Corpus exploration of discourse relations in RST

Mikel Iruskieta
mikel.iruskieta@ehu.eus

Ixa group for NLP
University of the Basque Country (UPV/EHU)

Valencia, January 18\textsuperscript{th}-22\textsuperscript{nd}, 2016
Structuring Discourse in Multilingual Europe

Training School: Methods and tools
for the analysis of discourse relational devices
PART 1 — Discourse relations in RST: method

PART 2 — Practice

PART 3 — Tools for corpus exploration

PART 4 — Resources
PART 1 — Discourse relations in RST: method

1. Introduction
2. Segmentation
3. Central Unit
4. Rhetorical relations
5. Signals of rhetorical relations
6. Corpora for corpus exploration
7. Applications

PART 2 — Practice

1. Signaling relational structures
2. An ambiguous RST analysis
3. Annotation in RST

PART 3 — Tools for corpus exploration

1. Segmenters
2. CU detector
3. Annotation tools for RST
4. Evaluation tools/methods of RST
5. Parsers

PART 4 — Resources

1. Projects
2. Resources
3. Workshops
About me

- Professor and researcher at **University of the Basque Country**
  - Member of the **ixa group for NLP** (mostly Basque)
    - Researchers from Comp. Science (32), Linguists (13)
    - More than 23 Ph-D, 60 projects, 20 applications
PART 1 — Discourse relations in RST: method

Introduction

About me

- Professor and researcher at **University of the Basque Country**
  - Member of the **Ixa group for NLP** (mostly Basque)
    - Researchers from Comp. Science (32), Linguists (13)
    - More than 23 Ph-D, 60 projects, 20 applications
Basque language (from Wikipedia 2012)

- Native speakers 720,000 out of 3,000,000
- An isolate language (indigenous to the Basque Country $42^\circ 52'55''\text{N} 1^\circ 55'01''\text{W}$). Listen to my Basque dialect
PART 1 — Discourse relations in RST: method

Introduction

Abstract

In the RST framework, there are several discourse-annotated corpora available in different languages, such as: English, Spanish, Brazilian Portuguese, German, and Basque, among others. Some of them can be consulted and several tools have been developed for corpus exploration. There is also a small multilingual aligned RST corpus, which can be explored for getting information about different linguistic phenomena. After the annotation process is over, evaluation is necessary to check reliability (precision and recall). In order to do so, a sound evaluation method and some search tools (which can be used in multilingual corpora) were developed:

i) to study whether the annotators were consistent when looking for the relations or signals in a kwic style,

ii) to check the aligned segments in different languages,

iii) to check a kind of macro-structure of RS-tree looking for the RST relations that are linked to the most salient unit, and

iv) to look for any information in the corpus based on part of speech.

In this session, I will present this method and the tools developed to consult the Multilingual RST TB we have developed in the Ixa group (UPV/EHU).
## Keywords

<table>
<thead>
<tr>
<th>Relational discourse structure</th>
<th>Rhetorical relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation</td>
<td>Indicators</td>
</tr>
<tr>
<td>Applications</td>
<td>Inference</td>
</tr>
<tr>
<td><strong>Central Unit</strong></td>
<td>Macro-structure</td>
</tr>
<tr>
<td>Coherence</td>
<td>Micro-structure</td>
</tr>
<tr>
<td>Context</td>
<td><strong>Nuclearity</strong></td>
</tr>
<tr>
<td>Corpus</td>
<td><strong>Nucleus</strong></td>
</tr>
<tr>
<td>Discourse markers</td>
<td>Parser</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Question-answering</td>
</tr>
<tr>
<td>Expl. relations</td>
<td><strong>Recursivity</strong></td>
</tr>
<tr>
<td><strong>Hierarchy</strong></td>
<td>Rhetorical analysis</td>
</tr>
<tr>
<td>Impl. relations</td>
<td>Structure</td>
</tr>
<tr>
<td></td>
<td>Signals</td>
</tr>
<tr>
<td></td>
<td>Segmenter</td>
</tr>
<tr>
<td></td>
<td>Segmentation</td>
</tr>
<tr>
<td></td>
<td>Sentiment analysis</td>
</tr>
<tr>
<td></td>
<td>Summaryization</td>
</tr>
</tbody>
</table>
– Other linguistic levels have been addressed:
  • Phonetics: AhoTSS (Hernaez et al., 2001)
  • Morphology: analysis with MORPHEUS (Aduriz et al., 1998) and disambiguation with EUSTAGGER (Aduriz et al., 2003)
  • Syntax: shallow syntax with IXAti and dependencies with MALTIXA (Bengoetxea and Gojenola, 2007)
  • Semantics: entities with EIHERA (Alegria et al., 2003) and synset disambiguation with ADIERAK prototype

– And what about discourse?
Other linguistic levels have been addressed:

- Phonetics: AhoTSS (Hernaez et al., 2001)
- Morphology: analysis with MORPHEUS (Aduriz et al., 1998) and disambiguation with EUSTAGGER (Aduriz et al., 2003)
- Syntax: shallow syntax with IXAti and dependencies with MALTIXA (Bengoetxea and Gojenola, 2007)
- Semantics: entities with EIHERA (Alegria et al., 2003) and synset disambiguation with ADIERAK prototype

And what about discourse?
Discourse

- Discourse types:
  - Monologue
  - Dialogue

- Discourse levels (van Dijk, 1980a)
  - Local level: between word level and sentence level
  - Global coherence: the structural relation between the main topic (central unit) with the other thematical units

- Discourse characteristics:
  - Structure (referential, relational)
  - Genre (context)
  - Intention (inter-level: phonetics, lexicon, syntax)
Discourse structure phenomena in CL

CL works on discourse structure:

- Referential: co-reference disambiguation (Mitkov, 2002; Recasens et al., 2010) in Basque (IXA group) (Goenaga et al., 2012; Ceberio et al., 2009; Soraluze et al., 2015)

- Relational: rhetorical annotation (Asher and Lascarides, 2003; Mann and Thompson, 1988) in Basque (Gomez, 1996; Barrutieto et al., 2002, 2001) and in IXA group (Iruskietu et al., 2011, 2013b)
  - Segmeter: EusEduSeg
  - Central Unit detector
  - Signal annotation
  - Applications: corpus exploration tools
Can we explain discourse structure with only explicit and semantic relations? Examples from van Dijk (1980b)

(1) I bought a ticket and went to my seat. (Macro-structure)
(2) #Peter went to the cinema. He has blue eyes. (Unlikely)
(3) John is sick. He has the flu. (Semantic)
(4) John can’t come. He is sick. (Semantic, Pragmatic)

– The relationship between the local and global coherence (the topic “cinema”) is necessary in (1)
– A lack of coherence in (2)
– ELABORATION in (3): sick > flu
– Can there be more than one interpretation in (4)?
  • CAUSE_{sem.}: sickness is the reason for not going
  • JUSTIFY_{pragm.}: an accepted situation for not working
Theories and annotation guidelines:

- RST (Mann and Thompson, 1987) and its annotation guidelines (Carlson and Marcu, 2001).
- SDRT (Asher and Lascarides, 2003) and its annotation guidelines (Reese et al., 2007).
- PDTB (Miltsakaki et al., 2004) and its annotation guidelines (Prasad et al., 2007).
Relational discourse structure

A rhetorical structure tree (RS-tree) is a hierarchical structure in which all the propositions of the text have a relationship in the structure.

In RST a hierarchical tree structure is composed with:
1. **Hierarchy**: i) nucleus and ii) satellite
2. **Relations**: i) presentational and ii) subject-matter
Rhetorical relations: definitions at the RST Web Site

The next music day is scheduled for July 21 (Saturday), noon-midnight.

I'll post more details later,
but this is a good time to reserve the place on your calendar.

---

**Conc.**
- on N: W has positive regard for N
- on S: W is not claiming that S does not hold;

Constitution on \( S \) or \( N \)

**Concession**
- W acknowledges a potential or apparent incompatibility between \( N \) and \( S \); recognizing the compatibility between \( N \) and \( S \) increases R’s positive regard for \( N \)

Constraints on \( S + N \)

**Just.**
- none

Intention of \( W \)

- R’s readiness to accept \( W \)’s right to present \( N \) is increased
Why annotate an RST TreeBank

- Linguistic description
  - Nuclearity
  - Recursive Rhetorical Relations
- Real texts in different languages
  - RST TB, SFU Corpus (Taboada and Renkema, 2011), RST Spanish TB (da Cunha et al., 2011), Potsdam Corpus (Stede, 2004), TCC (Pardo and Nunes, 2006), Rhetalho corpus (Pardo and Seno, 2005), spoken corpus (Antonio and Cassim, 2012), Basque RST Treebank (Iruskieta et al., 2013a),
- Many tools for annotation and for analysis
- Applications in NLP (Taboada and Mann, 2006)
Applications based on RST

- Automatic text creation (Bouayad-Agha, 2000; Agirrezabal et al., 2015),
- Automatic text summarization (Marcu, 2000b; Zipitria et al., 2013),
- Machine translation (Ghorbel et al., 2001),
- Assessment of written texts (Burstein et al., 2003),
- Information retrieval (Haouam and Marir, 2003),
- Automatic Discourse Analyzer (Pardo and Nunes, 2008; Soricut and Marcu, 2003)
- Question answering (Bosma, 2005)
- Polarity extractor (Alkorta et al., 2015)
Problems and solutions for RS annotation

- Discourse annotation is complex (Hovy, 2010)
  - Different types of ambiguity of RS (hierarchical segmentation, discourse markers, nuclearity, effect)
  - Structure shape: tree or graph (multiple relations, partial connectivity)
  - Implicit discourse relations

- Solution in Computational Linguistics: corpus annotation
  a) Consistent: enough to support machine learning
  b) Descriptive: enough to work with NLP advanced applications
Main goals

Our main goals:

i) To analyze typical cases of annotators’ disagreement

ii) To disseminate the results in a friendly environment for corpus exploration

iii) To describe a rhetorical structure of scientific abstract by means of corpus annotation (mainly Basque)

iv) To build a discourse parser

v) To evaluate the segmenter/parser in several NLP applications
The corpus

- The Basque RST TreeBank (Iruskieta et al., 2013a):
  - Short texts, but with complex RS
  - Abstracts: structured texts (Ripple et al., 2011)
  - Different domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Sub-corpus</th>
<th>Texts</th>
<th>EDUs</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine</td>
<td>GMB</td>
<td>20</td>
<td>283</td>
<td>3010</td>
</tr>
<tr>
<td>Terminology</td>
<td>TERM</td>
<td>20</td>
<td>584</td>
<td>5664</td>
</tr>
<tr>
<td>Science</td>
<td>ZTF</td>
<td>20</td>
<td>603</td>
<td>6892</td>
</tr>
<tr>
<td>Life</td>
<td>BIZ</td>
<td>20</td>
<td>569</td>
<td>5535</td>
</tr>
<tr>
<td>Health</td>
<td>OSA</td>
<td>20</td>
<td>475</td>
<td>4878</td>
</tr>
<tr>
<td>Informatics</td>
<td>INF</td>
<td>20</td>
<td>236</td>
<td>1860</td>
</tr>
<tr>
<td>Economy</td>
<td>EKO</td>
<td>20</td>
<td>216</td>
<td>2108</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>140</strong></td>
<td><strong>2966</strong></td>
<td><strong>29947</strong></td>
<td></td>
</tr>
</tbody>
</table>

- Parallel texts (da Cunha and Iruskieta, 2010; Iruskieta and da Cunha, 2010) and Multilingual RST TreeBank (Iruskieta et al., 2015a)
RST analysis styles

- **A reader view**: First segment and then link the discourse units without any restriction from left to right (Mann and Thompson, 1988)

- **A parser approach**: First segment and then link the discourse units following a modular way: sentential (E)DU first and paragraph DU after (Pardo, 2005)

- **An analyst style**: First segment and then choose the CU. After that, link the (E)DUs in a modular way taking into account the CU and genre constraints (Iruskieta, 2014)
Annotation method and automatic tasks

- **Segmentation:**
  - EusEduSeg, $F_1$: 0.83 (based on dependencies)
  - $F_1$: 0.82 (based on CG3 rules)

- **Central Unit (CU):**
  - Detection of the most important unit of the RS-tree: $F_1$: 0.44 (ongoing)

- **Rhetorical relations (RR):**
  - Annotation tool: RSTTool
  - Automatic evaluation: RSTEval
  - Queries of RRs in a corpus: Basque RST Treebank
  - Detection of the cause subgroup (ongoing)
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Signaling relational structures
   - An ambiguous RST analysis
   - Annotation in RST

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
Perfil del usuario de la zona ambulatoria del Servicio de Urgencias del Hospital de Galdakao

The profile of the users from the emergency department from Galdakao’s Hospital

I. Bengoetxea Martínez
Médico de Familia.

Introducción

El número de asistencias urgentes crece constantemente, en España el ritmo de crecimiento se ha establecido en torno al 4% anual. Se estima que el 80% de los usuarios acuden por iniciativa propia y que el 70% de las consultas son consideradas leves por el personal sanitario. Realizar estudios epidemiológicos que describan las características de los usuarios y los motivos de la sobreutilización de los servicios de urgencia hospitalarios pueden resultar interesante desde el punto de vista de la planificación sanitaria. Por lo que hemos creído oportuno realizar un estudio para conocer el perfil del usuario de urgencias del hospital de Galdakao.

Resultados: El perfil del usuario sería el de un varón (51,4%) de mediana edad (43,2 años) que consulta por patología traumática (50,5%) y procede de la comarca sanitaria cercana al hospital.

Palabras clave: Usuarios de urgencias, sobreutilización, perfil de usuario.

SUMMARY

The number of urgent cares grows continuously, the rate of growth in Spain has been set around the 4% annually. According to the estimates, the 80% of the users, go by their own initiative to the emergency department, and the 70% of the surgeries are considered slight by the health staff. It could be interesting from the sanitary planning point of view, to carry out epidemiological studies which describe the users characteristics, and the reasons for the overuse of the hospital emergency department. We have seen convenient to archive a study to know the profile of the users from the emergency department from Galdakao’s Hospital.

Results: The general profile of users would be, man (51.4%) of middle age (43.2 years) who consults because of traumatologic pathologies (50.5%) and who comes from the sanitary area near the hospital.

Key words: Emergency department users, overuse, users profile.
Perfiles del usuario de la zona ambulatoria del Servicio de Urgencias del Hospital de Galdakao

The profile of the users from the emergency department from Galdakao’s Hospital

I. Bengoetxea Martínez

Médico de Familia.

