise the annotation of the corpora
- Same procedure for both languages
- Inputs a raw text, outputs a XML

Design of the program

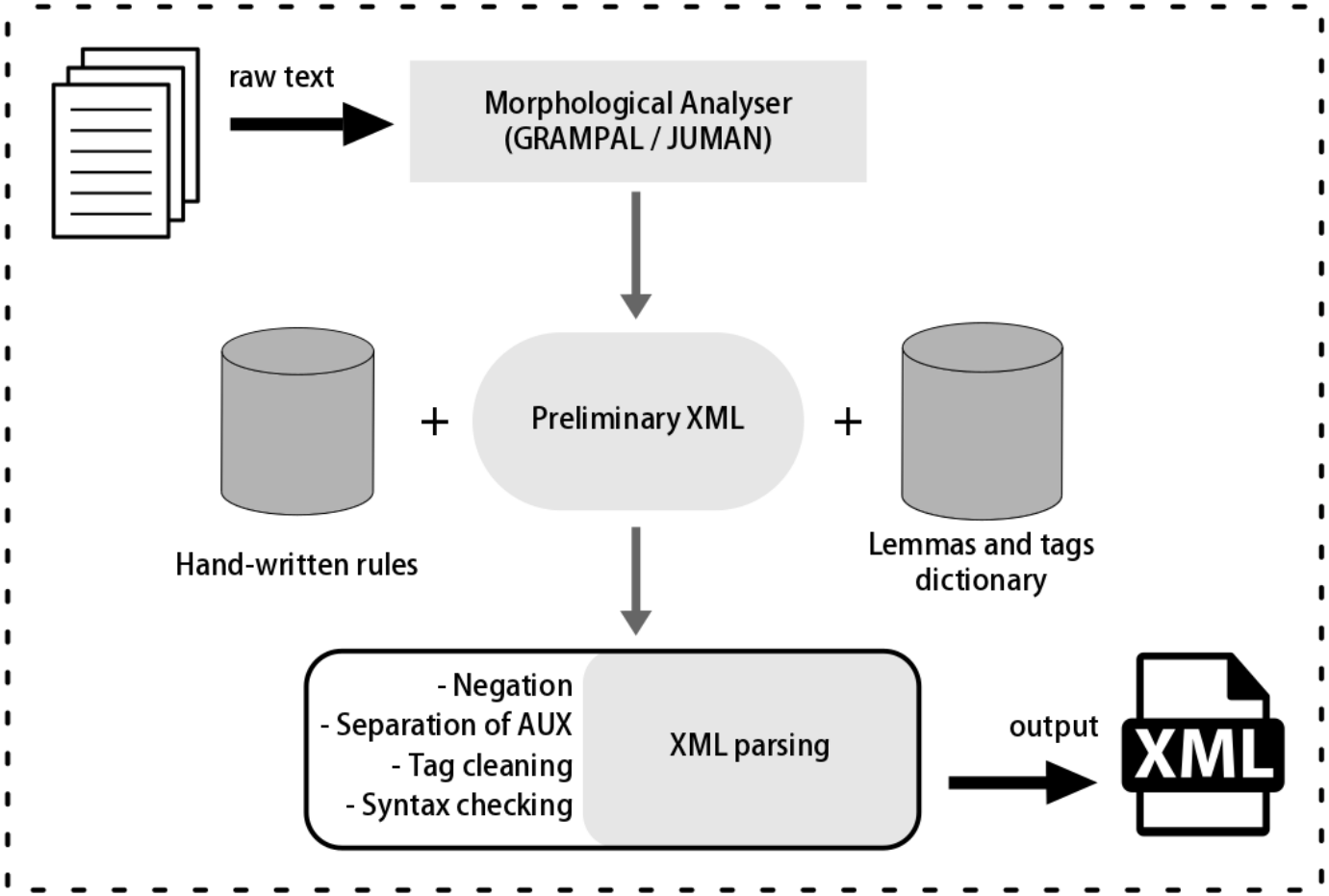
Mañana a lo mejor llueve

Modality: Necessity
