Learner and Error Corpora Based Computational Systems

University of the Basque Country

Itziar Aldabe, Leire Amoros, Bertol Arrieta, Arantza Díaz de Ilarraza, Montse Maritxalar, **Maite Oronoz**, and **Larraitz Uria**

Learner and Error Corpora Based Computational Systems

- Introduction
- The error classification
- The ‘error editor’ tool
- The ERREUS web application
- The IRAKAZI web application
- Union of two databases
- Conclusions and future work

Learner and Error Corpora Based Computational Systems

- **Introduction**
- The error classification
- The ‘error editor’ tool
- The ERREUS web application
- The IRAKAZI web application
- Union of two databases
- Conclusions and future work

Introduction

- We present learner and error corpora based computational systems for Computer-aided Error Analysis (CEA)
- Computer-aided Error Analysis (CEA):

  "a new type of computer corpus annotation"
- Computer Learner Corpus (CLC), definition:

  "collection of machine-readable natural language data produced by L2 learners" (Dagneaux *et al*., 1998)
- We have a broader view: not just learners' deviations but any error instance (for NLP)

Introduction

Our aims:

- to gather from real corpora error instances for the development of new Natural Language Processing (NLP) tools (e.g. Basque Grammar Checker)

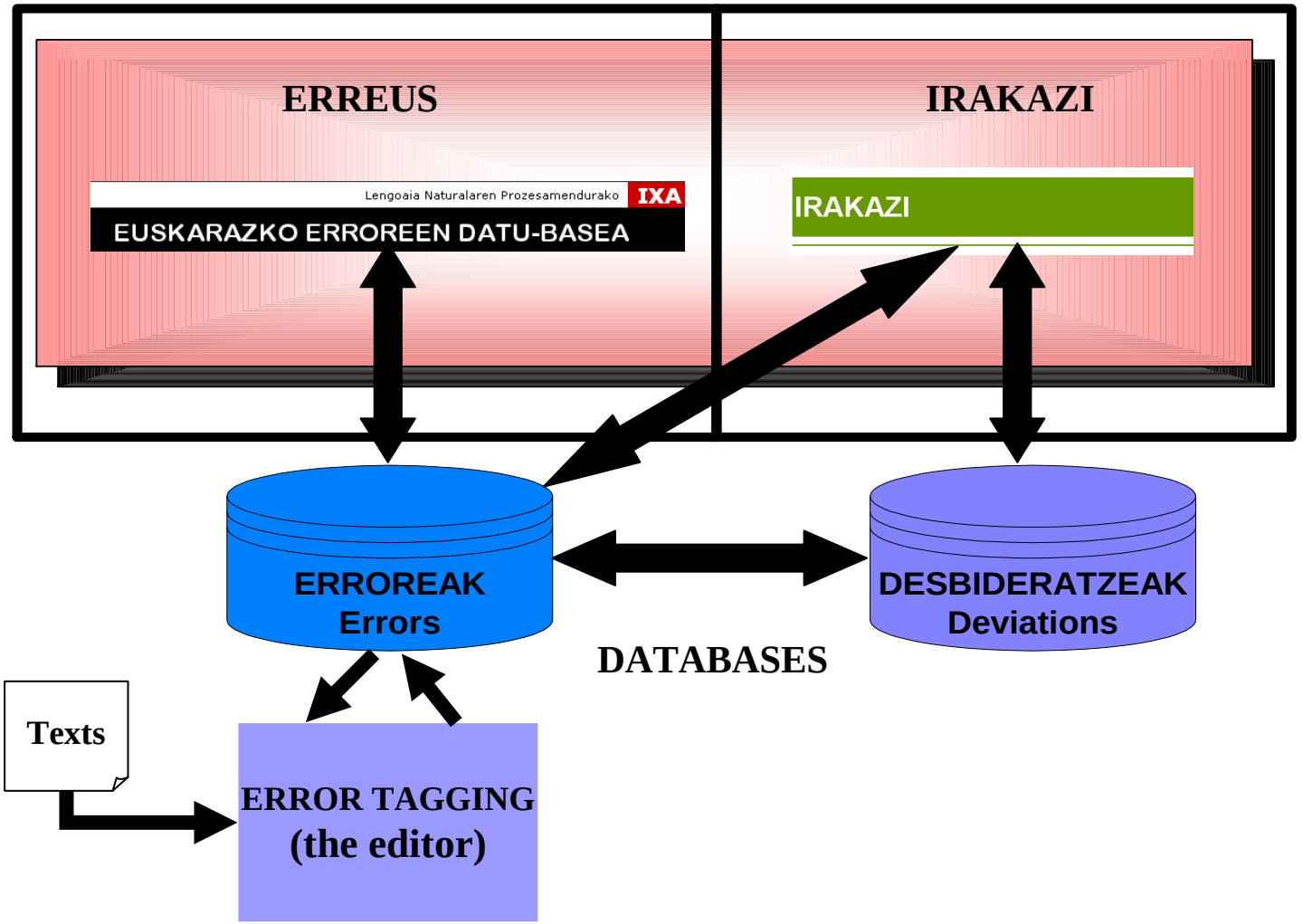
- to store Basque learner corpora for further systematic studies in the field of Computer Assisted Language Learning (CALL) and Second Language Learning (SLL)

Introduction

Computer-aided learner and error corpus research:

- is growing fast; is an interesting field
- overcomes some of the limitations attributed to Error Analysis (EA) and has many advantages
- makes error collection and analysis faster
- offers many possibilities: count errors, retrieve lists of specific error types, view errors in context…
- makes contributions to different fields (SLL, NLP…)