RESUMEN

El número de asistencias urgentes crece constantemente, en España el ritmo de crecimiento se ha establecido en torno al 4% anual. Se estima que el 80% de los usuarios acuden por iniciativa propia a los servicios de urgencia y que el 70% de las consultas son consideradas leves por el personal sanitario. Realizar estudios epidemiológicos que describan las características de los usuarios y los motivos de la sobreutilización de los servicios de urgencia hospitalarios puede resultar interesante desde el punto de vista de la planificación sanitaria. Por lo que hemos creído oportuno realizar un estudio para conocer el perfil del usuario de urgencias del hospital de Galdakao.

Resultados: El perfil del usuario sería el de un varón (51,4%) de mediana edad (43,2 años) que consulta por patología traumática (50,5%) y procede de la comarca sanitaria cercana al hospital.

Palabras clave: Usuarios de urgencias, sobreutilización, perfil de usuario.

SUMMARY

The number of urgent cares grows continuously, the rate of growth in Spain has been set around the 4% annually. According to the estimates, the 80% of the users, go by their own initiative to the emergency department, and the 70% of the surgeries are considered slight by the health staff. It could be interesting from the sanitary planning point of view, to carry out epidemiological studies which describe the users characteristics, and the reasons for the overuse of the hospital emergency department. We have seen convenient to archive a study to know the profile of the users from the emergency department from Galdakao’s Hospital.

Results: The general profile of users would be, man (51,4%) of middle age (43,2 years) who consults because of traumatologic pathologies (50,5%) and who comes from the sanitary area near the hospital.

Key words: Emergency department users, overuse, users profile.

LABURPENA

Larrialdi zerbitzuetako asistentzia medikoa kopuru gihitua da atongoako, estate espolalean izoga hau urteko ¼'akakukatzen da. Erabiltzaleak izaniko bideko erabilera batetako jatorria eta kontakua hauen %70a lantzaun geltukixat jotan dizkie zerbitzu haustako edoetako. Zerbitzu haun perfile alaitzen datu lerrekin epidemiologikoak egitea balagatua izan daiteke esanazkinak aitzindak, hau dela eta, Galdakako ospitaleko larrialdi zerbitzuetako erabiltzaileen perfile edoetako bat egitea apropan laufridin galtzeko.

Basic concepts of discourse segmentation

- A first step of any discourse parser is to identify the units
  - But what is an Elementary Discourse Unit (EDU) is controversial also in RST (van der Vliet, 2010b)
- Segmentation proposals are based on three basic concepts:
  - Linguistic “form” (or category)
  - “Function” (the function of the syntactic components)
  - “Meaning” (the coherence relation between propositions)
Segmentation guidelines: Basque

- Segmentation guidelines conflate RST and Basque clause combining constraints (Tofiloski et al., 2009; Salaburu, 2012; Artiagoitia et al., 2003)
  - Based on function (adjunct clauses) and form (which contain a verb)

<table>
<thead>
<tr>
<th>Clause type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perpaus independentea ‘an independent sentence’</td>
<td>[Whipple (EW) gaixotasunak hesteei eragiten die bereziki.] GMB0503</td>
</tr>
<tr>
<td>Perpaus nagusi koordinatua ‘a main clause, part of sentence’</td>
<td>[pT1 tumoreko 13 kasuetan ez zen gongoila inbasiorik hauteman;] [aldiz, pT1 101 tumoretatik 19 kasutan (18.6%) inbasioa hauteman zen, eta pT1c tumoreen artetik 93 kasutan (32.6%).] GMB0703</td>
</tr>
<tr>
<td>Aditz jokatudun adjuntu perpausa ‘finite adjunct clauses’</td>
<td>[Hainen sailkapena egiteko hormona hartzailen eta c-erb-B2 onkogenearen gabeziaz baliatu gara,] [ikerketa anatomopatologikoetan erabili ohi diren zehaztapenak direlako.] GMB0702</td>
</tr>
<tr>
<td>Aditz jokatugbedun adjuntu perpausa ‘non-finite adjunct clauses’</td>
<td>[Ohiko tratamendu motek porrot eginez gero,] [gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.] GMB0502</td>
</tr>
<tr>
<td>Erlatibo ez-murriztailea ‘non-restrictive relative clause’</td>
<td>[Dublin Hiriko Unibertsitateko atal bat da Fiontar,] [zeinak Ekonomia, Informatika eta Empresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.] TERM23</td>
</tr>
</tbody>
</table>
Adjunct verb clause-based segmentation (Tofiloski et al., 2009)

*English translation is ours
### Automatic segmentation based on rules (CG3)

<table>
<thead>
<tr>
<th>Segments</th>
<th>Correct</th>
<th>Missed</th>
<th>Excess</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>765</td>
<td>606</td>
<td>159</td>
<td>98</td>
<td>0.86</td>
<td>0.79</td>
</tr>
<tr>
<td>MAP:171</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP:358</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP:231</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP:180</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP:211</td>
<td>413</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP:148</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP:472</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results obtained with CG3 rule by rule:
Evaluation of the segmentation

Evaluation is performed based on the end-EDU. But following this, both segmentations have the same result, even if W2 and W4 are verbs.

A better evaluation is to use the WindowDiff (WD) (Pevzner and Hearst, 2002) or Deviation (D) (Cardoso et al., 2013), following this Automatic-1 is better than Automatic-2.
Some conclusions and topics to discuss: Granularity and RR

- Less agreement at intra-sentential agreement than at sentential one (−13.74%), but more agreement in relations (+14.19%) and more robust (RCA +9.5%) (Iruskieta et al., 2011)
  - Parallelism: syntax-discourse (Marcu and Echihabi, 2002)
  - Some relations (R) can be derived from syntax (Soricut and Marcu, 2003)
  - Simpler constituents (C) and fewer attachment points (A)
  - Parsers are more reliable (Pardo and Nunes, 2008; Soricut and Marcu, 2003)
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration
- Segmenters
- CU detector
- Annotation tools for RST
- Evaluation tools/methods of RS

PART 4 — Resources
- Projects
- Resources
- Workshops
Central Unit (CU), indicators and RST

- Texts ought to be coherent at local level and global level. But the coherence of CU with other units (or RRs) is not considered in RST
  - not in the annotation guidelines (Carlson et al., 2001)
  - not in the evaluation method (Marcu, 2000a)
- Central Unit (Stede, 2008)
  - Central proposition (Pardo et al., 2003), thesis statement (Burstein et al., 2001), and thematical sentence(s) (van Dijk, 1980a)
- Indicators of CU: nouns (paper, article, presentation, investigation, method, result…), verbs (discuss, introduce, present, examine, analy-, stud-…), demonstratives and determiners (this, the, a, some…) and pronouns (we, I)… (Paice, 1980)
  - Ambiguity: some of them are very vague, they could refer also to micro-structure (Paice, 1980, 179)
Central Unit (CU), indicators and RST

- Texts ought to be coherent at local level and global level. But the coherence of CU with other units (or RRs) is not considered in RST
  - not in the annotation guidelines (Carlson et al., 2001)
  - not in the evaluation method (Marcu, 2000a)
- Central Unit (Stede, 2008)
  - Central proposition (Pardo et al., 2003), thesis statement (Burstein et al., 2001), and thematical sentence(s) (van Dijk, 1980a)
- Indicators of CU: nouns (paper, article, presentation, investigation, method, result...), verbs (discuss, introduce, present, examine, analy-, stud-...), demonstratives and determiners (this, the, a, some...) and pronouns (we, I)... (Paice, 1980)
  - Ambiguity: some of them are very vague, they could refer also to micro-structure (Paice, 1980, 179)
An example of Central Unit (CU) annotated with RSTTool

Estomatitis Aftosa Recurrente (I):
Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.

"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da.

Honen etiologia eztabaidagarria da. Its etiology is controversial.

Ultzera mingarri batzu bezela agertzen da, It is characterised by the appearance of painful and recurrent ulcers,
tamainu, kokapena eta iraunkortasuna aldakorra izanik.

Hauek periodikoki beragertzen dira. These ulcers reappear periodically.

(5) [Lan honetan patologia arrunt honetan ezaugarri etiopatogeniko eta klinikopatologiko garrantzitsuenak analizatzen ditugu.]}_{7} [GMB0301]

[This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.]_{7}
Central Unit: harmonization

- CU annotation guidelines for scientific abstracts
  
  i) Topic or thesis statement
  ii) Purpose
  iii) Method
  iv) Results
  v) Conclusions
### Indicators from train dataset (Iruskieta et al., 2014a)

<table>
<thead>
<tr>
<th>EUS</th>
<th>ENG&lt;sub&gt;MCR&lt;/sub&gt;</th>
<th>EUS</th>
<th>ENG&lt;sub&gt;MCR&lt;/sub&gt;</th>
<th>Pronouns</th>
<th>Bonus words</th>
</tr>
</thead>
<tbody>
<tr>
<td>aztertu</td>
<td>examine&lt;sub&gt;1&lt;/sub&gt;</td>
<td>abiapuntu&lt;sub&gt;1&lt;/sub&gt;</td>
<td>starting_point&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Demonstrative Pronoun</td>
<td>garantzi importance</td>
</tr>
<tr>
<td>analizatu</td>
<td>examine&lt;sub&gt;1&lt;/sub&gt;</td>
<td>arlo&lt;sub&gt;1&lt;/sub&gt;</td>
<td>subject_field&lt;sub&gt;1&lt;/sub&gt;</td>
<td>hau this</td>
<td>nagusi main</td>
</tr>
<tr>
<td>oinarritu</td>
<td>base&lt;sub&gt;1&lt;/sub&gt;</td>
<td>artikulu&lt;sub&gt;1&lt;/sub&gt;</td>
<td>article&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Personal Pronouns</td>
<td>azpimarragarri</td>
</tr>
<tr>
<td>baloratu</td>
<td>value&lt;sub&gt;2&lt;/sub&gt;</td>
<td>asmo&lt;sub&gt;2&lt;/sub&gt;</td>
<td>purpose&lt;sub&gt;1&lt;/sub&gt;</td>
<td>gu we</td>
<td>remarcable</td>
</tr>
<tr>
<td>azaldu</td>
<td>recount&lt;sub&gt;1&lt;/sub&gt;</td>
<td>bide&lt;sub&gt;2&lt;/sub&gt;</td>
<td>means&lt;sub&gt;1&lt;/sub&gt;</td>
<td>-gu (inside the verb)</td>
<td>eskerga huge</td>
</tr>
<tr>
<td>aurkeztu</td>
<td>present&lt;sub&gt;2&lt;/sub&gt;</td>
<td>gai&lt;sub&gt;6&lt;/sub&gt;</td>
<td>topic&lt;sub&gt;1&lt;/sub&gt;</td>
<td></td>
<td>(gaur) egun nowadays</td>
</tr>
<tr>
<td>aipatu</td>
<td>present&lt;sub&gt;2&lt;/sub&gt;</td>
<td>ikerkuntza&lt;sub&gt;3&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>berri eman</td>
<td>present&lt;sub&gt;2&lt;/sub&gt;</td>
<td>ikerketa&lt;sub&gt;2&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jardun</td>
<td>present&lt;sub&gt;2&lt;/sub&gt;</td>
<td>azterlan&lt;sub&gt;3&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plazaratu</td>
<td>present&lt;sub&gt;2&lt;/sub&gt;</td>
<td>ikerlan&lt;sub&gt;3&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>erabili</td>
<td>use&lt;sub&gt;1&lt;/sub&gt;</td>
<td>arazo&lt;sub&gt;3&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ikertu</td>
<td>investigate&lt;sub&gt;1&lt;/sub&gt;</td>
<td>irtenbide&lt;sub&gt;2&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>komunikazio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hitzaldi&lt;sub&gt;2&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lan&lt;sub&gt;3&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lan-ildo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>lerro&lt;sub&gt;1&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ikerketa-lerro</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>proiektu&lt;sub&gt;2&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ikerketa-proiektu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>talde&lt;sub&gt;1&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ikerketa-talde</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>xede&lt;sub&gt;1&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>helburu&lt;sub&gt;2&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Heuristics to identify the Central Unit (test dataset)

- Difficulty to choose the CU: 0.032
- Agreement between 2 annotators: 0.89 F1

<table>
<thead>
<tr>
<th>Heuristics</th>
<th>C</th>
<th>E</th>
<th>M</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 Nouns and verbs</td>
<td>15</td>
<td>31</td>
<td>29</td>
<td>0.33</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>H2 Nouns and verbs + pronouns</td>
<td>22</td>
<td>68</td>
<td>22</td>
<td>0.24</td>
<td>0.50</td>
<td>0.33</td>
</tr>
<tr>
<td>H3 Bonus words</td>
<td>5</td>
<td>14</td>
<td>39</td>
<td>0.26</td>
<td>0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>H4 Title words</td>
<td>7</td>
<td>3</td>
<td>37</td>
<td>0.70</td>
<td>0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>H5 EDU position</td>
<td>40</td>
<td>711</td>
<td>4</td>
<td>0.05</td>
<td>0.91</td>
<td>0.10</td>
</tr>
<tr>
<td>H6 Main verb</td>
<td>41</td>
<td>721</td>
<td>3</td>
<td>0.05</td>
<td>0.93</td>
<td>0.10</td>
</tr>
<tr>
<td>H7 H1, H2 and H4</td>
<td>21</td>
<td>30</td>
<td>23</td>
<td>0.41</td>
<td>0.48</td>
<td>0.44</td>
</tr>
<tr>
<td>H8 H1, H2, H3, H4 and H5</td>
<td>23</td>
<td>48</td>
<td>21</td>
<td>0.32</td>
<td>0.52</td>
<td>0.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Machine Learning</th>
<th>C</th>
<th>E</th>
<th>M</th>
<th>Pre.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptron + postproc.</td>
<td>24</td>
<td>25</td>
<td>20</td>
<td>0.48</td>
<td>0.54</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Some conclusions and topics to discuss: the annotation of the Central Unit (Iruskieta et al., 2014b)

<table>
<thead>
<tr>
<th>Texts</th>
<th>Annotators</th>
<th>Measure</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burstein et al. (2001)</td>
<td>100</td>
<td>2 professionals</td>
<td>F-score</td>
</tr>
<tr>
<td>Basque</td>
<td>60</td>
<td>4 non-professionals</td>
<td>F-score</td>
</tr>
</tbody>
</table>

- Annotation of the CU (2 annotators):
  - Derived from RS-trees: 65% (GMB)
  - Annotating the CU first: 85% (in TERM and in ZTF)

- Agreement is bigger in relations, when annotators have annotated the same CU (+5.04%, T-test: 0.013)
- Agreement is bigger in RRs linked to the CU (+17.29%, T-test: 0.001)
Within the RRs linked to the CU, those with an IMRaD structure appear most frequently (except ELABORATION) (Iruskieta, 2014)