Subtype: Epistemic
Class: Adverb
Negated: No
Value: 50%

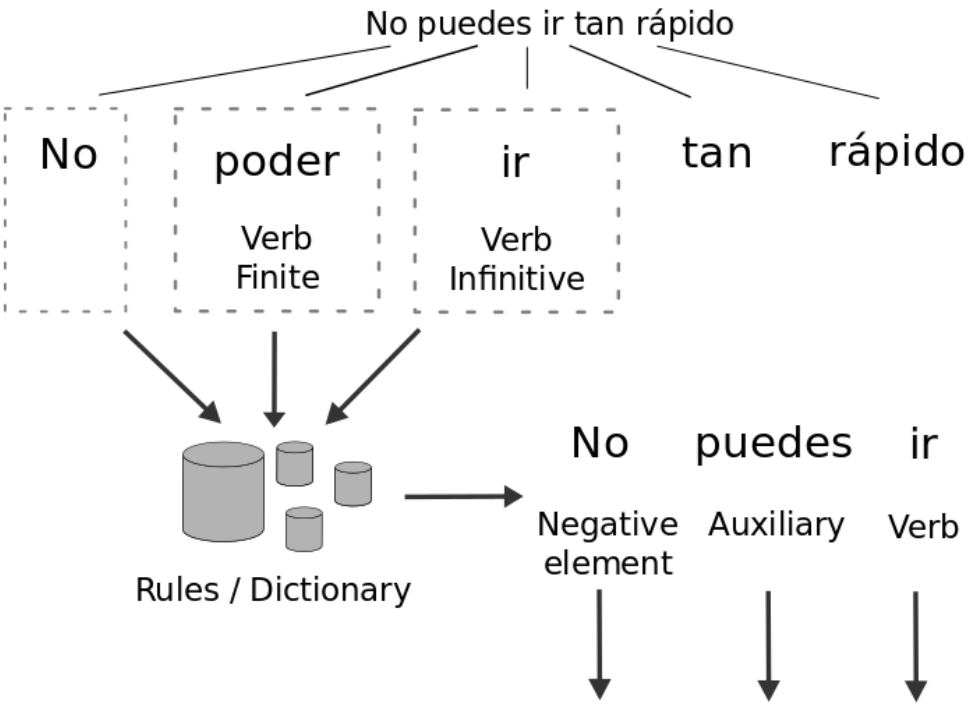
明日は多分雨が降るだろう

Modality type: Necessity
Subtype: Epistemic
Class: Auxiliary
Negated: No
Value: 50%