Introduction



Learner and Error Corpora Based Computational Systems

- Introduction
- **The error classification**
- The ‘error editor’ tool
- The ERREUS web application
- The IRAKAZI web application
- Union of two databases
- Conclusions and future work

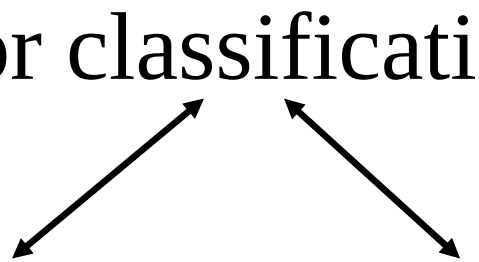
The error classification

- Objective:

  the error instances we gather to be organised and classified for computer-aided error analysis and treatment

The error classification

- worked on automatic error detection for developing a robust spelling checker, XUXEN (Aduriz *et al*., 1992)
- planned a grammar checker approach
- importance of collecting error corpora



The error classification

Based on different sources:

- IXA group's previous experience in CEA (Maritxalar 99; Gojenola 00)
- Basque grammars (Alberdi et al., 2001; Zubiri 1995)
- Language programs (Basque language schools)
- Corpus data
- Other classifications (Basque and other languages)
  - *A Grammar of Basque* (J. Ortiz de Urbina, 2001)
  - *A Brief Grammar of Euskera* (I. Laka)
  - Becker *et al*. (1999)
  - *Language learners and their errors* (Norrish, 1983)
  - *Errors in Language Learning and Use: Exploring Error Analysis* (C. James, 1998)
  - *Interlengua y Análisis de Errores* (S. Fernández, 1997)

The error classification

- Dynamic (modified depending on the corpora we collect)
- Hierarchical

1.- **Spelling** } omission, addition, misformation, misordering
automatically detectable

2.- **Lexis**
3.- **Morphology, syntax and morphosyntax**
4.- **Notions**
5.- **Semantics**
6.- **Punctuation marks**
7.- **Style**

} - general linguistic categories
- specific subcategories
(verbs, pronouns, declensions, etc.)

The error classification (example)

1.- Spelling errors

1.1.- Missing words:

1.1.2.- **LEKHH** (missing **H** in the beginning**)**: *erri

1.1.7.- **LEKTRE** (missing **RE** in the middle): *soldatakin

3.- Morphological, syntactic and morphosyntactic errors

3.1.- Declension:

3.1.1.- **DEKL** (wrong declension**)**: *Jonen autoaz etorri naiz

3.2.- Determinant:

3.2.2.- **DETLMG** (adding finite determiner when not necessary): *zer ordua da?

The error classification (example)

1.- Spelling:

1.1.- Letrak kendu (hitzari letra bat kendu):

1.1.1.- **LEKHE** (**LE**tra **K**endu **H**asieran, **E**): *ta

1.1.2.- **LEKHH** (**LE**tra **K**endu **H**asieran, **H**): *erri

1.1.3.- **LEKTA** (**LE**tra **K**endu **T**artean, **A**): *Donostitik

1.1.4.- **LEKTE** (**LE**tra **K**endu **T**artean, **E**): *oihanan

1.1.5.- **LEKTD** (**LE**tra **K**endu **T**artean, **D**): *euki

1.2.- Letrak gehitu (hitzari letra bat gehitu):

1.2.1.- **LEGHH** (**LE**tra **G**ehitu **H**asieran, **H**): *harrisku

1.2.2.- **LEGTA** (**LE**tra **G**ehitu **T**artean, **A**): *institutoako

1.2.3.- **LEGTE** (**LE**tra **G**ehitu **T**artean, **E**): *euskalduneen

1.2.4.- **LEGTI** (**LE**tra **G**ehitu **T**artean, **I**): *laister

1.2.5.- **LEGTZ** (**LE**tra **G**ehitu **T**artean, **Z**): *jasatzen

1.3.- Letrak ordezkatu (letrak ordezkatu):

1.3.1.- **LEOEAE** (**LE**tra **O**rdezk., **A-E**): *erreztasuna

1.3.2.- **LEOEEA** (**LE**tra **O**rdezk., **E-A**): *ospatsuak

1.3.3- **LEOEEI** (**LE**tra **O**rdezk., **E-I**): *erlien

1.3.4.- **LEOEIE** (**LE**tra **O**rdezk., **I-E**): *nere

1.3.5.- **LEOEOU** (**LE**tra **O**rdezk., **O-U**): *freskua

2.- Lexis:

2.1.- Maileguak: (beste hizkuntza batetik hartutako hitza)

2.1.1.- **LMA-M** (**L**e**MA**, **M**ailegua): *afamatu; moskeatu

2.2.- Konposizioa eta Eratorpena: (hitzak asmatzea)

2.2.1.- **KONERA** (**KON**posizioa-**ERA**torpena):*hautoki

2.3.- Generoa: (generoa gaizki erabiltzea)

2.3.1.- **GENE** (**GENEroa gaizki**): *aitak etorri dira

2.4.- Esamoldeak eta Kolokazioak

2.4.1.- **ESAMOL** (**ESAMOL**deak): *lur eta zur

2.5.- Bestelakoak

3.- Morphological, syntactic and morphosyntactical:

3.1.- Deklinabidea: (deklinabide erroreak)

3.1.1.- **DEKL (DEKL gaizki)**: *Jonen autoaz etorri naiz

3.1.2.- **INS** (**INS**trumentaltasuna): *haizkorarekin ebaki dugu

3.2.- Determinantea: (determinanteen erabilera okerra)

3.2.1.- **DETMK (DET M**ugatzailea **K**endu): *txokolate nahi dut

3.2.2.- **DETMG (DET M**ugatzailea **G**ehitu): *zer ordua da?

