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^{th}-22^{nd}, 2016
Structuring Discourse in Multilingual Europe

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

Outline

Outline

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

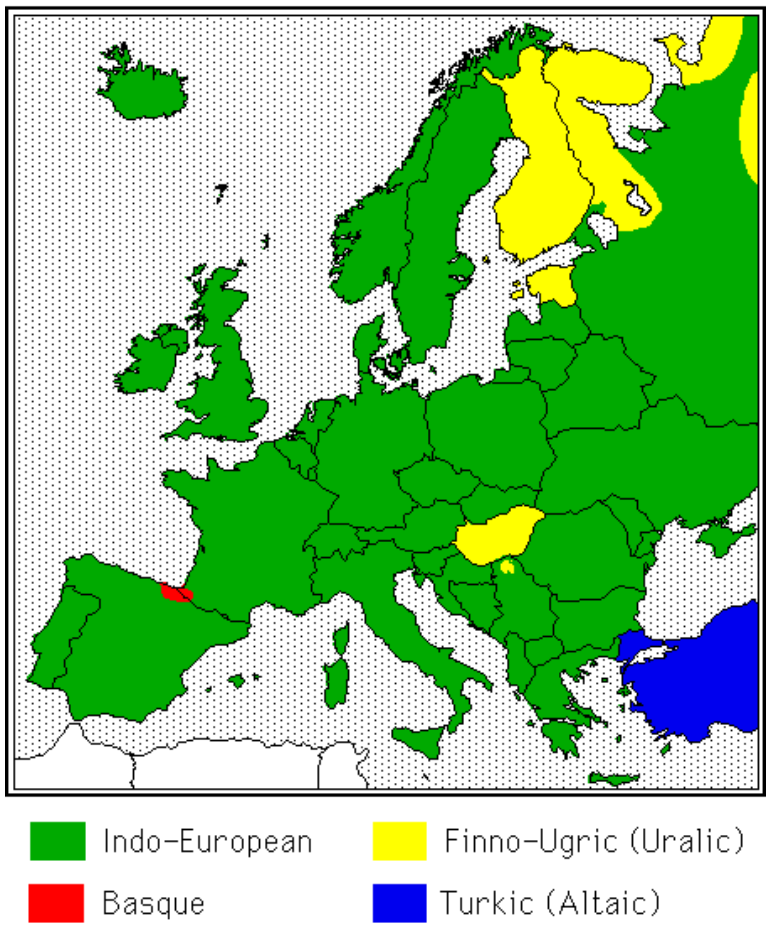
About me

- 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°52'55"N 1°55'01"W). Listen to my Basque dialect

Language Families in Europe

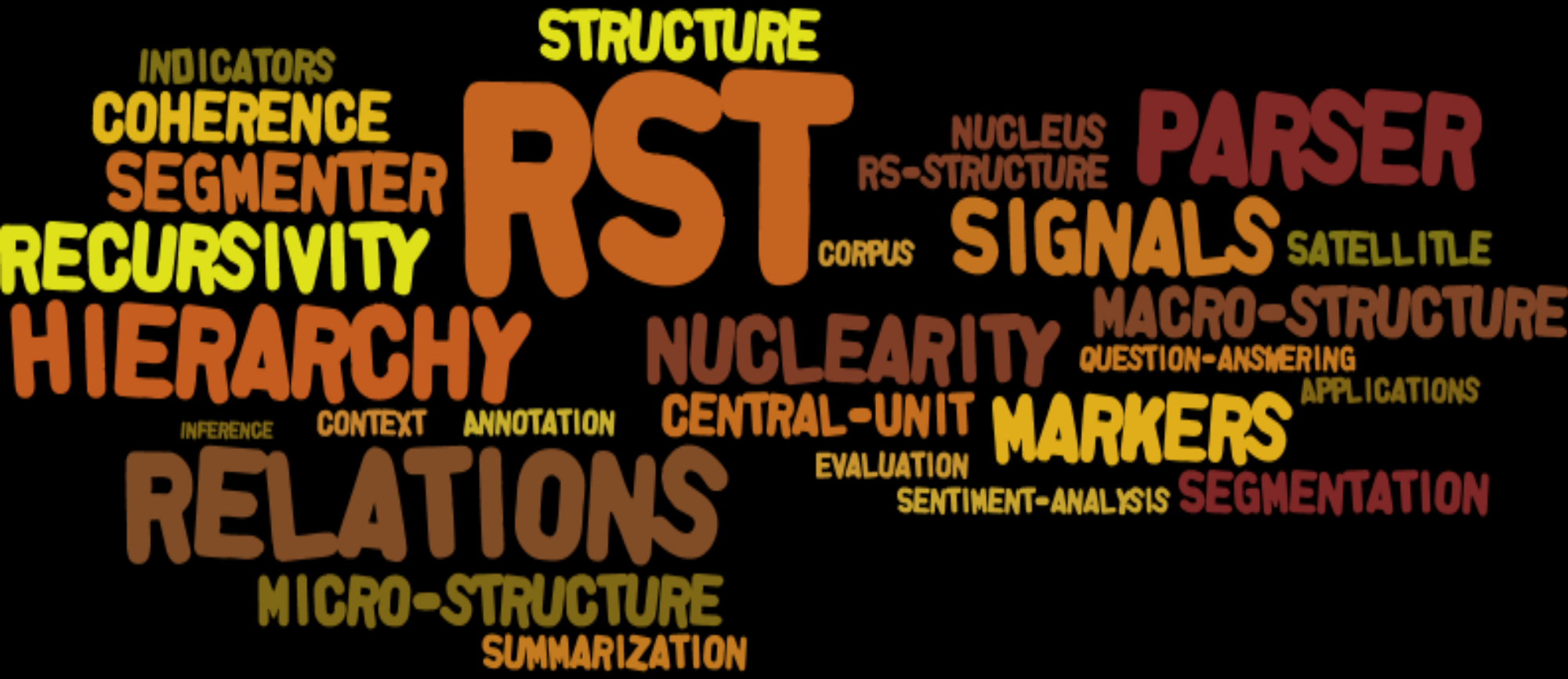


Abstract

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).


STRUCTURE
INDICATORS
COHERENCE
RST
NUCLEUS
RS-STRUCTURE
PARSER
SEGMENTER
RECURSIVITY
CORPUS
SIGNALS
SATELLITLE
MACRO-STRUCTURE
HIERARCHY
NUCLEARITY
QUESTION-ANSWERING
APPLICATIONS
INFERENCE
CONTEXT
ANNOTATION
CENTRAL-UNIT
MARKERS
EVALUATION
RELATIONS
SENTIMENT-ANALYSIS
SEGMENTATION
MICRO-STRUCTURE
SUMMARIZATION

Keywords

Relational discourse structure

Annotation
Applications
Central Unit
Coherence
Context
Corpus
Discourse markers
Evaluation
Expl. relations
Hierarchy
Impl. relations
Indicators
Inference
Macro-structure
Micro-structure
Nuclearity
Nucleus
Parser
Question-answering
Recursivity
Rhetorical analysis
Rhetorical relations
RS-structure
Satellite
Segmentation
Segmenter
Sentiment analysis
Signals
Structure
Summarization

Natural Language Processing of Basque

- 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?

Natural Language Processing of Basque

- 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; Barrutieta et al., 2002, 2001) and in IXA group (Iruskieta et al., 2011, 2013b)
  - Segmeter: EusEduSeg
  - Central Unit detector
  - Signal annotation
  - Applications: corpus exploration tools

Discourse structure phenomena in CL

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 of discourse structures in CL

– 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).

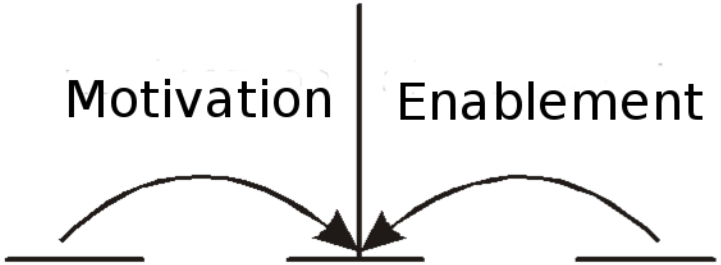
Relational discourse structure

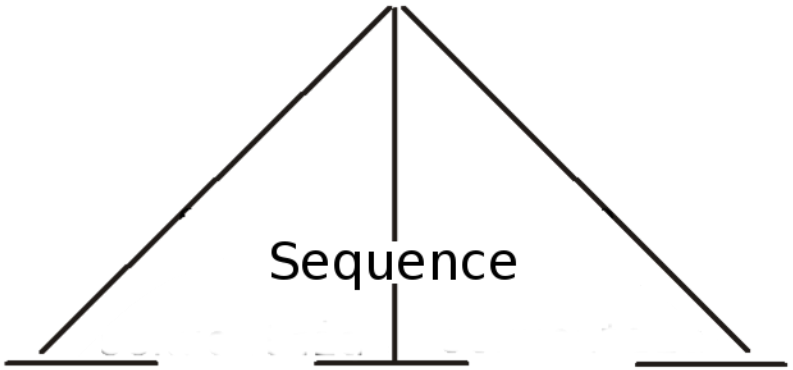
A rhetorical structure tree (RS-tree) is

a hierarchical structure in which all the propositions of the text have a relationship in the structure

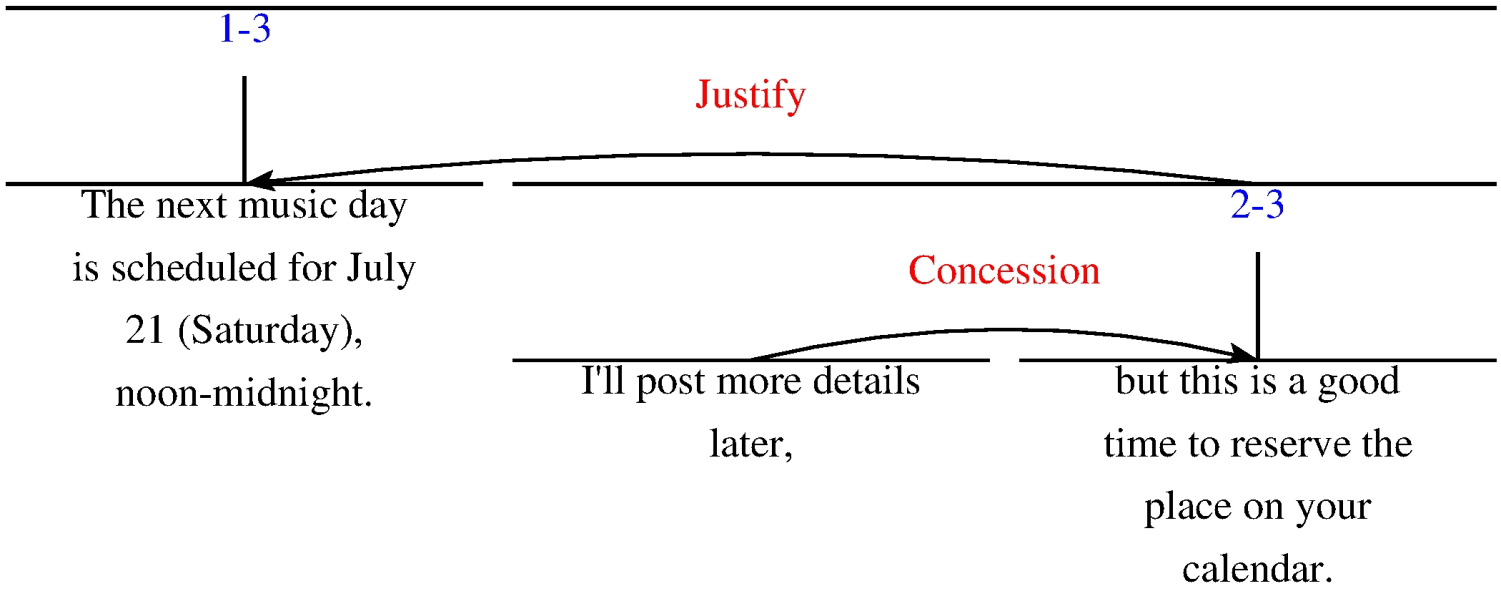
In RST a hierarchical tree structure is composed with:

1. **Hierarchy**: *i*) nucleus and *ii*) satellite
2. **Relations**: *i*) presentational and *ii*) subject-matter





Rhetorical relations: definitions at the RST Web Site



	Const. on *S* or *N*	Constraints on *S* + *N*	Intention of W
Conc.	on N: W has positive regard for N on S: W is not claiming that S does not hold;	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	R's positive regard for N is increased
Just.	none	R's comprehending S increases R's readiness to accept W's right to present N	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

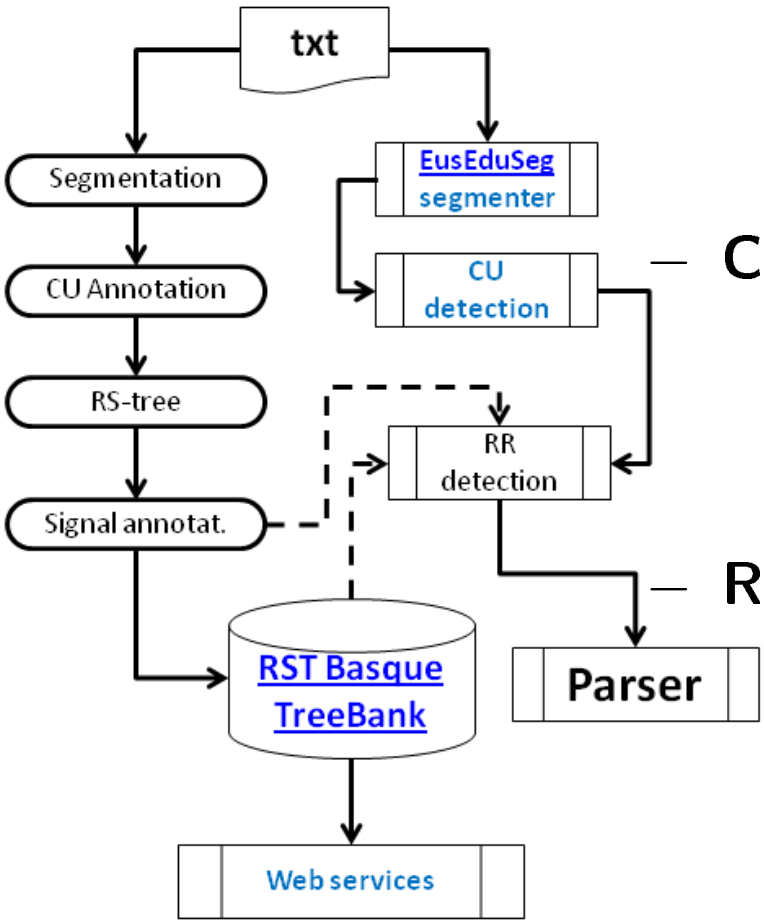
Domain	Sub-corpus	Texts	EDUs	Words
Medicine	GMB	20	283	3010
Terminology	TERM	20	584	5664
Science	ZTF	20	603	6892
Life	BIZ	20	569	5535
Health	OSA	20	475	4878
Informatics	INF	20	236	1860
Economy	EKO	20	216	2108
	Total	140	2966	29947

– 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

Abstracts of a scientific text [GMB0401]

ORIGINAL

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.

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 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 continuosly, 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 slights by the health staff. It could be interesting from the sanitary planning poin 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 archieve 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%) who consults because of traumatologic phatologies (50.5%) and who comes from the sanitary area near the hospital.
Key words: Emergency department users, overuse, users profile.

LABURPENA

Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe, estatu espainolean igoera hau urteko %4an kokatzen da. Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek. Zerbitzu hauen perfila azaltzen duten ikerketa epidemiologikoak egitea baliagarria izan daiteke osasun planifikazioaren aldetik, hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskriptibo bat egitea aproposa iruditu zaigu.
Emaitzak: Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (%50,5). Galdakao inguruko herrietatik datorrelarik gehiengoa.
Hitz garrantzitsuak: Larrialdi zerbitzuen erabiltzaileak, gainerabilpena, erabiltzaileen perfila.

Correspondencia:
Dra. Itsaso Bengoetxea Martínez
Atutxa Saiburua, 2 - 3º
48330 - LEMOA - Bizkaia
Enviado 23/01/2004. Aceptado 8/09/2004

Introducción

El número de asistencias urgentes crece constantemente. Se ha estimado que más de la mitad de la población utiliza alguna vez los servicios de urgencia a lo largo de un año (1). En España el ritmo de crecimiento se ha establecido en torno al 4% anual (2). Dicho crecimiento también queda patente en el territorio de la Comunidad Autónoma Vasca.
Los motivos propuestos para explicar este crecimiento constante son: el envejecimiento de la población, la accesibilidad a los servicios de urgencia, la confianza en la atención hospitalaria, la demora de la atención especializada y la cultura de la inmediatez entre otros (3).
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 procesos leves por el personal sanitario (4).
Diversos estudios han constatado que ciertos determinantes externos como el nivel socioeconómico, los cambios atmosféricos, las epidemias de gripe, los niveles de contaminación y/o polinización ambiental, los ciclos lunares o los eventos deportivos televisados condicionan una fluctuación de la demanda asistencial (5).
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. Hasta la fecha no se dispone de estudios similares en nuestro medio laboral, por lo que se ha creído oportuno realizar un estudio que describa las características de los usuarios que acuden a los servicios de urgencia y se etiquetan como " de poca gravedad" por el personal de triaje, ya que son en principio la causa del aumento asistencial anteriormente citado.
El objetivo general es conocer el perfil del usuario de la zona ambulatoria (pacientes etiquetados como "no graves" en el con-

Abstracts of a scientific text [GMB0401]

ORIGINAL

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.

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 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 continuosly, 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 slights by the health staff. It could be interesting from the sanitary planning poin 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 archieve 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%) who consults because of traumatologic phatologies (50.5%) and who comes from the sanitary area near the hospital.
Key words: Emergency department users, overuse, users profile.

LABURPENA

Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe, estatu espainolean igoera hau urteko %4an kokatzen da. Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek. Zerbitzu hauen perfila azaltzen duten ikerketa epidemiologikoak egitea baliagarria izan daiteke osasun planifikazioaren aldetik, hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskriptibo bat egitea aproposa iruditu zaigu.
Emaitzak: Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (%50,5). Galdakao inguruko herrietatik datorrelarik gehiengoa.
Hitz garrantzitsuak: Larrialdi zerbitzuen erabiltzaileak, gainerabilpena, erabiltzaileen perfila.

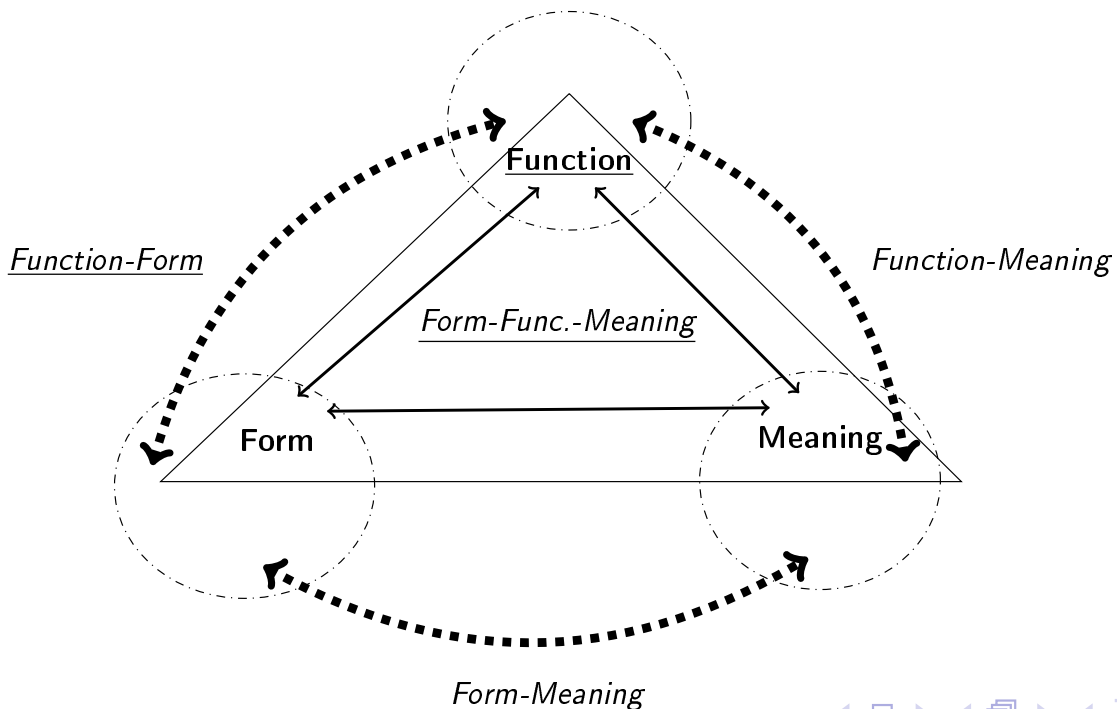
Correspondencia:
Dra. Itsaso Bengoetxea Martínez
Atutxa Saiburua, 2 - 3º
48330 - LEMOA - Bizkaia
Enviado 23/01/2004. Aceptado 8/09/2004

Introducción

El número de asistencias urgentes crece constantemente. Se ha estimado que más de la mitad de la población utiliza alguna vez los servicios de urgencia a lo largo de un año (1). En España el ritmo de crecimiento se ha establecido en torno al 4% anual (2). Dicho crecimiento también queda patente en el territorio de la Comunidad Autónoma Vasca.
Los motivos propuestos para explicar este crecimiento constante son: el envejecimiento de la población, la accesibilidad a los servicios de urgencia, la confianza en la atención hospitalaria, la demora de la atención especializada y la cultura de la inmediatez entre otros (3).
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 procesos leves por el personal sanitario (4).
Diversos estudios han constatado que ciertos determinantes externos como el nivel socioeconómico, los cambios atmosféricos, las epidemias de gripe, los niveles de contaminación y/o polinización ambiental, los ciclos lunares o los eventos deportivos televisados condicionan una fluctuación de la demanda asistencial (5).
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. Hasta la fecha no se dispone de estudios similares en nuestro medio laboral, por lo que se ha creído oportuno realizar un estudio que describa las características de los usuarios que acuden a los servicios de urgencia y se etiquetan como " de poca gravedad" por el personal de triaje, ya que son en principio la causa del aumento asistencial anteriormente citado.
El objetivo general es conocer el perfil del usuario de la zona ambulatoria (pacientes etiquetados como "no graves" en el con-

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)

Clause type	Example
Perpaus independentea 'an independent sentence'	[*Whipple (EW) gaixotasunak hesteei eragiten die bereziki.*]$_1$ GMB0503
Perpaus nagusi koordinatua 'a main clause, part of sentence'	[pT1 tumoreko 13 kasuetan ez *zen* gongoila inbasiorik *hauteman*;]$_1$ [aldiz, pT1 101 tumoretatik 19 kasutan (18.6%) inbasioa *hauteman zen*, eta pT1c tumoreen artetik 93 kasutan (32.6%).]$_2$ GMB0703
Aditz jokatudun adjuntu perpausa 'finite adjunct clauses'	[Haien sailkapena egiteko hormona hartzaileen eta c-erb-B2 onkogenearen gabeziaz baliatu gara,]$_1$ [*ikerketa anatomopatologikoetan erabili ohi diren zehaztapenak direlako.*]$_2$ GMB0702
Aditz jokatugabedun adjuntu perpausa 'non-finite adjunct clauses'	[Ohiko tratamendu motek porrot eginez gero,]$_1$ [gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.]$_2$ GMB0502
Erlatibo ez-murriztailea 'non-restrictive relative clause'	[Dublin Hiriko Unibertsitateko atal bat da Fiontar,]$_1$ [zeinak Ekonomia, Informatika eta Enpresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.]$_2$ TERM23

Segmentation of discourse units (EDUs) [GMB0401]

1	2	3	4	5	6	7	8	9	10
Galdakaoko ospitaleko larrialdi zerbitzuko erabiltzaileen perfila --- Outpatient department user profile for Galdakao Hospital's Emergency Services	Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe, --- The amount of medical attention provided is growing constantly;	estatu españolean igoera hau urteko %4an kokatzen da. --- in Spain, the growth rate has stabilized at about 4% annually.	Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea --- It is calculated that about 80% of users come to emergency services on their own initiative	eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek. --- and that 70% of visits are considered minor by health care personnel.	Zerbitzu hauen perfila azaltzen duten ikerketa epidemiologikoak egitea baliagarria izan daiteke osasun planifikazioaren aldetik, --- Carrying out epidemiological studies describing user characteristics and motives for over-use of emergency hospital services could prove interesting from the point of view of medical planning.	hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskriptibo bat egitea aproposa iruditu zaigu. --- Consequently, we believed it would be appropriate to carry out a study in order to better understand the profile for users of Galdakao Hospital's emergency services.	Emaitzak: --- Results:	Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoa gatik kontsultatzen duena (%50,5). --- The average user is as follows: male (51.3%), middle-aged (43.2 years old), and treated for trauma pathology (50.5%).	Galdakao inguruko herrietatik datorrelarik gehiengoa. --- He comes from the region surrounding the hospital

Adjunct verb clause-based segmentation (Tofiloski et al., 2009)