<table>
<thead>
<tr>
<th>RRs</th>
<th>GMB</th>
<th>TERM</th>
<th>ZTF</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN</td>
<td>NS</td>
<td>SN</td>
<td>NS</td>
</tr>
<tr>
<td>PREPARATION</td>
<td>22</td>
<td>24</td>
<td>22</td>
<td>68</td>
</tr>
<tr>
<td>ELABORATION</td>
<td>6</td>
<td>15</td>
<td>28</td>
<td>49</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>13</td>
<td>15</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td>MEANS</td>
<td>1</td>
<td>14</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>RESULT</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>CIRCUMSTANCE</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>INTERPRETATION</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAUSE</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>JUSTIFY</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CONCESSION</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SOLUTIONHOOD</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Total: 39 44 45 39 39 48 123 131
PART 1 — Discourse relations in RST: method

- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
  - Signals of rhetorical relations
  - Corpora for corpus exploration
  - Applications

PART 2 — Practice

- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration

- Segmenters
- CU detector
- Annotation tools for RST
- Evaluation tools/methods of RS

PART 4 — Resources

- Projects
- Resources
- Workshops
The extended RST relation set

<table>
<thead>
<tr>
<th>Type</th>
<th>Relation</th>
<th>Relation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Preparation</td>
<td>Elaboration</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td>Background</td>
<td>Means</td>
<td>SM</td>
</tr>
<tr>
<td></td>
<td><strong>Enablement and Motivation</strong></td>
<td>Circumstance</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td>Enablement</td>
<td>Solution-hood</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td>Motivation</td>
<td><strong>Conditional relations</strong></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Evidence</td>
<td>Condition</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td>Justify</td>
<td>Otherwise</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td><strong>Evidence and Justify</strong></td>
<td>Unless</td>
<td>SM</td>
</tr>
<tr>
<td></td>
<td><strong>Antithesis and Concession</strong></td>
<td>No-Conditional</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td>Antithesis</td>
<td><strong>Interpretation and Evaluation</strong></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Concession</td>
<td>Interpretation</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td><strong>Reformulation and Summary</strong></td>
<td>Evaluation</td>
<td>SM</td>
</tr>
<tr>
<td>P</td>
<td>Reformulation</td>
<td><strong>Cause subgroup</strong></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Summary</td>
<td>Cause</td>
<td>SM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Result</td>
<td>SM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Purpose</td>
<td>SM</td>
</tr>
<tr>
<td>N-N</td>
<td>List</td>
<td>Sequence</td>
<td>N-N</td>
</tr>
<tr>
<td>N-N</td>
<td>Disjunction</td>
<td>Contrast</td>
<td>N-N</td>
</tr>
<tr>
<td>N-N</td>
<td>Joint</td>
<td>Conjunction</td>
<td>N-N</td>
</tr>
<tr>
<td>N-N</td>
<td>Reformulation-NN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>⌀</td>
<td>Same-unit</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relations from the RST webpage at [http://www.sfu.ca/rst/](http://www.sfu.ca/rst/)
RSTTool annotation interface

- A TXT text and a relation set are necessary to annotate with the RSTTool
- The segmenter EusEduSeg has integrated the RS3 output and a Basque relation set
PART 1 — Discourse relations in RST: method

Rhetorical relations

Rhetorical structure of a text [GMB0401]

A modular and incremental annotation (Pardo, 2005)
PART 1 — Discourse relations in RST: method

Rhetorical relations

Different interpretations of [GMB0401]
PART 1 — Discourse relations in RST: method

Rhetorical relations

Different interpretations of [GMB0401]
Different interpretations of [GMB0401]
PART 1 — Discourse relations in RST: method

Rhetorical relations

Inter-annotator agreement in RST relations

– The **RST TreeBank** (Carlson et al., 2001)
  • from 0.5973 to 0.7921 $\kappa$ (2 annot., 30 texts: 1918 EDUs)
  • from 0.6017 $\kappa$ to 0.7555 $\kappa$ (3 trained professionals, 4/5 texts 515/343 EDUs)
– The **Spanish RST TreeBank** (da Cunha et al., 2010)
  • 77.64% $F_1$ (2 trained annot.: 84 texts, 694 EDUs)
– The **Dutch TreeBank** (van der Vliet et al., 2011)
  • 0.57 $\kappa$ (2 annotators, 4 texts)
– The **Basque RST TreeBank** (Iruskieta et al., 2013a)
  • 0.568 $\kappa$ or 61.47% $F_1$ (2 annot., 60 texts: 1470 EDUs)
An automatic evaluation of RS-trees with RSTeval (Maziero and Pardo, 2009) of GMB0701
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS
   - Parsers

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
Signalling the RRs

- Signalling in
  - Brazilian Portuguese (Pardo and Nunes, 2004),
  - Spanish (da Cunha, 2013)
  - English (Das et al., 2015)
  - Basque (where some tools to visualize signals were developed to improve RRs queries)

- Annotation tool: Rhetorical Database (Pardo, 2005)
  - Relation by relation
  - Searches can be done to maintain consistency

- Annotation tool: UAM CorpusTool
  - Different annotation levels
Signalling the RRs

- **What is signalling?**
  
  a) DM annotation (automatically)
  
  b) Annotation of the most frequent forms (and functions) (Taboada and Das, 2013)
  
  - to distinguish volitional/non-volitional relations of cause exploiting the information provided by verb tense (Antonio, 2012)
  
  - to have more explicit relations

- If signals can be from any linguistic form, is annotation more reliable?

- Is there any ground for the automatic signalling?
Signalling the RRs

- **What is signalling?**
  
  a) DM annotation (automatically)
  
  b) Annotation of the most frequent forms (and functions) (Taboada and Das, 2013)
  
  - to distinguish volitional/non-volitional relations of cause exploiting the information provided by verb tense (Antonio, 2012)
  
  - to have more explicit relations

- **If signals can be from any linguistic form, is annotation more reliable?**

- **Is there any ground for the automatic signalling?**
Signalling the RRs

- **What is signalling?**
  
  a) DM annotation (automatically)
  b) Annotation of the most frequent forms (and functions) (Taboada and Das, 2013)
  * to distinguish volitional/non-volitional relations of cause exploiting the information provided by verb tense (Antonio, 2012)
  * to have more explicit relations

- If signals can be from any linguistic form, is annotation more reliable?

- Is there any ground for the automatic signalling?
Criteria to annotate signals

- Annotate more than discourse markers (Iruskieta, 2014)
- Check every discourse units of the relation (nucleus or satellite)
- Look for more than one signal and not always one after another
- Check different categories (coordinators, nouns, verbs, particles...) and language levels (semantic: synonym, syntactic: question-answer...)

<table>
<thead>
<tr>
<th>Signals</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinators</td>
<td>however, therefore, in fact</td>
</tr>
<tr>
<td>Morphology</td>
<td>-ing, non-finite verbs</td>
</tr>
<tr>
<td>Lexical</td>
<td>concede, cause</td>
</tr>
<tr>
<td>Entity</td>
<td>entities</td>
</tr>
<tr>
<td>Semantic</td>
<td>synonyms, antonyms, hyponyms</td>
</tr>
<tr>
<td>Syntax</td>
<td>question-answer,</td>
</tr>
<tr>
<td>Graphic-numeric</td>
<td>1. (...) 2., a) (...) b)</td>
</tr>
<tr>
<td>Complex signals</td>
<td>...</td>
</tr>
</tbody>
</table>
Signal annotation with Rhetorical Database

- A tool to annotate signals and extract statistics
Signals of cause subgroup

How reliable is the annotation of signals, is it equal in every relation?

<table>
<thead>
<tr>
<th>Annotators</th>
<th>CAUSE%</th>
<th>RESULT%</th>
<th>PURPOSE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁-A₂</td>
<td>71.43</td>
<td>59.70</td>
<td>90.00</td>
</tr>
<tr>
<td>A₁-A₄</td>
<td>67.86</td>
<td>50.75</td>
<td>80.91</td>
</tr>
<tr>
<td>A₂-A₄</td>
<td>73.21</td>
<td>37.31</td>
<td>78.18</td>
</tr>
<tr>
<td>A₁-A₂-A₄</td>
<td>58.93</td>
<td>37.31</td>
<td>75.45</td>
</tr>
</tbody>
</table>

How reliable is the annotation of signals, which is complex (multiple) and with different levels/categories?

- Signals are much more ambiguous than discourse markers (at least in the cause subgroup)
  - Mean inter-annotator disagreement in discourse markers 15.27%
  - Mean inter-annotator disagreement in other signals 68.13%
## Results of the RR}s and their signals

<table>
<thead>
<tr>
<th>Rhetorical Relations</th>
<th>Signals%</th>
<th>DU₁</th>
<th>DU₂</th>
<th>DU₁/₂</th>
<th>N</th>
<th>S</th>
<th>S/N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Presentational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(pragmatic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PREPARATION</td>
<td>110</td>
<td>2</td>
<td>1.82</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>75</td>
<td>16</td>
<td>21.33</td>
<td>12</td>
<td>4</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>ENABLEMENT</td>
<td>6</td>
<td>6</td>
<td>100.00</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>MOTIVATION</td>
<td>5</td>
<td>5</td>
<td>100.00</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVIDENCE</td>
<td>11</td>
<td>7</td>
<td>63.64</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUSTIFY</td>
<td>14</td>
<td>13</td>
<td>92.86</td>
<td>1</td>
<td>11</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>ANTITHESIS</td>
<td>5</td>
<td>4</td>
<td>80.00</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CONCESSION</td>
<td>40</td>
<td>39</td>
<td>97.50</td>
<td>11</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESTATMENT</td>
<td>10</td>
<td>7</td>
<td>70.00</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUMMARY</td>
<td>10</td>
<td>5</td>
<td>50.00</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELABORATION</td>
<td>286</td>
<td>84</td>
<td>29.37</td>
<td>82</td>
<td>2</td>
<td>82</td>
<td>2</td>
</tr>
<tr>
<td>MEANS</td>
<td>93</td>
<td>81</td>
<td>87.10</td>
<td>19</td>
<td>62</td>
<td></td>
<td>81</td>
</tr>
<tr>
<td>CIRCUMSTANCE</td>
<td>57</td>
<td>53</td>
<td>92.98</td>
<td>44</td>
<td>9</td>
<td>1</td>
<td>52</td>
</tr>
<tr>
<td>SOLUTIONHOOD</td>
<td>10</td>
<td>9</td>
<td>90.00</td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>CONDITION</td>
<td>20</td>
<td>19</td>
<td>95.00</td>
<td>12</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>UNCONDITIONAL</td>
<td>1</td>
<td>1</td>
<td>100.00</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERPRETATION</td>
<td>28</td>
<td>22</td>
<td>78.57</td>
<td>3</td>
<td>17</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>EVALUATION</td>
<td>11</td>
<td>10</td>
<td>90.91</td>
<td>10</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>CAUSE</td>
<td>56</td>
<td>53</td>
<td>94.64</td>
<td>23</td>
<td>21</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>RESULT</td>
<td>67</td>
<td>57</td>
<td>85.07</td>
<td>1</td>
<td>55</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>110</td>
<td>109</td>
<td>99.09</td>
<td>40</td>
<td>68</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><strong>Subject-matter</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(semantic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIST</td>
<td>166</td>
<td>87</td>
<td>52.41</td>
<td>3</td>
<td>53</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>SEQUENCE</td>
<td>32</td>
<td>21</td>
<td>65.63</td>
<td>2</td>
<td>15</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>CONJUNCTION</td>
<td>50</td>
<td>38</td>
<td>76.00</td>
<td>2</td>
<td>37</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CONTRAST</td>
<td>40</td>
<td>33</td>
<td>82.50</td>
<td>2</td>
<td>23</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>DISJUNCTION</td>
<td>2</td>
<td>2</td>
<td>100.00</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1315</td>
<td>783</td>
<td>59.54</td>
<td>180</td>
<td>532</td>
<td>71</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>550</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27</td>
</tr>
</tbody>
</table>
Relations and signals: interpretation of the results

- The 4 most annotated relations 48.44% are not so signalled 29.20%. General relations (not very informative relations)
  - ELABORATION, LIST, PREPARATION, BACKGROUND
- The other 22 relations are highly signalled: 86.28%. Signalling trends:
  - **Low** ($\leq % 25$): PREPARATION, BACKGROUND
  - **Middle** ($\geq % 25$ and $\leq % 75$): EVIDENCE, RESTATEMENT, SUMMARY, ELABORATION, LIST, SEQUENCE
  - **High** ($\geq % 75$): ENABLEMENT, MOTIVATION, JUSTIFY, ANTITHESIS, CONCESSION, MEANS, CIRCUMSTANCE, CONDITION, SOLUTIONHOOD, UNCONDITIONAL, INTERPRETATION, EVALUATION, CAUSE, RESULT, PURPOSE, CONTRAST, CONJUNCTION, DISJUNCTION
Signals and relations: ambiguity ($\geq 3$ occurrences)

<table>
<thead>
<tr>
<th>Ambiguous signals</th>
<th>Non-ambiguous signals and RRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>Translation</td>
</tr>
<tr>
<td>eta</td>
<td>and</td>
</tr>
<tr>
<td>-nez</td>
<td>given</td>
</tr>
<tr>
<td>-tuz</td>
<td>-ing</td>
</tr>
<tr>
<td>baina</td>
<td>but</td>
</tr>
<tr>
<td>bait-</td>
<td>because</td>
</tr>
<tr>
<td>ba-</td>
<td>if</td>
</tr>
<tr>
<td>bestalde</td>
<td>moreover</td>
</tr>
<tr>
<td>era berean</td>
<td>likewise</td>
</tr>
<tr>
<td>izan ere</td>
<td>in fact</td>
</tr>
<tr>
<td>gainera</td>
<td>furthermore</td>
</tr>
<tr>
<td>berriz</td>
<td>whereas</td>
</tr>
<tr>
<td>alde batetik</td>
<td>on the one hand</td>
</tr>
<tr>
<td>-ta</td>
<td>-ed</td>
</tr>
<tr>
<td>lortutako emaitzek</td>
<td>the results obtained</td>
</tr>
<tr>
<td>baietzatzen dute</td>
<td></td>
</tr>
<tr>
<td>hau da</td>
<td></td>
</tr>
</tbody>
</table>

- Are these signals unambiguous in a larger corpus?
- Can we detect Cause subgroup relations automatically, for question-answering tasks?
- And EVALUATION and INTERPRETATION for sentiment analysis?

