Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean

A description of pragmatics rhetorical structure and its evaluation in computational linguistics

Candidate: Mikel Iruskieta Quintian

Advisors: Arantza Díaz de Ilarraza Sánchez
Mikel Lersundi Ayestaran

University of the Basque Country (UPV/EHU)

February 26, 2014

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

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)
  - Relational: rhetorical annotation (Asher and Lascarides, 2003; Mann and Thompson, 1988) in Basque (Barrutieta et al., 2002, 2001) and in IXA group (Iruskieta et al., 2013b, 2011b)
- Can we explain discourse structure with only explicit and semantic relations? Examples from van Dijk (1980b)

(1) Tiketa erosi dut eta nire aulkira joan naiz.
I bought a ticket and went to my seat. (Macro-structure)

(2) #Peter zinemara joan zen. Berak begi urdinak ditu.
#Peter went to the cinema. He has blue eyes. (Improvable)

(3) John gaixorik dago. Gripea dauka.
John is sick. He has the flu. (Semantic)

(4) Johnek ezin du etorri. Gaixorik dago.
John can't come. He is sick. (~~Semantic,~~ Pragmatic)

Theories of discourse structures in CL (Stede, 2008)

a. Strong formalization based on syntactic or semantic theories
    - Based on sentence level and few analysis of corpus with real texts
        - SDRT (Asher and Lascarides, 2003)
        - D-LTAG (Forbes et al., 2003)
        - LDM (Polanyi, 1988)

b. Real text corpora and analysis of different phenomena
    - Shortcomings in formalization
        - RST (Mann and Thompson, 1987)
        - PDTB (Miltsakaki et al., 2004)

Why an RST TreeBank for Basque?

- General reasons (Taboada and Mann, 2006)
  - Linguistic description
  - Real texts in different languages
    - RST TB, SFU Corpus (Taboada and Renkema, 2011), RST Spanish TB (da Cunha et al., 2011a), Potsdam Corpus (Stede, 2004), TCC (Pardo and Nunes, 2006) and Rhetalho corpus (Pardo and Seno, 2005)
  - Several applications based on RST:
    - automatic text creation (Bouayad-Agha, 2000),
    - automatic text summarization (Marcu, 2000b),
    - machine translation (Ghorbel et al., 2001),
    - assessment of written texts (Burstein et al., 2003),
    - information retrieval (Haouam and Marir, 2003),
- Specific to this thesis:
  - No annotation needed at other linguistic levels
  - Free and available tools for annotation and evaluation
    - (RSTTool, RhetDB, RSTeval)
  - Building (automatically) (Marcu, 2000b) and evaluating RS-trees is easier than graphs

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Main goals

Three main goals:

i) To describe a rhetorical structure of Basque texts by means of corpus annotation

ii) To establish an annotation method

iii) To validate the annotation method and analyze typical cases of annotators' disagreement

Other aims

- Methodological decisions:
  - to analyze influence of the macro-structure in micro-structure
  - to avoid circularity
  - to study a qualitative evaluation
  - to propose some guidelines for the resolution of annotation disagreements
- Gold Standard:
  - in segmentation:
    - for a Basque segmenter
  - in macro-structure:
    - to analyze indicators
  - in rhetorical relations:
    - to signal annotation
- Disseminating the results

Outline

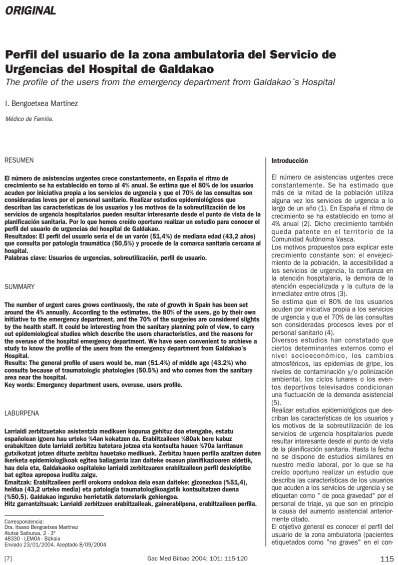
1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Rhetorical structure in Basque

	Explicit RRs	Implicit RRs
Formal	Euskaltzaindia (1990, 1994); Aierbe (2008); Urrutia (2008)	
Discourse	Esnal (2008); Alberdi and Landa (2013); García (2010); Ibarra (2013); Larringan (1995)	

- Little attention to implicit RRs in Basque
  - Implicit RRs are necessary to describe RS and carry out some tasks in CLs
  - Most of the RRs are implicit (Taboada, 2006)
    - 66.67% implicit RRs in GMB0401
- In CLs is very important to describe all the RRs to apply in several applications (main goal of IXA group)

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 españolean 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 - 3°
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).