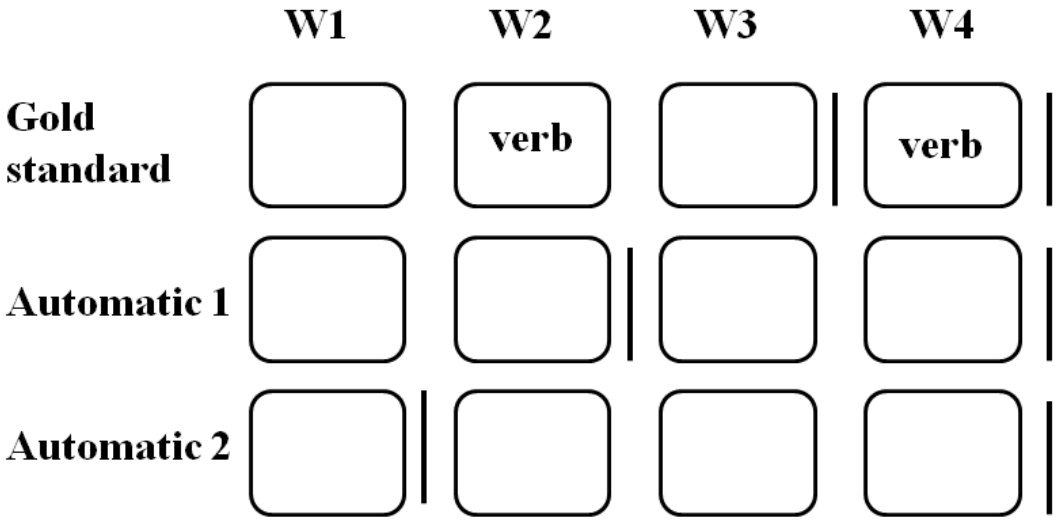
∗English translation is ours

Automatic segmentation based on rules (CG3)

MAP:171	MAP (}EDU) TARGET (PUNT_BI_PUNT) (1 ADI OR ADT BARRIER PUNTUAZIOA) (NOT -1 OSAGARRIAK BARRIER PUNTUAZIOA) (NOT 1 OSAGARRIAK BARRIER PUNTUAZIOA);
MAP:358	MAP (}EDU) TARGET ("bide") IF (-1 ("<dela>"))(NOT 1 PUNTUAZIOA);
MAP:231	MAP (}EDU) TARGET (PUNT_PUNT_KOMA) (1 ADI OR ADT BARRIER PUNTUAZIOAG) (-1 ADI OR ADT BARRIER PUNTUAZIOAG)
MAP:180	MAP (}EDU) TARGET (PUNT_GALD) IF (NOT 1 (PUNT_GALD) OR (PUNT_ESKL) OR (PUNT_PUNT) OR (PUNT_KOMA) OR BEREIZ);
MAP:211	MAP (}EDU) TARGET (PUNT_PUNT) IF (0 &ESALDI_BUK_1) (NOT -1 (LAB) OR (ERROM) OR (ZEN)) (NOT 1 PUNTUAZIOA);
MAP:131	MAP (}EDU) TARGET (PUNT_KOMA) IF (1 ADI OR ADT BARRIER PUNTUAZIOA) (-1 ADI OR ADT BARRIER PUNTUAZIOA);
MAP:472	MAP (}EDU) TARGET ("bitarte") IF (-1 (ADL) OR (ADT) OR (PART)) (NOT 1 PUNTUAZIOA);

Segments	Correct	Missed	Excess	Recall	Precision	F-measure
765	606	159	98	0.86	0.79	0.82
MAP:171	31					
MAP:358	1					
MAP:231	120		89			
MAP:180	25					
MAP:211	413					
MAP:148	15		9			
MAP:472	1					
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)

Go to Exercises: 80

Outline

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)

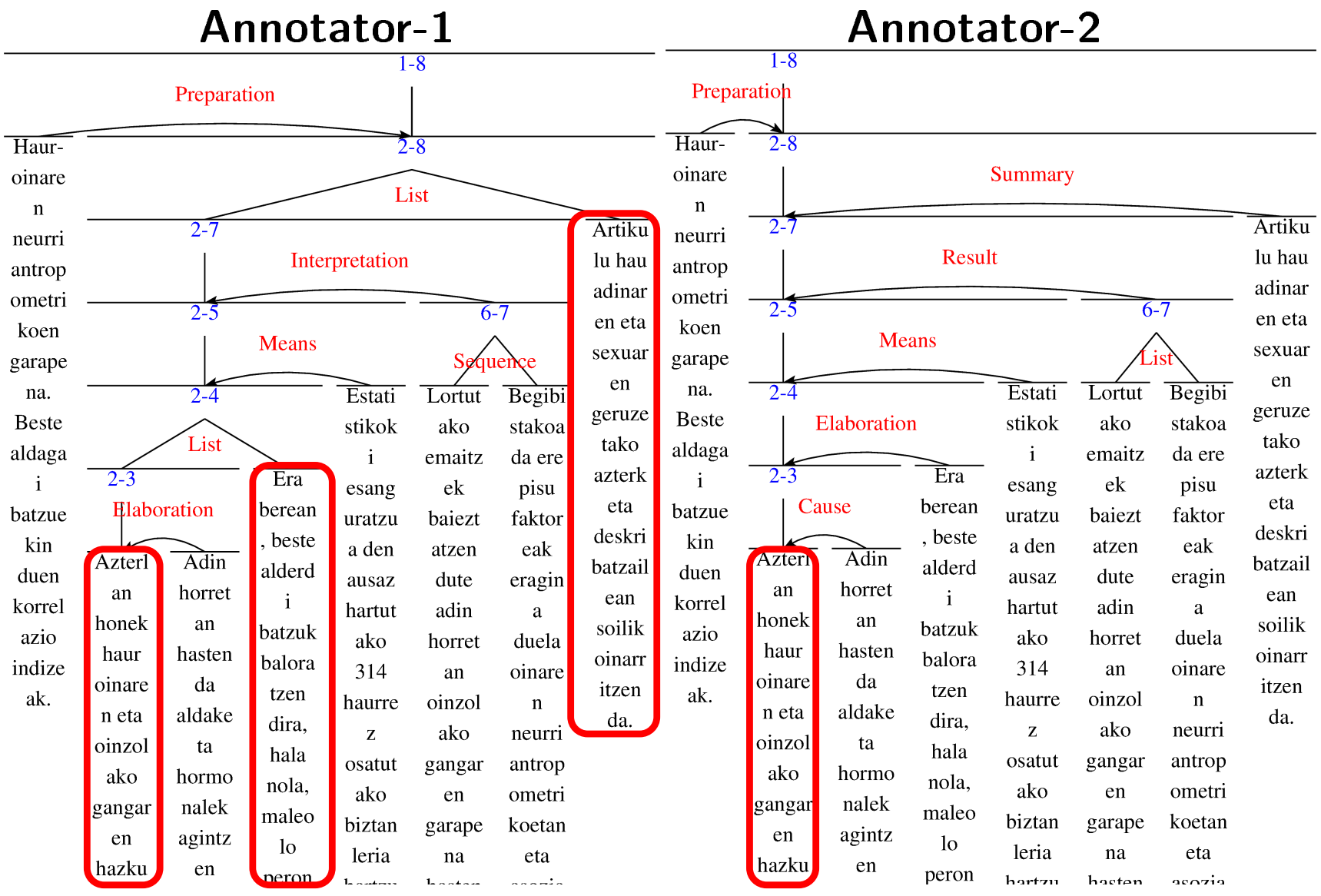
An example of Central Unit (CU) annotated with RSTTool

Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak . --- Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.	"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da. --- Recurrent aphtous stomatitis is one of the most frequent oral conditions.	Honen etiologia eztabaidagarria da. --- Its etiology is controversial.	Ultzera mingarri batzu bezela agertzen da, --- It is characterised by the appearance of paintful and recurrent ulcers,	tamainu, kokapena eta iraunkortasuna aldakorra izanik. --- whose sizez, locations, and durations vary.	Hauek periodiki beragertzen dira. --- These ulcers reappear periodically.	Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu. --- This paper analyzes the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.

(5) [**Lan honetan** patologia arrunt honetan ezaugarri etiopatogeniko eta klinikopatologiko **garrantzitsuenak analizatzen** ditu**gu**.]$_7$ [GMB0301]

[**This paper analyzes the most important** epidemiological, etiological, pathological and clinical features of this common oral pathology.]$_7$

Different Central Units in some RS-structure [GMB0203]



Central Unit: harmonization

– CU annotation guidelines for scientific abstracts

i) Topic or thesis statement
ii) Purpose
iii) Method
iv) Results
v) Conclusions

An enlarged list of indicators proposed by Paice (1980)

Indicators from train dataset (Iruskieta et al., 2014a)

Verbs		Nouns		Pronouns	Bonus words
EUS	**ENG$_{MCR}$**	**EUS**	**ENG$_{MCR}$**		
aztertu	examine$_1$	abiapuntu$_1$	starting_point$_1$	**Demonstrative Pronoun**	garrantzi *importance*
analizatu	examine$_1$	arlo$_1$	subject_field$_1$	hau *this*	nagusi *main*
oinarritu	base$_1$	artikulu$_7$	article$_1$	**Personal Pronouns**	azpimarragarri *remarcable*
baloratu	value$_2$	asmo$_2$	purpose$_1$	gu *we*	eskerga *huge*
azaldu	recount$_1$	bide$_2$	means$_1$	**-gu** (inside the verb)	(gaur) egun *nowadays*
aurkeztu	present$_2$	gai$_6$	topic$_1$		
aipatu	present$_2$	ikerkuntza$_3$			
berri eman	present$_2$	ikerketa$_2$	research$_2$		
jardun	present$_2$	azterlan$_3$			
plazaratu	present$_2$	ikerlan$_3$			
erabili	use$_1$	arazo$_3$	problem$_2$		
ikertu	investigate$_1$	irtenbide$_2$	resolution$_4$		
		komunikazio	paper$_5$		
		hitzaldi$_2$	speech$_1$		
		lan$_3$	work$_2$		
		lan-ildo	— —		
		lerro$_{11}$ ikerketa-lerro	line$_8$		
		proiektu$_2$ ikerketa-proiektu	project$_2$		
		talde$_1$ ikerketa-talde	group$_1$		
		xede$_1$ helburu$_2$	goal$_1$		

Heuristics to identify the Central Unit (test dataset)

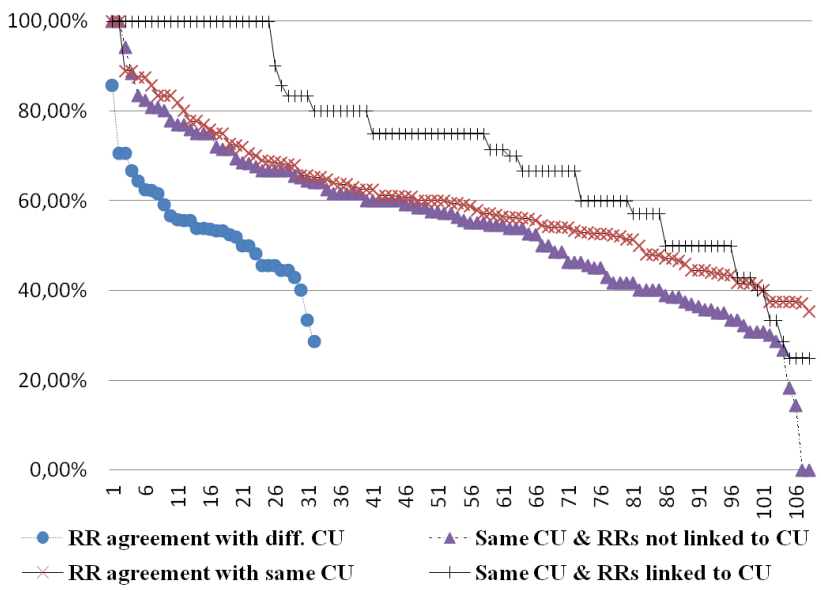
- Difficulty to choose the CU: 0.032
- Agreement between 2 annotators: 0.89 F1

	Heuristics	C	E	M	Pre.	Rec.	F_1
H1	Nouns and verbs	15	31	29	0.33	0.34	0.33
H2	Nouns and verbs + pronouns	22	68	22	0.24	0.50	0.33
H3	Bonus words	5	14	39	0.26	0.11	0.16
H4	Title words	7	3	37	**0.70**	0.16	0.26
H5	EDU position	40	711	4	0.05	0.91	0.10
H6	Main verb	41	721	3	0.05	**0.93**	0.10
H7	H1, H2 and H4	21	30	23	0.41	0.48	**0.44**
H8	H1, H2, H3, H4 and H5	23	48	21	0.32	0.52	0.40
	Machine Learning	C	E	M	Pre.	Rec.	F_1
	Perceptron + postproc.	24	25	20	0.48	0.54	0.51

Some conclusions and topics to discuss: the annotation of the Central Unit (Iruskieta et al., 2014b)

	Texts	Annotators	Measure	Results
Burstein et al. (2001)	100	2 professionals	F-score	71%
Basque	60	4 non-professionals	F-score	61%

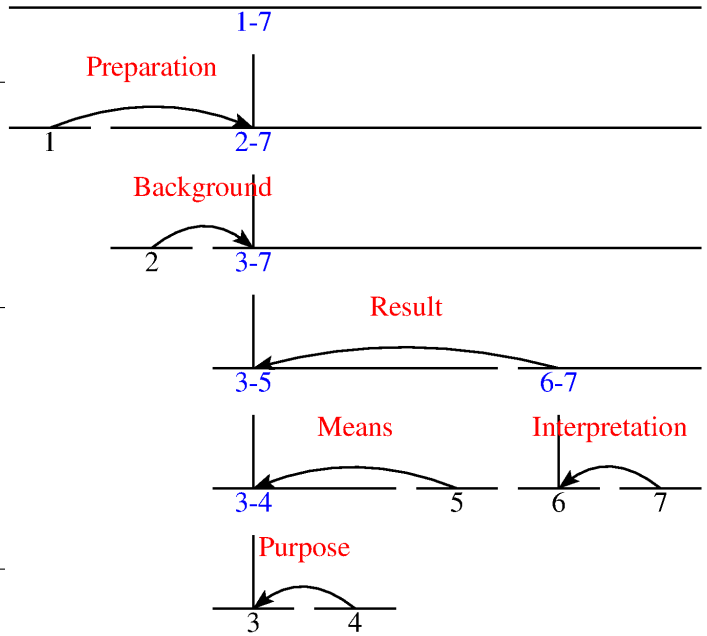
- Annotation of the CU (2 annotators):
  - Derived from RS-trees: 65% (GMB)
  - Annotating the CU first: 85% (in TERM and in ZTF)
- Agreement is bigger in relations, when annotators have annotated the same CU (+5.04%, T-test: 0.013)
- Agreement is bigger in RRs linked to the CU (+17.29% T-test: 0.001)



CU and RRs: the IMRaD structure (Swales, 1990)

Within the RRs linked to the CU, those with an IMRaD structure appear most frequently (except ELABORATION) (Iruskieta, 2014)

	GMB		TERM		ZTF		Corpus	
RRs	SN	NS	SN	NS	SN	NS	SN	NS
PREPARATION	22		24		22		68	
ELABORATION		6		15		28		49
BACKGROUND	13		15		16		44	
MEANS	1	14		5		6	1	25
PURPOSE	2		1	6		9	3	15
RESULT		10		2				12
SUMMARY		4		3				7
CIRCUMSTANCE		2		3		1		6
INTERPRETATION		5						5
CAUSE		2		1		1		4
JUSTIFY		1		2				3
CONCESSION			1	2			1	2
SOLUTIONHOOD			3				3	
Total	**39**	**44**	**45**	**39**	**39**	**48**	**123**	**131**



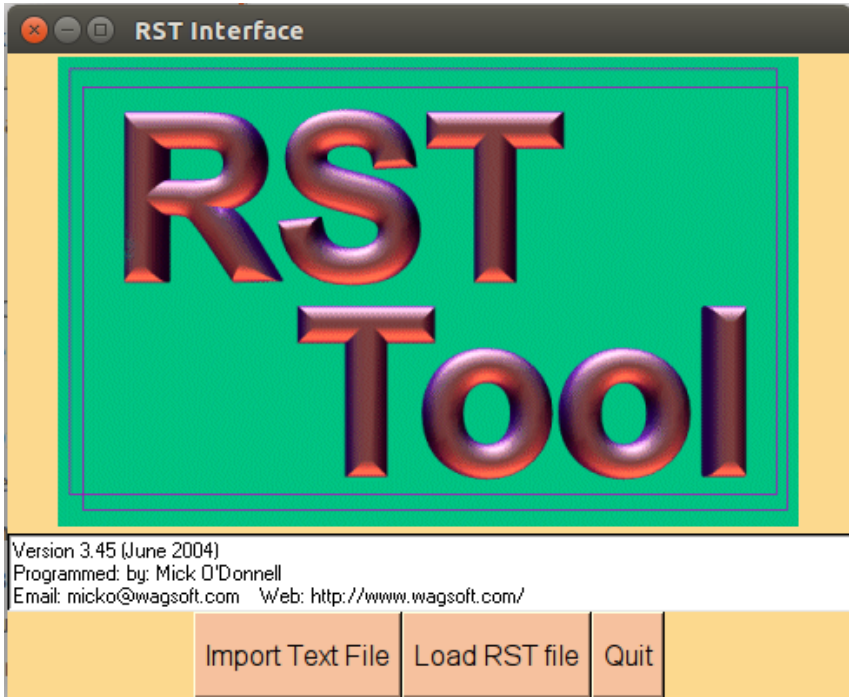
Outline

The extended RST relation set

Type	Relation	Relation	Type
P	Preparation	Elaboration	SM
P	Background	Means	SM
	Enablement and Motivation	Circumstance	SM
P	Enablement	Solution-hood	SM
P	Motivation	**Conditional relations**	
	Evidence and Justify	Condition	SM
P	Evidence	Otherwise	SM
P	Justify	Unless	SM
	Antithesis and Concession	No-Conditional	SM
P	Antithesis	**Interpretation and Evaluation**	
P	Concession	Interpretation	SM
	Reformulation and Summary	Evaluation	SM
P	Reformulation	**Cause subgroup**	
P	Summary	Cause	SM
		Result	SM
		Purpose	SM
N-N	List	Sequence	N-N
N-N	Disjunction	Contrast	N-N
N-N	Joint	Conjunction	N-N
N-N	Reformulation-NN		
∅	Same-unit		

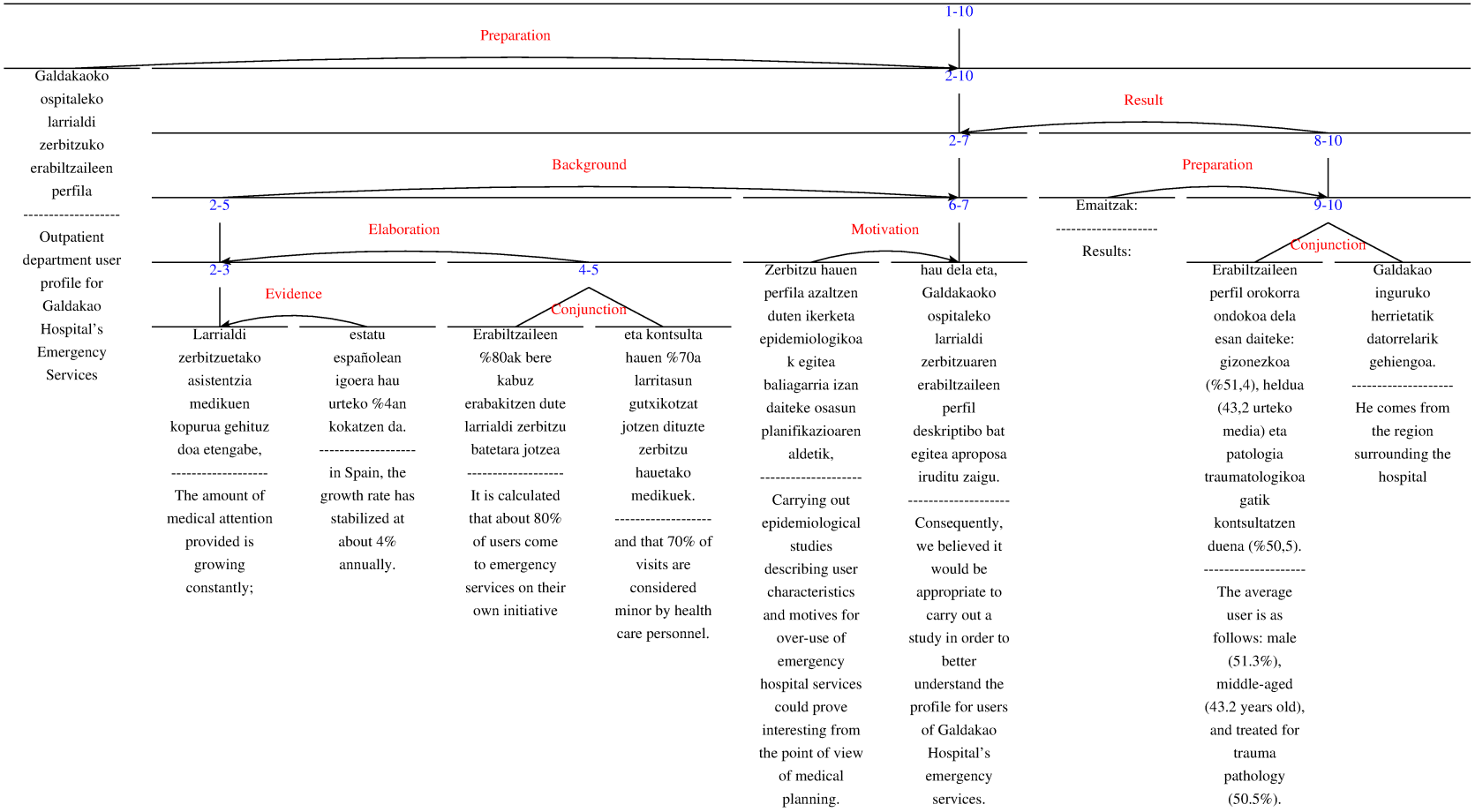
Relations from the RST webpage at http://www.sfu.ca/rst/

RSTTool annotation interface



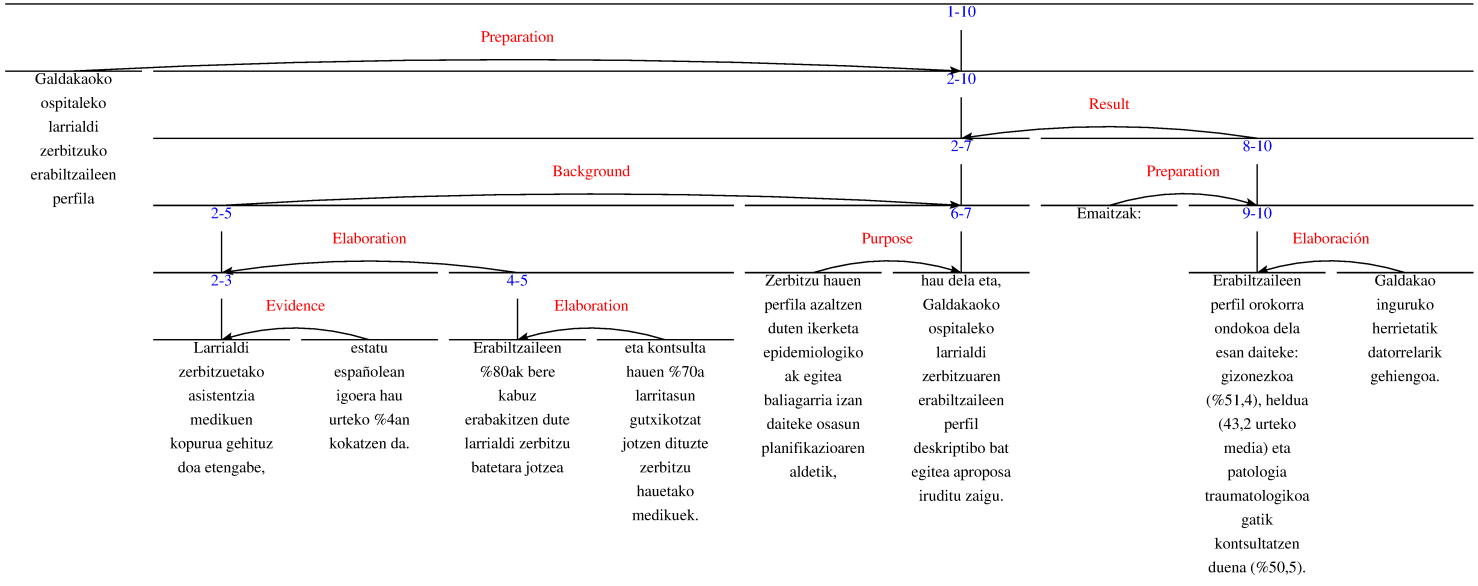
- A TXT text and a relation set are necessary to annotate with the RSTTool
- The segmenter EusEduSeg has integrated the RS3 output and a Basque relation set