Design of the program



Spanish program



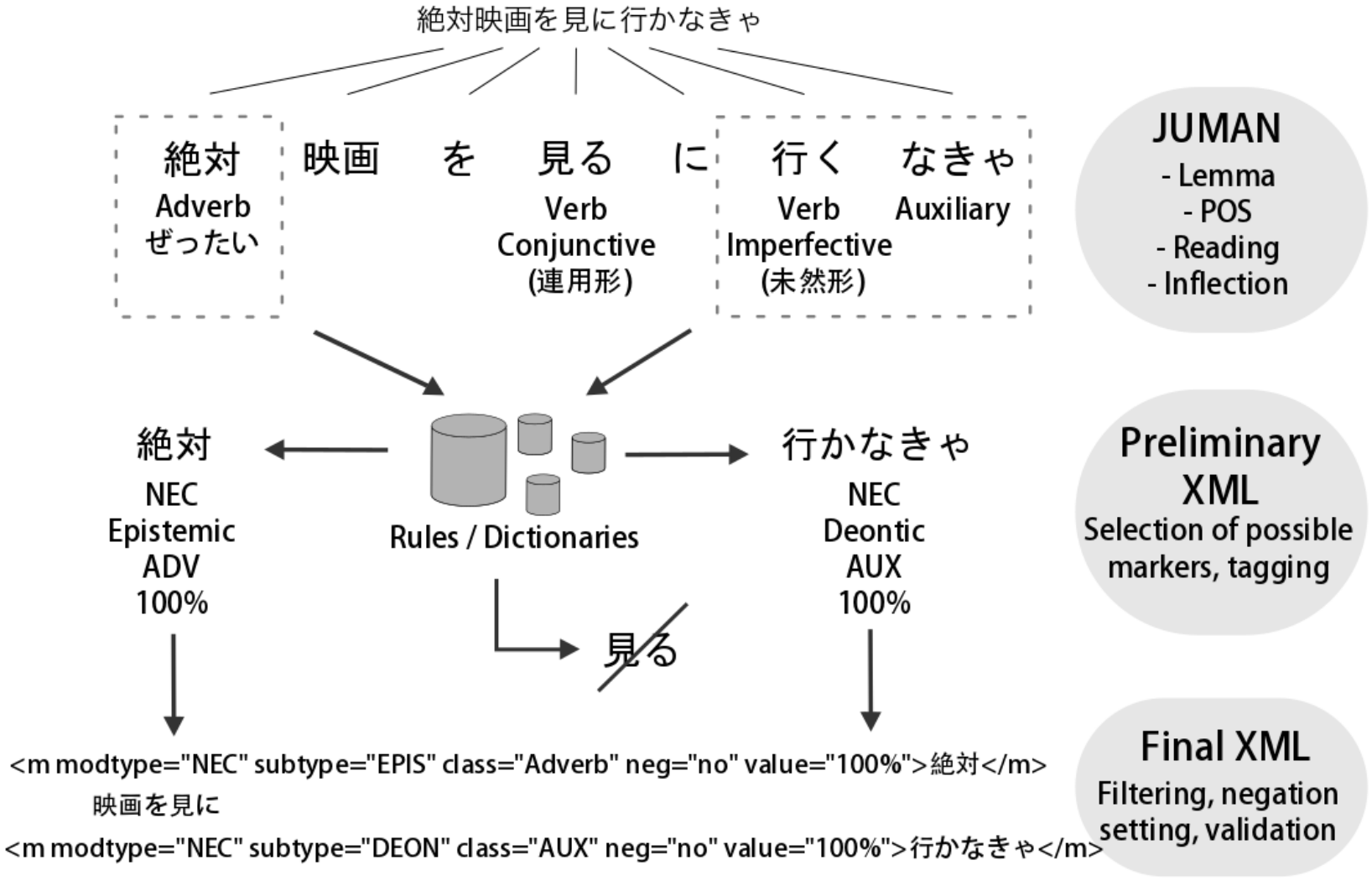
GRAMPAL
- POS
- Inflection

Preliminary XML
Selection of possible markers, tagging

Final XML
Filtering, negation setting, validation

```
<w neg="yes">No</w>
<m modtype="NEC" subtype="DEON" class="mood_SUBJ"
neg="yes" value="0%">vayas</m>
```

Japanese program



Examples

Input	Output
Quizás lo retrasen un poco	<text> <s> <m class="Adverb" modtype="POSS" subtype="EPIS" neg="no" value="70%"> Quizás</m> lo retrasen un poco. </s> </text>
結構見られない	<text> <s> 結構 <m class="mood_POT" modtype="NEC" neg="yes" subtype="DEON" value="0%"> 見られない </m> </s> </text>

Conclusions

- About modality
  - A dual selection between Necessity and Possibility allows us an objective handling of modality avoiding ambiguity.
  - Using a syntax and logic-based approach can be easily formalised into rules.
  - Allows us to perform a cross-linguistic study.
  - Can deal with negation.

Conclusions

- Corpus study
  - Modality is significally related to type of interaction, social restrictions.
  - Necessity used freely in Spanish, possibility similar in both languages.
  - High level of ambiguity in Spanish, makes the Epistemic/Deontic classification less reliable.

Conclusions

- Automatic processing
  - Two very different languages: the program must adapt to the different challenges.
  - Multiword expressions are the most problematic. Separation and ellipsis is not very high, but may decrease precision of the tagger.
  - Negation is very frequent and must be taken into account for its role in changing the classification.

Future work

- Modality classification

Include more markers, iteraction with past tense, interrogatives.

- Corpus

Further studies in different discourses.

- Automatic processing

Evaluation of the program.

Thank you!

carlos.herrero@uam.es