3.7.- Aditza: (aditzen erabilera okerra)

3.7.1.- **ADAM** (**D**enbora,**A**spektua,**M**odua): *goaz mendira?

3.7.2.- Aditz-paradigmen nahasketa: (aditz-paradigmak nahastea)

3.7.2.1.- **PARADIG_N_N-NK (...)**: *ez da funtzionatzen

3.7.2.2.- **PARADIG_N_N-NI (...)**: *nagusiari zuzendu da

3.7.2.3.- **PARADIG_N-NI_N-NK (...)**: *niri hori ez zait molestat

3.7.2.4.- **PARADIG_N-NK_N-NI-NK (...)**: *Joni ikusi diot

3.7.2.5.- **PARADIG_N-NI_N-NI-NK (...)**: *gustatzen dit

3.8.- Komunztadura (komunztadura erroreak)

3.8.1.- **KOMSIN** (**SIN**tagma barruko **KOM**unztadura eza): *guk geu

3.8.2.- **KOMAPOS** (**APO**sizioan): *zure laguna, Dublinen bizi denari

3.8.3.- **KOMP** (**KOM**unztadura eza **P**erpausean)

3.8.3.1.- Aditza – Subjektua

3.8.3.1.1.- **KOMPAS-NUM (...)**: *aurrerapen haundia daude

3.8.3.1.2.- **KOMPAS-KAS (...)**: *zuk etorri zara

3.8.3.2.- Aditza – Objektua

3.8.3.2.1.- **KOMPAO (...)**: *eman dizut liburuak

3.8.3.3.- Aditza – Zehar-objektua

3.8.3.3.1.- **KOMPAZO (...)**: *pertsonei dagokion izena da

3.8.3.4.- Aditza – Predikatua

3.8.3.4.1.- **KOMPAP (...)**: *gure erleak oso soziablea dira

The error classification

- Classifying and categorising errors not an easy task: errors can be classified in more than one category

  → subjective

- A handout with the decisions made to make the classifying easier and coherent
- Assessment of the classification; tested by some:

- Proofreaders
- Basque language teachers
- Linguists of the IXA group

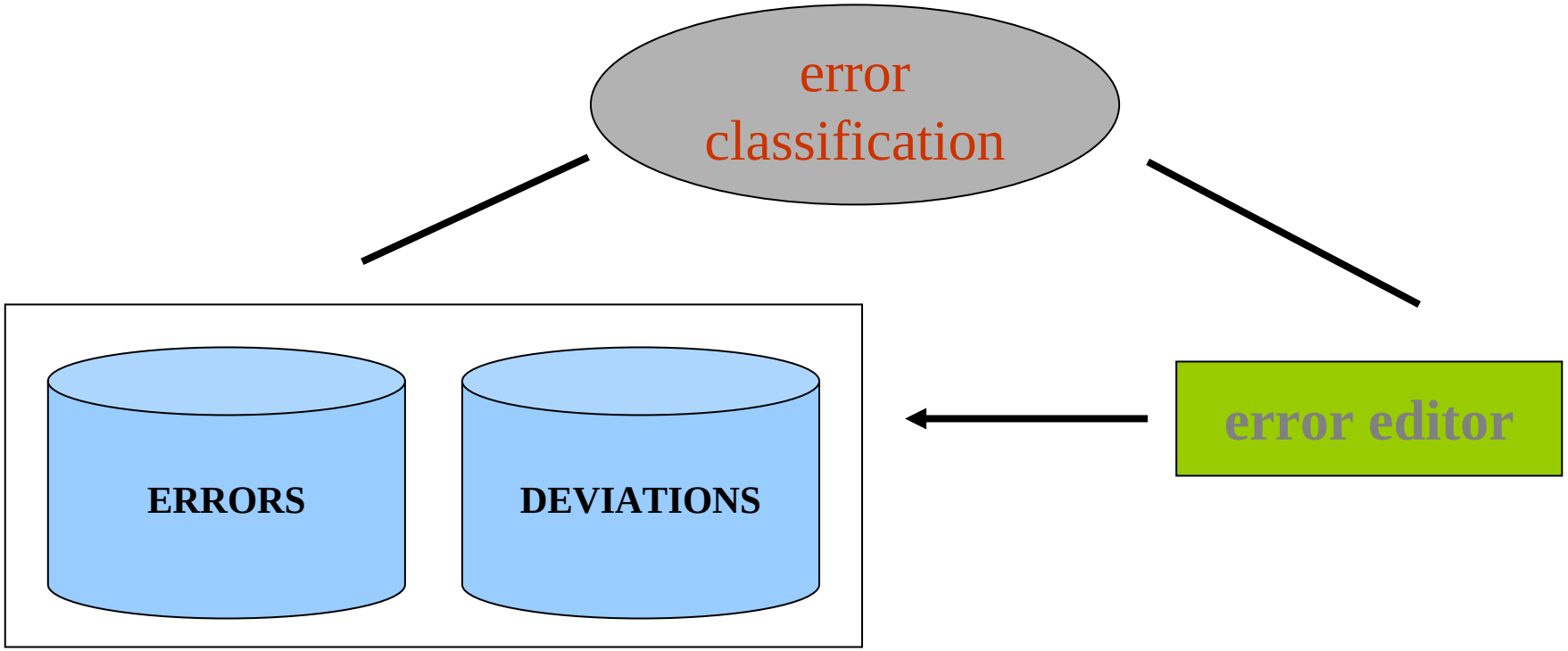
→ Accurate Complete

- The error classification is accessible via Internet (from ERREUS and IRAKAZI)