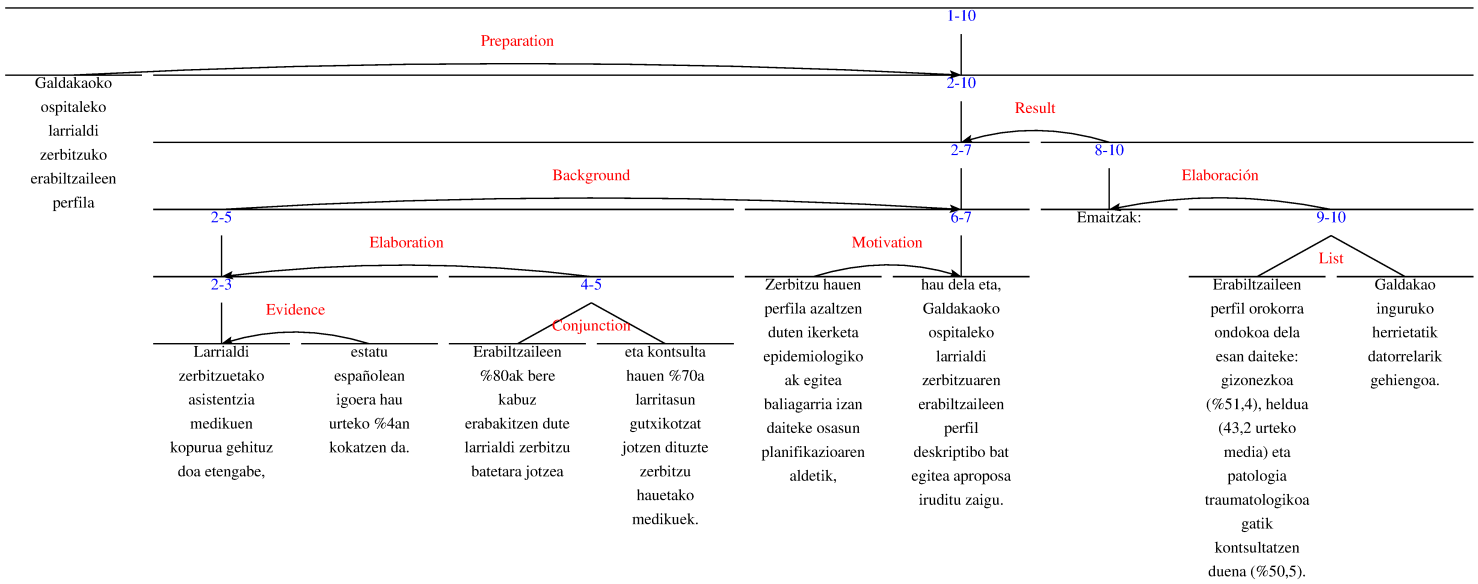
Rhetorical structure of a text [GMB0401]



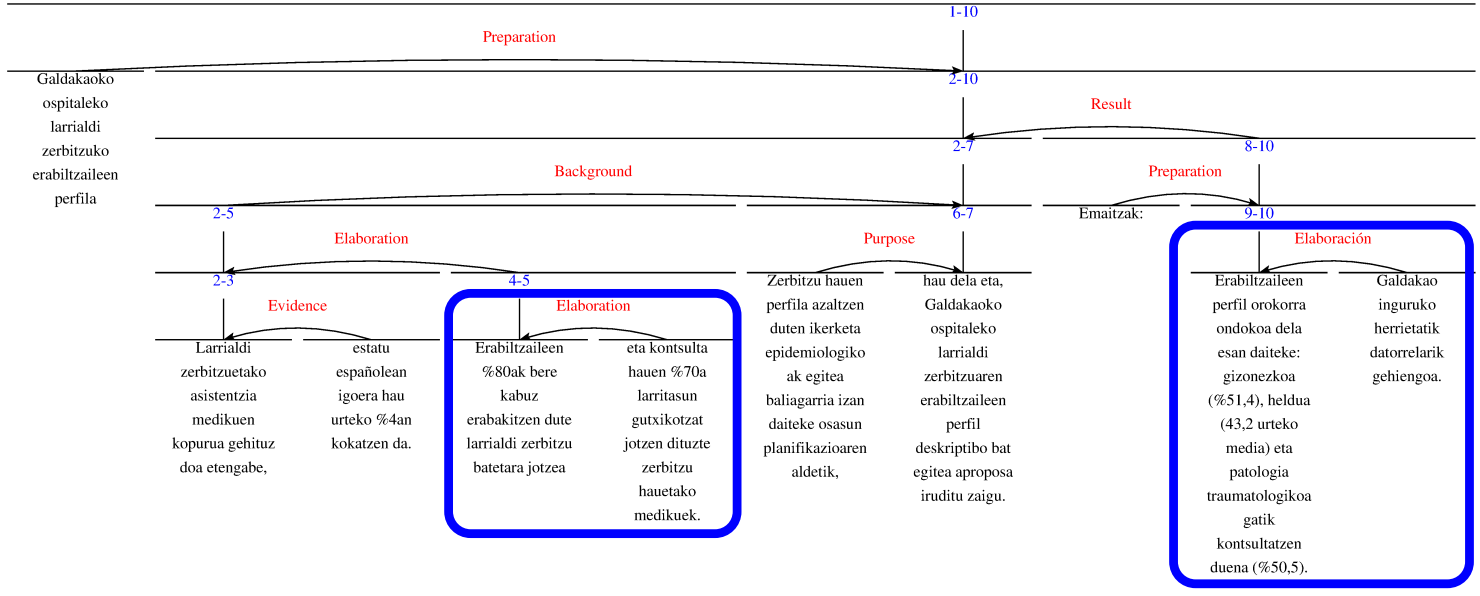
– A modular and incremental annotation (Pardo, 2005)

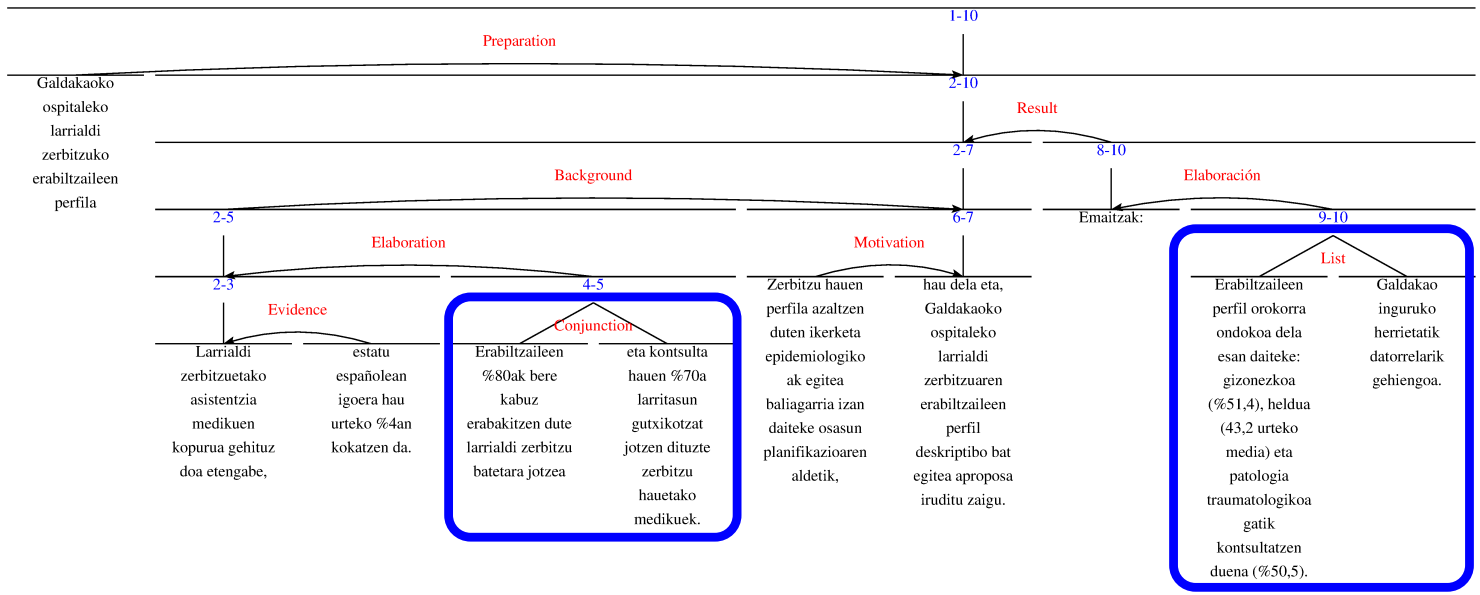
Different interpretations of [GMB0401]



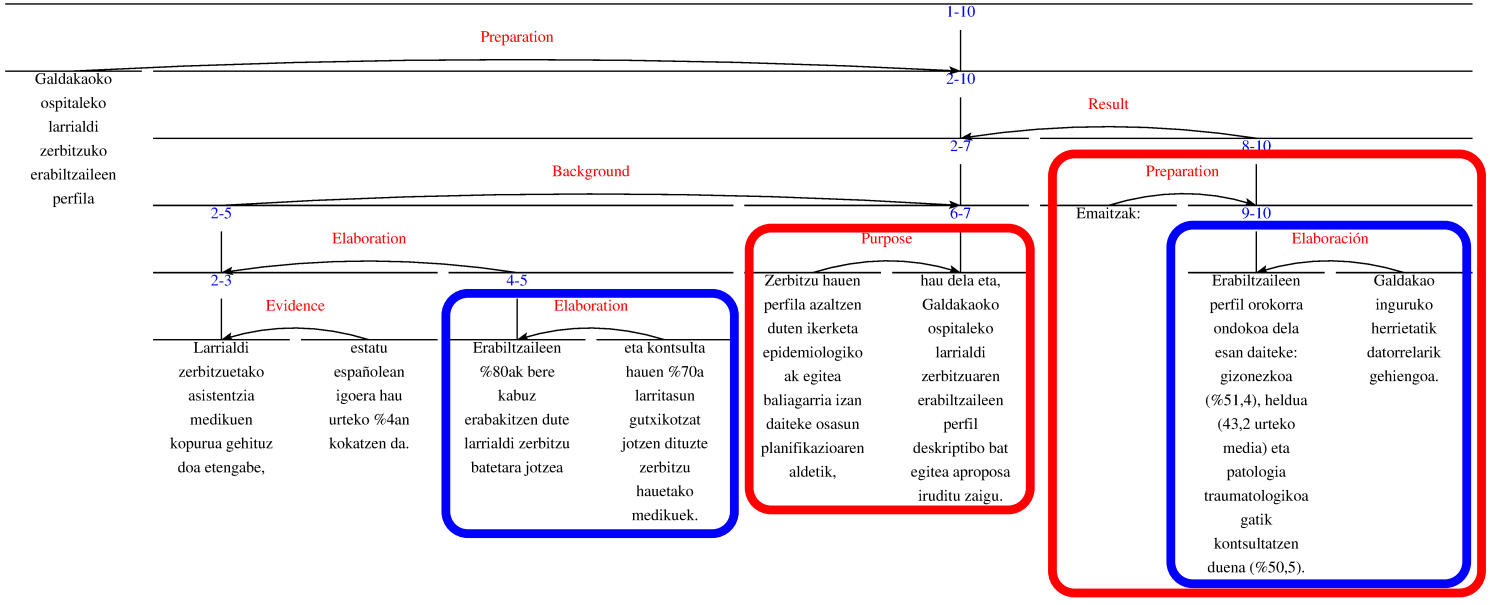


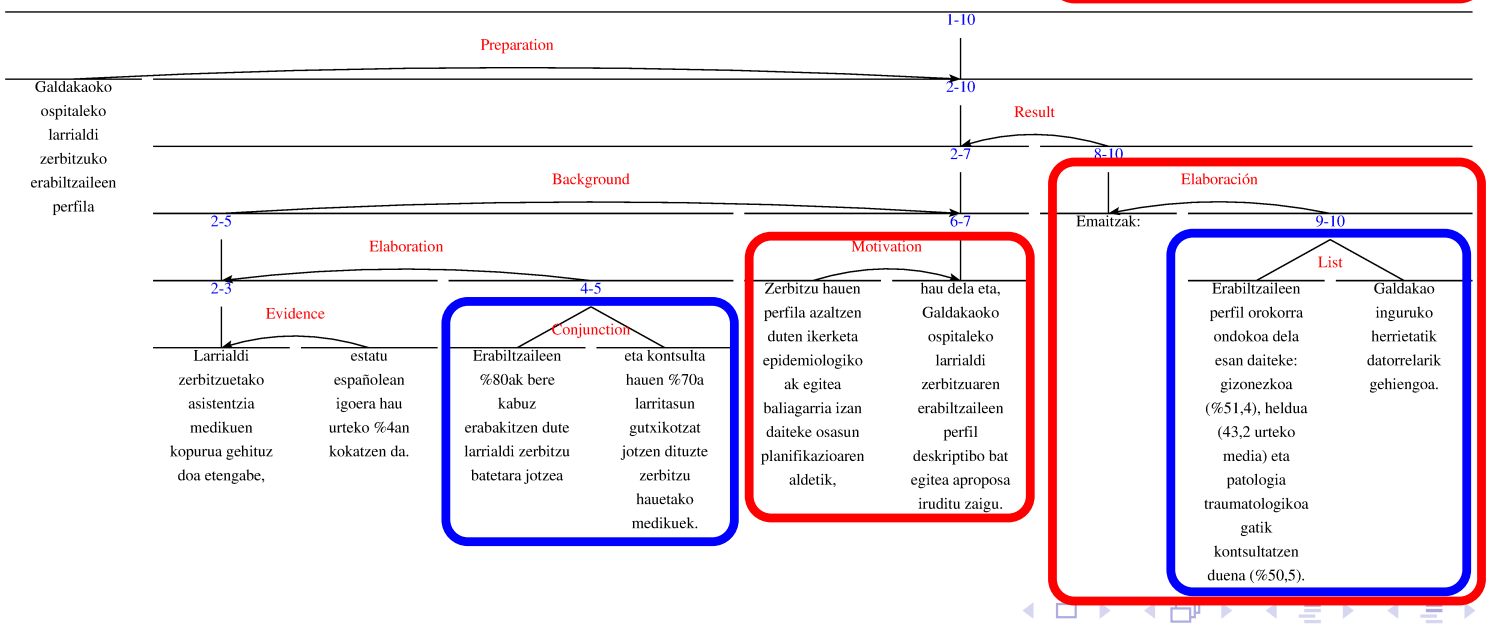
Different interpretations of [GMB0401]





Different interpretations of [GMB0401]





Inter-annotator agreement in RST relations

- The RST TreeBank (Carlson et al., 2001)
  - from 0.5973 to 0.7921 κ (2 annot., 30 texts: 1918 EDUs)
  - from 0.6017 κ to 0.7555 κ (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 κ (2 annotators, 4 texts)
- The Basque RST TreeBank (Iruskieta et al., 2013a)
  - 0,568 κ or 61.47% F_1 (2 annot., 60 texts: 1470 EDUs)

N	RCA	RC	RA	R	RR agreement
81.73%	47.76%	6.27%	3.41%	4.03%	**61.47%**
No-Match	**Nuclearity**	**N/N-N/S**	**Attachment**	**Constituent**	
0.23%	6.73%	8.90%	0.08%	0.15%	**RR disagreement**
Relation	**R-Similar**	**R-MissMatch**	**R-Specificy**	**Segmentation**	**38.53%**
13.62%	5.88%	2.01%	0.93%	0.15%	

An automatic evaluation of RS-trees with RSTeval (Maziero and Pardo, 2009) of GMB0701

RSTeval | Tool for discourse parsing evaluation

This tool provides an automatic method to compare two RST structures,
one made by a human being (the ideal structure)
and another made by an automatic system.

Evaluation ID: Euskara

Text	Units			Span			Nuclearity			Relation		
ID	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision
GMB07	10 of 10	1	1	17 of 19	0.894736842105263	0.894736842105263	16 of 19	0.842105263157895	0.842105263157895	16 of 19	0.842105263157895	0.842105263157895

Evaluation Table

Constituent	Units		Spans		Nuclearity		Relations	
	Manual	Auto	Manual	Auto	Manual	Auto	Manual	Auto
1 to 4 (Larritasunezko_irizpide...onkologian)	x	x	x	x	s	s	prestatzea	prestatzea
5 to 15 (Ikerketa_Pierre...aztertu)	x	x	x	x	n	n	span	span
16 to 22 (Basurtoko_Ospitaleko...gaixok)	x	x	x	x	n	n	span	span
23 to 31 (Pierre_Martyren...asmoz)	x	x	x	x	s	s	helburua	helburua
32 to 35 (elkarrizketa_zitzaien...guztiei)	x	x	x	x	n	n	span	span
!23 to 35 (Pierre_Martyren...guztiei)			x	x	s	n	elaborazioa	span
36 to 38 (7_itemak...aztertuta)	x	x	x	x	s	s	metodoa	metodoa
39 to 50 (estatistikoki_desberdintasun...05)	x	x	x	x	n	n	span	span
!36 to 50 (7_itemak...05)			x	x	n	n	lista	lista
51 to 57 (Horrez_item...bereizten)	x	x	x	x	n	n	lista	lista
58 to 60 (horiei_balorazio...orokorra)	x	x	x	x	n	n	lista	lista
!51 to 60 (Horrez_item...orokorra)			x	x	n	n	lista	lista
61 to 65 (prozesuaren_igurkapenen...dizkigute)	x	x	x	x	n	n	lista	lista
!51 to 65 (Horrez_item...dizkigute)			x	x	n	n	lista	lista
!36 to 65 (7_itemak...dizkigute)			x	x	s	s	ondorioa	ondorioa
!23 to 65 (Pierre_Martyren...dizkigute)				x		s		elaborazioa
!16 to 65 (Basurtoko_Ospitaleko...dizkigute)				x		s		metodoa
!5 to 65 (Ikerketa_Pierre...dizkigute)			x	x	n	n	span	span
!1 to 65 (Larritasunezko_irizpide...dizkigute)			x	x	r	r	span	span
!16 to 35 (Basurtoko_Ospitaleko...guztiei)			x		s		metodoa	
!5 to 35 (Ikerketa_Pierre...guztiei)			x		n		span	

Outline

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

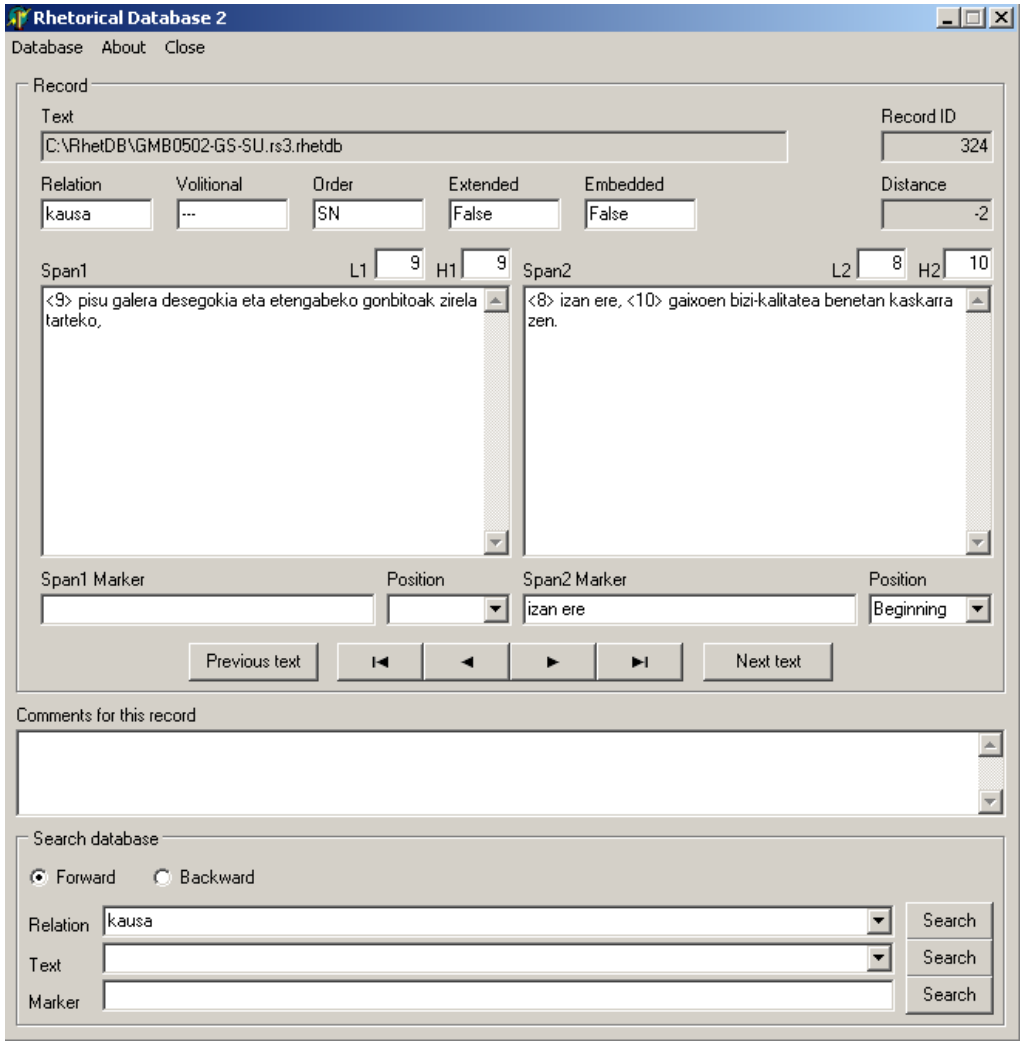
- **What is signalling?**
  - *a*) DM annotation (automatically)
  - *b*) Annotation of the most frequent forms (and functions) (Taboada and Das, 2013)
    - to distinguish volitional/non-volitional relations of cause exploiting the information provided by verb tense (Antonio, 2012)
    - to have more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

Signals	**Examples**
Coordinators	however, therefore, in fact
Morphology	-ing, non-finite verbs
Lexical	concede, cause
Entity	entities
Semantic	synonyms, antonyms, hyponyms
Syntax	question-answer,
Graphic-numeric	1. (...) 2., a) (...) b)
Complex signals	...

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?

Annotators	CAUSE%	RESULT%	PURPOSE%
A_1-A_2	71.43	59.70	90.00
A_1-A_4	67.86	50.75	80.91
A_2-A_4	73.21	37.31	78.18
A_1-A_2-A_4	58.93	37.31	75.45

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 RRs and their signals

Rhetorical Relations			Signals	%	DU_1	DU_2	$DU_{1/2}$	N	S	S/N
Presentational (pragmatic)	PREPARATION	110	2	1.82	2				2	
	BACKGROUND	75	16	21.33	12	4		4	12	
	ENABLEMENT	6	6	100.00		6		1	5	
	MOTIVATION	5	5	100.00		3	2		3	2
	EVIDENCE	11	7	63.64	1	6		1	6	
	JUSTIFY	14	13	92.86	1	11	1		12	1
	ANTITHESIS	5	4	80.00	1	1	2		2	2
	CONCESSION	40	39	97.50	11	26	2	7	30	2
	RESTATEMENT	10	7	70.00		7			7	
	SUMMARY	10	5	50.00		5			5	
Subject-matter (semantic)	ELABORATION	286	84	29.37		82	2		82	2
	MEANS	93	81	87.10	19	62			81	
	CIRCUMSTANCE	57	53	92.98	44	9		1	52	
	SOLUTIONHOOD	10	9	90.00	3	3	3	3	3	3
	CONDITION	20	19	95.00	12	5	2		17	2
	UNCONDITIONAL	1	1	100.00		1			1	
	INTERPRETATION	28	22	78.57	3	17	2		20	2
	EVALUATION	11	10	90.91		10			10	
	CAUSE	56	53	94.64	23	21	9	3	41	9
	RESULT	67	57	85.07	1	55	1	2	54	1
	PURPOSE	110	109	99.09	40	68	1	3	105	1
Multinuclear	LIST	166	87	52.41	3	53	31			
	SEQUENCE	32	21	65.63	2	15	4			
	CONJUNCTION	50	38	76.00		37	1			
	CONTRAST	40	33	82.50	2	23	8			
	DISJUNCTION	2	2	100.00		2				
	Total	1315	783	59.54	180	532	71	25	550	27

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 (≥3 occurrences)

Ambiguous signals			Non-ambiguous signals and RRs			
Signal	**Translation**	**#**	**Signal**	**Translation**	**#**	**RR**
eta	and	34	-tzeko	Purpose morpheme	27	PURPOSE
-nez	given	15	erabiliz	used	8	MEANS
-tuz	-ing	11	-tzean	-ing	8	CIRCUMSTANCE
baina	but	11	helburu	purpose	8	PURPOSE
bait-	because	10	adibidez	for example	6	ELABORATION
ba-	if	10	ondoren	then	6	SEQUENCE
bestalde	moreover	9	hala ere	however	6	CONCESSION
era berean	likewise	8	-ela eta	cause morpheme	5	CAUSE
izan ere	in fact	8	arazo	problem	4	SOLUTIONHOOD
gainera	futhermore	6	izan arren	despite	4	CONCESSION
berriz	whereas	5	-tu ondoren	then	4	CIRCUMSTANCE
alde batetik	on the one hand	5	-nean	when	4	CIRCUMSTANCE
-ta	-ed	5	nahiz eta	although	3	CONCESSION
			lortutako emaitzek baieztatzen dute	the results obtained confirm	3	INTERPRETATION
			hau da	that is to say	3	RESTATEMENT
			1.	1.	3	LIST

- 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

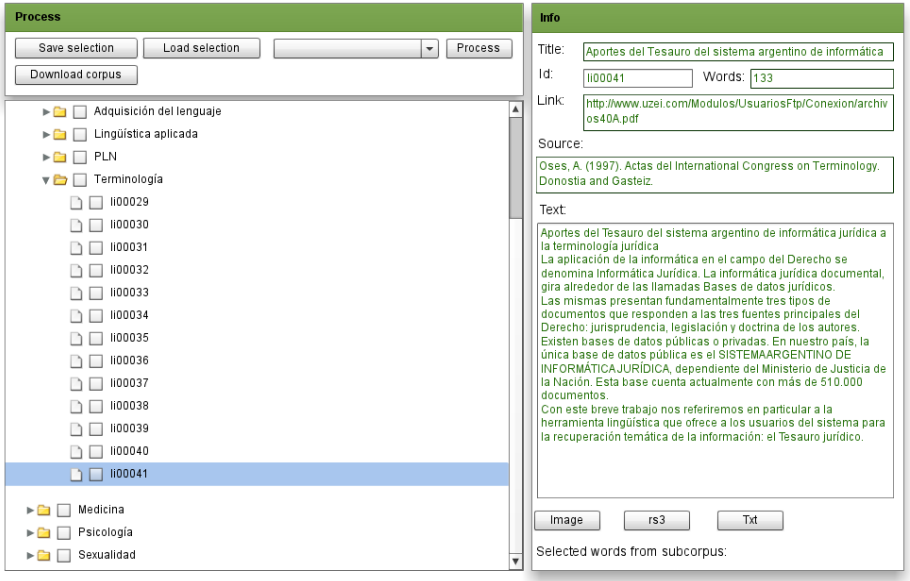
Outline

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
  - 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 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

RST Spanish Treebank (da Cunha et al., 2011)

- 9 different domains, 267 texts. A double annotation of test-set (84 texts) and 10 different annotators.
- Different queries for the first time:
  - *i*) Consult statistics
  - *ii*) 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)

SEARCH section: queries based on POS features

– Queries based on word-form, lemma and POS features

	Doc.	EDU Id	Word	CU	EDU
1	TERM50	sent2	taldeek / helburua	BAI	[...] Hitzaldi honek azken hiru urteotan lau unibertsitate hauen *talde*ek egindako ikerkuntzaren ondorioetako batzuk azaltzeko *helburua* izango luke.
			groups / aim	YES	[...] 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.
2	ZTF13	sent1	taldearen / helburu	BAI	[...] Gure *ikerkuntza talde*aren *helburu* nagusia, [...]
			group's / aim	YES	[...] Our research group's principal aim, [...]
3	ZTF13	sent17	taldearen / helburu	EZ	Alor honetan, gure *ikerkuntza talde*aren *helburu* nagusiak bi dira.
			group's / aim	NO	In this field, our research group has two main aims.
1	ZTF15	sent7	helburu / talde	EZ	[...] bestelako galdera zailagoei ere erantzutea dute *helburu*, hala nola, espezieen biogeografia, *talde*aren filogenia, eta abar.
			aim / group	NO	[...] the aim is to answer other such difficult questions, such as species biogeography, group phylogeny, etc.