Go to Exercises: 95
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Signaling relational structures
   - An ambiguous RST analysis
   - Annotation in RST

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
Free RST Treebanks

- Brazilian Portuguese corpora:
  - RST corpus *Rhetalho* (Pardo and Seno, 2005) and *Corpus TCC* (Pardo and Nunes, 2006)
  - CST & RST corpus
    - [http://www.nilc.icmc.usp.br/CSTNews](http://www.nilc.icmc.usp.br/CSTNews)
  - Spoken corpus analysed with RST (Antonio and Cassim, 2012)

- English: The Discourse Relations Reference Corpus (Taboada and Renkema, 2011), available at [http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html](http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html) and the SFU Corpus

- German Potsdam Commentary Corpus (Stede, 2004): a corpus of 220 newspaper commentaries, downloadable from: [http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html](http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html)
- 9 different domains, 267 texts. A double annotation of test-set (84 texts) and 10 different annotators.

- Different queries for the first time:
  
  1) Consult statistics
  2) Check for all the instances of a rhetorical relation in the corpus
The Basque RST TreeBank (Iruskieta et al., 2013a)

- The Basque RST TreeBank is the first corpus annotated with coherence relations in Basque.
- Its delivery phase has followed Ide and Pustejovsky (2010).
- Innovations: a number of operations can be carried out with this annotated corpus.
Queries in a KWIC style of different annotation levels

- All the occurrences of any relation in the corpus (distinguishing annotators)
  - Signals are underlined in colour in the gold standard files
- Relations of a chosen text
  - CU is underlined in colour
- Linear segmentation of a text and its CU
  - Relations that are linked to the CU in the RS-tree
- Check whether a signal is in only a relation or whether it is in more than one
- Any information based on part of speech in the corpus
  - Or in a specific domain of the corpus
Basics of the Basque RST Treebank

- Supported languages: **Basque** (fully developed), Spanish, English, Brazilian Portuguese, (Chinese very soon)
  - The Basque RST Treebank
  - Multilingual RST Treebank (with Taboada & da Cunha)
  - Brazilian Portuguese RST Treebank (with Antonio)

- Read from different programs:
  - Automatic parsing (POS tagging)
  - Maltixa dependency parser (basis of the segmenter)
  - EusEduSeg (a Basque segmenter)
  - RSTTool (to create the relational discourse structure)
  - RhetDB (to annotate signals)
Queries based on word-form, lemma and POS features

<table>
<thead>
<tr>
<th>Doc.</th>
<th>EDU Id</th>
<th>Word</th>
<th>CU</th>
<th>EDU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TERM50</td>
<td>sent2</td>
<td>BAI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>taldeek / helburua</td>
<td></td>
<td>[..] Hitzaldi honek azken hiru urteotan lau unibertsitate hauen taldeek egindako ikerkuntzaren ondorioetako batzuk azaltzeko helburua izango luke. groups / aim</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[..] The aim of this talk is to present some of the results of the research carried out by groups from these four universities over the last three years.</td>
</tr>
<tr>
<td>2</td>
<td>ZTF13</td>
<td>sent1</td>
<td>BAI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>taldearen / helburu</td>
<td></td>
<td>[..] Gure ikerkuntza taldearen helburu nagusia, [..] group’s / aim</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[..] Our research group’s principal aim, [..]</td>
</tr>
<tr>
<td>3</td>
<td>ZTF13</td>
<td>sent17</td>
<td>EZ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>taldearen / helburu</td>
<td></td>
<td>Alor honetan, gure ikerkuntza taldearen helburu nagusia bi dira. group’s / aim</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>In this field, our research group has two main aims.</td>
</tr>
<tr>
<td>1</td>
<td>ZTF15</td>
<td>sent7</td>
<td>EZ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>helburu / talde</td>
<td></td>
<td>[..] bestelako galdera zailagoi ere erantzutea dute helburu, hala nola, espezieen biogeografia, taldearen filogenia, eta abar. aim / group</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[..] the aim is to answer other such difficult questions, such as species biogeography, group phylogeny, etc.</td>
</tr>
</tbody>
</table>
### Multilingual SEARCH section: POS queries

<table>
<thead>
<tr>
<th>Doc.</th>
<th>EDU Id</th>
<th>Word</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>TERM38_A1.txt</td>
<td>seg2</td>
<td>paper / look</td>
<td>This paper is intended to look at the challenges faced by neology in terminology at the present time.</td>
</tr>
<tr>
<td>TERM19_A1.txt</td>
<td>seg12</td>
<td>paper / looks</td>
<td>This paper looks, on the basis of experience in the standardisation of terminology in Catalan, at the social need for standardisation of terminology.</td>
</tr>
<tr>
<td>TERM23_A1.txt</td>
<td>seg13</td>
<td>paper / groups</td>
<td>Our paper will discuss the methodology used by both groups in term creation.</td>
</tr>
<tr>
<td>TERM30_A1.txt</td>
<td>seg27</td>
<td>paper / groups</td>
<td>This paper will discuss challenges encountered, opportunities identified and solutions suggested for managing terminology of specialist languages in multilingual environments where at least one language belongs to the lesser used category on numerical groups.</td>
</tr>
<tr>
<td>TERM50_A1.txt</td>
<td>seg2</td>
<td>paper / groups</td>
<td>The purpose of this paper is to set forth some of the results of research by working groups at the above universities over the last three years.</td>
</tr>
<tr>
<td>TERM30_A1.txt</td>
<td>seg25</td>
<td>used / groups / and</td>
<td>Over the last ten years we have been building terminology collections in languages used by numerically larger groups of people, like English, German and Spanish,</td>
</tr>
<tr>
<td>TERM31_A1.txt</td>
<td>seg6</td>
<td>divided / groups / and</td>
<td>Their areas of application can be divided into two main groups: information indexing and the making-up of terminological glossaries.</td>
</tr>
</tbody>
</table>

- **Lemma “paper”** + a word which begins with “look”
- **Lemma “paper”** + lemma “group”
- **Word which ends with “-ed”** + a word which begins with “group” + a connector
EDUs and CUs in RS-trees: SEGMENTS section

- CU and RRs linked to CU
- Annotator’s info

<table>
<thead>
<tr>
<th>EDU</th>
<th>Segment</th>
<th>GMB0301-GS.rs3 (7)</th>
<th>Tagger</th>
<th>CU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.</td>
<td>GS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>“Estomatitis aftosa recurrente” deritzon patologia, ahoan agertzen den ugarientetako bat da.</td>
<td>GS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>tamainu, kokapena eta iraunkortasuna aldakorra izanik.</td>
<td>GS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Honen etiologia eztabaidagarria da.</td>
<td>GS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Ultzera mingarri batzu bezela agertzen da,</td>
<td>GS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Hauek periodiki beragertzen dira.</td>
<td>GS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantzitsuenak analizatzen ditugu. In this paper we analyze the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.</td>
<td>GS</td>
<td>See</td>
<td></td>
</tr>
</tbody>
</table>

Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.

“Recurrent aphthous stomatitis” is one of the most frequent oral pathologies.

having a variable size, location and duration.

It has a controversial etiology.

It is characterized by the apparition of painful ulcers,

These ulcers appear recurrently.
### Relations linked to the CU

**GMB0301-GS.rs3: CU and relations**

**CU:** Lan honetan patologia arrunt honetan ezaugarri ... garrantitsuenak analizatzen ditugu.

*In this paper we analyze the most important ... features of this common oral pathology.*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.</td>
<td>preparation</td>
<td>“Recurrent aphthous stomatitis” is one of the most frequent oral pathologies having a variable size, location and duration. It has a controversial etiology. It is characterized by the apparition of painful ulcers, these ulcers appear recurrently. In this paper we analyze the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.</td>
</tr>
</tbody>
</table>

---

**In this paper we analyze the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.**
Multilingual EDUs section

- Check the harmonized segmentation of the Multilingual RST Treebank

<table>
<thead>
<tr>
<th>Id</th>
<th>Segment</th>
<th>Tagger</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>General trends in standardization of scientific terminology in Serbian: a critical analysis of the state of affairs</td>
<td>A1</td>
</tr>
<tr>
<td>2</td>
<td>Building the terminology of any scientific area is a long and laborious process.</td>
<td>A1</td>
</tr>
<tr>
<td>3</td>
<td>In the recent past, a trend has been noted, and reported by many researchers in the area of Serbian scientific terminology, of importing borrowings of lexical and larger structural units from English into specific scientific registers, rather that to opt for translations, calques, etc.</td>
<td>A1</td>
</tr>
<tr>
<td>4</td>
<td>This corresponds closely to the fact that a consensus has been reached among Serbian scientists of various orientations regarding the status of English as the only language of scientific communication in the last several decades.</td>
<td>A1</td>
</tr>
<tr>
<td>5</td>
<td>In this paper, an attempt is made to critically evaluate the above outlined trend, from both inherently linguistic perspectives.</td>
<td>A1</td>
</tr>
<tr>
<td>1</td>
<td>Tendencias generales de la normalización en la terminología científicotécnica de la lengua serbia: análisis crítico de la situación</td>
<td>A2</td>
</tr>
<tr>
<td>2</td>
<td>La construcción terminológica de cualquier área científica es un proceso largo y laborioso.</td>
<td>A2</td>
</tr>
<tr>
<td>3</td>
<td>En décadas precedentes se ha puesto de manifiesto, y así lo han atestiguado muchos investigadores de la terminología científica serbia, una tendencia a importar préstamos de unidades estructurales tanto léxicas como otras mayores del inglés a una serie de registros científicos específicos, en lugar de optar por la traducción, el calco, etc.</td>
<td>A2</td>
</tr>
<tr>
<td>4</td>
<td>Empleamos un enfoque abierto y multidisciplinar desarrollado por Bugarski (1988; 1996) y adaptado a los fines de esta ponencia, para contrastarlo con una serie de datos provenientes de varios campos científicos como la ingeniería, el transporte, etc.</td>
<td>A2</td>
</tr>
<tr>
<td>5</td>
<td>Izan ere, iritsi ezeberinetako zientzialari serbiarrek adostasuna hasi du eta aurreko pertsonak aurreko pertsonak</td>
<td>A3</td>
</tr>
</tbody>
</table>
**RELATIONS section**

- Specific RRs queries where signals are underlined

<table>
<thead>
<tr>
<th>Relation: Kausa ‘Cause’ (27)</th>
<th>Left span</th>
<th>NS</th>
<th>Right span</th>
<th>Relation</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unita[...]</td>
<td>&lt; −</td>
<td>Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko hamarkadetan ingeleseari emandiote [...]</td>
<td>Cause</td>
<td>TERM18</td>
<td></td>
</tr>
<tr>
<td>In recent decades, many Serbian researchers working in different scientific fields have noticed a tendency and this is outlined here: the English unit [...]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terminologiak berak ere, uztartu egin behar ditu joera orokor horiek, eransten zaizkien beste batzuekin batera, hala no-la: teknologien [...]</td>
<td>&lt; −</td>
<td>gizartearekin lotuta dagoen jarduera denez,</td>
<td>Cause</td>
<td>TERM19</td>
<td></td>
</tr>
<tr>
<td>Terminology itself must seek to unite these general trends, along with others related to them, for example: technology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## CIRCUMSTANCE relation in three languages

<table>
<thead>
<tr>
<th>Left unit</th>
<th>Sense</th>
<th>Right unit</th>
<th>Relation name</th>
<th>Document</th>
<th>Tagger</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focussing on less widely used and taught languages (LWUTLs) including Irish,</td>
<td>-&gt;</td>
<td>the VOCALL partners are compiling multilingual glossaries of technical terms in the areas of computers, office skills and electronics and this involves the creation of a large number of new Irish terms in the above areas.</td>
<td>circumstance</td>
<td>TERM23</td>
<td>A1</td>
<td></td>
</tr>
<tr>
<td>Ever since information technology first made it possible to store and then process linguistic data,</td>
<td>-&gt;</td>
<td>terminology has had to adapt constantly to technological innovations.</td>
<td>circumstance</td>
<td>TERM29</td>
<td>A1</td>
<td></td>
</tr>
<tr>
<td>Desde que la informática hizo posible el almacenamiento de datos lingüísticos y posteriormente su tratamiento,</td>
<td>-&gt;</td>
<td>la terminología no ha cesado de adaptarse a las innovaciones tecnológicas,</td>
<td>circumstance</td>
<td>TERM29</td>
<td>A2</td>
<td></td>
</tr>
<tr>
<td>Informatikak hizkuntzako datuak gorde eta, aurrerago, tratatzeko aukera eman zigunetik,</td>
<td>-&gt;</td>
<td>terminologiak teknologi berrikuntzetara egokitu behar izan du etengabe.</td>
<td>circumstance</td>
<td>TERM29</td>
<td>A3</td>
<td></td>
</tr>
</tbody>
</table>
### SIGNALS section

Queries based on signals to detect which of them are ambiguous *baina* ‘but’ or unambiguous *erabiliz* ‘using’

<table>
<thead>
<tr>
<th>Signal: <em>baina</em> ‘but’</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gainerakoan, prokasu adierazle egokiak daude,</td>
<td>Kontzesioa <em>baina</em> altan dagoen gaixoaren ahalmen funtzionalaren erregistro urria antzematen da,</td>
</tr>
<tr>
<td><em>baina</em> <code>but</code></td>
<td>GMB0504</td>
</tr>
<tr>
<td>With respect to the other aspects, the indicators of process are good</td>
<td>Concession <em>but</em> there is poor recording of the patient’s functional capacity on discharge,</td>
</tr>
<tr>
<td>Bestalde, Euskaltzaindiak hitz elkartuen bidea (1995eko urtarrilaren 27an onartutako araua) proposatzen du adjektibo erreferentzialak itzultzeak,</td>
<td>Kontrastea <em>baina</em> arauan bertan esaten denez, “…ahal den guztian…”,</td>
</tr>
<tr>
<td>Euskaltzaindia proposed a mechanism of compound words (in a standard approved on January 27th 1995) for the translation of referential adjectives.</td>
<td>Contrast <strong>However</strong> the academy also confirmed,</td>
</tr>
<tr>
<td></td>
<td>…“whenever possible”,</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signal: <em>erabiliz</em> ‘using’</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Komunikazio honekin, hauxe frogatu nahi da: halako kasurik gehien-gehienetan, proposamen autoktonoa baztertzeko emandako arrazoia ez direla ez hizkuntzarenak ez semantikoak, soziologikoak baizik,</td>
<td>metodoa adibide paraleloak <strong>erabiliz</strong>,</td>
</tr>
<tr>
<td><em>erabiliz</em> <code>using</code></td>
<td>TERM21</td>
</tr>
<tr>
<td>The purpose of this paper is to show that in the vast majority of cases the local word is not rejected out of any linguistic or semantic reason but merely on sociological grounds which are sometimes implicitly acknowledged.</td>
<td>method <strong>through</strong> parallel examples,</td>
</tr>
<tr>
<td>Horretarako edo nagusiak lortu behar dira.</td>
<td>metodoa dauden hiztegi teknikoetan oinarritu, eta</td>
</tr>
<tr>
<td>To that end, principal models must be obtained.</td>
<td>metodoa teknika estatistikoak <strong>erabiliz</strong>,</td>
</tr>
<tr>
<td></td>
<td>TERM31</td>
</tr>
<tr>
<td></td>
<td><em>erabiliz</em>,</td>
</tr>
<tr>
<td></td>
<td>basing work on existing technical dictionaries and <strong>using</strong> statistical techniques,</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PART 1 — Discourse relations in RST: method

