Abar-Hitz: An Annotation Tool for the Basque Dependency Treebank

Díaz de Ilarraza A., Garmendia A., Oronoz M.

IXA Research Group on NLP (http://ixa.si.ehu.es)
Faculty of Computer Science - University of the Basque Country

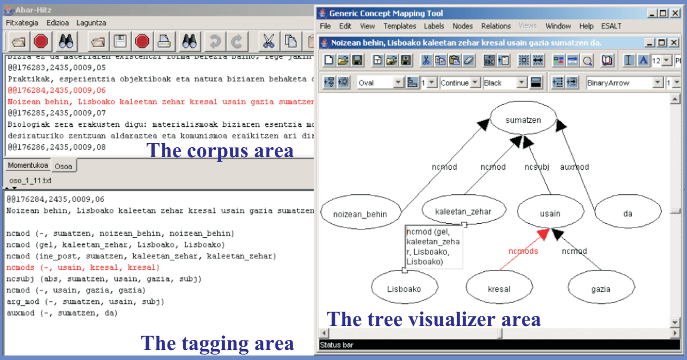
Introduction

- Objective: creation and management of the

  Basque Dependency Treebank

- Main characteristics:
  - Makes the annotation process faster
  - Avoids possible mistakes
  - Implemented in Java and multiplatform
  - Friendly interface and language independent
  - Main areas:

    The corpus area
    The tagging area
    The tree visualizer area



```
@@176283,2435,0009,05
Praktikak, esperientzia objektiboak eta natura biziaren behaketa c
@@176284,2435,0009,06
Noizean behin, Lisboako kaleetan zehar kresal usain gazia sumatzen
@@176285,2435,0009,07
Biologiak zera erakusten digu: materialismoak biziaren esentzia mo
desiraturiko zentzuan aldaraztea eta komunismoa eraikitzen ari dir
@@176286,2435,0009,08
```

The corpus area

```
@@176284,2435,0009,06
Noizean behin, Lisboako kaleetan zehar kresal usain gazia sumatzen

ncmod (-, sumatzen, noizean_behin, noizean_behin)
ncmod (gel, kaleetan_zehar, Lisboako, Lisboako)
ncmod (ine_post, sumatzen, kaleetan_zehar, kaleetan_zehar)
ncmods (-, usain, kresal, kresal)
ncsubj (abs, sumatzen, usain, gazia, subj)
ncmod (-, usain, gazia, gazia)
arg_mod (-, sumatzen, usain, subj)
auxmod (-, sumatzen, da)
```

The tagging area

The tree visualizer area

The Basque Dependency Treebank

- General project: annotation of corpora at syntactic, semantic and pragmatic levels in Catalan, Spanish and Basque (http://www.dlsi.ua.es/projectes/3lb)
- Grammatical relations specifying dependencies between modifiers and their nucleus
- Tagset:
  - Adaptation of Carroll *et al.* (1998, 1999)
  - Difference: Arguments not lexicalised (phonetically empty *pro*)

The Corpus

- Name: Eus3LB
- Characteristics: standard written Basque
- Already tagged:
  - 25.000 words from EPEC (Aduriz *et al.*, 2003)
  - 25.000 word-forms from newspapers (equivalent to Catalan and Spanish)
- Near future: 300.000 word-forms

Abar-Hitz

Previously analysed tools

Annotation tools:
- *WordFreak* (Morton and LaCivita, 2003)
  - ➤ *Our annotation formalism not supported*

Tree management tools:
- *Treetrans* (AGTK) (Bird *et al.*, 2002)
  - ➤ *Based in constituents*
- *TrEd* (Prague Dependency Treebank)
  - ➤ *Dependency tags in nodes (as attribute) but not in the connectors between nodes*
- *The Graph Tree Editor Tool*
  - ➤ *Dependency tags in nodes but not in the connectors between nodes*
- TreeScape
  - ➤ *Draws not editable trees*
- *CM-ED* (Arruarte *et al.*, 2001*)*
  - ➤ *Concept map editor adapted into ESALT, a tree visualizer that follows a dependency-based formalism*

Example

Noizean behin, Lisboako kaleetan zehar kresal usain gazia sumatzen da.

"From time to time, the salty scent of seawater can be perceived in the streets of Lisbon"

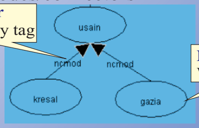
The Interface

The corpus area

Two formats:
i) Whole file (in first figure)
ii) Sets of 3 sentences in context

The tree visualizer area

- ESALT interpretes the relation tags and draws the tree
- Checks possible errors and it marks them in red. If an error is corrected in the tree, the correction is carried to the tagging area
- Manipulation of the tree
  - Change of tags and fields
  - Roll up of subtrees
  - Removal/addition of nodes/connectors



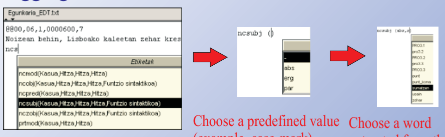
Connector
dependency tag

Node
Word or multiword

The tagging area

Two options:

- Tagging of a new sentence from raw text:



```
@@00,06,1,0000600,7
Noizean behin, Lisboako kaleetan zehar kres
ncs
```

```
ncmod(Kasua,Hitza,Hitza,Hitza)
ncobj(Kasua,Hitza,Hitza,Hitza,Funtzio sintaktikoa)
ncpred(Kasua,Hitza,Hitza,Hitza)
ncsubj(Kasua,Hitza,Hitza,Hitza,Funtzio sintaktikoa)
nczobj(Kasua,Hitza,Hitza,Hitza,Funtzio sintaktikoa)
prtmod(Kasua,Hitza,Hitza)
```

```
ncsubj ()
```

```
ncsubj (abs,s
```

Choose a dependency tag

Choose a predefined value (example, case-mark)
Definition in an XML doc.

Choose a word extracted from the sentence

Avoids mistakes and saves time

- Revision of an already annotated corpus:
  - When opening a sentence, the correctness of the tags and slots is automatically checked by the ✓ button and mistakes are marked in red
  - Some results when revising 181 tagged sentences:

Sentences	Mistakes	Total	Percents
Wrong	Label	30	16,57 %
	Number of slots	12	6,63%
	Label + Number of slots	10	5,52%
	Total	52	28,73%
Correct		129	71,27%
Total		181	100,00%

Mistakes that can be avoided

- Wrong number of slots in a concrete dependency tag

  (e.g., *assigning 4 slots to a 'ncsubj' which needs 5)*
- Wrong type of slot

  (e.g., giving a word instead of a case-mark)
- Misspell the name of the tag or the word-form

  (e.g., writing 'ncmods' instead of 'ncmod')

  (e.g., writing 'usan' instead of 'usain')

Conclusions and Future work

Conclusions

- Makes the annotation process faster and avoids mistakes
- Massively used by three linguists in the annotation of a treebank of 50.000 word-forms
  - One half of the corpus revised with Abar-Hitz
  - The other half tagged with Abar-Hitz

Future Work

- EULIA (Artola *et al.*, 2004) a tool for creating, consulting, visualizing and modifying documents in XML will be integrated in Abar-Hitz
- Abar-Hitz will give the output, the syntactic analysis, in an XML document that will be compared to the document produced by the parser