Multilingual SEARCH section: POS queries

	Doc.	EDU Id	Word	Segment	
1	TERM38_A1.txt	seg2	paper / look	This paper is intended to look at the challenges faced by neology in terminology at the present time .	Context
2	TERM19_A1.txt	seg12	paper / looks	This paper looks , on the basis of experience in the standardisation of terminology in Catalan , at the social need for standardisation of terminology .	Context
1	TERM23_A1.txt	seg13	paper / groups	Our paper will discuss the methodology used by both groups in term creation .	Context
2	TERM30_A1.txt	seg27	paper / groups	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 .	Context
3	TERM50_A1.txt	seg2	paper / groups	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 .	Context
1	TERM30_A1.txt	seg25	used / groups / and	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 ,	Context
2	TERM31_A1.txt	seg6	divided / groups / and	Their areas of application can be divided into two main groups : information indexing and the making-up of terminological glossaries .	Context

- 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

EDU	Segment	GMB0301-GS.rs3 (7)	Tagger	CU
1	Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.		GS	
	Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.			
2	"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da.		GS	
	"Recurrent aphthous stomatitis" is one of the most frequent oral pathologies.			
3	tamainu, kokapena eta iraunkortasuna aldakorra izanik.		GS	
	having a variable size, location and duration.			
4	Honen etiologia eztabaidagarria da.		GS	
	It has a controversial etiology.			
5	Ultzera mingarri batzu bezela agertzen da,		GS	
	It is characterized by the apparition of painful ulcers,			
6	Hauek periodiki beragertzen dira.		GS	
	These ulcers appear recurrently.			
7	**Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.**		GS	See
	In this paper we analyze the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.			

Relations linked to the CU

GMB0301-GS.rs3: CU and relations

CU: Lan honetan patologia arrunt honetan ezaugarri . . . garrantsitsuenak analizatzen ditugu.

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

Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.	prestatzea –>	"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da. Honen etiologia eztabaidagarria da. Ultzera mingarri batzu bezela agertzen da, tamainu, kokapena eta iraunkortasuna aldakorra izanik. Hauek periodiki beragertzen dira. Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.
Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.	preparation –>	"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.
"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da. Honen etiologia eztabaidagarria da. Ultzera mingarri batzu bezela agertzen da, tamainu, kokapena eta iraunkortasuna aldakorra izanik. Hauek periodiki beragertzen dira.	testuingurua –>	Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.
"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.	preparation –>	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

TERM18% (24)

Id	Segment	Tagger
1	General trends in standardization of scientific terminology in Serbian: a critical analysis of the state of affairs	A1
2	Building the terminology of any scientific area is a long and laborious process.	A1
3	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.	A1
4	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.	A1
5	In this paper, an attempt is made to critically evaluate the above outlined trend, from both inherently linguistic	A1

Id	Segment	Tagger
1	Tendencias generales de la normalización en la terminología científicotécnica de la lengua serbia: análisis crítico de la situación	A2
2	La construcción terminológica de cualquier área científica es un proceso largo y laborioso.	A2
3	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.	A2
4	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	A2

Id	Segment	Tagger
1	Zientzia-arloko terminologiaren normalizazio-joera orokorrak serbiera: egoeraren analisi kritikoa	A3
2	Edozein zientzia-arlotako terminologia eraikitzea lan luzea eta neketsua da.	A3
3	Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unitate lexikalen maileguak eta unitate-egitura luzeagoen maileguak hartzen dira zientzia-erregistro zehatz baterako, itzulpenak edo kalkoak egin ordez.	A3
4	Horien arabera, ingelesetik hartutako maileguei lehentasuna ematen zaie "serbieraren zientzia-barietate modernoa" izeneko hizkuntzakodean, itzulpena eta egitura-kalkoaren aurretik.	A3
5	Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko	A3

RELATIONS section

– Specific RRs queries where signals are underlined

Relation: Kausa 'Cause' (27)				
Left span	**NS**	**Rigth span**	**Relation**	**Ref.**
Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unita[...] In recent decades, many Serbian researchers working in different scientific fields have noticed a tendency and this is outlined here: the English unit [...]	< –	Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko hamarkadetan ingelesari eman diote [...] Indeed, Serbian scientists from different schools of thought have reached a consensus and have given English [...]	Cause	TERM18
Terminologiak berak ere, uztartu egin behar ditu joera orokor horiek, eransten zaizkien beste batzuekin batera, hala nola: teknologien [...] Terminology itself must seek to unite these general trends, along with others related to them, for example: technology	< –	gizartearekin lotuta dagoen jarduera denez, since it is an activity linked to society,	Cause	TERM19

Multilingual RELATIONS section

– CIRCUMSTANCE relation in three languages

Left unit	Sense	Right unit	Relation name	Document	Tagger	Notes
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.	circumstance	TERM23	A1	
Ever since information technology first made it possible to store and then process linguistic data,	–>	terminology has had to adapt constantly to technological innovations.	circumstance	TERM29	A1	
Desde que la informática hizo posible el almacenamiento de datos lingüísticos y posteriormente su tratamiento,	–>	la terminología no ha cesado de adaptarse a las innovaciones tecnológicas,	circumstance	TERM29	A2	
Informatikak hizkuntzako datuak gorde eta, aurrerago, tratatzeko aukera eman zigunetik,	–>	terminologiak teknologi berrikuntzetara egokitu behar izan du etengabe.	circumstance	TERM29	A3	
– : – : –	– : – : –	– : – : –	– : – : –	– : – : –	– : – : –	

SIGNALS section

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

Signal: *baina* 'but'			
Gainerakoan, prokasu adierazle egokiak daude,	Kontzesioa	baina altan dagoen gaixoaren ahalmen funtzionalaren erregistro urria antzematen da,	GMB0504
With respect to the other aspects, the indicators of process are good	Concession	**but** there is poor recording of the patient's functional capacity on discharge,	
Bestalde, Euskaltzaindiak hitz elkartuen bidea (1995eko urtarrilaren 27an onartutako araua) proposatzen du adjektibo erreferentzialak itzultzeko,	Kontrastea	baina arauan bertan esaten denez, "...ahal den guztian...",	TERM22
Euskaltzaindia proposed a mechanism of compound words (in a standard approved on January 27th 1995) for the translation of referential adjectives.	Contrast	**However** the academy also confirmed, ..."whenever possible",	
Signal: *erabiliz* 'using'			
Komunikazio honekin, hauxe frogatu nahi da: halako kasurik gehien-gehienetan, proposamen autoktonoa baztertzeko emandako arrazoiak ez direla ez hizkuntzarenak ez semantikoak, soziologikoak baizik,	metodoa	adibide paraleloak erabiliz,	TERM21
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.	method	**through** parallel examples,	
Horretarako eredu nagusiak lortu behar dira.	metodoa	dauden hiztegi teknikoetan oinarritu, eta teknika estatistikoak erabiliz,	TERM31
To that end, principal models must be obtained.	method	basing work on existing technical dictionaries and **using** statistical techniques,	

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)

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

- 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

Outline

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

Outline

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 (instalation is needed)

Different segmentations of modified GMB0301

- Compare this segmentations:

Text	GS	SEG1	SEG2	CODRA
Recurrent aphtous stomatitis is one of the most frequent oral conditions.	EDU2	EDU2	EDU2	EDU2
Its etiology is controversial and	EDU3	EDU3-B	EDU3-B	EDU3
it is characterised by the appearance of painful and recurrent ulcers,	EDU4-B	EDU3-E	EDU3-M	EDU4
whose sizes, locations, and durations vary.	EDU4-E	EDU4	EDU3-E	EDU5
These ulcers reappear periodically.	EDU5	EDU5	EDU4	EDU6
This paper analyses the most important epidemiological, etiological, pathological	EDU6	EDU6	EDU5	EDU7
and clinical features of this common oral pathology.	EDU7	EDU7	EDU6	EDU8

- 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:

Text	GS	SEG1	SEG2	CODRA
Recurrent aphtous stomatitis is one of the most frequent oral conditions.	EDU2	EDU2	EDU2	EDU2
Its etiology is controversial and	EDU3	EDU3-B	EDU3-B	EDU3
it is characterised by the appearance of painful and recurrent ulcers,	EDU4-B	EDU3-E	EDU3-M	EDU4
whose sizes, locations, and durations vary.	EDU4-E	EDU4	EDU3-E	EDU5
These ulcers reappear periodically.	EDU5	EDU5	EDU4	EDU6
This paper analyses the most important epidemiological, etiological, pathological	EDU6	EDU6	EDU5	EDU7
and clinical features of this common oral pathology.	EDU7	EDU7	EDU6	EDU8

- 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

Outline

Nuclearity and summarization: GMB0301

1	2	3	4	5	6	7
Estomatitis aftosa recurrente (I): epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.	“Estomatitis aftosa recurrente” deritzon patologia, ahoan agertzen den ugarienetako bat da.	Honen etiologia eztabaidagarria da.	Ultzera mingarri batzu bezela agertzen da,	tamainu, kokapena eta iraunkortasuna aldakorra izanik.	Hauek periodiki beragertzen dira.	Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.
Recurrent aphtous stomatitis (I): epidemiologic, etiologic and clinical features.	Recurrent aphtous 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.

— Summarize the text above choosing 3 or 4 discourse units:

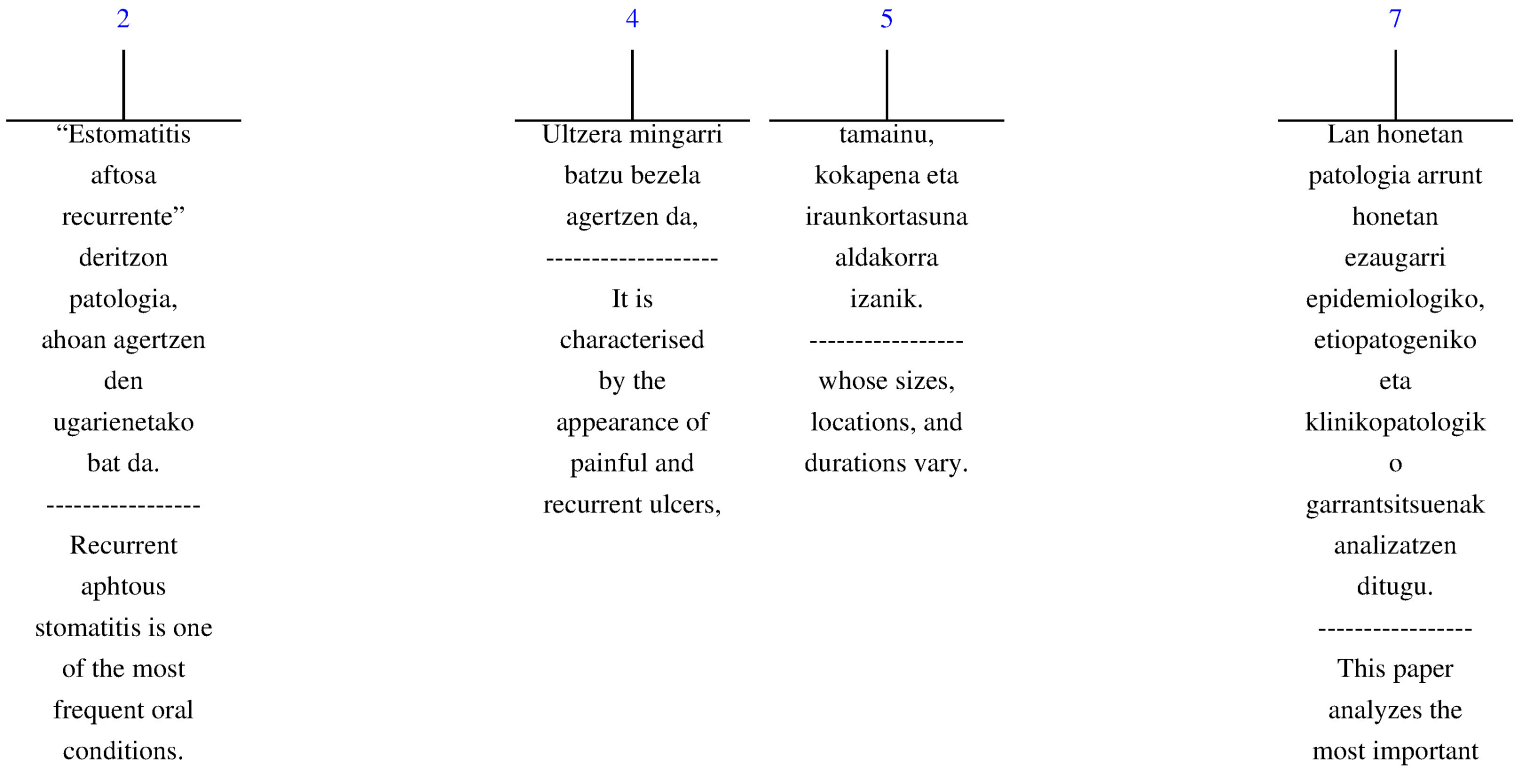
Nuclearity and summarization: GMB0301

1	2	3	4	5	6	7
Estomatitis aftosa recurrente (I): epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.	"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da.	Honen etiologia eztabaidagarria da.	Ultzera mingarri batzu bezela agertzen da,	tamainu, kokapena eta iraunkortasuna aldakorra izanik.	Hauek periodiki beragertzen dira.	Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.
Recurrent aphtous stomatitis (I): epidemiologic, etiologic and clinical features.	Recurrent aphtous 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.

– Summarize the text above choosing 3 or 4 discourse units:

Nuclearity and summarization: GMB0301

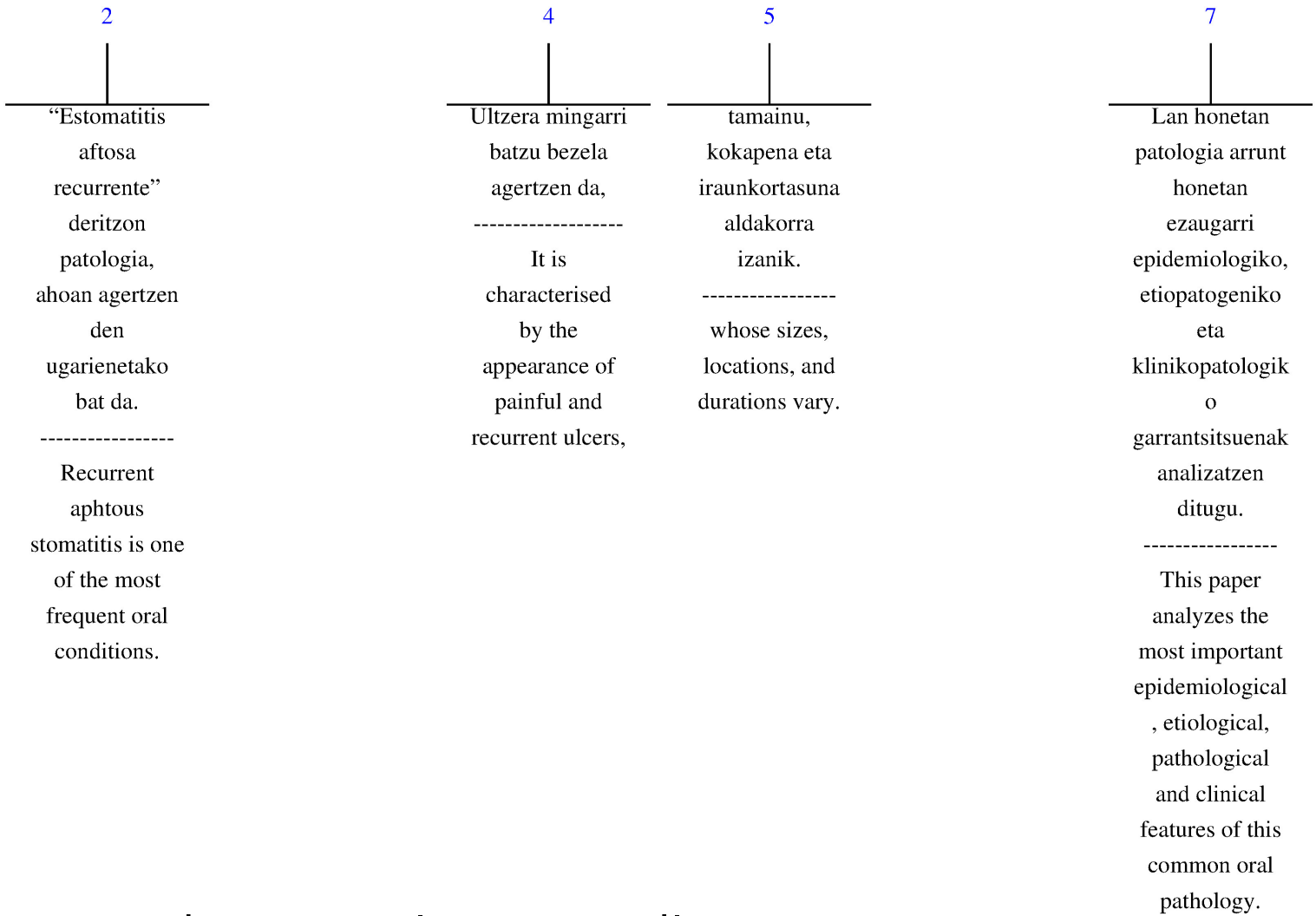
– Has the created summary any sense?



– Choose now the 2 most important discourse segments

Nuclearity and summarization: GMB0301

– Has the created summary any sense?



– Choose now the 2 most important discourse segments

Nuclearity and summarization: GMB0301

– Has the created summary any sense?

2

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

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

7

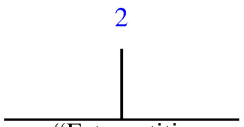
Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.

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:

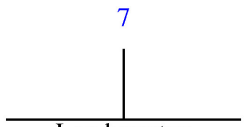
Nuclearity and summarization: GMB0301

– 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 garrantsitsuenak 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:

Nuclearity and summarization: GMB0301

– Has the central unit any topic indicator?

– *This paper analyzes the most important...*

7

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

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

Nuclearity and summarization: GMB0301

- Has the central unit any topic indicator?
  - *This paper analyzes the most important...*

7

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

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

Summarization: based on discourse structure: GMB0401

- 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).

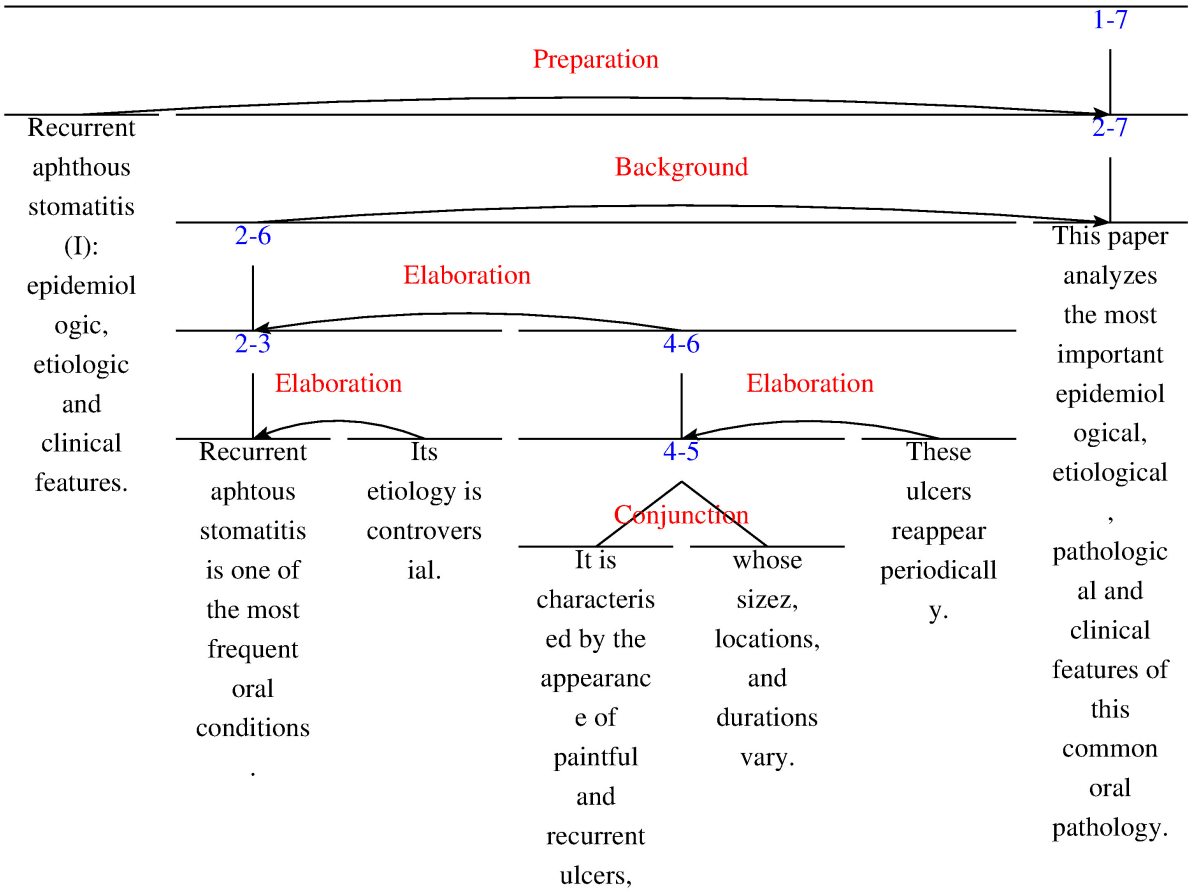
(7) **Recurrent aphtous stomatitis is one of the most frequent oral conditions.** *It is characterised by the appearance of paintful and recurrent ulcers, whose sizes, locations, and durations vary.* **This paper analyzes the most important epidemiological, etiological, pathological and clinical features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

"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da. *Ultzera mingarri batzu bezela agertzen da, tamainu, kokapena eta iraunkortasuna aldakorra izanik. Hauek periodiki beragertzen dira.* **Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.** GMB0301

A simple summary based on rhetorical structure. GMB0301

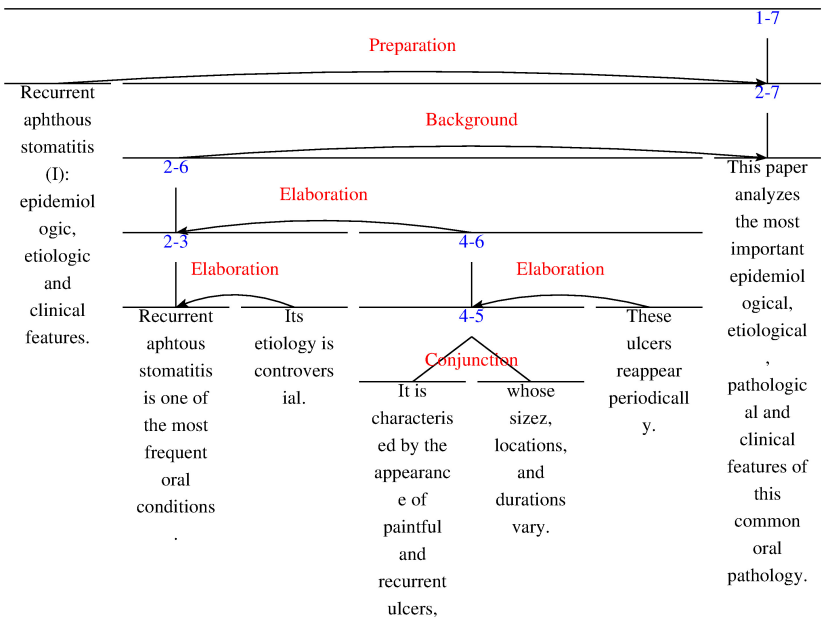
(8) Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.