Corpora for corpus exploration

TREE section

- Some statistics and a lot of different file formats for the scientific community: TXT (plain text), XML (RS-tree), RS3 (RS-tree RSTTool format), RHETBD (annotation of signals), KAF (POS format)

<table>
<thead>
<tr>
<th>Files (88)</th>
<th>EDUs</th>
<th>RRs</th>
<th>P</th>
<th>SM</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMB0001-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>22</td>
<td>10</td>
<td>2</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>GMB0002-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GMB0201-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>37</td>
<td>12</td>
<td>3</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>GMB0202-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>20</td>
<td>13</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>GMB0203-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GMB0204-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GMB0301-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GMB0302-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GMB0401-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GMB0402-GS.rs3 segments figure XML text rs3 rhetdb kaf</td>
<td>17</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

- Statistics:
  - RRs: Different rhetorical relations
  - P: Presentational
  - SM: Subject-matter
  - Multi: Multinuclear
RST Discourse Treebank

- The RST Discourse Treebank (Carlson et al., 2002): https://catalog.ldc.upenn.edu/LDC2002T07
  - A corpus of 385 WSJ texts annotated with RST
- RST Signalling Corpus (Das et al., 2015): https://catalog.ldc.upenn.edu/LDC2015T10
  - The signalling annotation of 385 WSJ texts
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations
- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST

PART 3 — Tools for corpus exploration
- Segmenters
- CU detector
- Annotation tools for RST
- Evaluation tools/methods of RS
- Parsers

PART 4 — Resources
- Projects
- Resources
- Workshops
Applications based on RST

- Question answering
  - Improve the relevance of the questions (nuclearity, Central Unit)
  - Locate answers, create distractors with the same relation
  - Improve existing question answering tools (Lopez-Gazpio and Marichalar Anglada, 2013; Aldabe, 2011)

- Polarity extractor
  - Improve existing QWN-PPV polarity tool
  - Select relevant segments for sentiment analysis (Alkorta et al., 2015)
Outline

1. PART 1 — Discourse relations in RST: method
2. PART 2 — Practice
3. PART 3 — Tools for corpus exploration
4. PART 4 — Resources
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations
   - Signaling relational structures
   - An ambiguous RST analysis
   - Annotation in RST

3. PART 3 — Tools for corpus exploration
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
Segmentation. Modified GMB0301

– Segment all the EDUs of this text (with RSTweb or RSTTool):

(6) Recurrent aphtous stomatitis (I): epidemiologic, etiologic and clinical features.
Recurrent aphtous stomatitis is one of the most frequent oral conditions. Its etiology is controversial and it is characterised by the appearance of painful and recurrent ulcers, whose sizes, locations, and durations vary. These ulcers reappear periodically. This paper analyses the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

– Try online the segmenter of CODRA (Joty et al., 2015)
– Or try the SLSeg English segmenter (installation is needed)
Different segmentations of modified GMB0301

— Compare this segmentations:

<table>
<thead>
<tr>
<th>Text</th>
<th>GS</th>
<th>SEG1</th>
<th>SEG2</th>
<th>CODRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrent aphthous stomatitis is one of the most frequent oral conditions.</td>
<td>EDU2</td>
<td>EDU2</td>
<td>EDU2</td>
<td>EDU2</td>
</tr>
<tr>
<td>Its etiology is controversial and it is characterised by the appearance of painful and recurrent ulcers, whose sizes, locations, and durations vary.</td>
<td>EDU3</td>
<td>EDU3-B</td>
<td>EDU3-B</td>
<td>EDU3</td>
</tr>
<tr>
<td>These ulcers reappear periodically.</td>
<td>EDU4-B</td>
<td>EDU3-E</td>
<td>EDU3-M</td>
<td>EDU4</td>
</tr>
<tr>
<td>This paper analyses the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.</td>
<td>EDU4-E</td>
<td>EDU4</td>
<td>EDU3-E</td>
<td>EDU5</td>
</tr>
</tbody>
</table>

— Explain the errors of each segmentation (SEG1, SEG2 and CODRA) in terms of missed (M) and excess (E) EDUs:

— SEG1: 1M and 1E
— SEG2: 1M
— CODRA: 1E
Different segmentations of modified GMB0301

- Compare this segmentations:

<table>
<thead>
<tr>
<th>Text</th>
<th>GS</th>
<th>SEG1</th>
<th>SEG2</th>
<th>CODRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrent aphtous stomatitis is one of the most frequent oral</td>
<td>EDU2</td>
<td>EDU2</td>
<td>EDU2</td>
<td>EDU2</td>
</tr>
<tr>
<td>conditions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Its etiology is controversial and</td>
<td>EDU3</td>
<td>EDU3-B</td>
<td>EDU3-B</td>
<td>EDU3</td>
</tr>
<tr>
<td>it is characterised by the appearance of painful and recurrent</td>
<td>EDU4-B</td>
<td>EDU3-E</td>
<td>EDU3-M</td>
<td>EDU4</td>
</tr>
<tr>
<td>ulcers, whose sizes, locations, and durations vary.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>These ulcers reappear periodically.</td>
<td>EDU5</td>
<td>EDU5</td>
<td>EDU4</td>
<td>EDU6</td>
</tr>
<tr>
<td>This paper analyses the most important epidemiological, etiological,</td>
<td>EDU6</td>
<td>EDU6</td>
<td>EDU5</td>
<td>EDU7</td>
</tr>
<tr>
<td>pathological and clinical features of this common oral pathology.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Explain the errors of each segmentation (SEG1, SEG2 and CODRA) in terms of missed (M) and excess (E) EDUs:
  - SEG1: 1M and 1E
  - SEG2: 1M
  - CODRA: 1E
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration
- Segmenters
- CU detector
- Annotation tools for RST
- Evaluation tools/methods of RS

PART 4 — Resources
- Projects
- Resources
- Workshops

PART 2 — Practice
- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrent aphtous stomatitis (I): epidemiologic, etiologic and clinical features.</td>
<td>Recurrent aphtous stomatitis is one of the most frequent oral conditions.</td>
<td></td>
<td>It is characterised by the appearance of painful and recurrent ulcers, whose sizes, locations, and durations vary.</td>
<td></td>
<td></td>
<td>This paper analyzes the most important epidemiological, etiological, pathological, and clinical features of this common oral pathology.</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Recurrent aphtous stomatitis (I): epidemiologic, etiologic and clinical features.</td>
<td>Recurrent aphtous stomatitis is one of the most frequent oral conditions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.</td>
</tr>
</tbody>
</table>

— Summarize the text above choosing 3 or 4 discourse units:
− Has the created summary any sense?

2

“Estomatitis aftosa recurrente” deritzon patologia, ahoan agertzen den ugarientako bat da.

Recurrent aphtous stomatitis is one of the most frequent oral conditions.

4

Ultzera mingarri batzu bezela agertzen da,

It is characterised by the appearance of painful and recurrent ulcers,

5

tamainu, kokapena eta iraunkortasuna aldakorra izanik.

whose sizes, locations, and durations vary.

7

Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantzitsuenak analizatzen ditugu.

This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

− Choose now the 2 most important discourse segments
Has the created summary any sense?

2

“Estomatitis aftosa recurrente”
deritzon patologia, ahoan agertzen den ugarienetako bat da.

Recurrent aphthous stomatitis is one of the most frequent oral conditions.

4

Ultzera mingarri batzu bezela agertzen da,

It is characterised by the appearance of painful and recurrent ulcers,

5

tamainu, kokapena eta iraunkortasuna aldakorra izanik.

whose sizes, locations, and durations vary.

7

Lan honetan patologia arrunt honetan ezaguarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantzitsuenak analizatzen ditugu.

This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

Choose now the 2 most important discourse segments
Has the created summary any sense?

“Estomatitis aftosa recurrente”
deritzon patologia,
ahoan agertzen
den ugarienetako
bat da.

Recurrent aphtous stomatitis is one of the most frequent oral conditions.

Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantzitsuenak analizatzen ditugu.

This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

Choose now the central unit or the most salient discourse unit:
Has the created summary any sense?

“Estomatitis aftosa recurrente”
deritzon patologia,a
dehoan agertzen
den ugarienetako
bat da.
-----------------
Recurrent aphtous
stomatitis is one
of the most
frequent oral
conditions.

Choose now the central unit or the most salient discourse unit:

Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiapatogeniko eta klinikopatologiko garrantzitsuenak analizatzen ditugu.
------------------
This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.
Has the central unit any topic indicator?

This paper analyzes the most important...
Nuclearity and summarization: GMB0301

- Has the central unit any topic indicator?
  - *This paper analyzes the most important...*
Delete the satellites, deletion macro-rule (van Dijk, 1983):

- After the deletion of these propositions, the core of the text is still coherent

- If we maintain the nuclear units (units: 2, 4, 5 and 7) the text GMB0301 is summarized as in Example (7).

(7) Recurrent aphtous stomatitis is one of the most frequent oral conditions. It is characterised by the appearance of painful and recurrent ulcers, whose sizes, locations, and durations vary. This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

Recurrent aphthous stomatitis is one of the most frequent oral conditions. Its etiology is controversial. It is characterised by the appearance of painful and recurrent ulcers, whose sizes, locations, and durations vary. These ulcers reappear periodically. This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.
A simplification of the RS-tree. GMB0301

After deleting the satellite units the text part is still coherent.
A simplification of the RS-tree. GMB0301

After deleting the satellite units, the text part is still coherent.
The text obtained with satellites is incoherent or it fails describing the global meaning

- The representation of the RS-tree is different

(9) # [Recurrent aphthous stomatitis (l): epidemiologic, etiologic and clinical features.]_1_ [Its etiology is controversial.]_3_ [These ulcers reappear periodically.]_6_ GMB0301
### Basic heuristics based on nuclearity

<table>
<thead>
<tr>
<th>Heuristics</th>
<th>Example</th>
<th>EDUs</th>
<th>Words</th>
<th>Summ. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>The text</td>
<td>(6)</td>
<td>1, 2, 3, 4, 5, 6, 7</td>
<td>53</td>
<td>% 0,00</td>
</tr>
<tr>
<td>All the Ns</td>
<td>(10)</td>
<td>2, 4, 5, 7</td>
<td>36</td>
<td>% 32,08</td>
</tr>
<tr>
<td>CU + another N</td>
<td>(11)</td>
<td>2, 7</td>
<td>24</td>
<td>% 54,72</td>
</tr>
<tr>
<td>The CU of the text (the principal N)</td>
<td>(12)</td>
<td>7</td>
<td>13</td>
<td>% 75,47</td>
</tr>
<tr>
<td>The incoherent text</td>
<td>(9)</td>
<td>1, 3, 6</td>
<td>17</td>
<td>% 67,92</td>
</tr>
</tbody>
</table>

(10) Recurrent aphtous stomatitis is one of the most frequent oral conditions. It is characterised by the appearance of painful and recurrent ulcers, whose size, locations, and durations vary. This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

(11) Recurrent aphtous stomatitis is one of the most frequent oral conditions. This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

(12) This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.
Automatic summarization in Basque

- Automatic summarization is a well known task in NLP
  - Works based on RST (Ono et al., 1994; O’Donnell, 1997; Bosma, 2008)
  - There is not any proposal for Basque
- Our aim is to study whether some features can help to select the most important discourse units
  - Discourse units not related to the central unit and satellites of CU as ELABORATION, BACKGROUND, PREPARATION can be omitted from extractive summaries
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops

- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST
1. Secondly, we must make it clear that the prefix-core/base-complement of the romance languages and English has a corresponding feature in Basque in base-complement/suffix-core. ← This is an important contribution to modern lexicography.

2. Key words are extracted from parsing such definitions so that literal translation of English key words into Chinese can be achieved. ←→ Then the Chinese key word translations are processed in the coiner making use of Chinese morpheme database and Chinese word formation rules.

3. In recent years work has begun to develop instruments in several languages for automatic terminology extraction in technical texts, ← though human intervention is still required to make the final selection from the terms automatically chosen.
1. Secondly, we must make it clear that the prefix-core/base-complement of the romance languages and English has a corresponding feature in Basque in base-complement/suffix-core. ← This is an important contribution to modern lexicography. INTERPRETATION

2. Key words are extracted from parsing such definitions so that literal translation of English key words into Chinese can be achieved. ←→ Then the Chinese key word translations are processed in the coiner making use of Chinese morpheme database and Chinese word formation rules. SEQUENCE

3. In recent years work has begun to develop instruments in several languages for automatic terminology extraction in technical texts, ← though human intervention is still required to make the final selection from the terms automatically chosen. CONCESSION
1. Focussing on less widely used and taught languages (LWUTLs) including Irish, the VOCALL partners are compiling multilingual glossaries of technical terms in the areas of computers, office skills and electronics and this involves the creation of a large number of new Irish terms in the above areas. –> With the help of the Terminology Committee for the Irish Language (An Coiste Tarmaochta) Fiontar and VOCALL are addressing the terminological needs of both Irish-medium third level education and Irish-medium vocational training.

2. Once all this is correctly organised in a single text we can mould the “legal discourse” of Basque. –> To attain this goal we have been translating doctrinal texts in law at the University of Deusto since 1994.
Choosing relations: SOLUTIONHOOD or PURPOSE

1. Focussing on less widely used and taught languages (LWUTLs) including Irish, the VOCALL partners are compiling multilingual glossaries of technical terms in the areas of computers, office skills and electronics and this involves the creation of a large number of new Irish terms in the above areas. —> With the help of the Terminology Committee for the Irish Language (An Coiste Tarmachta) Fiontar and VOCALL are addressing the terminological needs of both Irish-medium third level education and Irish-medium vocational training. SOLUTIONHOOD

2. Once all this is correctly organised in a single text we can mould the “legal discourse” of Basque. —> To attain this goal we have been translating doctrinal texts in law at the University of Deusto since 1994. PURPOSE
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration
- Segmenters
- CU detector
- Annotation tools for RST
- Evaluation tools/methods of RS
- Parsers

PART 4 — Resources
- Projects
- Resources
- Workshops

Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST
Mention what the signal is and where (N or S) it is:

1. While these tools are being prepared, we must work on the modelling of technical terms, i.e. we must reduce their characteristics.