Learner and Error Corpora Based Computational Systems

- Introduction
- The error classification
- **The ‘error editor’ tool**
- The ERREUS web application
- The IRAKAZI web application
- Union of two databases
- Conclusions and future work

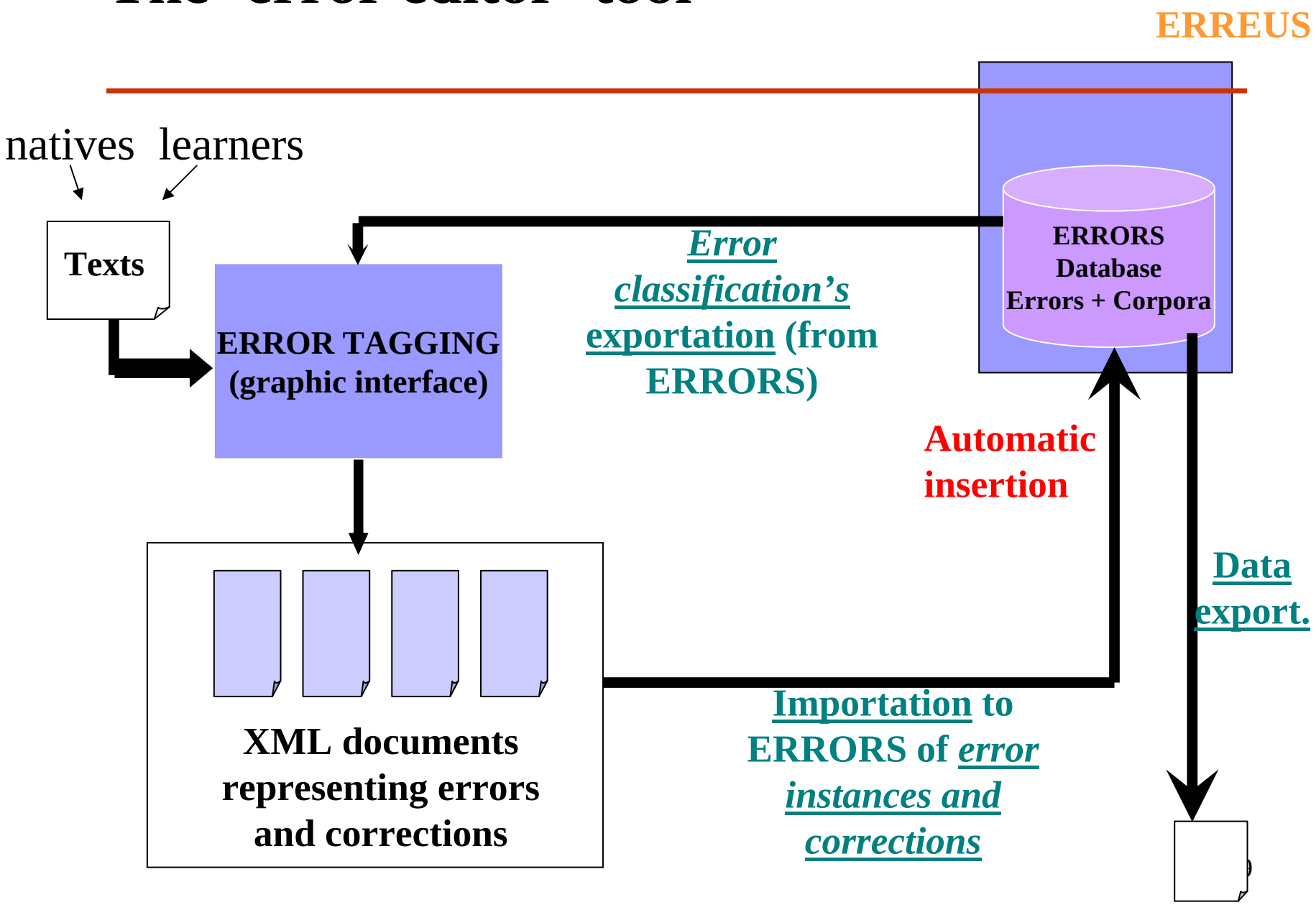
The ‘error editor’ tool



The ‘error editor’ tool

- Purpose: to tag and classify errors and their possible corrections in an easy, fast and intuitive way
- Not in use yet, but the design is finished
- A user-friendly interface
- The error classification available at any time
- Exportation/importation of data between the editor and the databases
- The results of the tagging represented in XML documents

The ‘error editor’ tool



The ‘error editor’ tool

***Nire bizilagunek ondartzara nirekin joan nahi du, aspertuta bait dago.**

1 2 1 3

***My neighbours wants to come to the bech with me, be cause he is bored.**

1 1 2 3

1 no- agreement between the subject and the verb

2 spelling error

3 spelling error

The ‘error editor’ tool

* Nire bizilagunek ondartzara nirekin joan nahi du, aspertuta bait dago.