Recurrent aphtous stomatitis is one of the most frequent oral conditions. Its etiology is controversial. It is characterised by the appearance of paintful 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. **GMB0301**



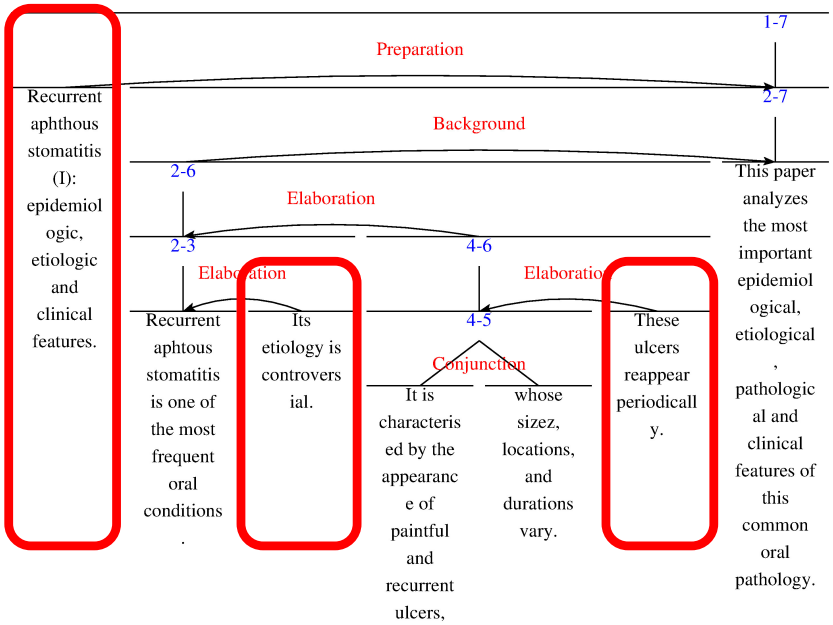
A simplification of the RS-tree. GMB0301

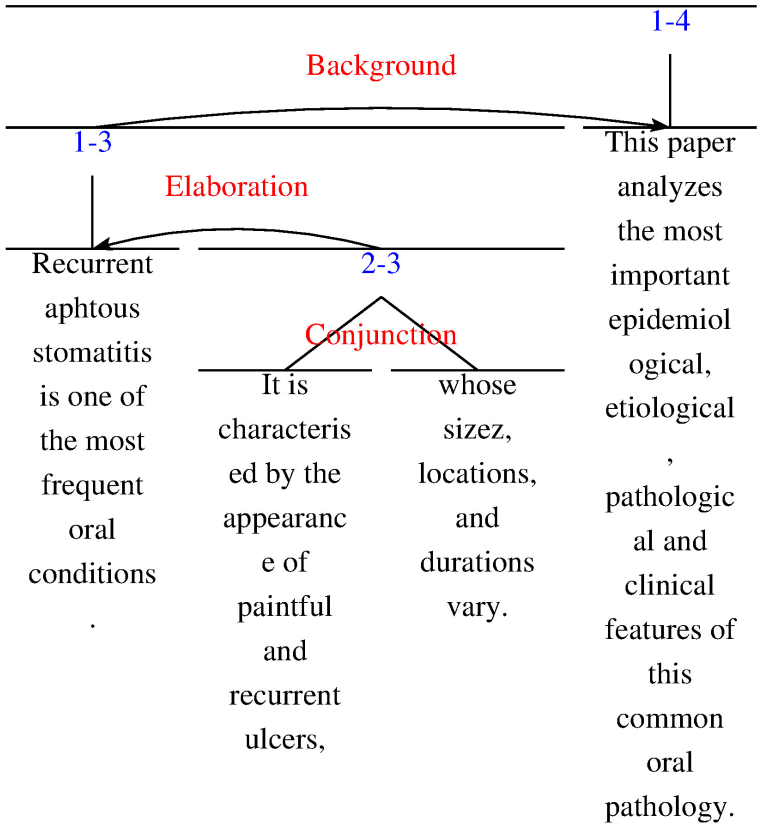
– After deleting the satellite units the text part is still coherent



A simplification of the RS-tree. GMB0301

– After deleting the satellite units the text part is still coherent





No-coherent summary of GMB0301

– 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 (I): epidemiologic, etiologic and clinical features.]$_1$ [Its etiology is controversial.]$_3$ [These ulcers reappear periodically.]$_6$ GMB0301

Basic heuristics based on nuclearity

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

(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

Go to CU: 34

Outline

Choosing relations: SEQUENCE or CONCESSION or INTERPRETATION

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.

Choosing relations: SEQUENCE or CONCESSION or INTERPRETATION

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

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 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 Tarmaochta) Fiontar and VOCALL are addressing the terminological needs of both Irishmedium 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

Outline

CIRCUMSTANCE: signals

– **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.

CIRCUMSTANCE: signals II

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, –> 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, 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, ahaztuxe uzten da normalizazioaren eta gizartearen artean egon behar den lotura estua.

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.
4. Halako tresna bat euskararako garatu nahi badugu, –> eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik: bateratze-prozesua bukatzeke izateagatik, egindako ikerketak murritzak direlako eta hizkuntza eranskaria izateagatik.

CONDITION: signals II

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.
4. Halako tresna bat euskararako garatu nahi **ba**dugu, –> eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik: bateratze-prozesua bukatzeke izateagatik, egindako ikerketak murritzak direlako eta hizkuntza eranskaria izateagatik.

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.
3. Testu juridikoen itzulpenari ekiteko, ezinbestekoa da terminologia bera lantzea. <– Euskararen kasuan beharrizan hori areagotu egin da, esparru horretan gure hizkuntzaren egoera ez baita primerakoa, ezta hurrik eman ere.

ELABORATION: Signals II

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.
3. Testu juridikoen itzulpenari ekiteko, ezinbestekoa da terminologia bera lantzea. <– Euskara**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 **bait**a 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,

CAUSE: signals IV

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 dire**nez**, –> 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

An ambiguous example: The text

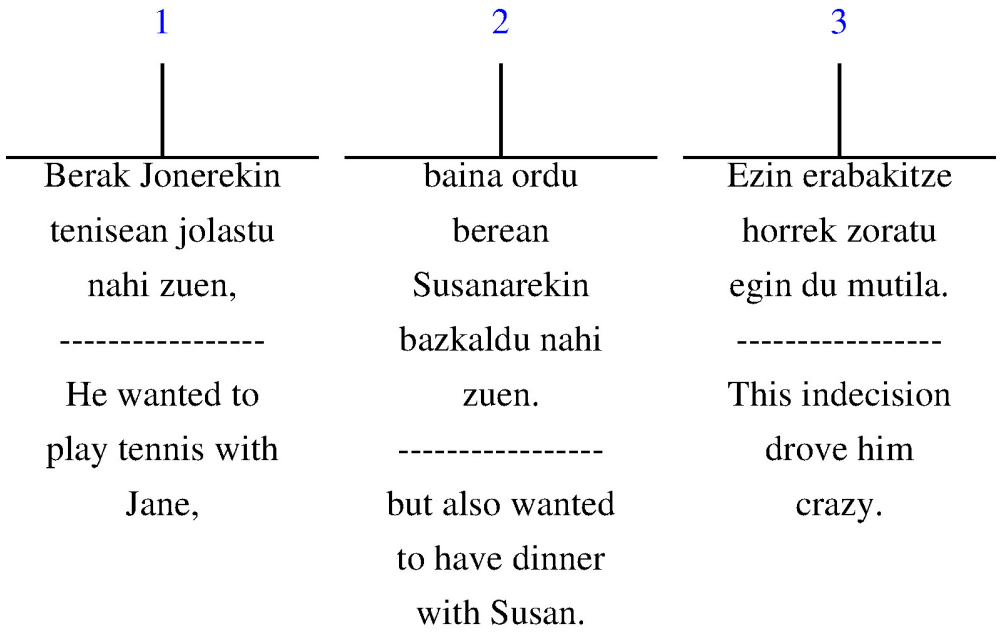
(13) He wanted to play tennis with Jane, but also wanted to have dinner with Susan. This indecision drove him crazy.

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

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

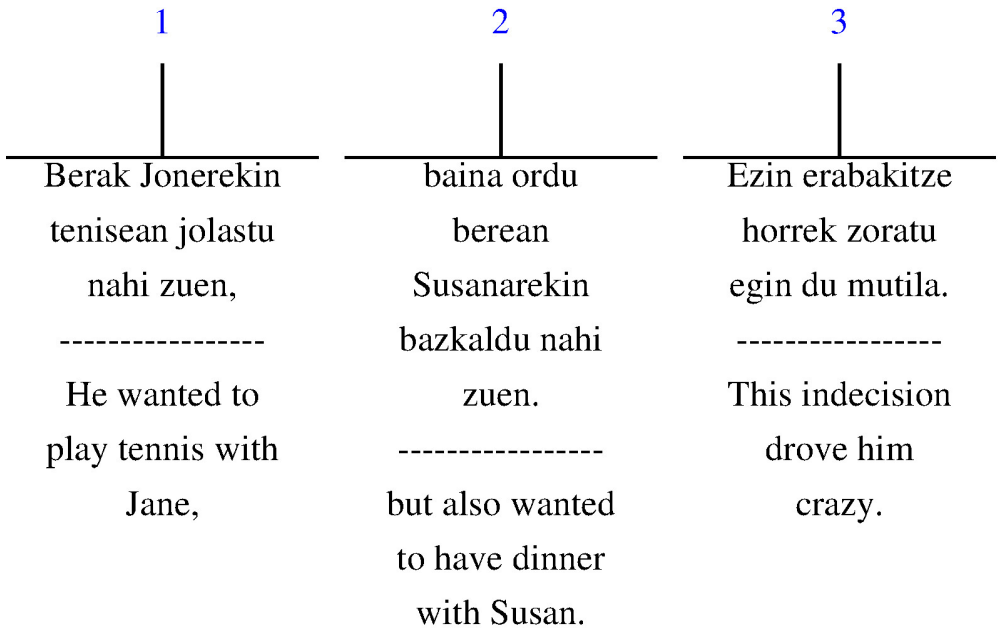
– How many propositions or discourse units are there in the example?

An ambiguous example: segments



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

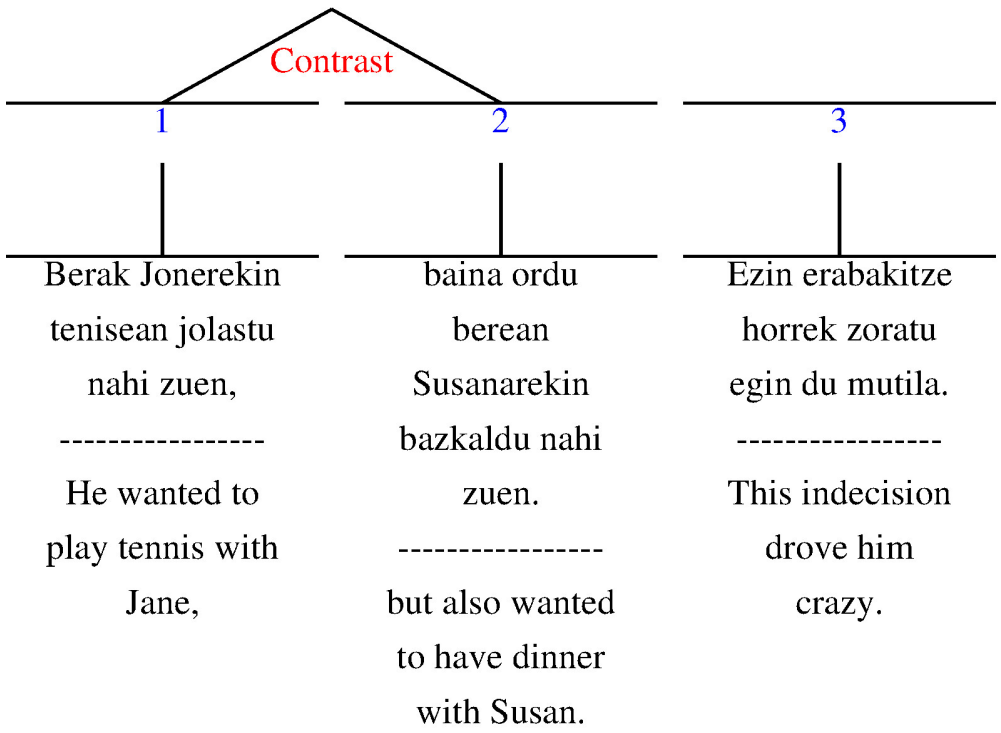
An ambiguous example: segments



– 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

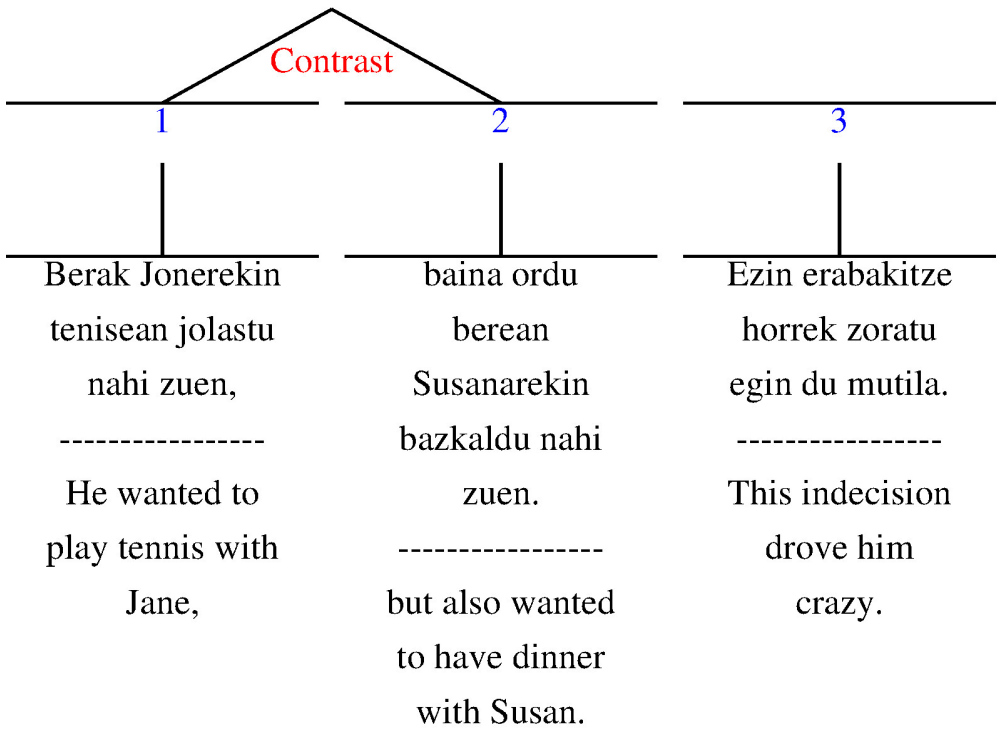
– Linking the intrasentential segments



– Now link the (inter)sentential span or segments

An ambiguous example: the incremental way

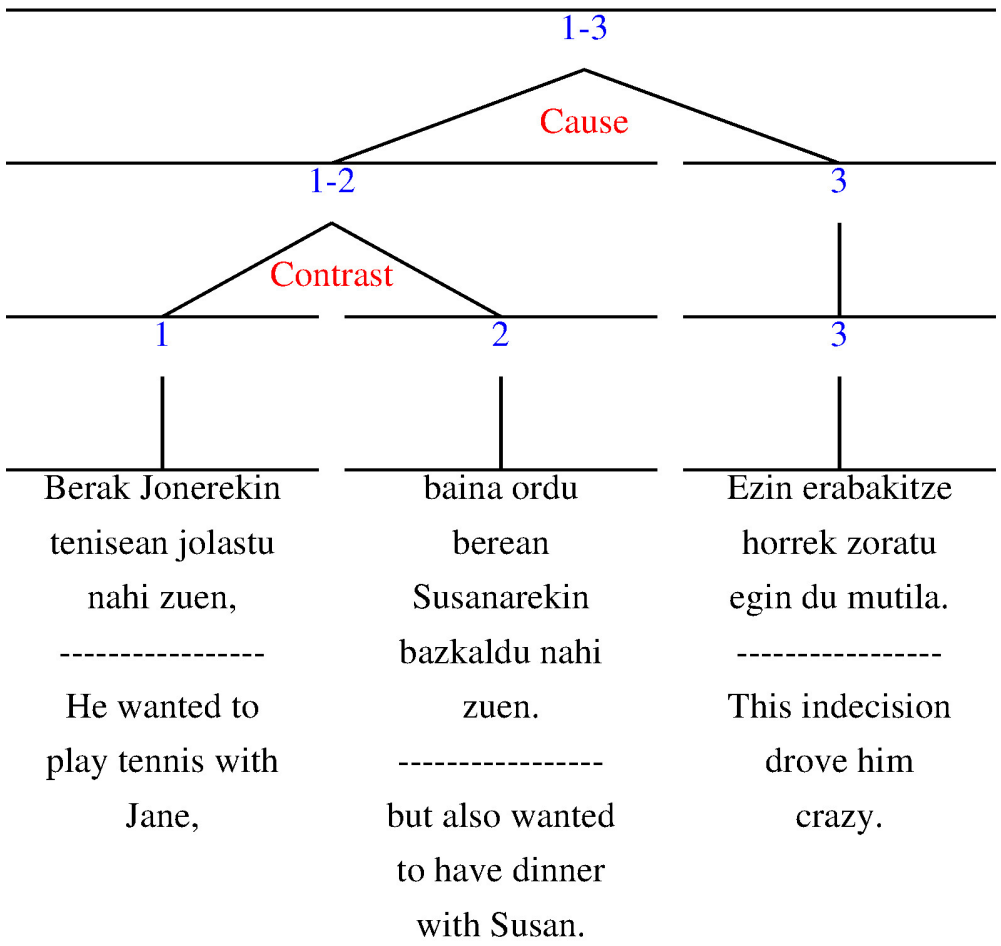
– Linking the intrasentential segments



– Now link the (inter)sentential span or segments

An ambiguous example: the incremental way

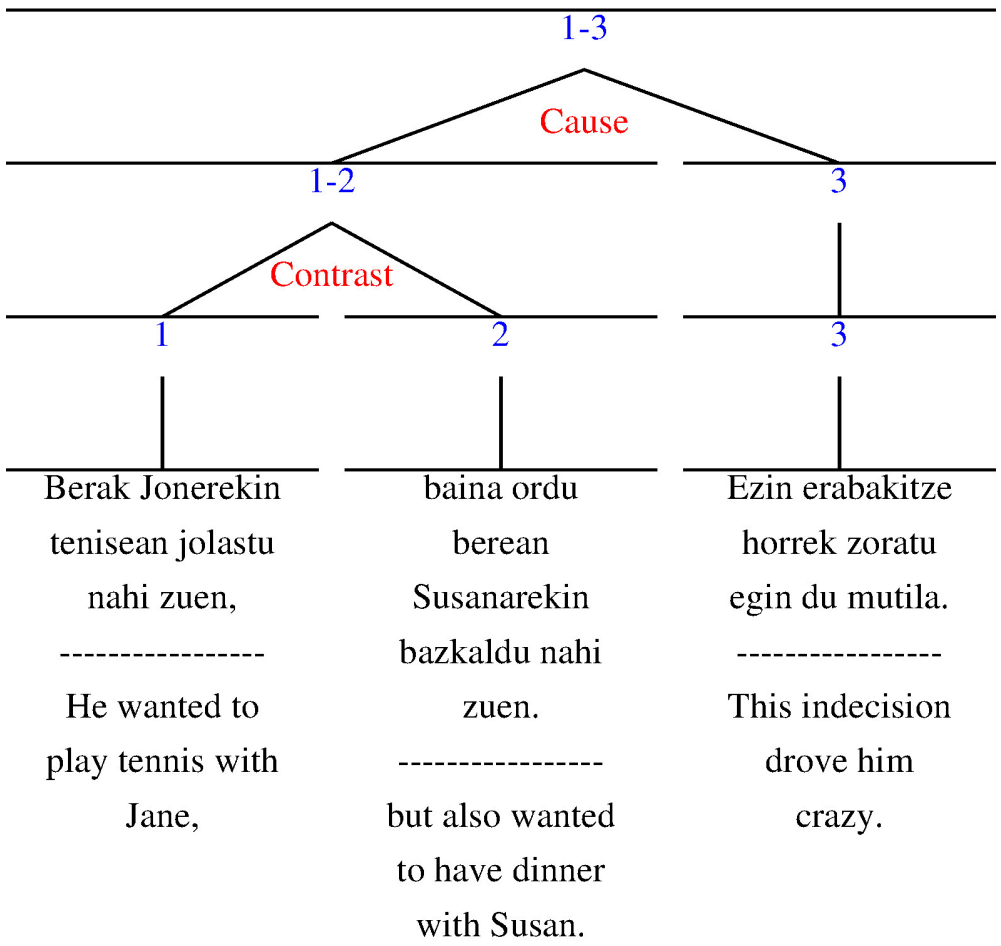
– Linking the (inter)sentential span or segments



– Now choose the nucleus (N) and satellite (S) units at intrasentential level

An ambiguous example: the incremental way

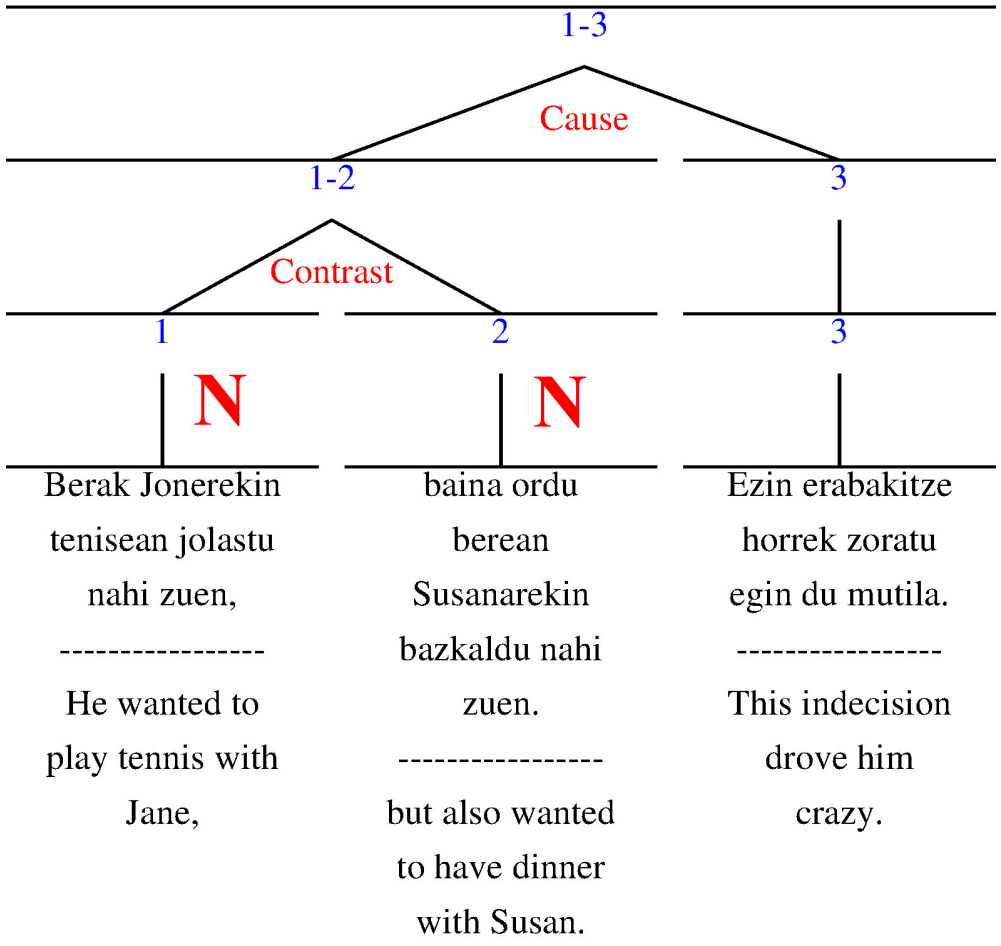
– Linking the (inter)sentential span or segments



– Now choose the nucleus (N) and satellite (S) units at intrasentential level

An ambiguous example: nuclearity of intrasentential units

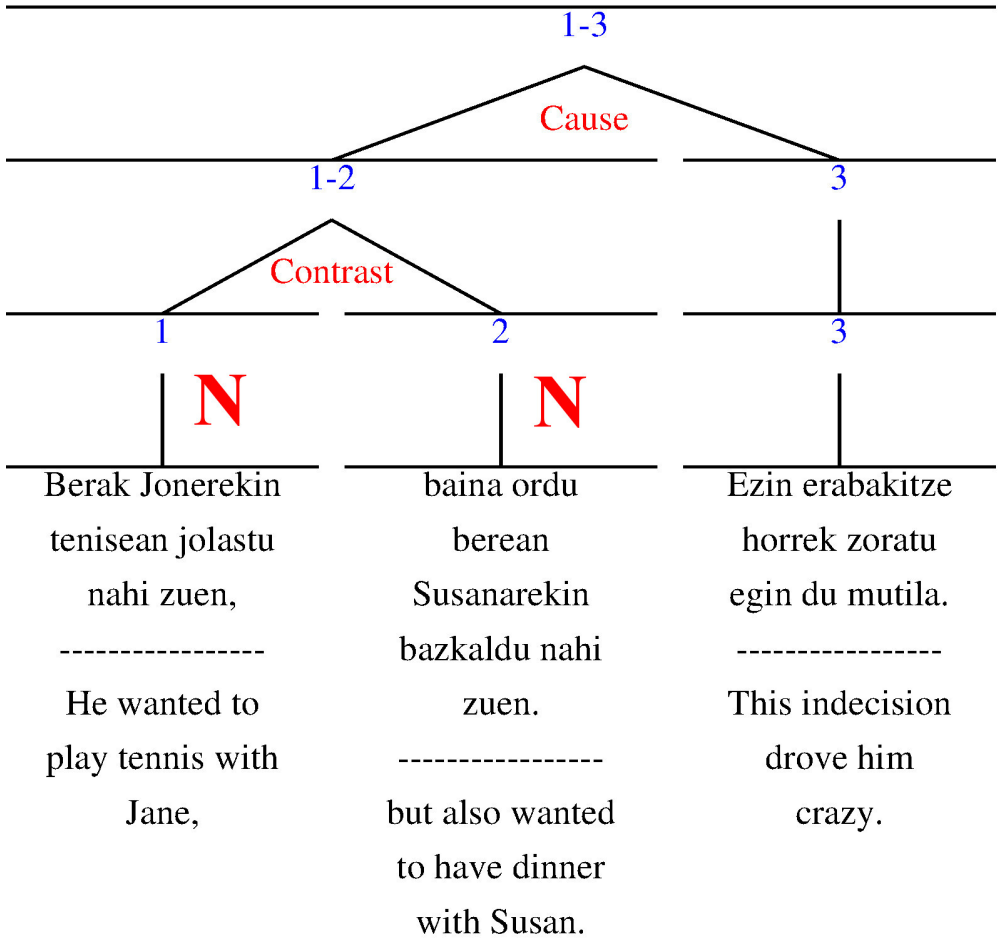
– Choosing the nucleus (N) and satellite (S) units of intrasentential units