2. Mientras se preparan dichas herramientas, habremos de trabajar sobre la modelización de los términos técnicos, es decir, hemos de reducir las características de los mismos.

3. Tresna horiek prest dauden bitartean — termino teknikoen modelizazioari ekin behar diogu, hau da murriztu behar ditugu termino teknikoen ezaugarriak.
1. **While** these tools are being prepared, –> we must work on the modelling of technical terms, i.e. we must reduce their characteristics.

2. **Mientras** se preparan dichas herramientas, –> habremos de trabajar sobre la modelización de los términos técnicos, es decir, hemos de reducir las características de los mismos.

3. Tresna horiek prest dauden **bitartean** –> termino teknikoen modelizazioari ekin behar diogu, hau da murriztu behar ditugu termino teknikoen ezaugarriak.
CONCESSION: signals

- Mention what the signal is and where (N or S) it is:
  1. The basic principles of standardisation, such as consensus between the sectors of society involved, remain fully valid in guaranteeing specialist communication, \( \rightarrow \) but in practical terminological work the close relationship which must exist between standardisation and society is sometimes neglected.
  2. Nahiz eta gaur egun normalizazioko oinarrizko printzipioek balio osoa gorde komunikazio espezialduaren bermearen bidez (eta elkarrekin zerikusia duten gizarteko sektoreen arteko adostasuna da printzipio horietako bat), \( \rightarrow \) terminologiako lan praktikoan, batzuetan, ahaztuxe uzten da normalizazioaren eta gizartearen artean egon behar den lotura estua.
CONCESSION: signals II

1. The basic principles of standardisation, such as consensus between the sectors of society involved, remain fully valid in guaranteeing specialist communication, → **but** in practical terminological work the close relationship which must exist between standardisation and society is sometimes neglected.

2. **Nahiz eta** gaur egun normalizazioko oinarrizko printzipioek balio osoa gorde komunikazio espezialduaren bermearen bidez (eta elkarrekin zerikusia duten gizarteko sektoreen arteko adostasuna da printzipio horietako bat), → terminologiako lan praktikoan, batzuetan, ahatxute uzten da normalizazioaren eta gizartearen artean egon behar den lotura estua.
PART 2 — Practice  Signaling relational structures

CONDITION: signals

- Mention what the signal is and where (N or S) it is:

1. We wish to indicate the difficulties we have had over the years and also our achievements, ← if there can be said to be any.
2. halakorik izanez gero, → lorpenak ere azaldu nahi ditugu.
3. If a similar instrument is to be developed for Basque → we shall come up against more major drawbacks, because the unifying process of the language has not been completed, research carried out is limited and Basque is an agglutinative language.
1. We wish to indicate the difficulties we have had over the years and also our achievements, ← if there can be said to be any.

2. halakorik izan ez gero, → lorpenak ere azaldu nahi ditugu.

3. If a similar instrument is to be developed for Basque → we shall come up against more major drawbacks, because the unifying process of the language has not been completed, research carried out is limited and Basque is an agglutinative language.

ELABORATION: Signals

- Mention what the signal is and where (N or S) it is:

1. For the translation of legal texts it is absolutely necessary to study terminology. ← In the case of Basque the need is even greater, as our language is not in a good situation in the field of law.

2. Para la traducción de textos jurídicos es totalmente necesario el estudio de la terminología ← y en el caso del euskera esa necesidad es aún más acentuada, ya que en el ámbito jurídico nuestra lengua no se encuentra todavía en una buena situación.

1. For the translation of legal texts it is absolutely necessary to study terminology. \(\text{\textless}\) In the case of Basque the need is even greater, as our language is not in a good situation in the field of law.

2. Para la traducción de textos jurídicos es totalmente necesario el estudio de la terminología \(\text{\textless}\) y en el caso del euskera esa necesidad es aún más acentuada, ya que en el ámbito jurídico nuestra lengua no se encuentra todavía en una buena situación.

3. Testu juridikoen itzulpenari ekiteko, ezinbestekoa da terminologia bera lantzea. \(\text{\textless}\) Euskara\text{\texttt{ren kasuan}} beharrizan hori areagotu egin da, esparru horretan gure hizkuntzaren egoera ez baita primerakoa, ezta hurrik eman ere.

— Check more multilingual examples at http://ixa2.si.ehu.es/rst/ (Iruskieta et al., 2015a)
CAUSE: signals

- Mention what the signal is and where (N or S) it is:

1. In the case of Basque the need is even greater, ← as our language is not in a good situation in the field of law.
2. y en el caso del euskera esa necesidad es aún más acentuada, ← ya que en el ámbito jurídico nuestra lengua no se encuentra todavía en una buena situación.
3. Euskararen kasuan beharrizan hori areagotu egin da, ← esparru horretan gure hizkuntzaren egoera ez baita primerakoa, ezta hurrik eman ere.
CAUSE: signals II

1. In the case of Basque the need is even greater, as our language is not in a good situation in the field of law.

2. y en el caso del euskera esa necesidad es aún más acentuada, ya que en el ámbito jurídico nuestra lengua no se encuentra todavía en una buena situación.

3. Euskararen kasuan beharrizan hori areagotu egin da, esparru horretan gure hizkuntzaren egoera ez baita primerakoa, ezta hurrik eman ere.
CAUSE: signals III

- **Mention what the signal is and where (N or S) it is:**

  1. we based our study on those originals and then found their Basque equivalents, ← in the sure knowledge that legal terminology in Spanish is sufficiently well consolidated and set down in dictionaries.

  2. Habida cuenta de que las versiones en euskera son traducciones de las originales en castellano, → nos hemos basado en estas últimas para luego poder encontrar los equivalentes vascos, en la seguridad de que la terminología jurídica en castellano está suficientemente consolidada y recopilada en sus correspondientes diccionarios.

  3. Legeen euskal bertsioak gaztelaniazko jatorrizko testuen itzulpenak direnez, → erdal testuez baliatu gara,
1. we based our study on those originals and then found their Basque equivalents, \textit{in the sure knowledge that} legal terminology in Spanish is sufficiently well consolidated and set down in dictionaries.

2. \textit{Habida cuenta de que} las versiones en euskera son traducciones de las originales en castellano, \textit{nos hemos basado} en estas últimas para luego poder encontrar los equivalentes vascos, en la seguridad de que la terminología jurídica en castellano está suficientemente consolidada y recopilada en sus correspondientes diccionarios.

3. Legeen euskal bertsioak gaztelaniazko jatorrizko testuen itzulpenak dire\textit{ne}z, \textit{erdal testuez baliatu} gara,
Relations in Portuguese

— Mention what the signal is and where (N or S) it is:

1. A internet se tornou um recurso tecnológico fundamental para nossa vida, ← porém, em alguns casos ela se torna nociva.

2. Jogos de carros que atropelam as pessoas e de armas altamente destrutivas estimulam a juventude ao mundo do crime, ← já que os criminosos do mundo virtual nunca são punidos.

3. Portanto a internet deve sim ser utilizada no cotidiano, mas seu uso deve ser moderado e restrito, ← para que os jovens e crianças cresçam conscientes de que seu uso indevido não os favorece em nada, somente acarreta o surgimento de anomalias na sociedade, tais como a criminalidade.
Relations in Portuguese

1. A internet se tornou um recurso tecnológico fundamental para nossa vida, ←> **porém**, em alguns casos ela se torna nociva. **CONTRAST**

2. Jogos de carros que atropelam as pessoas e de armas altamente destrutivas estimulam a juventude ao mundo do crime, ← **já que** os criminosos do mundo virtual nunca são punidos. **CAUSE**

3. Portanto a internet deve sim ser utilizada no cotidiano, mas seu uso deve ser moderado e restrito, ← **para que** os jovens e crianças cresçam conscientes de que seu uso indevido não os favorece em nada, somente acarreta o surgimento de anomalias na sociedade, tais como a criminalidade. **PURPOSE**

— Check more Brazilian Portuguese examples at

http://ixa2.si.ehu.es/rst/pt/ (Antonio and Iruskieta, 2014)
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS
   - Parsers

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops

- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST
An ambiguous example: The text

(13) He wanted to play tennis with Jane, but also wanted to have dinner with Susan. This indecision drove him crazy.

[Adapted example from Pardo et al. (2004)]

Berak Jonerekin tenisean jolastu nahi zuen, baina ordu berean Susanarekin bazkaldunahi zuen. Ezin erabakitze horrek zoratu egin du mutila.

— How many propositions or discourse units are there in the example?
An ambiguous example: segments

1

Berak Jonerek

tenisean jolastu

nahi zuen,

--------------------

He wanted to

play tennis with

Jane,

2

baina ordu

berean

Susanarekin

bazkaldunahi

zuen.

--------------------

but also wanted

to have dinner

with Susan.

3

Ezin erabakitze

horrek zoratu

egin du mutila.

--------------------

This indecision
drove him
crazy.

— How many possibilities are there to link these 3 segments? Explain your choices.

- Link first the intrasentential EDUs (the incremental way)
### An ambiguous example: segments

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berak Jonerekintensean jolastunahi zuen,</td>
<td>baina orduberean</td>
<td>Ezin erabakitzehorrek zoratu egin du mutila.</td>
</tr>
<tr>
<td>He wanted to play tennis with Jane,</td>
<td>bazkaldunahi zuen.</td>
<td>This indecision drove him crazy.</td>
</tr>
<tr>
<td></td>
<td>but also wanted to have dinner with Susan.</td>
<td></td>
</tr>
</tbody>
</table>

- How many possibilities are there to link these 3 segments? Explain your choices.
  - Link first the intrasentential EDUs (the incremental way)
An ambiguous example: the incremental way

– Linking the intrasentential segments

1. Berak Jonerekin tenisean jolastu nahi zuen, He wanted to play tennis with Jane,

2. baina ordu berean Susanarekin bazkaldunahi zuen. but also wanted to have dinner with Susan.

3. Ezin erabakitze horrek zoratu egin du mutila.

   This indecision drove him crazy.
An ambiguous example: the incremental way

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Berak Jonerekin</td>
<td>baina ordu</td>
<td>Ezin erabakitze</td>
</tr>
<tr>
<td></td>
<td>tenisean jolastu</td>
<td>berean</td>
<td>horrek zoratu</td>
</tr>
<tr>
<td></td>
<td>nahi zuen,</td>
<td>Susanarekin</td>
<td>egin du mutila.</td>
</tr>
<tr>
<td></td>
<td>----------------</td>
<td>bazkaldu nahi</td>
<td>----------------</td>
</tr>
<tr>
<td>2</td>
<td>He wanted to</td>
<td>zuen.</td>
<td>This indecision</td>
</tr>
<tr>
<td></td>
<td>play tennis with</td>
<td>but also wanted</td>
<td>drove him</td>
</tr>
<tr>
<td></td>
<td>Jane,</td>
<td>to have dinner</td>
<td>crazy.</td>
</tr>
</tbody>
</table>

— Linking the intrasentential segments

— Now link the (inter)sentential span or segments
An ambiguous example: the incremental way

- Linking the (inter)sentential span or segments

```
1-3
  ├── Cause
  │    ├── 1-2
  │    │    └── Contrast
  │    │        └── 1
  │    │            ├── Berak Jonerekin
tenisean jolastu
  │    │            │   nahi zuen,
  │    │            └── He wanted to play tennis with Jane,
  │    │            └── ---------------------
  │    │            └── baina ordu berean
  │    │            └── Susanarekin
  │    │            └── bazkaldu nahi zuen.
  │    │            └── ---------------------
  │    │            └── but also wanted to have dinner with Susan.
  │    │            └── Ezin erabakitze horrek zoratu
egin du mutila.
  │    │            └── ---------------------
  │    │            └── This indecision drove him crazy.
  │    └── 3
  └── 3
```

- Now choose the nucleus (N) and satellite (S) units at intrasentential level
An ambiguous example: the incremental way

- Linking the (inter)sentential span or segments

1. Berak Jonerekin tenisean jolastu nahi zuen, He wanted to play tennis with Jane,
2. baina ordu berean Susanarekin bazkaldu nahi zuen. but also wanted to have dinner with Susan.
3. Ezin erabakitze horrek zoratu egin du mutila. This indecision drove him crazy.

- Now choose the nucleus (N) and satellite (S) units at intrasentential level
An ambiguous example: nuclearity of intrasentential units

Choosing the nucleus (N) and satellite (S) units of intrasentential units

1. Berak Jonerek
   tenisean jolastu
   nahi zuen,
   ----------------
   He wanted to
   play tennis with
   Jane,

2. baina ordu
   berean
   Susanarekin
   bazkaldunahi
   zuen.
   ----------------
   but also wanted
   to have dinner
   with Susan.

3. Ezin erabakitze
   horrek zoratu
   egin du mutila.
   ----------------
   This indecision
drove him
crazy.

Now choose the nucleus (N) and satellite (S) units of sentential units
An ambiguous example: nucleus and satellite units of intrasentential units

Choosing the nucleus (N) and satellite (S) units of intrasentential units

- Berak Jonerekin tenisean jolastu nahi zuen,
- He wanted to play tennis with Jane,
- baina ordu berean Susanarekin bazkaldunahi zuen.
- but also wanted to have dinner with Susan.
- Ezin erabakitzehorrekezoratu egin du mutila.
- This indecision drove him crazy.

Now choose the nucleus (N) and satellite (S) units of sentential units
An ambiguous example: nuclearity of sentential units

- Choosing the nucleus (N) and satellite (S) units of sentential units

```
[Diagram showing RST relationships]
```

Berak Jonerekin
tenisean jolastu
nahi zuen,
--------------
He wanted to
play tennis with
Jane,
baina ordu
berean
Susanarekin
bazkaldunahi
zuen.
--------------
but also wanted
to have dinner
with Susan.
Ezin erabakitze
horrek zoratu
egin du mutila.
--------------
This indecision
drove him
crazy.
The importance of nuclearity in relational discourse structure

- S-N relations (nuclear) are represented with arrows in RST

```
Berak Jonerekin tenisean
jolastu nahi zuen, ------
He wanted to play tennis
with Jane,

Contrast

baina ordu berean
Susanarekin bazkaldunahi zuen. --------------
but also wanted to have
dinner with Susan.
```

- What is the difference between these two structures?

```
Berak Jonerekin tenisean
jolastu nahi zuen, ------
He wanted to play tennis
with Jane,

Contrast

baina ordu berean
Susanarekin bazkaldunahi zuen. --------------
but also wanted to have
dinner with Susan.
```

Ezin erabakitze horrek zoratu egin du mutila.
This indecision drove him crazy.
The importance of nuclearity in relational discourse structure

We can not choose anyone, if we do not decide first the CU

- In real text examples more context is often available
Problems of the tree structure representation, for discussion

- The modularity principle is sometimes violated in real texts
  - Are paragraphs always attached to the CU? (depends on genre)
  - Do all written texts follow the idea of “1 paragraph = 1 idea”?