* My neighbours wants to come to the bech with me, be cause he is bored.

```
<p id='linkGrp'>
<linkGrp type="w-err" tagOrder="yes">
 1 <link id="errInst1" targets="Xw1 Xw2 Xw3 Xw4 Xw5 Xw6 Xw7"/>
 2 <link id="errInst2" targets="Xw3"/>
 3 <link id="errInst3" targets="Xw9 Xw10"/>
</linkGrp>
```

.err.xml

The ‘error editor’ tool

*** My neighbours wants to come to the bech with me, be cause he is bored.**

```
<p id='linkGrp'>
<linkGrp type="w-err" tagOrder="yes">
 1 <link id="errInst1" targets="Xw1 Xw2 Xw3 Xw4 Xw5 Xw6 Xw7"/>
 2 <link id="errInst2" targets="Xw3"/>
 3 <link id="errInst3" targets="Xw9 Xw10"/>
</linkGrp>
```

.err.xml

```
<p id='linkGrp'>
<linkGrp type='w-err' tagOrder='y'>
 1 <link targets='Xerr20 XerrInst1'/>
 2 <link targets='Xerr7 XerrInst2'/>
 3 <link targets='Xerr26 XerrInst3'/>
</linkGrp>
```

.errlnk.xml

The ‘error editor’ tool

*** My neighbours wants to come to the bech with me, be cause he is bored.**

```
<p id='linkGrp'>
<linkGrp type="w-err" tagOrder="yes">
 1 <link id="errInst1" targets="Xw1 Xw2 Xw3 Xw4 Xw5 Xw6 Xw7"/>
 2 <link id="errInst2" targets="Xw3"/>
 3 <link id="errInst3" targets="Xw9 Xw10"/>
</linkGrp>
```

.err.xml

```
<p id='linkGrp'>
<linkGrp type='w-err' tagOrder='y'>
 1 <link targets='Xerr20 XerrInst1'/>
 2 <link targets='Xerr7 XerrInst2'/>
 3 <link targets='Xerr26 XerrInst3'/>
</linkGrp>
```

.errlnk.xml

```
<p id='linkGrp'>
<linkGrp type='w-err-zuz' tagOrder='y'>
 1 <link targets='XerrInst1 Xcorrect1'/>
 1 <link targets='XerrInst1 Xcorrect2'/>
 2 <link targets='XerrInst2 Xcorrect3'/>
 3 <link targets='XerrInst3 Xcorrect4'/>
</linkGrp>
```