– Now choose the nucleus (N) and satellite (S) units of sentential units

An ambiguous example: nuclearity of intrasentential units

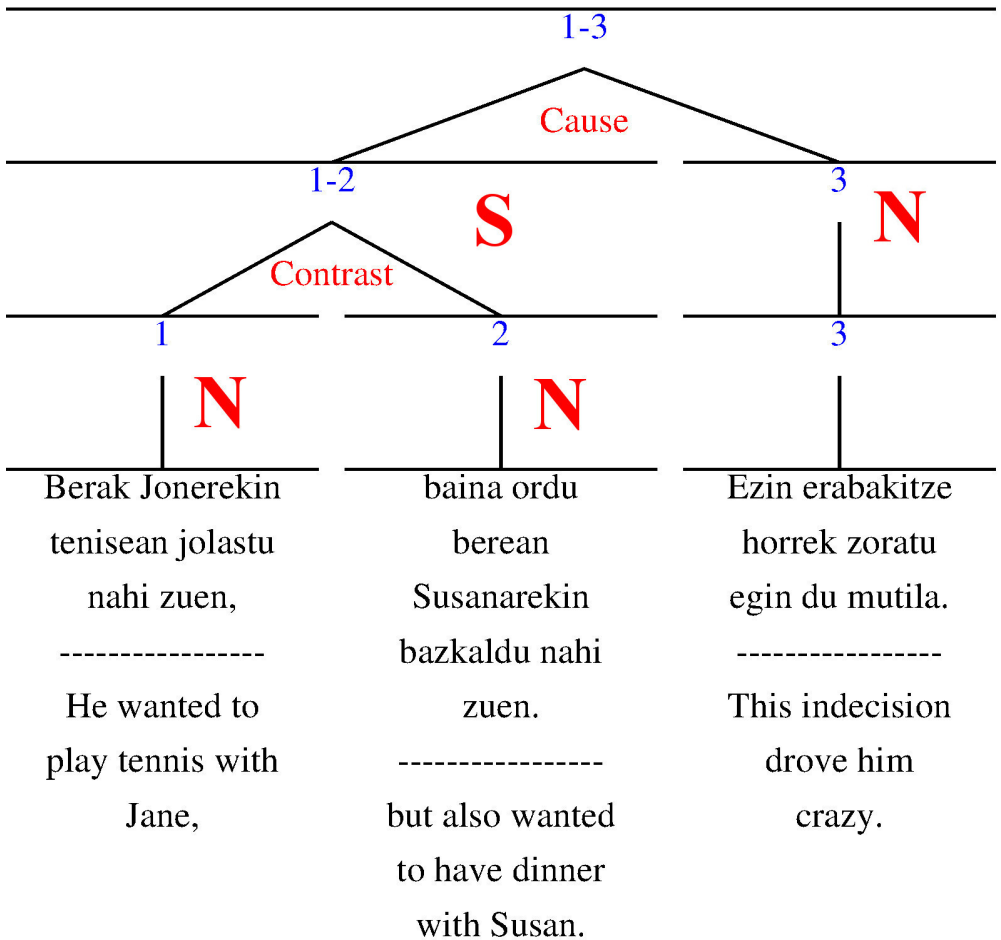
– Choosing the nucleus (N) and satellite (S) units of intrasentential units



– Now choose the nucleus (N) and satellite (S) units of sentential units

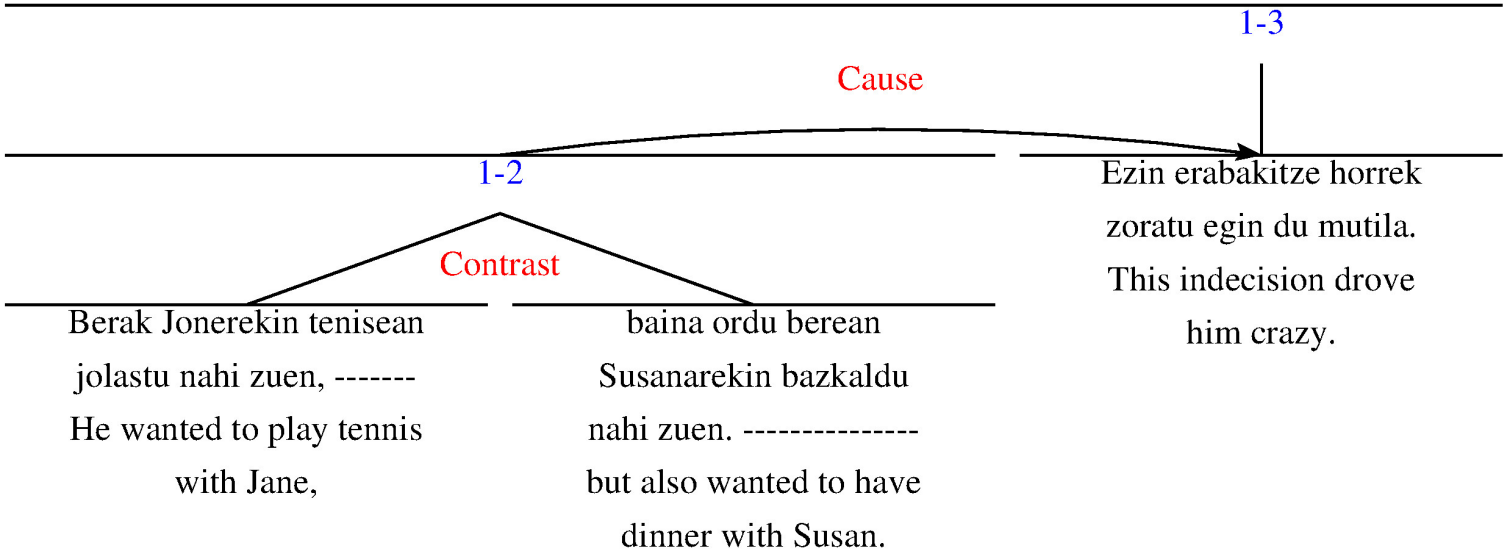
An ambiguous example: nuclearity of sentential units

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

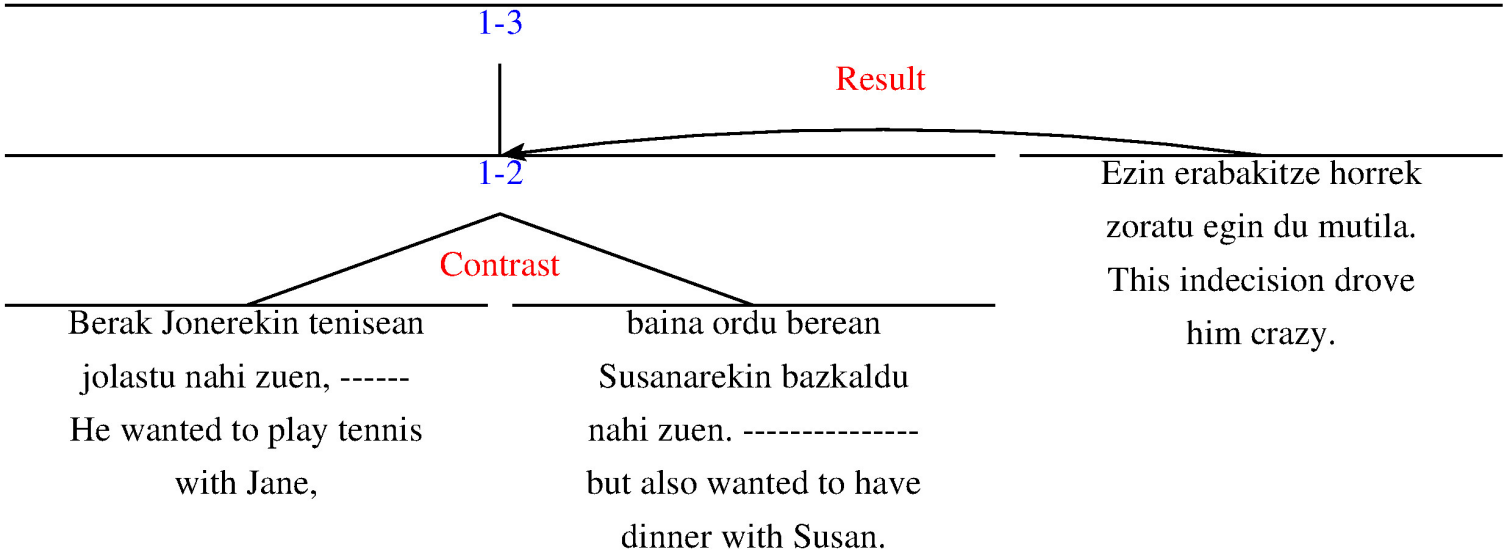


The importance of nuclearity in relational discourse structure

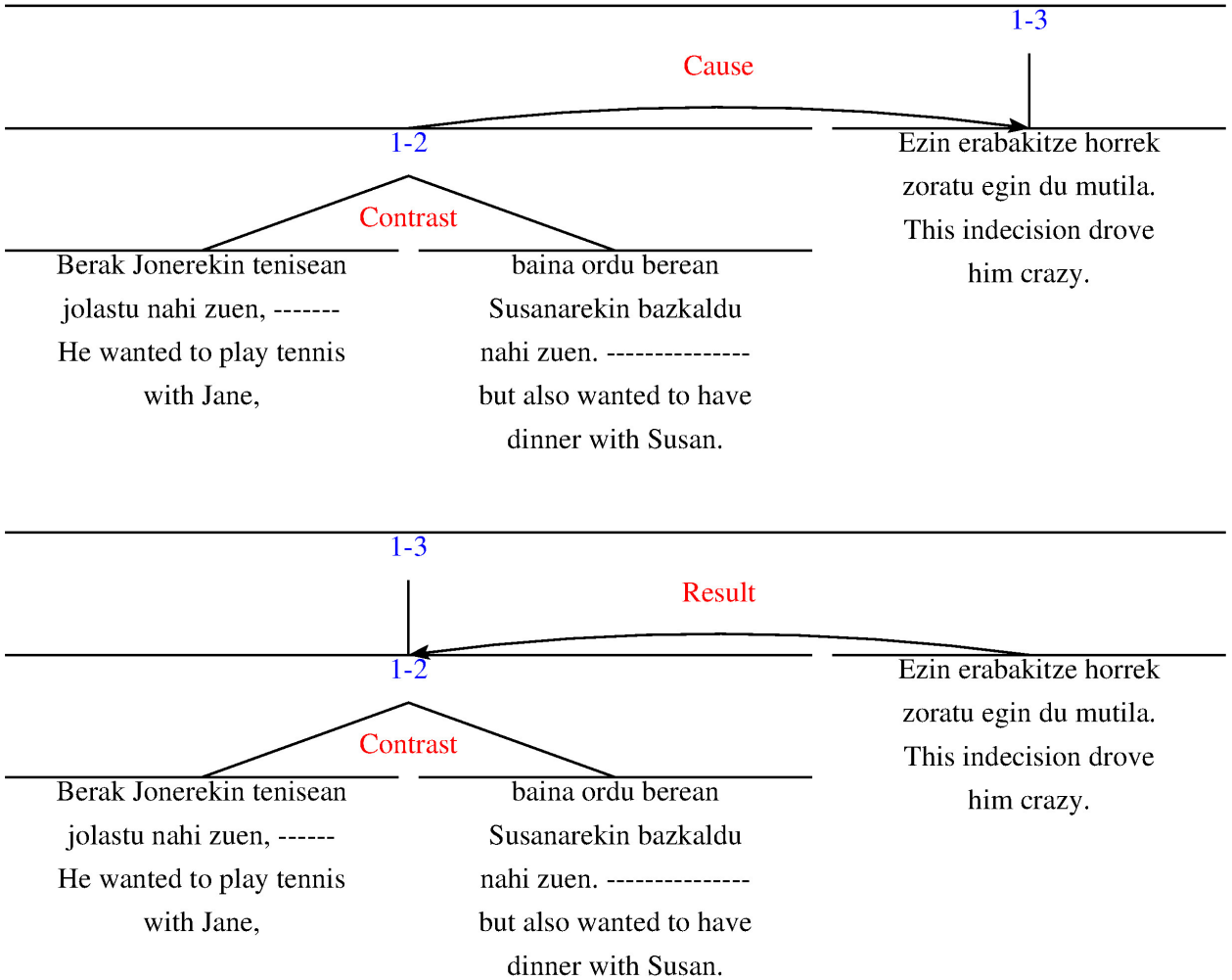
– S-N relations (nuclear) are represented with arrows in RST



– What is the difference between these two structures?



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

Outline

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/~taspardo/RhetDB_Install.zip
   - First get the appropriate format with RSTToolkit: http://www.icmc.usp.br/~taspardo/RSTToolkit_Install.rar

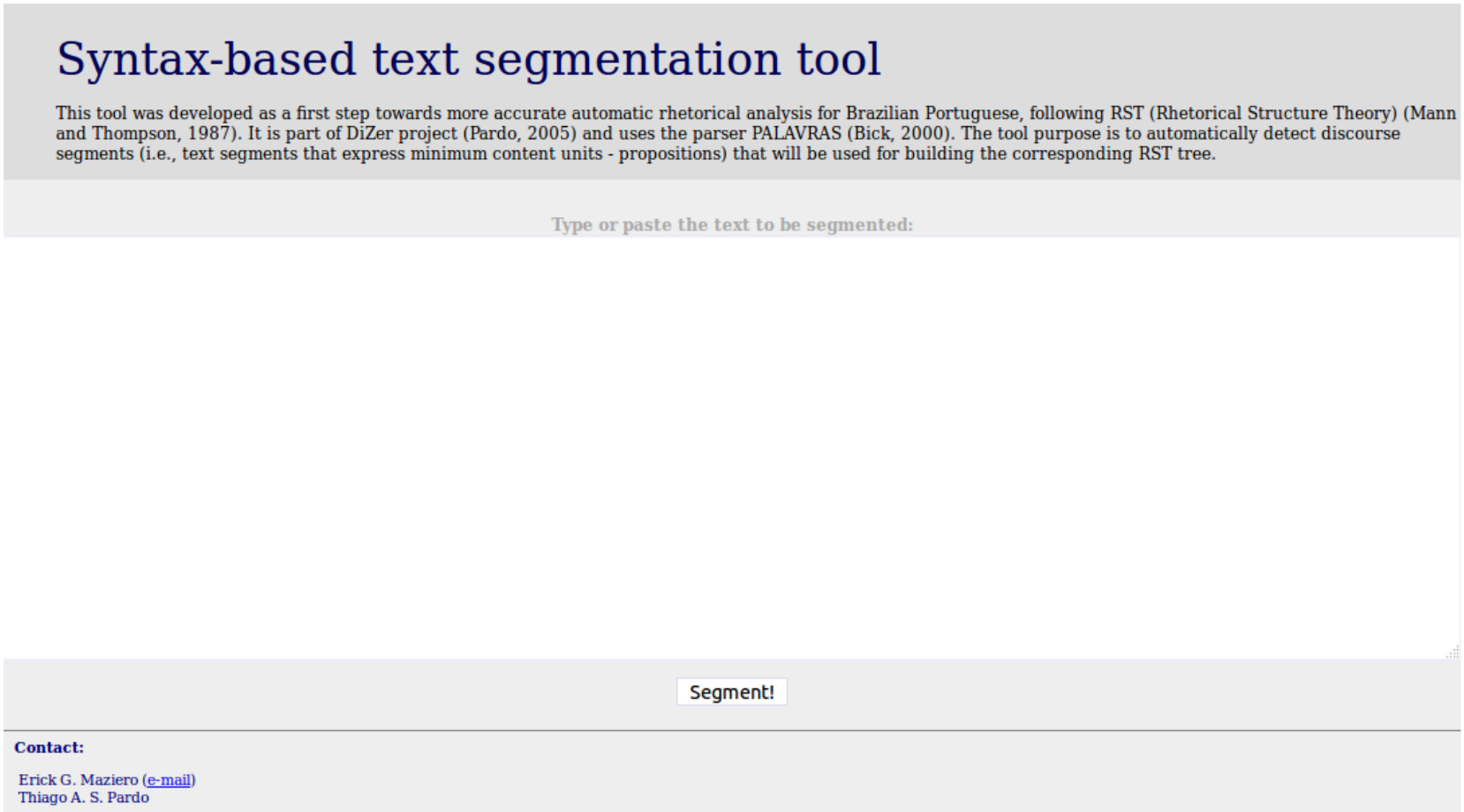
Go to corpus exploration: 61

Outline

Outline

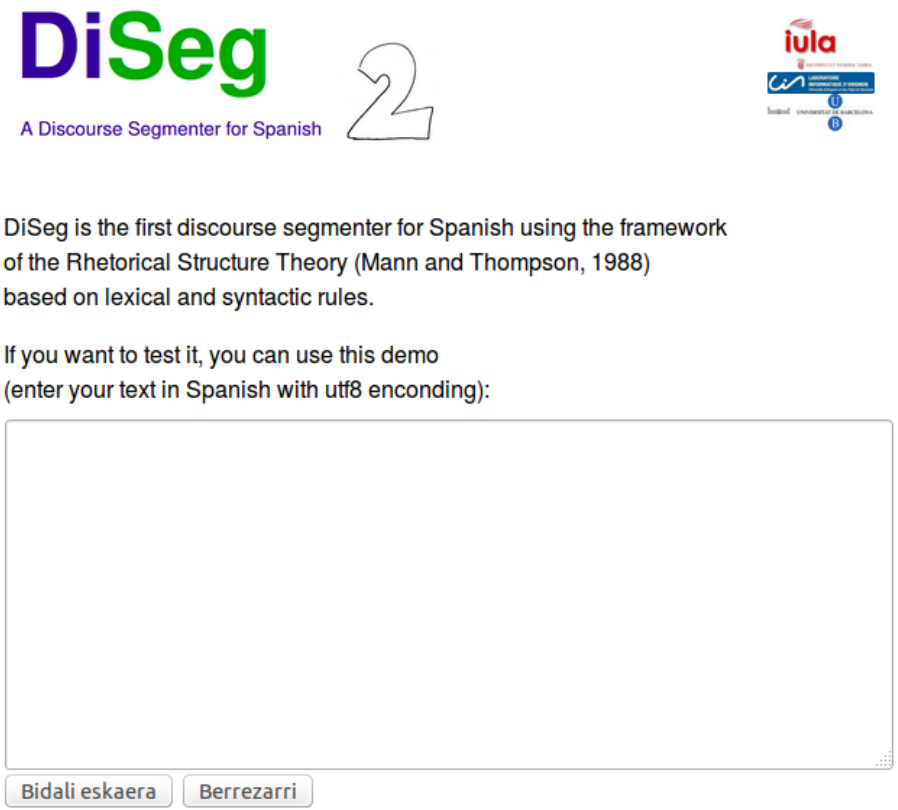
SENTER: Brazilian Portuguese segmenter

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



DiSeg: Spanish RST segmenter

– DiSeg is an intrasentential rule based segmenter
  - Rules based on lexical and syntactic rules
  - http://diseg2.termwatch.es/



EusEduSeg: Basque RST Segmenter (ongoing)

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 Ilarraza 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 aplications.

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

format: rsttool

text:

send

References:
Mann, W.C. Thompson, S.A. 1987. Rhetorical Structure Theory: A Theory of Text Organization. Text 8.243-281.
Diaz de Ilarraza, A. Gojenola, K. Oronoz, M. 2005. Design and Development of a System for the Detection of Agreement Errors in Basque. In Computational Linguistics and Intelligent Text Processing, 793-802. Springer.
Iruskieta, M. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean. Doktore-tesia. EHU. Informatika Fakultatea.

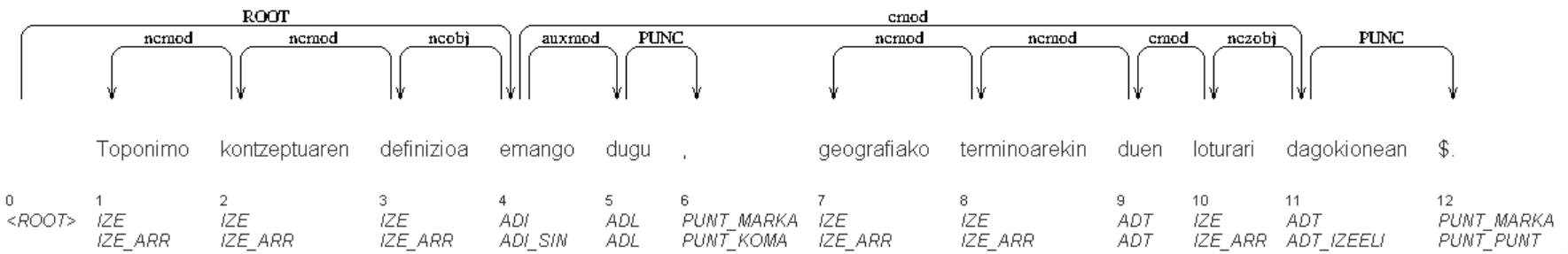
EusEduSeg: System background

– Our segmenter is based on MALTI-XA an automatic dependency analyzer (Diaz de Ilarraza et al., 2005)

(14) a. [Toponimo kontzeptuaren definizioa emango dugu,]$_1$ [geografiako terminoarekin duen loturari dagokionean.]$_2$ TERM51

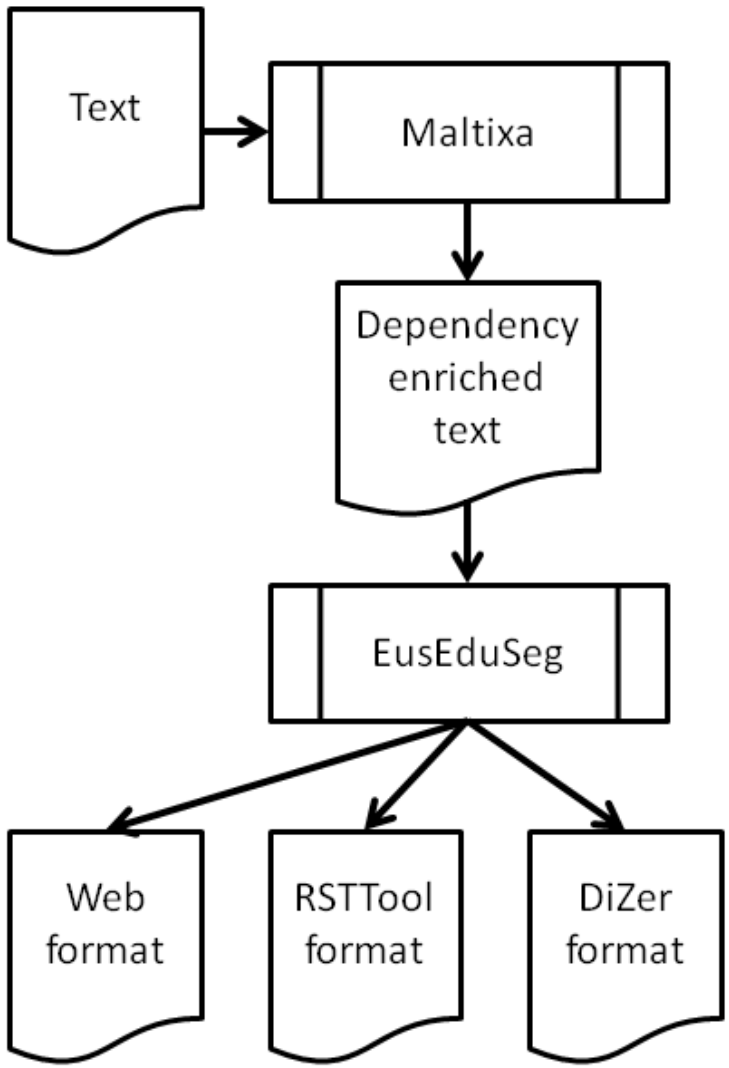
b. [We present the definition of a toponym,]$_1$ [regarding to geographical terms.]$_2$

Token	Word	Head	Rel.	Lemma
1	toponimo	2	ncmod	toponimo
2	kontzeptuaren	3	ncmod	kontzeptu
3	definizioa	4	ncobj	definizio
4	emango	0	ROOT	eman
5	dugu	4	auxmod	*edun
6	,	5	PUNC	,
7	geografiako	8	ncmod	geografia
8	terminoarekin	9	ncmod	termino
9	duen	10	cmod	ukan
10	loturari	11	nczobj	lotura
11	dagokionean	4	cmod	egon
12	.	11	PUNC	.