- Multiple relations:
  - Has the reader/writer on her mind multiple relations when she reads/writes a text?
  - Hierarchy sometimes is a simplification of all the possible relation structures, but a macrostructure (a high level representation of a text) can be achieved

- Linking units at the top level of a tree is sometimes difficult
Tools and exercises

a) RST annotation with RSTTool:
   http://www.wagsoft.com/RSTTool/
   − Segment this text TERM18 and build the RS-tree
   − Compare analyses among pairs and comment on the annotations
     • Is there any way to harmonize them?
   − Compare the harmonized RS-tree with the annotation at the multilingual RST Treebank at TERM18

b) Annotate signals with Rhetorical Database: http://www.icmc.usp.br/~tasparдо/RhetDB_Install.zip
   − First get the appropriate format with RSTToolkit:
     http://www.icmc.usp.br/~tasparдо/RSTToolkit_Install.rar

Go to corpus exploration: 61
Outline

1. PART 1 — Discourse relations in RST: method
2. PART 2 — Practice
3. PART 3 — Tools for corpus exploration
4. PART 4 — Resources
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Signaling relational structures
   - An ambiguous RST analysis
   - Annotation in RST
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS
   - Parsers

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
SENDER: Brazilian Portuguese segmenter

- SENDER is a fine graned intrasentential segmenter for RST
  - First step for DiZer (Automatic Discourse Analyzer)
  - http://143.107.183.175:21480/segmenter/

Syntax-based text segmentation tool

This tool was developed as a first step towards more accurate automatic rhetorical analysis for Brazilian Portuguese, following RST (Rhetorical Structure Theory) (Mann and Thompson, 1987). It is part of DiZer project (Pardo, 2005) and uses the parser PALAVRAS (Bick, 2000). The tool purpose is to automatically detect discourse segments (i.e., text segments that express minimum content units - propositions) that will be used for building the corresponding RST tree.

Type or paste the text to be segmented:

Contact:
Erick G. Maziero (e-mail)
Thiago A. S. Pardo
DiSeg: Spanish RST segmenter

- DiSeg is an intrasentential rule based segmenter
  - Rules based on lexical and syntactic rules
  - http://diseg2.termwatch.es/
EusEduSeg: syntax-based text segmentation tool for Basque

Contact: mikel.iruskieta at ehu.es

In the framework of the Rhetorical Structure Theory (RST by Mann and Thompson, 1987), this segmenter was developed as a first step towards an automatic rhetorical analysis for Basque. The segmenter uses the parser MALTIXA (Diaz de Illarraz et al. 2005) and our purpose is to automatically detect the Elementary Discourse Units (EDUs) or discourse segments (propositions). EDU segmentation is defined in Iruskieta (2014). In future, this segmentation will be the basis for building automatically the corresponding RST tree or other many NLP applications.

NOTE: With the aim of preserving the paragraphs, this tool considers every line break as a paragraph.

References:
Our segmenter is based on MALTIXA an automatic dependency analyzer (Díaz de Ilarraza et al., 2005)

(14)  a. [Toponimo kontzeptuaren definizioa emango dugu,]₁ [geografiako terminoarekin duen loturari dagokionean.]₂

b. [We present the definition of a toponym,]₁ [regarding to geographical terms.]₂
EusEduSeg: System architecture

- Entirely based on dependency and linguistic rules
- A versatile tool with different outputs:
  a) to use in different NLP tasks: a line-break format
  b) to manually annotate text with RSTTool: RS3 format
  c) to use in a discourse parser: DiZer format
- http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration
- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST

PART 4 — Resources
- Projects
- Resources
- Workshops
Central Unit detector for Basque and B. Portuguese

- Ongoing projects
  J.D. Antonio  G. Labaka

- Detection of the Central Unit of
  - Science abstracts by researchers (Basque)
  - Argumentative answers by students (B. Portuguese)

- Heuristics based on nouns, verbs, pronouns, bonus words, title words, EDU position in the document, main verb
  - Results for Basque: $F_1$ of 0.51
  - Results for B. Portuguese: $F_1$ of 0.55

- Machine learning techniques
  - Results for Basque: $F_1$ of 0.51 (ongoing)
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration
- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST

PART 4 — Resources
- Projects
- Resources
- Workshops
rstWeb Tool (a collaborative RS-tree)

- Collaborative web annotation at https://corpling.uis.georgetown.edu/rstweb/info/
RSTTool (to structure trees)

- Manual segmentation and rhetorical annotation

- Further tools based on RSTTool output format (RS3)
  - Rhetorical Database for signal annotation
  - Web resources for corpus exploration: the Basque RST Treebank and the Multilingual RST Treebank
- To change the format for Rhetorical Database
- To extracts tree structure statistics
- [Link](http://www.icmc.usp.br/~taspardo/RSTToolkit_Install.rar)
Rhetorical DataBase (to signal)

- To annotate signals
  - Relation by relation
  - Check consistency
- Extract statistics of the signals
- http://www.icmc.usp.br/~taspadro/RhetDB_Install.zip
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Signaling relational structures
   - An ambiguous RST analysis
   - Annotation in RST

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
RSTeval input (to compare RS-trees)

- Compares an (automatic) annotation with a gold standard annotation (BP, ENG, SPA, BSQ) (Maziero and Pardo, 2009)
RSTeval output (for comparing RS-trees)

- A quantitative evaluation method based on Marcu (2000a)
The aim of the qualitative evaluation is to describe the (dis)similarities of two RS-trees (Iruskieta et al., 2015a)

- Understand annotator decisions
- Describe translation strategies
Our evaluation method

- Quantitative RS-tree evaluation method (Marcu, 2000a) by means of EDUs, spans, nuclearity and RRs
  - Automatic comparison (Maziero and Pardo, 2009)
  - Not independent factors (NS and RRs) (van der Vliet, 2010a)
  - RRs are not (well) compared (Iruskieta et al., 2013b)

- A more accurate comparison
  - Independent factors
  - Qualitative description of agreement and disagreement

- Measurement of RS
  - in Basque-Basque (Iruskieta et al., 2013a)
  - in Basque-Spanish (da Cunha and Iruskieta, 2010) and in Basque-English-Spanish (Iruskieta et al., 2015a)
Our evaluation method: decision trees

- Qualitative agreement

- Qualitative disagreement
### RR confusion matrix (BSQ vs BSQ)

|       | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z | Total |
| ENABLEMENT | a | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 3 |
| ANTITHESIS    | b | 1 | 1 | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 5 |
| SOLUTIONHOOD | c | 14 | 2 |   |   | 1 | 3 | 9 | 13 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 23 |
| CONDITION     | d |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| JOINT         | e |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| RESTATEMENT   | f | 4 | 2 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 8 |
| DISJUNCTION   | g |   | 1 |   |   | 3 | 1 | 8 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 2 |
| EVALUATION    | h |   | 1 |   | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 8 |
| EVIDENCE      | i |   |   |   |   |   | 3 | 3 | 1 | 1 | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   |   | 10 |
| ELABORATION   | j | 8 | 1 | 162 | 2 | 5 | 1 | 6 | 18 | 2 | 14 | 13 | 2 | 15 | 4 | 49 |   |   |   |   |   |   |   |   | 302 |
| UNCONDITIONAL | k |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 1 |
| NO-EDU        | l |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 2 |
| PURPOSE       | ll | 10 | 88 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 108 |
| INTERPRETATION| m | 4 | 9 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 25 |
| JUSTIFY       | n | 1 | 2 | 11 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 18 |
| CAUSE         | ñ | 1 | 4 | 24 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 8 |
| CONJUNCTION   | o | 2 | 3 |   |   |   |   | 27 | 1 | 14 | 5 | 3 | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   | 37 |
| CONTRAST      | p | 2 | 5 |   |   |   | 12 | 5 | 5 | 2 | 1 | 2 |   |   |   |   |   |   |   |   |   |   |   |   |   | 35 |
| CONCESSION    | q | 3 | 1 | 1 | 3 | 2 | 26 | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 38 |
| SUMMARY       | r | 3 |   |   |   |   | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 5 |
| LIST          | s | 2 | 12 | 1 | 15 | 2 | 1 | 125 | 1 | 2 | 2 | 3 |   |   |   |   |   |   |   |   |   |   |   |   | 166 |
| MEANS         | t | 17 | 1 | 3 |   |   | 1 | 63 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 92 |
| MOTIVATION    | u | 1 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 3 |
| RESULT        | v | 1 | 12 | 3 | 1 | 1 | 1 | 1 | 1 | 39 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   | 60 |
| PREPARATION   | w | 12 |   |   |   |   |   |   |   |   |   | 1 | 79 | 15 |   |   |   |   |   |   |   |   |   |   |   | 107 |
| SEQUENCE      | x | 1 | 2 | 1 | 4 |   |   | 9 | 3 | 16 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   | 37 |
| BACKGROUND    | y | 1 | 4 | 2 | 2 | 5 | 1 | 1 | 54 | 1 |   |   |   |   |   |   |   |   |   |   |   |   |   | 71 |
| CIRCUMSTANCE  | z | 1 | 2 | 3 | 1 | 4 | 1 |   | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 56 |
| **Total**     | | 4 | 15 | 17 | 4 | 1 | 2 | 6 | 267 | 91 | 30 | 3 | 52 | 74 | 19 | 32 | 6 | 171 | 95 | 4 | 99 | 80 | 27 | 145 | 48 | 1292 |
### Reliability of RRs, agreement: Fleiss (1971) Kappa

<table>
<thead>
<tr>
<th>RRs</th>
<th>Kappa</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PURPOSE</td>
<td>0.872</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>PREPARATION</td>
<td>0.836</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>CIRCUMSTANCE</td>
<td>0.772</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>CONCESSION</td>
<td>0.743</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>CONDITION</td>
<td>0.733</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>LIST</td>
<td>0.710</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>DISJUNCTION</td>
<td>0.666</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>RESTATEMENT</td>
<td>0.665</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>MEANS</td>
<td>0.633</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>SEQUENCE</td>
<td>0.556</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>CAUSE</td>
<td>0.527</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>RESULT</td>
<td>0.458</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>ELABORATION</td>
<td>0.448</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>0.448</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>CONTRAST</td>
<td>0.416</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>CONJUNCTION</td>
<td>0.404</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>EVIDENCE</td>
<td>0.371</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>INTERPRETATION</td>
<td>0.313</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>ANTITHESIS</td>
<td>0.220</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>EVALUATION</td>
<td>0.178</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>0.178</td>
<td>&gt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RRs</th>
<th>Kappa</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>JUSTIFY</td>
<td>-0.008</td>
<td>0.760</td>
</tr>
<tr>
<td>JOINT</td>
<td>-0.007</td>
<td>0.803</td>
</tr>
<tr>
<td>SOLUTIONHOOD</td>
<td>-0.005</td>
<td>0.857</td>
</tr>
<tr>
<td>MOTIVATION</td>
<td>-0.003</td>
<td>0.923</td>
</tr>
<tr>
<td>ENABLEMENT</td>
<td>-0.001</td>
<td>0.967</td>
</tr>
<tr>
<td>UNCONDITIONAL</td>
<td>0.001</td>
<td>0.989</td>
</tr>
</tbody>
</table>

- Strong agreement (above average) in 9 RRs
- Weak agreement (below average) in 7 RRs
- Bad agreement in 5 RRs (with red color)
- No enough data for 6 RRs
Relevant RR disagreement: confusion matrix

<table>
<thead>
<tr>
<th>RRs</th>
<th>#</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELABORATION</td>
<td>50</td>
<td>183</td>
</tr>
<tr>
<td>MEANS</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>LIST</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>ELABORATION</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>ELABORATION</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>CONJUNCTION</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>INTERPRETATION</td>
<td>13</td>
<td>69</td>
</tr>
<tr>
<td>PREPARATION</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>PURPOSE</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>JUSTIFY</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>SEQUENCE</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>MEANS</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>SOLUTIONHOOD</td>
<td>9</td>
<td>60</td>
</tr>
<tr>
<td>ELABORATION</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>ELABORATION</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>CONJUNCTION</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>CAUSE</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CONTRAST</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CONTRAST</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>CONTRAST</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>312</strong></td>
</tr>
</tbody>
</table>

- One of them is the most widely used RR: 47.21% (ELABORATION-X)
- Similar RRs: 4.1% (LIST-CONJUNCTION, JUSTIFY-CAUSE, INTERPRETATION-RESULT)
  - Different nuclearity: 0.54% (CAUSE-RESULT)
- Not used by one of the annotators: 0.7% (SOLUTIONHOOD-BACKGROUND)
A confusion matrix between three annotators: Multilingual RST TreeBank

- A comparison among 3 different languages/annotators: 0.484 Fleiss kappa (Fleiss, 1971) (300 RRs, 15 texts) (*moderate*)