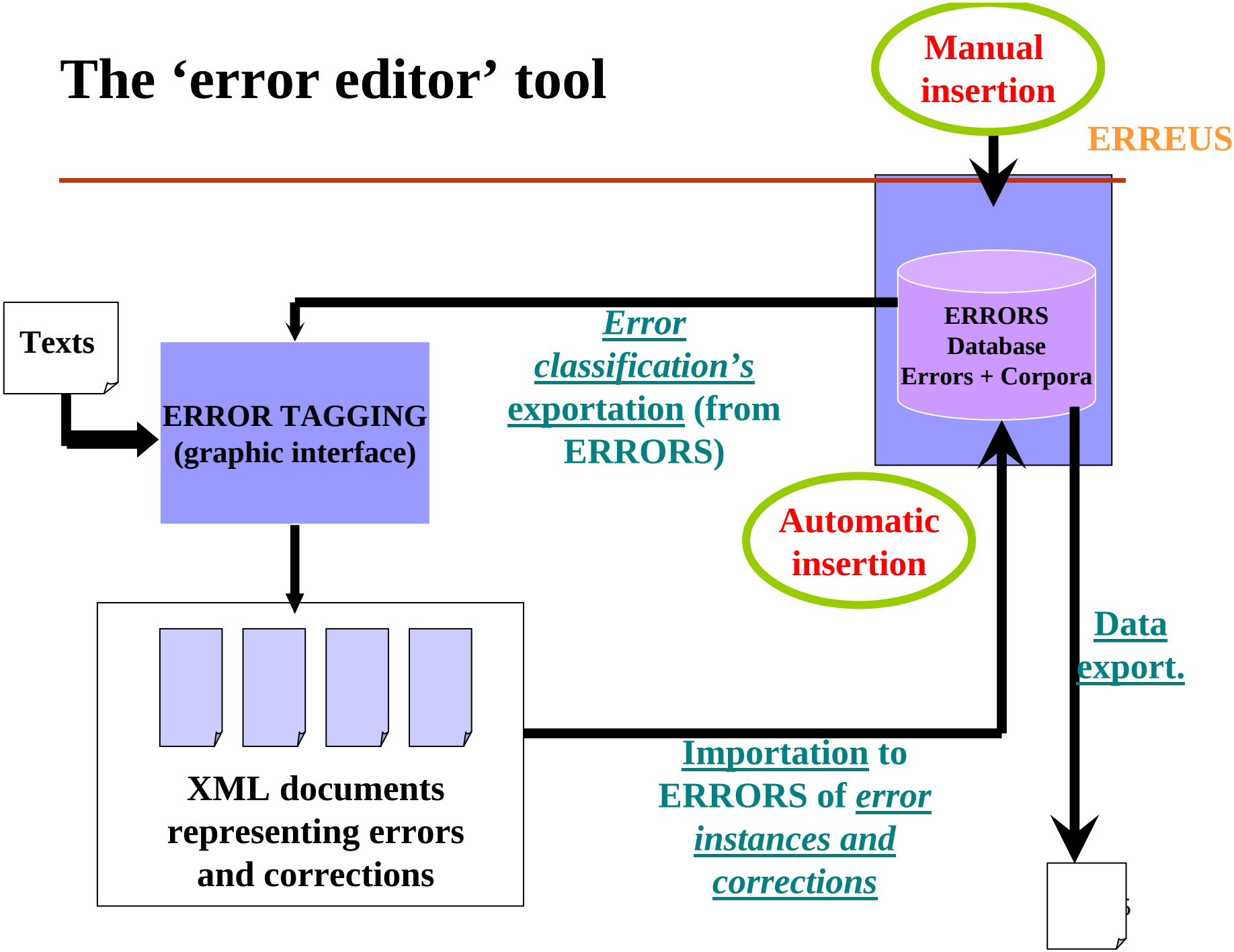
.errzcorrectlnk.xml

The ‘error editor’ tool

```
<body>        *My neighbours wants to come to the bech with me be cause he is bored.
    <p>
     <fs id="correct1" type="correctionType">
      <f name="description">
           <str>Singular subject to agree with the verb.</str>
          </f>
          <f name="correction"><str>My neighbour wants to come to the bech
                                    with me be cause he is bored </str></f>
     </fs>
    </p>
    <p>
     <fs id="correct2" type="correctionType">
      <f name="description">
           <str> Plural subject to agree with the verb.</str>
          </f>
      <f name="correction"><str> My neighbours want to come to the bech
                                    with me be cause he is bored </str></f>
     </fs>
   </p>                    …
```

.correct.xml

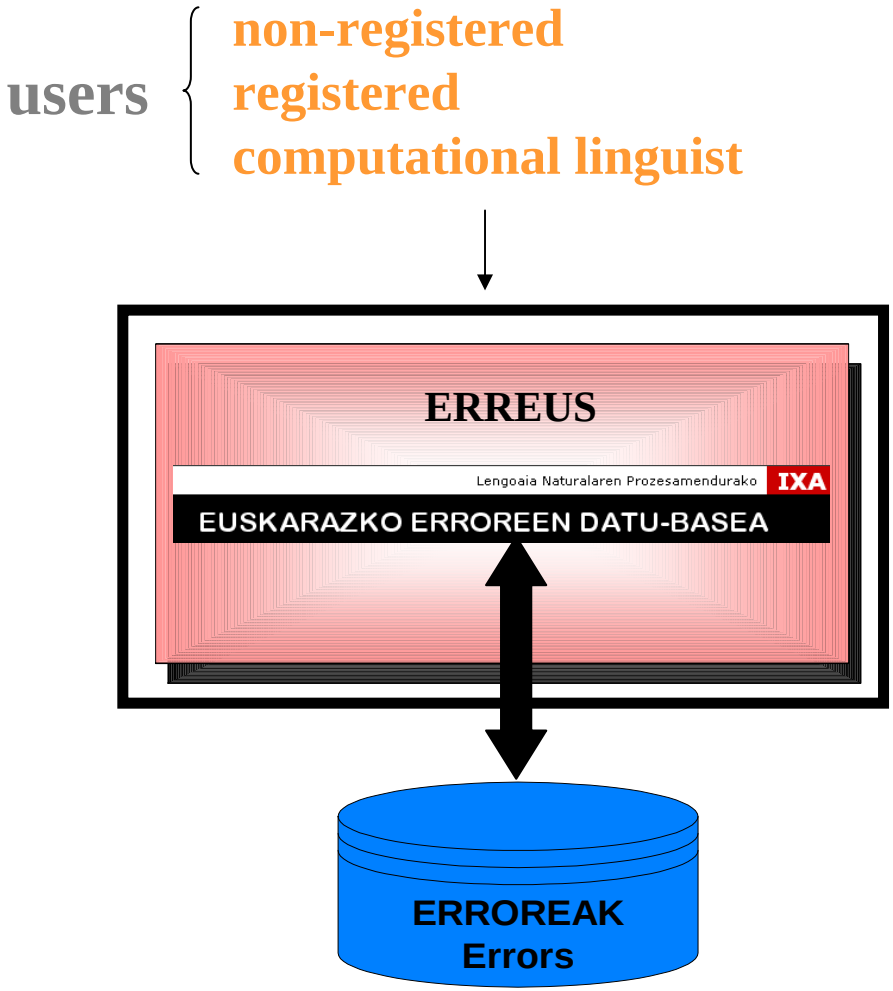
The ‘error editor’ tool



Learner and Error Corpora Based Computational Systems

- Introduction
- The error classification
- The ‘error editor’ tool
- **The ERREUS web application**
- The IRAKAZI web application
- Union of two databases
- Conclusions and future work