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



Outline

Central Unit detector for Basque and B. Portuguese

– Ongoing projects

J.D. Antonio G. Labaka


K. Bengoetxea

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

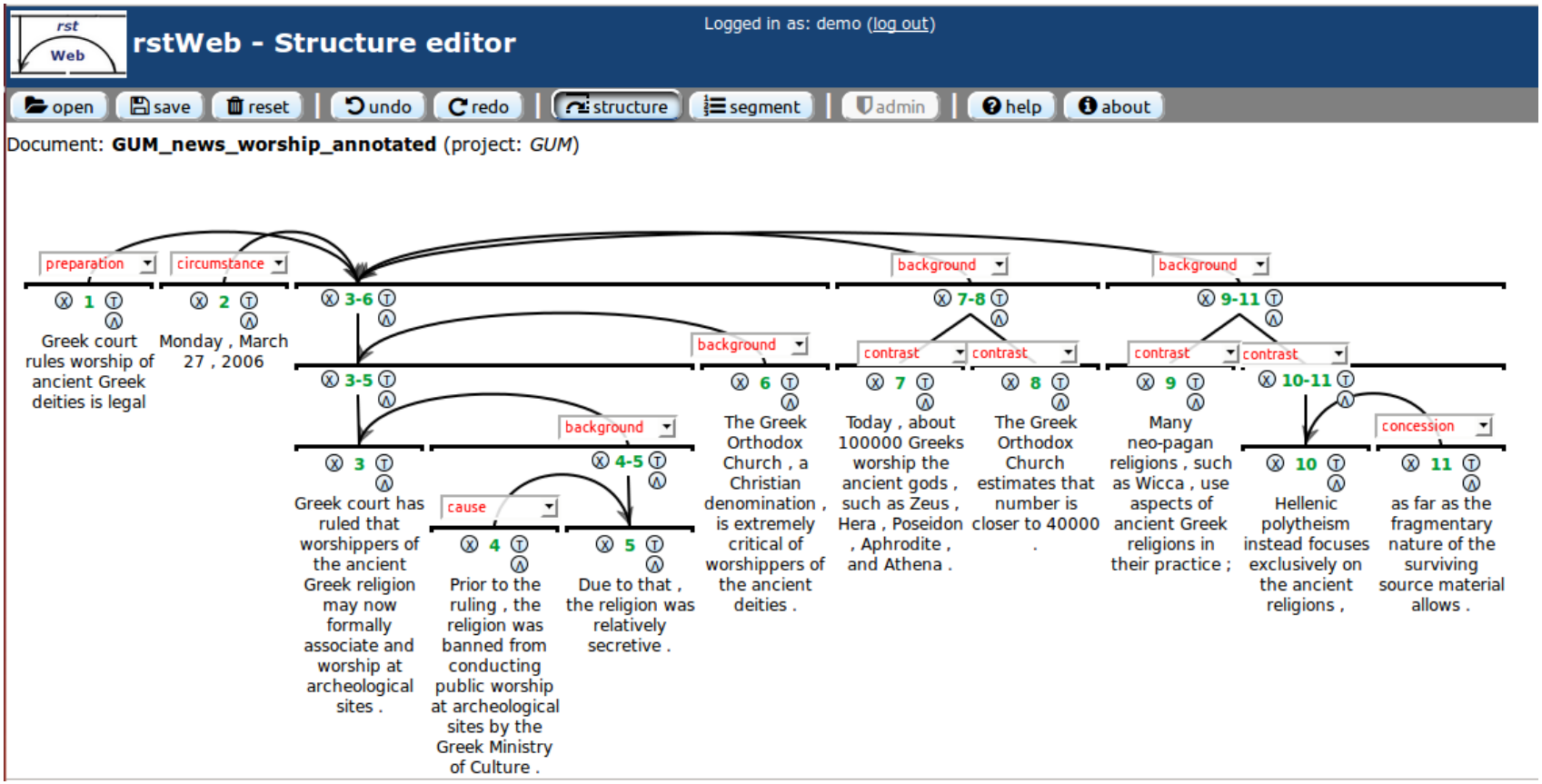
– Heuristics based on nouns, verbs, pronouns, bonus words, title words, EDU position in the document, main verb
  - Results for Basque: F_1 of 0.51
  - Results for B. Portuguese: F_1 of 0.55

– Machine learning techniques
  - Results for Basque: F_1 of 0.51 (ongoing)

Outline

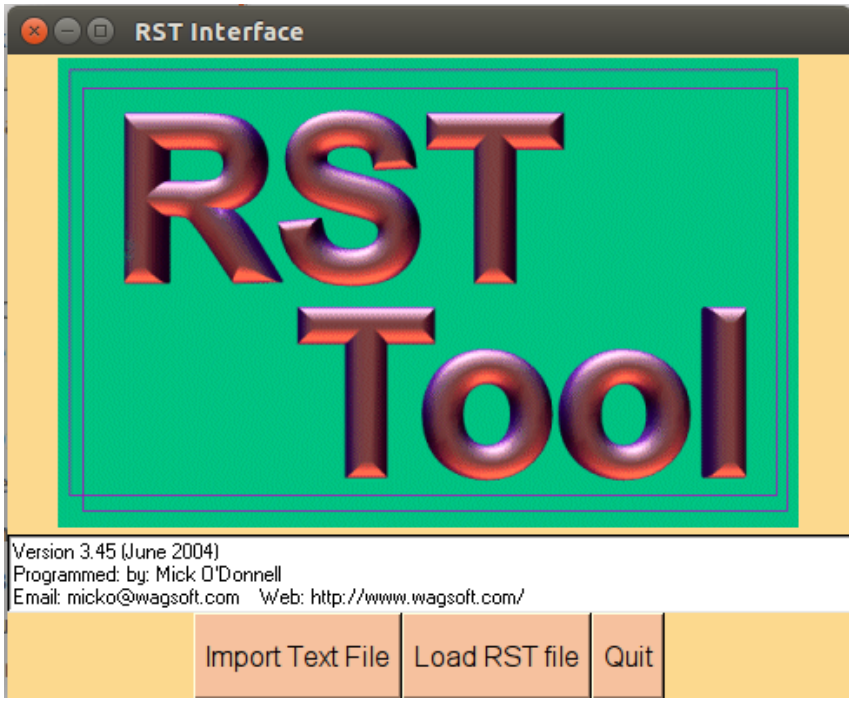
rstWeb Tool (a collaborative RS-tree)

– Collaborative web annotation at
https://corpling.uis.georgetown.edu/rstweb/info/



RSTTool (to structure trees)

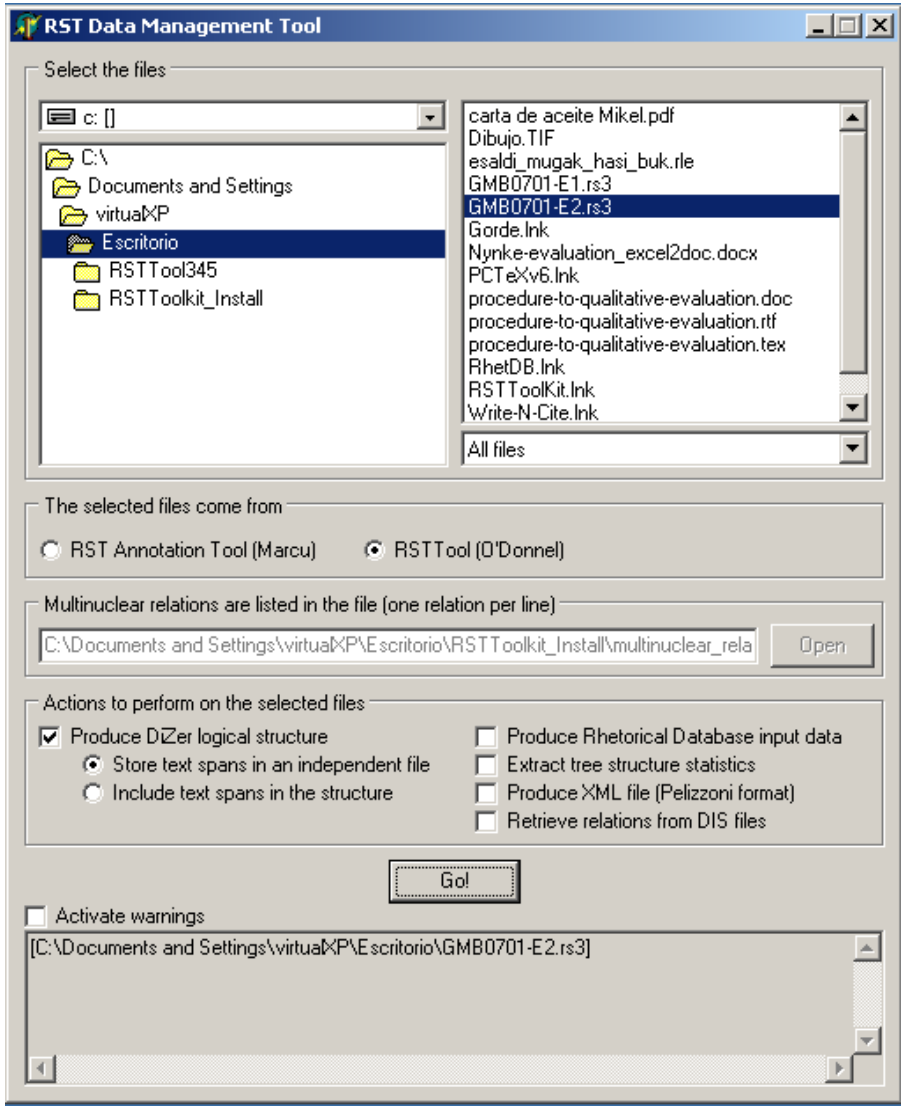
- Manual segmentation and rhetorical annotation
- http://www.wagsoft.com/RSTTool/



- 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

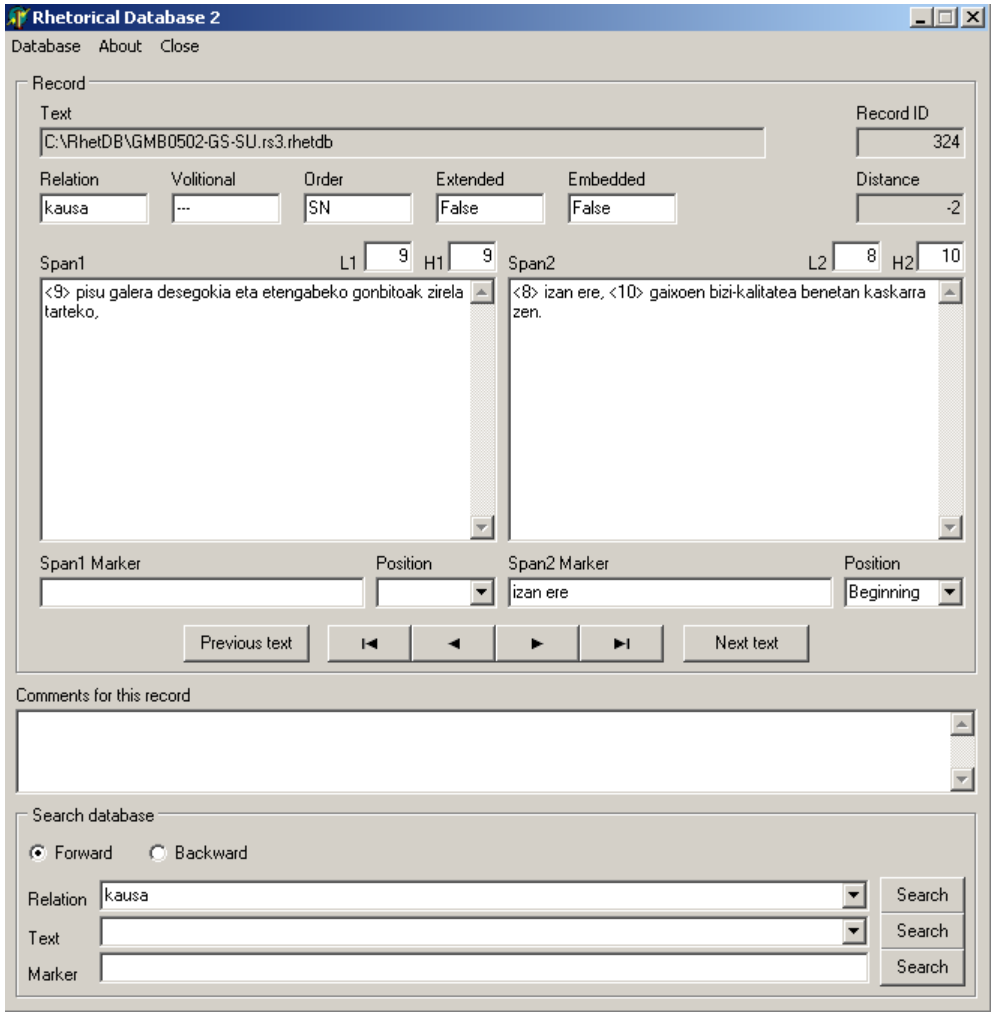
RSTToolkit

- To change the format for Rhetorical Database
- To extracts tree structure statistics
- http://www.icmc.usp.br/~taspardo/RSTToolkit_Install.rar



Rhetorical DataBase (to signal)

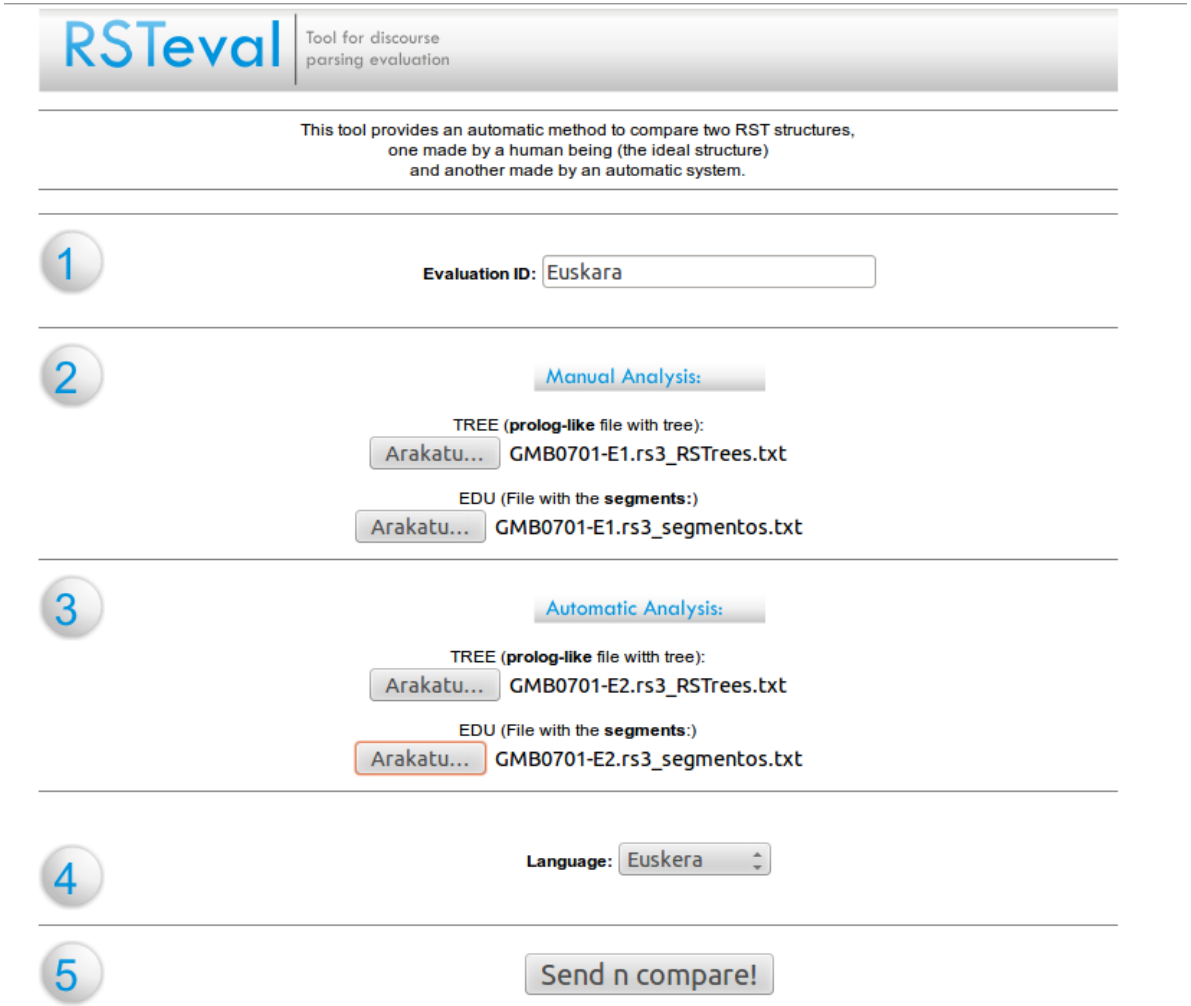
- To annotate signals
  - Relation by relation
  - Check consistency
- Extract statistics of the signals
- http://www.icmc.usp.br/~taspardo/RhetDB_Install.zip



Outline

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)

RSTeval | Tool for discourse parsing evaluation

This tool provides an automatic method to compare two RST structures,
one made by a human being (the ideal structure)
and another made by an automatic system.

Evaluation ID: Euskara

Text	Units			Span			Nuclearity			Relation		
ID	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision
GMB07	10 of 10	1	1	17 of 19	0.894736842105263	0.894736842105263	16 of 19	0.842105263157895	0.842105263157895	16 of 19	0.842105263157895	0.842105263157895

Evaluation Table

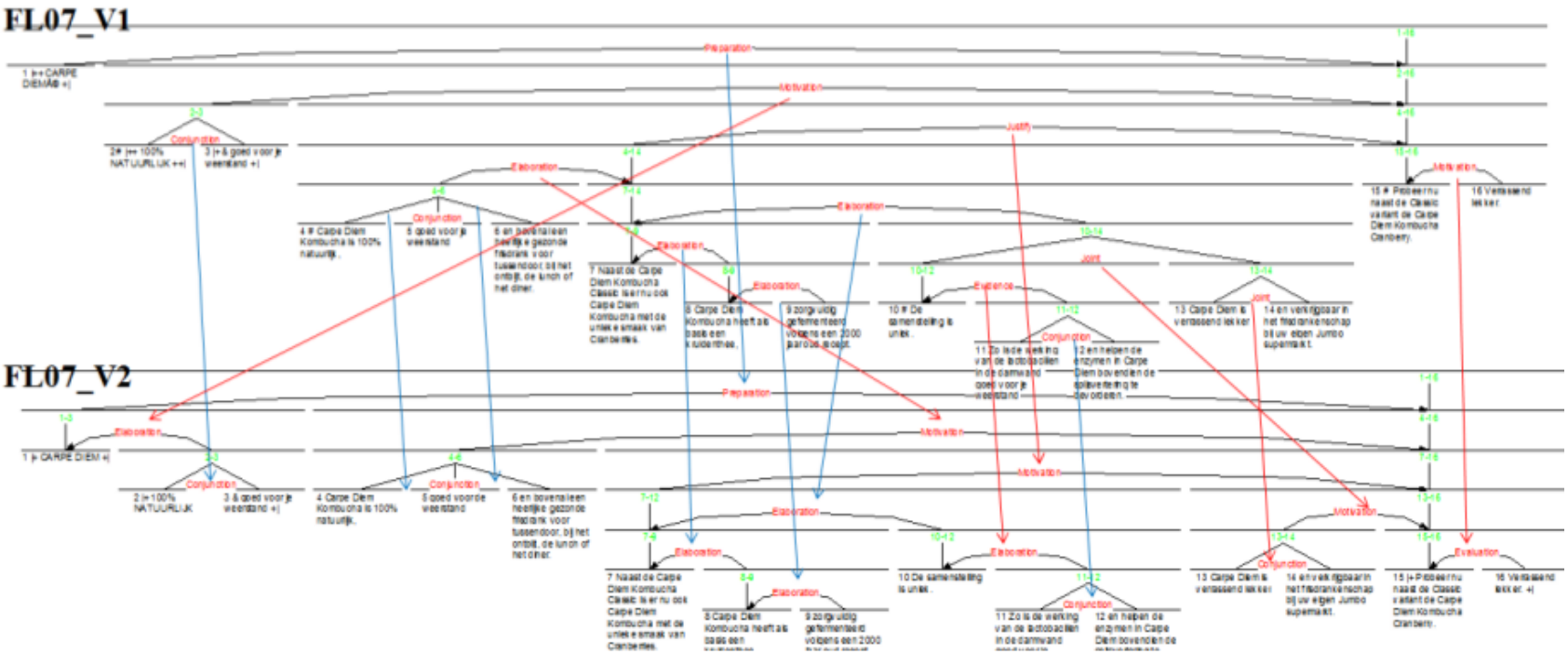
Constituent	Units		Spans		Nuclearity		Relations	
	Manual	Auto	Manual	Auto	Manual	Auto	Manual	Auto
1 to 4 (Larritasunezko_irizpide...onkologian)	x	x	x	x	s	s	prestatzea	prestatzea
5 to 15 (Ikerketa_Pierre...aztertu)	x	x	x	x	n	n	span	span
16 to 22 (Basurtoko_Ospitaleko...gaixok)	x	x	x	x	n	n	span	span
23 to 31 (Pierre_Martyren...asmoz)	x	x	x	x	s	s	helburua	helburua
32 to 35 (elkarrizketa_zitzaien...guztiei)	x	x	x	x	n	n	span	span
!23 to 35 (Pierre_Martyren...guztiei)			x	x	s	n	elaborazioa	span
36 to 38 (7_itemak...aztertuta)	x	x	x	x	s	s	metodoa	metodoa
39 to 50 (estatistikoki_desberdintasun...05)	x	x	x	x	n	n	span	span
!36 to 50 (7_itemak...05)			x	x	n	n	lista	lista
51 to 57 (Horrez_item...bereizten)	x	x	x	x	n	n	lista	lista
58 to 60 (horiei_balorazio...orokorra)	x	x	x	x	n	n	lista	lista
!51 to 60 (Horrez_item...orokorra)			x	x	n	n	lista	lista
61 to 65 (prozesuaren_igurkapenen...dizkigute)	x	x	x	x	n	n	lista	lista
!51 to 65 (Horrez_item...dizkigute)			x	x	n	n	lista	lista
!36 to 65 (7_itemak...dizkigute)			x	x	s	s	ondorioa	ondorioa
!23 to 65 (Pierre_Martyren...dizkigute)				x		s		elaborazioa
!16 to 65 (Basurtoko_Ospitaleko...dizkigute)				x		s		metodoa
!5 to 65 (Ikerketa_Pierre...dizkigute)			x	x	n	n	span	span
!1 to 65 (Larritasunezko_irizpide...dizkigute)			x	x	r	r	span	span
!16 to 35 (Basurtoko_Ospitaleko...guztiei)			x		s		metodoa	
!5 to 35 (Ikerketa_Pierre...guztiei)			x		n		span	

Qualitative evaluation method (Iruskieta et al., 2015a)