<table>
<thead>
<tr>
<th></th>
<th>Kappa</th>
<th>z</th>
<th>p.value</th>
<th></th>
<th>Kappa</th>
<th>z</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation</td>
<td>0.851</td>
<td>25.528</td>
<td>0.000</td>
<td>Purpose</td>
<td>0.335</td>
<td>10.057</td>
<td>0.000</td>
</tr>
<tr>
<td>Summary</td>
<td>0.712</td>
<td>21.36</td>
<td>0.000</td>
<td>Result</td>
<td>0.301</td>
<td>9.017</td>
<td>0.000</td>
</tr>
<tr>
<td>Concession</td>
<td>0.705</td>
<td>21.155</td>
<td>0.000</td>
<td>Means</td>
<td>0.221</td>
<td>6.617</td>
<td>0.000</td>
</tr>
<tr>
<td>List</td>
<td>0.554</td>
<td>16.629</td>
<td>0.000</td>
<td>Conjunction</td>
<td>0.172</td>
<td>5.151</td>
<td>0.000</td>
</tr>
<tr>
<td>Elaboration</td>
<td>0.531</td>
<td>15.933</td>
<td>0.000</td>
<td>Motivation</td>
<td>0.136</td>
<td>4.084</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Interpretation</td>
<td>0.080</td>
<td>2.390</td>
<td>0.017</td>
</tr>
<tr>
<td>Condition</td>
<td>0.525</td>
<td>15.763</td>
<td>0.000</td>
<td>Unless</td>
<td>-0.001</td>
<td>-0.033</td>
<td>0.973</td>
</tr>
<tr>
<td>Sequence</td>
<td>0.499</td>
<td>14.966</td>
<td>0.000</td>
<td>Disjunction</td>
<td>-0.001</td>
<td>-0.033</td>
<td>0.973</td>
</tr>
<tr>
<td>Restatement</td>
<td>0.424</td>
<td>12.723</td>
<td>0.000</td>
<td>Evaluation</td>
<td>-0.003</td>
<td>-0.100</td>
<td>0.920</td>
</tr>
<tr>
<td>Circumstance</td>
<td>0.420</td>
<td>12.586</td>
<td>0.000</td>
<td>Evidence</td>
<td>-0.008</td>
<td>-0.235</td>
<td>0.814</td>
</tr>
<tr>
<td>Background</td>
<td>0.420</td>
<td>12.589</td>
<td>0.000</td>
<td>Antithesis</td>
<td>-0.008</td>
<td>-0.235</td>
<td>0.814</td>
</tr>
<tr>
<td>Cause</td>
<td>0.352</td>
<td>10.552</td>
<td>0.000</td>
<td>Justify</td>
<td>-0.009</td>
<td>-0.269</td>
<td>0.788</td>
</tr>
<tr>
<td>Contrast</td>
<td>0.376</td>
<td>11.272</td>
<td>0.000</td>
<td>Solutionhood</td>
<td>-0.011</td>
<td>-0.337</td>
<td>0.736</td>
</tr>
</tbody>
</table>
Confusion matrix by pairs: Multilingual RST TreeBank
Translation strategies: Multilingual RST TreeBank

1) Different relation signalling: Marker Change (MC)
   i) inclusion of a marker
   ii) exclusion of a marker
   iii) changing a marker

2) Clause Structure Change (CSC):
   i) hierarchical downgrading
   ii) hierarchical upgrading

3) Punctuation is used differently: Unit Shift (US):
   i) an independent sentence is downgraded
   ii) a clause is translated in an independent sentence

<table>
<thead>
<tr>
<th></th>
<th>ENG&gt;SPA</th>
<th>ENG&gt;BSQ</th>
<th>SPA&gt;ENG</th>
<th>SPA&gt;BSQ</th>
<th>BSQ&gt;ENG</th>
<th>BSQ&gt;SPA</th>
<th>Different Language Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>1.45%</td>
<td></td>
<td>4.35%</td>
<td>7.25%</td>
<td>10.14%</td>
<td>11.59%</td>
<td>14.49%</td>
</tr>
<tr>
<td>CSC</td>
<td>1.45%</td>
<td>1.45%</td>
<td>2.90%</td>
<td>4.35%</td>
<td>4.35%</td>
<td>1.45%</td>
<td>2.90%</td>
</tr>
<tr>
<td>US</td>
<td>2.90%</td>
<td>2.90%</td>
<td>2.90%</td>
<td>1.45%</td>
<td>4.35%</td>
<td>2.90%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>68.12%</td>
<td></td>
<td>31.88%</td>
</tr>
</tbody>
</table>
Exclusion of a marker (translation strategy)

(15) a. [Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados; de ahí, por ejemplo, los apartados que encontramos en muchos Webs en que se difunden glosarios de términos sobre Internet o en que se exponen propuestas denominativas que los usuarios pueden incluso votar.]

b. [Furthermore, terms can be compiled, discussed and assessed anywhere: many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them.]

c. [Are gehiago, edozein tokitatik biltzen dira terminoak, baita komentatu eta haiztatu ere; adibidez, Interneti buruzko terminoen glosarioak zabaltzen dira Web askotan, eta izendegietarako proposamenak egin ere bai, eta erabiltzaileek botoa eman ahal izaten diete.]
Clause Structure Change (translation strategy)

(16) a. [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología,]$_6$N $\{$que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos $(\ldots)$]$_{7−11}$S $−$ ELABORATION

b. [All these factors lead to an increase in the number of specialist terms which enrich terminology]$_6$N $−$ CONTRAST [but also call into question some of its basic concepts $(\ldots)$]$_7$N $−$ CONTRAST

c. [Alderdi horiek guztiek, espezialitateko terminologiaren gehikuntza kuantitatiboa eragiteaz gain, terminologia lanen ikuspegia ere zabaldu egin dute;]$_6$N $−$ LIST $[\text{eta, egia bada ere ikuspegi berri horrek terminologia aberastu egin duela esatea, zalantzan }\text{jarri ditu}\text{ terminologiaren oinarrizko zenbait kontzeptu $(\ldots)$}]_{7}N−\text{LIST TERM19}_\text{SPA}
Unit Shift or different punctuation (translation strategy)

(17) a. [En esta comunicación, a partir de la experiencia en trabajos de normalización de terminología catalana, se planteará la necesidad social de la normalización terminológica,]_{N12−LIST} [se comentarán algunas de las dificultades con que se enfrenta y se apuntarán ideas para su enfoque dentro de la sociedad actual.]_{N13−14−LIST}

b. [This paper looks, on the basis of experience in the standardisation of terminology in Catalan, at the social need for standardisation of terminology.]_{N12−LIST} [Some of the difficulties faced will be discussed, and ideas will be given for approaching this field in present day society.]_{S13−14−ELABORATION TERM19 SPA}
Open questions for the qualitative evaluation

- Can we automate this evaluation method for different languages?
- Weighted or unweighted measures for:
  - RR linked to CU and RR not linked to CU?
  - RRs inside the sentence and RRs at the top of the RS-tree?
  - Least frequent RRs and more frequent RRs?
- Should evaluation method (and measures) be determined by the genre/task?
PART 1 — Discourse relations in RST: method
  • Introduction
  • Segmentation
  • Central Unit
  • Rhetorical relations
  • Signals of rhetorical relations
  • Corpora for corpus exploration
  • Applications

PART 2 — Practice
  • Segmentation
  • Nuclearity
  • Choosing relations

PART 3 — Tools for corpus exploration
  • Signaling relational structures
  • An ambiguous RST analysis
  • Annotation in RST

PART 4 — Resources
  • Projects
  • Resources
  • Workshops
RST parsers

- RST parsers
  - CODRA parser (Joty et al., 2015)
  - A Linear-Time Bottom-Up Discourse Parser (Feng and Hirst, 2014)
  - DIZER parser (Pardo and Nunes, 2006)
Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.

Recurrent aphthous stomatitis is one of the most frequent oral conditions. Its etiology is controversial. It is characterised by the appearance of painful and recurrent ulcers, whose sizes, locations, and durations vary. These ulcers reappear periodically. This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.
DiZer: an online customizable parser (BP, ENG, SPA) (Pardo and Nunes, 2006)

- One can build its own parser by incorporating discourse knowledge (based on rules and corpus statistics)
Outline

1 PART 1 — Discourse relations in RST: method
2 PART 2 — Practice
3 PART 3 — Tools for corpus exploration
4 PART 4 — Resources
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops

- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST
- Evaluation tools/methods of RS
Topics and collaborations

- **Automatic Discourse Analyzer (ADA) for Basque**: Mikel Iruskieta, Arantza Diaz de Ilarrazoa, Mikel Lersundi, Maxux Aranzabe, Oier Lopez de Lacalle, Beñat Zapirain, Gorka Labaka, Kepa Bengoetxea, Aitziber Atutxa
  - Corpus annotation
  - Segmenter
  - Central Unit detector: Juliano Desiderato (BP)
  - Detection of cause subgroup coherence relations
  - Automatic evaluation system: Maite Taboada
  - Tools for corpus exploration

- **Sentiment analysis**: Jon Alkorta, Koldo Gojenola

- **Automatic summarization (RST and CST)**: Unai Atutxa

- **Resources for (automatic) translation from Chinese to Spanish**: Shuyuan Cao, Iria da Cunha
Topics and collaborations

- **Automatic Discourse Analyzer (ADA) for Basque**: Mikel Iruskieta, Arantza Diaz de Ilarraza, Mikel Lersundi, Maxux Aranzabe, Oier Lopez de Lacalle, Beñat Zapirain, Gorka Labaka, Kepa Bengoetxea, Aitziber Atutxoa
  - Corpus annotation
  - Segmenter
  - Central Unit detector: Juliano Desiderato (BP)
  - Detection of cause subgroup coherence relations
  - Automatic evaluation system: Maite Taboada
  - Tools for corpus exploration
- **Sentiment analysis**: Jon Alkorta, Koldo Gojenola
- **Automatic summarization (RST and CST)**: Unai Atutxa
- **Resources for (automatic) translation from Chinese to Spanish**: Shuyuan Cao, Iria da Cunha
Topics and collaborations

- **Automatic Discourse Analyzer (ADA) for Basque**: Mikel Iruskieta, Arantza Diaz de Illarraza, Mikel Lersundi, Maxux Aranzabe, Oier Lopez de Lacalle, Beñat Zapirain, Gorka Labaka, Kepa Bengoetxea, Aitziber Atutxa
  - Corpus annotation
  - Segmenter
  - Central Unit detector: Juliano Desiderato (BP)
  - Detection of cause subgroup coherence relations
  - Automatic evaluation system: Maite Taboada
  - Tools for corpus exploration

- **Sentiment analysis**: Jon Alkorta, Koldo Gojenola

- **Automatic summarization (RST and CST)**: Unai Atutxa

- **Resources for (automatic) translation from Chinese to Spanish**: Shuyuan Cao, Iria da Cunha
Topics and collaborations

- **Automatic Discourse Analyzer (ADA) for Basque**: Mikel Iruskieta, Arantza Diaz de Ilarraza, Mikel Lersundi, Maxux Aranzabe, Oier Lopez de Lacalle, Beñat Zapirain, Gorka Labaka, Kepa Bengoetxea, Aitziber Atutxa
  - Corpus annotation
  - Segmenter
  - Central Unit detector: Juliano Desiderato (BP)
  - Detection of cause subgroup coherence relations
  - Automatic evaluation system: Maite Taboada
  - Tools for corpus exploration

- **Sentiment analysis**: Jon Alkorta, Koldo Gojenola

- **Automatic summarization** (RST and CST): Unai Atutxa

- **Resources for (automatic) translation** from Chinese to Spanish: Shuyuan Cao, Iria da Cunha
PART 1 — Discourse relations in RST: method
- Introduction
- Segmentation
- Central Unit
- Rhetorical relations
- Signals of rhetorical relations
- Corpora for corpus exploration
- Applications

PART 2 — Practice
- Segmentation
- Nuclearity
- Choosing relations

PART 3 — Tools for corpus exploration
- Segmenters
- CU detector
- Annotation tools for RST
- Evaluation tools/methods of RS
- Parsers

PART 4 — Resources
- Projects
- Resources
- Workshops

- Signaling relational structures
- An ambiguous RST analysis
- Annotation in RST
Resources

- Annotation tools:
  - RS-tree: a) **RSTTool** (tutorial: 1, 2), b) **rstWEB**
  - Signaling: a) **Rhetorical Database**, b) **UAM Corpus Tool**

- Segmenters: a) **EusEduSeg**\(_{(EUS)}\), b) **SLSeg**\(_{(ENG)}\), c) **DiSeg**\(_{(SP)}\), d) **Senter**\(_{(BP)}\)

- Automatic Discourse Analyzers: **DIZER**\(_{(ENG, POR, SP)}\) (Pardo and Nunes, 2006) and **CODRA** (Joty et al., 2015)

- Automatic evaluation: **EvalRST**\(_{(ENG, POR, SP, EUS)}\)

- Corpora
  - **Basque RST TreeBank**\(_{(EUS)}\)
  - **Multilingual RST TB**\(_{(EUS, SP, ENG)}\)
  - **Brazilian RST TreeBank**\(_{(BP)}\)
  - **RST Spanish TreeBank**\(_{(SP)}\)
  - **German Potsdam Commentary Corpus**
Outline

1. PART 1 — Discourse relations in RST: method
   - Introduction
   - Segmentation
   - Central Unit
   - Rhetorical relations
   - Signals of rhetorical relations
   - Corpora for corpus exploration
   - Applications

2. PART 2 — Practice
   - Segmentation
   - Nuclearity
   - Choosing relations

3. PART 3 — Tools for corpus exploration
   - Segmenters
   - CU detector
   - Annotation tools for RST
   - Evaluation tools/methods of RS
   - Parsers

4. PART 4 — Resources
   - Projects
   - Resources
   - Workshops
Workshops and Web Site

- Workshops:
  - 2007 - 1\textsuperscript{st} workshop in São Paulo, Brazil.
  - 2009 - 2\textsuperscript{nd} workshop “Brazilian RST Meeting” in São Carlos, Brazil.
  - 2011 - 3\textsuperscript{rd} workshop “RST and Discourse Studies” in Cuiabá, Brazil.
  - 2013 - 4\textsuperscript{th} workshop “RST and Discourse Studies” in Fortaleza, Brazil.
  - 2015 - 5\textsuperscript{th} workshop “RST and Discourse Studies” in Alicante, Spain.

- Website

  The RST Web Site:
  
  \url{http://www.sfu.ca/rst/index.html}
Publications and Projects

<table>
<thead>
<tr>
<th>Papers</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iruskieta and Zapirain (2015)</td>
<td>EusEduSeg: A Dependency-Based EDU Segmentation for Basque</td>
</tr>
<tr>
<td>Iruskieta et al. (2015b)</td>
<td>The Detection of Central Units in Basque scientific abstracts</td>
</tr>
<tr>
<td>Iruskieta et al. (2015a)</td>
<td>A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora</td>
</tr>
<tr>
<td>Iruskieta et al. (2013a)</td>
<td>The RST Basque TreeBank</td>
</tr>
</tbody>
</table>

- Basque discourse segmenter:  
  http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl

- Annotated Basque corpus (fully developed):  
  http://ixa2.si.ehu.es/diskurtsoa/

- Annotated multilingual corpus (English, Spanish, Basque):  
  http://ixa2.si.ehu.es/rst/

- Presentation of “Corpus exploration of discourse relations in RST” is available at  
  http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1452904951/publikoak/LTPS2016_Valena.pdf
Thanks

– for interesting comments and discussion to
  • Maite Taboada
  • Juliano A. Desiderato
  • Arantza Diaz de Ilarraza
– for English corrections to
  • Larraitz Uria


References II


References V


References VII


Corpus exploration of discourse relations in RST

Feel free to contact me for any doubt or particular interest on RST

Mikel Iruskieta
mikel.iruskieta@ehu.eus

Ixa group for NLP
University of the Basque Country (UPV/EHU)

Valencia, January 18th-22nd, 2016
Structuring Discourse in Multilingual Europe

Training School: Methods and tools for the analysis of discourse relational devices