The ERREUS web application



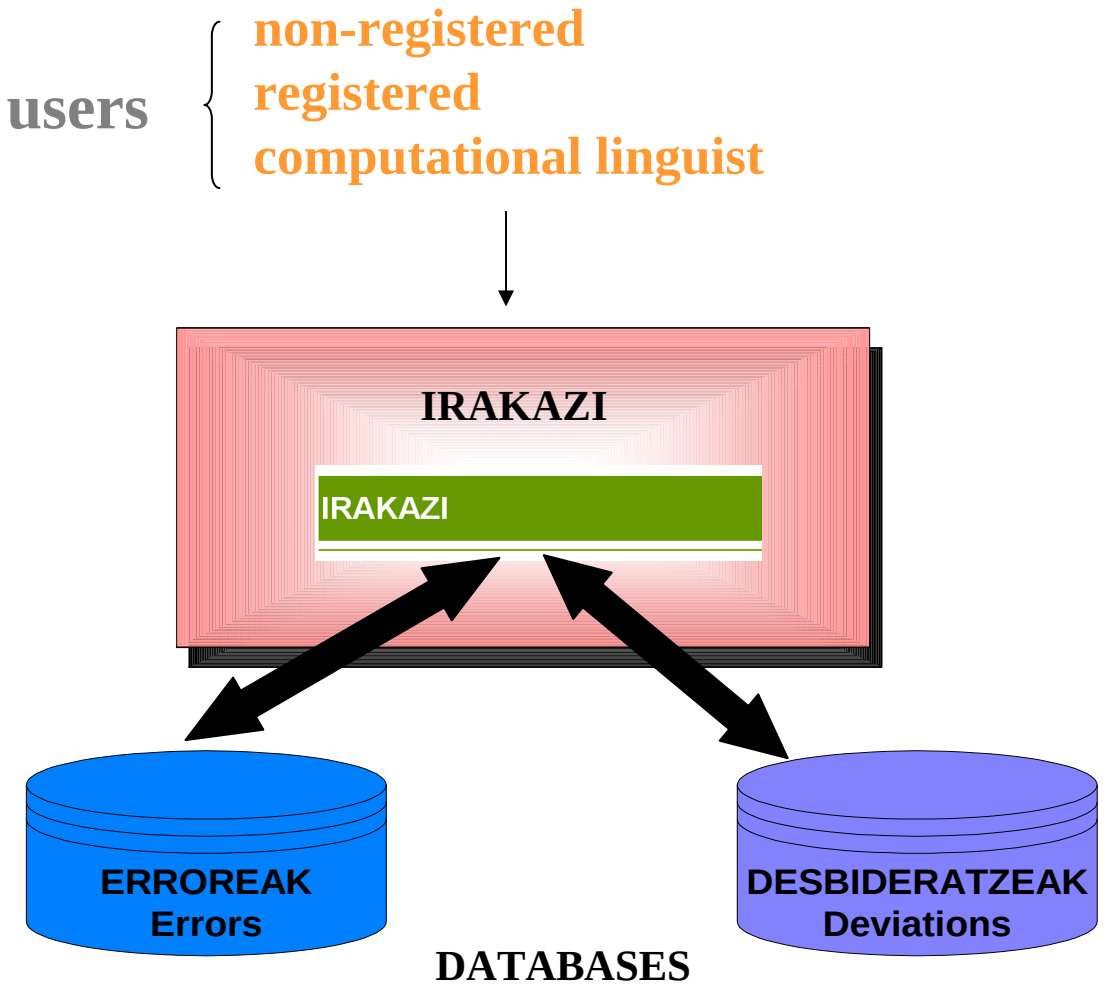
The ERREUS web application

- Aim: to store technical and linguistic information about ungrammatical instances for automatic error treatment and, more specifically, to be a repository of error corpora for the development of a robust grammar and style checker

Learner and Error Corpora Based Computational Systems

- Introduction
- The error classification
- The ‘error editor’ tool
- The ERREUS web application
- **The IRAKAZI web application**
- Union of two databases
- Conclusions and future work

The IRAKAZI web application



The IRAKAZI web application

- Designed within an ICALL environment for storing:
  - specific data about learners: age, language background, language level, learning context, etc.
  - students' deviations (in the error classification)

Important source of the psycholinguistic information of learners' interlanguage

- to better understand the process of SLL/FLL
- for further research in this field
- to provide language learners:
  - a more individual help
  - more appropriate tools

taking into account their specific needs and difficulties

Learner and Error Corpora Based Computational Systems

- Introduction
- The error classification
- The ‘error editor’ tool
- The ERREUS web application
- The IRAKAZI web application
- **Union of two databases**
- Conclusions and future work

Union of two databases

- **ERRORS**: a repository of error corpora to develop a robust grammar and style checker (technical and linguistic information of ungrammatical instances)

- **DEVIATIONS**: a rich source of psycholinguistic information of Basque learners' learning process and their deviations

DEVIATIONS database

Deviation

-Sentence: *Hura igeriketa maite du *(*He love swimming)*

-Category: AGREEMENT SUBJ VERB

-Deep reason: Generalization of a rule

Text

-Number of words: 245

-Type of text: exercise

-Reference: 122

Student

Name: Ana Berazadi

-Age: 25

-School: Iazki

-Language knowledge level:

+Spanish (speak,understand,write, read): 5, 5, 5, 5

+French (speak,understand,write, read): 3, 3, 3, 4

-Mother language: Spanish

-Learning history

ERRORS database

Error containing text

-Sentence: *Hura igeriketa maite du *(*He love swimming)*

-Correction: Hark igeriketa maite du *(He loves swimming)*

-Text reference: 122

Error

-Description: loss of the letter 's' in third singular person, in present tense

-IsDetected? Yes

-DetectionTool: Constraint Grammar

-IsCorrected? No

-CorrectionTool: -

Category and subcategory levels

-FirstCategory: Morphosyntactic

-SecondCategory: Agreement

-ThirdCategory: Agreement between subject and verb

Union of two databases

- Despite both were built with different purposes:

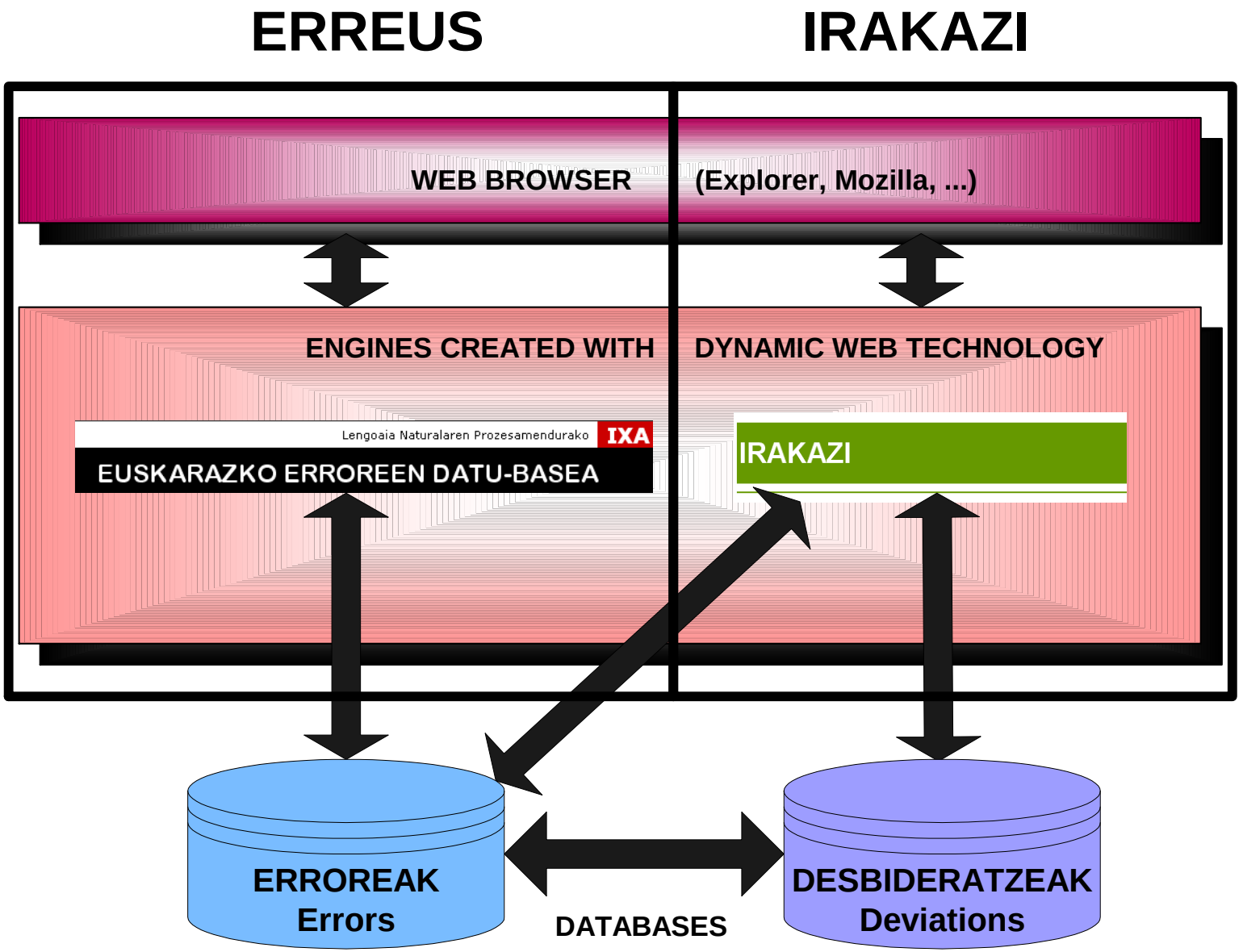
- are focused on computer-aided error analysis and treatment
- share some data concerning the error/deviation, its category and its corresponding corrections

Union of two databases

- Therefore, both databases could be complementarily connected, because:
  - they have some data in common
  - both viewpoints have in the end an underlying equivalence which is compatible
  - we avoid duplicated information
  - maintenance reasons

- For each error-containing-text, we have its technical-linguistic information in the ERRORS database as well as its corresponding psycholinguistic information in DEVIATIONS

Union of two databases



Union of two databases

- Both, ERREUS and IRAKAZI, have two sites: public and private

The *public site*, two sections:

- One is available for those users who just want to consult the stored information
- One is available to add examples and enrich it with new data (registration needed)

Union of two databases

The *private site*, controlled by a computational linguist who:

- verifies that the data introduced via the public site are right
- corrects the examples if necessary
- completes the specific information related to the techniques for automatic error treatment

As the classification is dynamic, it can be modified and updated (from the ERREUS' private site)

Union of two databases

- We considered the union of the two databases very interesting because this way we get a double perspective: the psycholinguistic and computational approaches of each error instance

- Despite we collect data related to Basque errors, the design of the databases is transferable to other languages

Learner and Error Corpora Based Computational Systems

- Introduction
- The error classification
- The ‘error editor’ tool
- The ERREUS web application
- The IRAKAZI web application
- Union of two databases
- **Conclusions and future work**

Conclusions and future work

- ERREUS and IRAKAZI, accessible via Internet to collect as much error instances and information related to them as possible
- The data stored will allow us:
  - to carry out a deeper research on automatic error treatment for the development of a robust grammar and style checker
  - to study Basque learners' interlanguage from authentic learner data → study of Basque learners' learning processes, needs, strategies, materials…
- Possibility to design IRAKAZI and ERREUS as multilingual web applications

Conclusions and future work

Based on the data collected in IRAKAZI, we want to develop a complete and helpful ICALL environment for both Basque teachers and learners:

- HIKAS, a student oriented learning application where NLP tools developed in IXA will be available and adapted to students' needs
- IRAKAZI, improved and adapted to teachers' needs
- students' writings automatic assessment
- a tool for automatic generation of language exercises

The web applications' URLs:

- **ERREUS**:

  http://ixa.si.ehu.es/Erreus

- **IRAKAZI**:

  http://ixa.si.ehu.es/Irakazi

Eskerrik asko!!

Thank you!!