M. Taboada

I. da Cunha

- 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

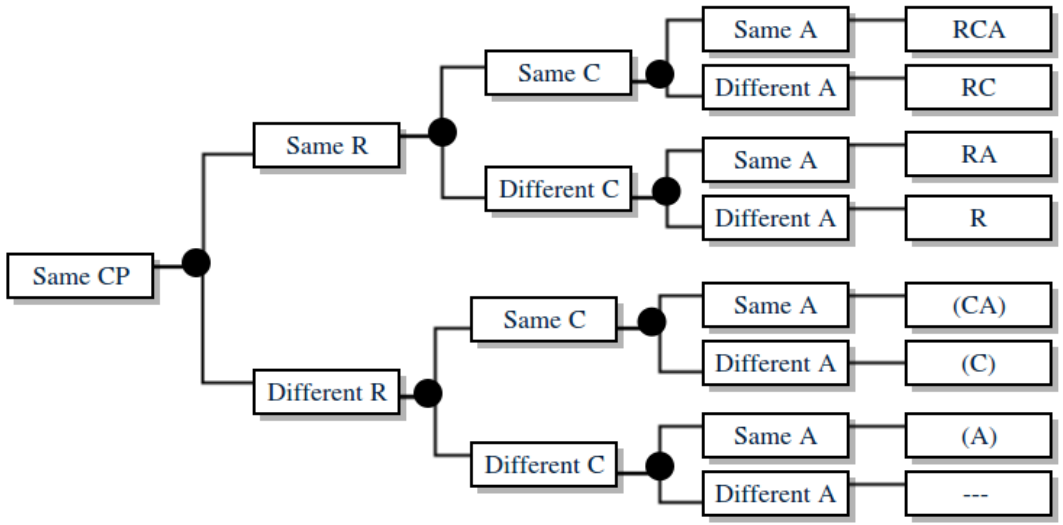


Our evaluation method

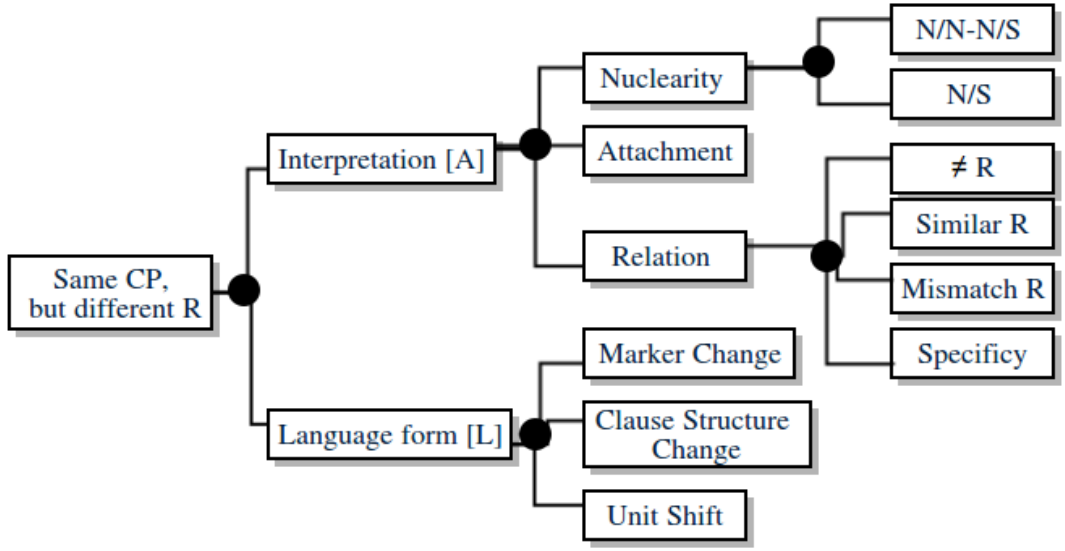
- Quantitative RS-tree evaluation method (Marcu, 2000a) by means of EDUs, spans, nuclearity and RRs
  - Automatic comparison (Maziero and Pardo, 2009)
  - Not independent factors (NS and RRs) (van der Vliet, 2010a)
  - RRs are not (well) compared (Iruskieta et al., 2013b)
- A more accurate comparison
  - 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	ll	m	n	ñ	o	p	q	r	s	t	u	v	w	x	y	z	
ENABLEMENT	a										1														2					3
ANTITHESIS	b		1								1				1					1									1	5
SOLUTIONHOOD	c														1										3			9		13
CONDITION	d				14						2											1	3						3	23
JOINT	e																													
RESTATEMENT	f						4				2				1						1									8
DISJUNCTION	g							1											1											2
EVALUATION	h								1						3		1								2			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																			1
NO-EDU	l																1					1								2
PURPOSE	ll										10			88		1	1	1				1	2	1	1			2		108
INTERPRETATION	m										4				9								1		10			1		25
JUSTIFY	n										1				2		11						1	1	2					18
CAUSE	ñ					1					4						24								8					37
CONJUNCTION	o					2					3							27	1			14			5		3	1	1	57
CONTRAST	p					2					5							1	12	5		5			2		1	2		35
CONCESSION	q		3		1					1	3								2	26					1			1		38
SUMMARY	r										3										1							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				1												3
RESULT	v								1		12				3		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		4	1					4						41	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

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

RRs	Kappa	p.value
JUSTIFY	-0.008	0.760
JOINT	-0.007	0.803
SOLUTIONHOOD	-0.005	0.857
MOTIVATION	-0.003	0.923
ENABLEMENT	-0.001	0.967
UNCONDITIONAL	0.001	0.989

- 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

RRs		#	Total
ELABORATION	BACKGROUND	50	
MEANS	ELABORATION	30	
LIST	CONJUNCTION	29	183
ELABORATION	RESULT	27	
ELABORATION	LIST	26	
ELABORATION	CONJUNCTION	21	
INTERPRETATION	RESULT	13	
PREPARATION	ELABORATION	12	
PURPOSE	ELABORATION	12	69
JUSTIFY	CAUSE	11	
SEQUENCE	LIST	11	
MEANS	BACKGROUND	10	
SOLUTIONHOOD	BACKGROUND	9	
ELABORATION	INTERPRETATION	9	
ELABORATION	JOINT	8	
CONJUNCTION	RESULT	8	60
CAUSE	RESULT	7	
CONTRAST	CONCESSION	7	
CONTRAST	LIST	7	
CONTRAST	ELABORATION	5	
Total			**312**

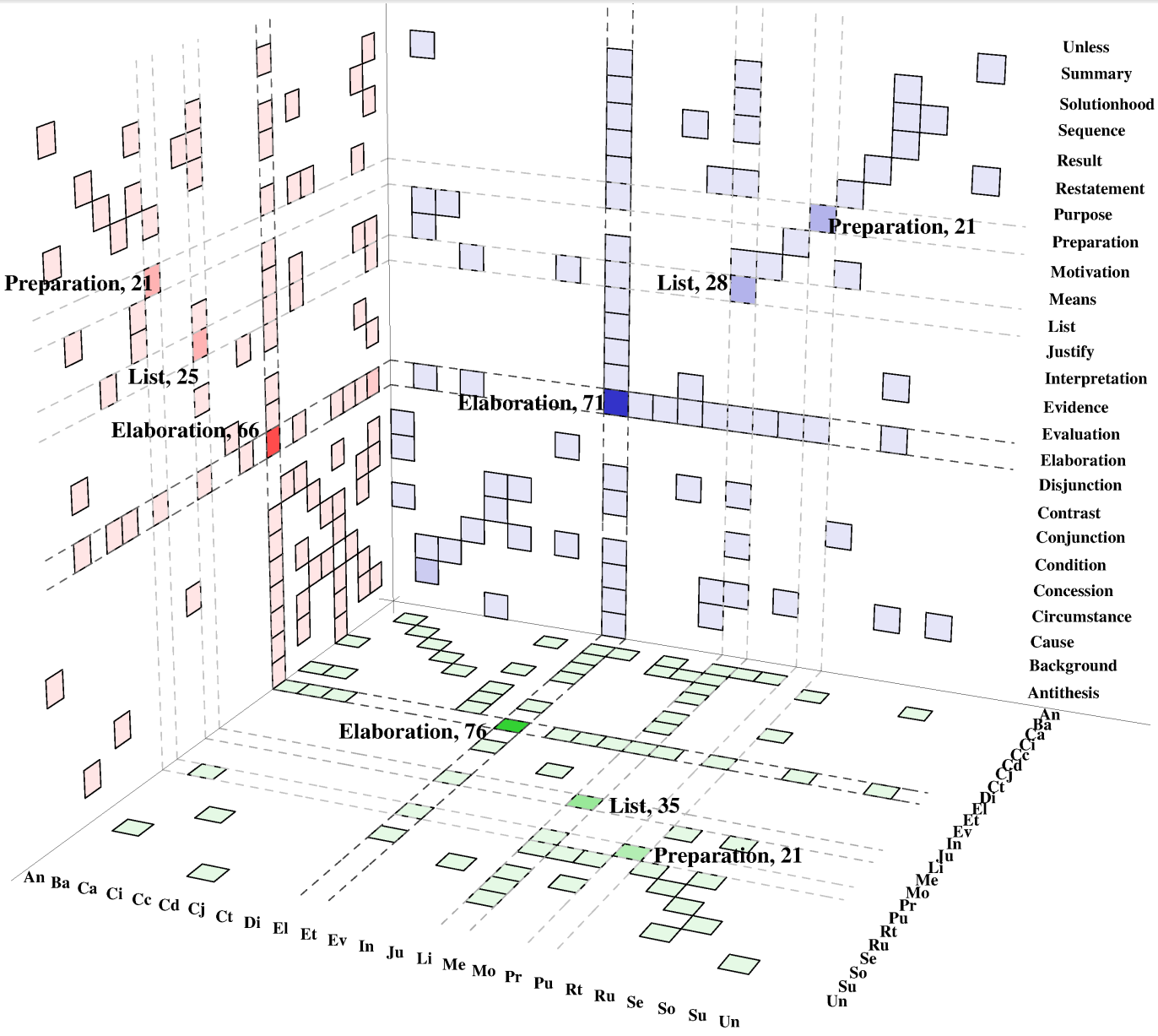
- One of them is the most widely used RR: 47.21% (ELABORATION-*X*)
- Similar RRs: 4.1% (LIST-CONJUNCTION, JUSTIFY-CAUSE, INTERPRETATION-RESULT)
  - Different nuclearity: 0.54% (CAUSE-RESULT)
- Not used by one of the annotators: 0.7% (SOLUTIONHOOD-BACKGROUND)

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*)

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

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

	Translation Strategies						Different Language Forms		
	ENG>SPA	ENG>BSQ	SPA>ENG	SPA>BSQ	BSQ>ENG	BSQ>SPA	ENG-SPA	ENG-BSQ	SPA-BSQ
MC	1.45%	–	4.35%	7.25%	10.14%	11.59%	14.49%	4.35%	1.45%
CSC	1.45%	1.45%	2.90%	4.35%	4.35%	1.45%	2.90%	1.45%	–
US	2.90%	2.90%	2.90%	1.45%	4.35%	2.90%	0.00%	4.35%	2.90%
Total			**68.12%**					**31.88%**	

Exclusion of a marker (translation strategy)

(15) a. [Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados;]$_{9N}$ [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.]$_{10S-EVIDENCE}$

b. [Furthermore, terms can be compiled, discussed and assessed anywhere:]$_{9N}$ [∅ many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them.]$_{10S-ELABORATION}$

c. [Are gehiago, edozein tokitatik biltzen dira terminoak, baita komentatu eta haztatu ere;]$_{9N}$ [∅ adibidez, Interneti buruzko terminoen glosarioak zabaltzen dira Web askotan, eta izendegietarako proposamenak egin ere bai, eta erabiltzaileek botoa eman ahal izaten diete.]$_{10S-ELABORATION}$
TERM38_SPA

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,$]_{6N}$ {que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos $(\ldots)]_{7-11S-ELABORATION}$

b. [All these factors lead to an increase in the number of specialist terms which enrich terminology$]_{6N-CONTRAST}$ [but also call into question some of its basic concepts $(\ldots)]_{7N-CONTRAST}$

c. [Alderdi horiek guztiek, espezialitateko terminologiaren gehikuntza kuantitatiboa eragiteaz gain, terminologia lanen ikuspegia ere zabaldu egin dute;$]_{6N-LIST}$ [eta, egia bada ere ikuspegi berri horrek terminologia aberastu egin duela esatea, zalantzan jarri ditu terminologiaren oinarrizko zenbait kontzeptu $(\ldots)]_{7N-LIST}$ TERM19_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.]$_{N12}$ [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?

Outline

RST parsers

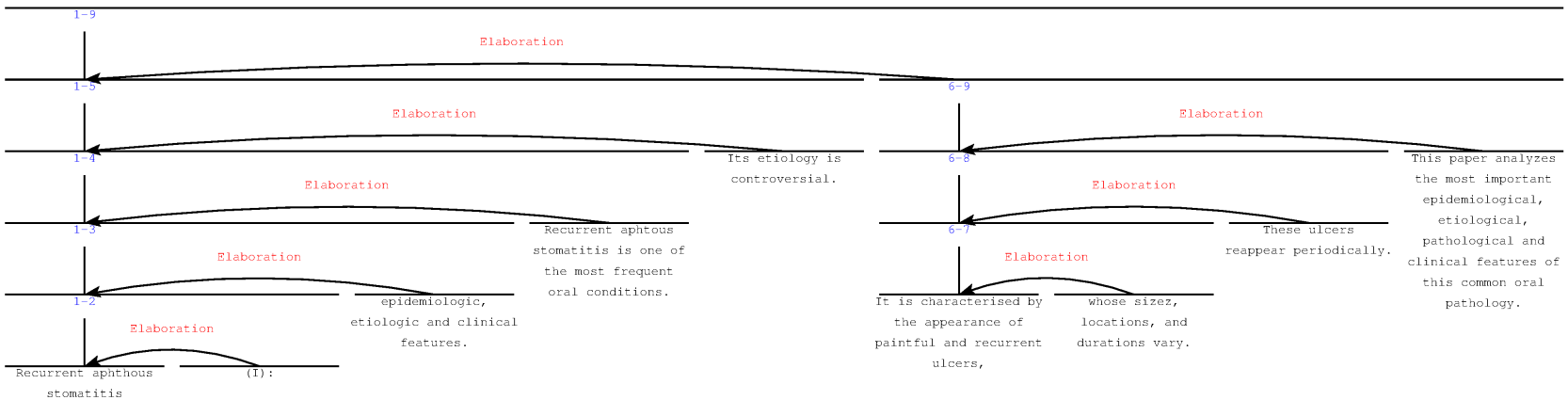
– RST parsers
  - CODRA parser (Joty et al., 2015)
  - A Linear-Time Bottom-Up Discourse Parser (Feng and Hirst, 2014)
  - DIZER parser (Pardo and Nunes, 2006)

CODRA parser (Joty et al., 2015)

– Input text

(18) **Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.**
Recurrent aphtous stomatitis is one of the most frequent oral conditions. Its etiology is controversial. It is characterised by the appearance of paintful and recurrent ulcers, whose sizez, 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.

– Output of the CODRA parser a la RST



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

Outline

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

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

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

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

Outline

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

Workshops and Web Site

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

  The RST Web Site:
  http://www.sfu.ca/rst/index.html

Publications and Projects

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

- 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_Valencia.pdf

Thanks

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

References I

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J., Artola, X., Diaz de Ilarraza, A., and Ezeiza, N. (1998). A framework for the automatic processing of basque. In *First International Conference on Language Resources and Evaluation*, Granada, Spain.

Aduriz, I., Aldezabal, I., Alegria, I., Arriola, J., Diaz de Ilarraza, A., Ezeiza, N., and Gojenola, K. (2003). Finite state applications for basque. In *EACL 2003 Workshop on Finite-State Methods in Natural Language Processing*, Budapest, Hungary.

Agirrezabal, M., Gonzalez-Dios, I., and Lopez-Gazpio, I. (2015). Euskararen sorkuntza automatikoa: lehen urratsak. In *IkerGazte*.

Aldabe, I. (2011). Automatic exercise generation based on corpora and natural language processing techniques. Unpublished doctoral dissertation, UPV/EHU, Donostia, Basque Country.

Alegria, I., Balza, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named entity recognition and classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información*, pages 1–8, Madrid.

Alkorta, J., Gojenola, K., Iruskieta, M., and Perez, A. (2015). Using relational discourse structure information in Basque sentiment analysis. In *5th Workshop "RST and Discourse Studies", in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Alicante, Espana.

Antonio, J. D. (2012). Expression of cause, evidence, justify and motivation rhetorical relations by causal hypotactic clauses in brazilian portuguese. *Acta Scientiarum: Language & Culture*, 34(2):253–268.

Antonio, J. D. and Cassim, F. T. R. (2012). Coherence relations in academic spoken discourse. *Linguistica*, 52:323–336.

Antonio, J. D. and Iruskieta, M. (2014). *A RST e suas aplicaçoes na linguistica e no processamento de linguas naturais*, pages 1–32. Estudos de descriçao sociofuncionalista: objetos e abordagens. Lincom-Europa.

Artiagoitia, X., Oyharçabal, B., Hualde, J. I., and de Urbina, J. O. (2003). *Subordination*, pages 632–844. A grammar of Basque. Mounton de Gruyter, Berlin-New York.

References II

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge Univ Pr, Cambridge.

Barrutieta, G., Abaitua, J., and Díaz, J. (2001). Grossgrained RST through XML metadata for multilingual document generation. In *MT Summit VIII*, pages 39–42, Santiago de Compostela, Spain.

Barrutieta, G., Abaitua, J., and Díaz, J. (2002). An XML/RST-based approach to multilingual document generation for the web. *Procesamiento del lenguaje natural*, 29:247–253.

Bengoetxea, K. and Gojenola, K. (2007). Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera. *Procesamiento del lenguaje natural*, 39:5–12.

Bosma, W. E. (2005). Query-based summarization using Rhetorical Structure Theory. In *15th Meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, pages 29–44, Amsterdam. LOT.

Bosma, W. E. (2008). Discourse oriented summarization. Doktore-tesia, University of Twente.

Bouayad-Agha, N. (2000). Using an abstract rhetorical representation to generate a variety of pragmatically congruent texts. In *38th Annual Meeting ACL*, volume 38, pages 16–22, Hong Kong.

Burstein, J. C., Marcu, D., Andreyev, S., and Chodorow, M. S. (2001). Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, pages 98–105. Association for Computational Linguistics.

Burstein, J. C., Marcu, D., and Knight, K. (2003). Finding the write stuff: Automatic identification of discourse structure in student essays. *Ieee Intelligent Systems*, 18(1):32–39.

Cardoso, P. C., Taboada, M., and Pardo, T. A. (2013). Subtopics annotation in a corpus of news texts: steps towards automatic subtopic segmentation. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report.

Carlson, L., Marcu, D., and Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, page 10, Aalborg, Denmark. Association for Computational Linguistics.

References III

Carlson, L., Okurowski, M. E., and Marcu, D. (2002). *RST Discourse Treebank, LDC2002T07 [Corpus]*. PA: Linguistic Data Consortium, Philadelphia.

Ceberio, K., Aduriz, I., Diaz de Ilarraza, A., and Garcıa, I. (2009). Empirical study of the relevance of semantic information for anaphora resolution: the case of adverbial anaphora. In *7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC09)*, pages 56–63, Goa, India.

da Cunha, I. (2013). A symbolic corpus-based approach to detect and solve the ambiguity of discourse markers. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*, Samos, Greece.

da Cunha, I. and Iruskieta, M. (2010). Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5):563–598.

da Cunha, I., SanJuan, E., Torres-Moreno, J.-M., Lloberes, M., and Castellón, I. (2010). Diseg: Un segmentador discursivo automatico para el español. *Procesamiento de Lenguaje Natural*, 45.

da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011). On the Development of the RST Spanish Treebank. In *5th Linguistic Annotation Workshop (LAW V '11)*, pages 1–10, Portland, USA. Association for Computational Linguistics.

Das, D., Taboada, M., and McFetridge, P. (2015). RST Signalling Corpus.

Diaz de Ilarraza, A., Gojenola, K., and Oronoz, M. (2005). Design and Development of a System for the Detection of Agreement Errors in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 793–802. Springer.

Feng, V. W. and Hirst, G. (2014). A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA, June*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Ghorbel, H., Ballim, A., and Coray, G. (2001). Rosetta: Rhetorical and semantic environment for text alignment. In *Corpus Linguistics*, pages 224–233, Lancaster University (UK).

References IV

Goenaga, I., Arregi, O., Ceberio, K., Diaz de Ilarraza, A., and Jimeno, A. (2012). Automatic Coreference Annotation in Basque. In *Eleventh International Workshop on Treebanks and Linguistic Theories*, Portugal.

Gomez, I. (1996). Euskararen zatiketa informazionalaren eredu baterantz. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 30(1):195–218.

Haouam, K. and Marir, F. (2003). SEMIR: Semantic indexing and retrieving web document using Rhetorical Structure Theory. In *4th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 596–604, Hong Kong.

Hernaez, I., Navas, E., Murugarren, J. L., and Etxebarria, B. (2001). Description of the AhoTTS conversion system for the Basque language. In *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, pages 151–154.

Hovy, E. (2010). Annotation: A tutorial. In *48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Ide, N. and Pustejovsky, J. (2010). W @articleRefWorks:1337, author=Juliano D. Antonio and Fernanda T. R. Cassim, year=2012, title=Coherence relations in academic spoken discourse, journal=Linguistica, volume=52, pages=323-336 hat Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. In *2nd Int. Conf. Global Interoperability Lang. Res*, Hong Kong.

Iruskieta, M. (2014). Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). Phd-thesis, Euskal Herriko Unibertsitatea, Donostia. http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena.pdf.

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., and de la Calle, O. L. (2013a). The rst basque treebank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, Brasil.

References V

Iruskieta, M. and da Cunha, I. (2010). Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera. In *XXVIII Congreso Internacional AESLA: Analizar datos > Describir variación*, pages 13–159, Vigo. Servicio de Publicaciones.

Iruskieta, M., da Cunha, I., and Taboada, M. (2015a). A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49:263–309.

Iruskieta, M., de Ilarraza, A. D., Labaka, G., and Lersundi, M. (2015b). The detection of central units in basque scientific abstracts. In *5th Workshop "RST and Discourse Studies"in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural*. SEPLN.

Iruskieta, M., de Ilarraza, A. D., and Lersundi, M. (2014a). The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *COLING*, pages 466–475. Dublin City University and ACL.

Iruskieta, M., de Ilarraza, A. D., and Lersundi, M. (2014b). The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *COLING*, pages 466–475. Dublin City University and ACL.

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2011). Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. *Procesamiento del Lenguaje Natural*, 47:144.

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2013b). Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 0(0):1–32.

Iruskieta, M. and Zapirain, B. (2015). EusEduSeg: A Dependency-Based EDU Segmentation for Basque. In *SEPNL*, Alicante.

Joty, S., Carenini, G., and Ng, R. T. (2015). Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics, page*, 41(3):385–435.

References VI

Lopez-Gazpio, I. and Marichalar Anglada, M. (2013). Web application for reading practice. In *IADAT-e2013: Proceedings of the 6th IADAT International Conference on Education*, pages pp–74. IADAT-e2013. ISBN: 978-84-935915-3-3.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical Structure Theory: A Theory of Text Organization. *Text*, 8(3):243–281.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Marcu, D. (2000a). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Marcu, D. (2000b). *The theory and practice of discourse parsing and summarization*. The MIT press, Cambridge.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics.

Maziero, E. G. and Pardo, T. A. S. (2009). Metodologia de avaliação automática de estruturas retóricas. In *7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. L. (2004). Annotating discourse connectives and their arguments. In *HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, USA.

Mitkov, R. (2002). *Anaphora resolution*, volume 134. Longman London.

O'Donnell, M. (1997). Variable-length on-line document generation. In *6th European Workshop on Natural Language Generation, Gerhard-Mercator University, Duisburg, Germany*.

Ono, K., Sumita, K., and Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 344–348. Association for Computational Linguistics.

References VII

Paice, C. D. (1980). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *3rd annual ACM conference on Research and development in information retrieval*, pages 172–191, Cambridge. Butterworth and Co.

Pardo, T. A. S. (2005). Métodos para análise discursiva automática. Master's thesis.

Pardo, T. A. S. and Nunes, M. G. V. (2004). Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em português do brasil [rhetorical relations and its surface markers: an analysis of scientific texts corpus in portuguese of brazil]. Technical Report NILC-TR-04-03.

Pardo, T. A. S. and Nunes, M. G. V. (2006). Review and Evaluation of DiZer–An Automatic Discourse Analyzer for Brazilian Portuguese. In *International Workshop on Computational Procesing of Written and Spoken Portuguese*, pages 180–189. Springer.

Pardo, T. A. S. and Nunes, M. G. V. (2008). On the development and evaluation of a brazilian portuguese discourse parser. *Revista de Informática Teórica e Aplicada*, 15(2):43–64.

Pardo, T. A. S., Nunes, M. G. V., and Rino, L. H. M. (2004). DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. *Advances in Artificial Intelligence–SBIA 2004*, pages 224–234.

Pardo, T. A. S., Rino, L. H. M., and Nunes, M. G. V. (2003). GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, pages 196–196.

Pardo, T. A. S. and Seno, E. R. M. (2005). Rhetalho: um corpus de referência anotado retoricamente. *Anais do V Encontro de Corpora*, pages 24–25.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). The Penn Discourse TreeBank 2.0: Annotation manual. Technical report.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *5th International Workshop on Semantic Evaluation*, pages 1–8, Sweden. Association for Computational Linguistics.

References VIII

Reese, B., Denis, P., Asher, N., Baldridge, J., and Hunter, J. (2007). Reference manual for the analysis and annotation of rhetorical structure (version 1.0). Technical report, Technical Report.< http://comp. ling. utexas. edu/discor/manual. pdf>(Mai 2008).

Ripple, A. M., Mork, J. G., Knecht, L. S., and Humphreys, B. L. (2011). A retrospective cohort study of structured abstracts in medline, 1992–2006. *Journal of the Medical Library Association: JMLA*, 99(2):160.

Salaburu, P. (2012). Menderakuntza eta menderagailuak (Sareko Euskal Gramatika: SEG). http://www.ehu.es/seg/morf/5/2/2/2.

Soraluze, A., Arregi, O., and eta Arantza Díaz de Ilarraza, X. A. (2015). Korreferentzia-ebazpena euskaraz idatzitako testuetan. In *IkerGazte*.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156. Association for Computational Linguistics.

Stede, M. (2004). The Potsdam Commentary Corpus. In *2004 ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Stede, M. (2008). *RST revisited: Disentangling nuclearity*, pages 33–57. 'Subordination' versus 'coordination' in sentence and text. John Benjamins, Amsterdam and Philadelphia.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press, Cambridge, UK.

Taboada, M. and Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*, 4(2):249–281.

Taboada, M. and Mann, W. C. (2006). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Taboada, M. and Renkema, J. (2011). Discourse relations reference corpus. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

References IX

Tofiloski, M., Brooke, J., and Taboada, M. (2009). A syntactic and lexical-based discourse segmenter. In *47th Annual Meeting of the Association for Computational Linguistics*, pages 77–80, Suntec, Singapore. ACL.

van der Vliet, N. (2010a). Inter annotator agreement in discourse analysis. http://www.let.rug.nl/ñerbonne/teach/rema-stats-meth-seminar/.

van der Vliet, N. (2010b). Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*, pages 203–210, Ljubljana, Slovenia.

van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated Dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171.

van Dijk, T. A. (1980a). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. L. Erlbaum Associates Hillsdale, NJ.

van Dijk, T. A. (1980b). The semantics and pragmatics of functional coherence in discourse. *Speech act theory: Ten years later, Versus*, 26(27):49–65.

van Dijk, T. A. (1983). *La ciencia del texto: un enfoque interdisciplinario*. Paidos, Barcelona.

Zipitria, I., Arruarte, A., and Elorriaga, J. (2013). Discourse measures for basque summary grading. *Interactive Learning Environments*, 21(6):528–547.

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 18^{th}-22^{nd}, 2016
Structuring Discourse in Multilingual Europe

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