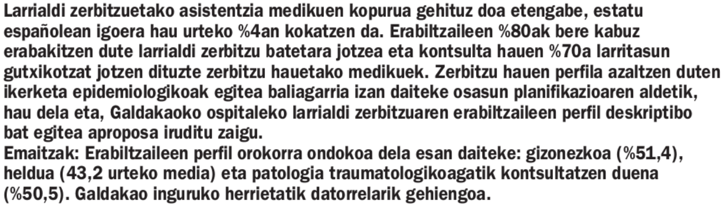
Diversos estudios han constatado que ciertos determinantes externos como el nivel socioeconómico, los cambios atmosféricos, las epidemias de gripe, los niveles de contaminación y/o polinización ambiental, los ciclos lunares o los eventos deportivos televisados condicionan una fluctuación de la demanda 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-

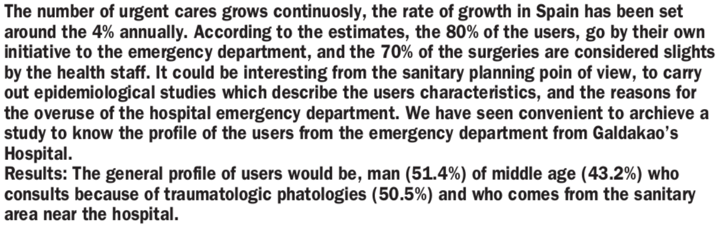
[7] Gac Med Bilbao 2004; 101: 115-120 115

Multilingual texts extraction [GMB0401]



Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe, estatu españolean 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.

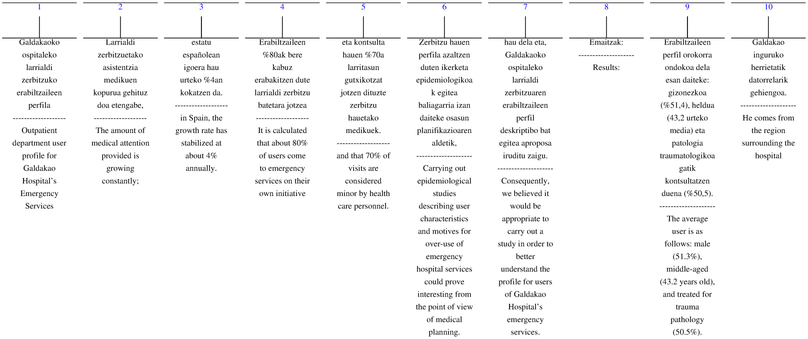
Emaitzak: Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (%50,5). Galdakao inguruko herrietatik datorrelarik gehiengoa.



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.

Segmentation of discourse units (EDUs) [GMB0401]



1
Galdakaoko ospitaleko larrialdi zerbitzuko erabiltzaileen perfila
----
Outpatient department user profile for Galdakao Hospital's Emergency Services

2
Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe,
----
The amount of medical attention provided is growing constantly;

3
estatu españolean igoera hau urteko %4an kokatzen da.
----
in Spain, the growth rate has stabilized at about 4% annually.

4
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

5
eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.
----
and that 70% of visits are considered minor by health care personnel.

6
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.

7
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.

8
Emaitzak:

Results:

9
Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (%50,5).

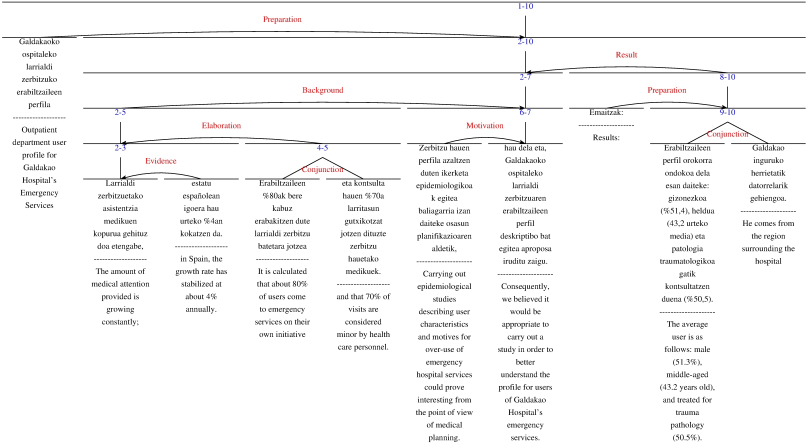
The average user is as follows: male (51.3%), middle-aged (43.2 years old), and treated for trauma pathology (50.5%).

10
Galdakao inguruko herrietatik datorrelarik gehiengoa.

He comes from the region surrounding the hospital

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

Rhetorical structure of a text [GMB0401]



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

- A modular and incremental annotation (Pardo, 2005)
- **Is there any correlation between the CU and the RRs?**

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Outline

1 Introduction
- Aims
- RS-structure in Basque studies and in CLs

2 Methodology
- Preparation phase
- Segmentation
- Central unit
- Rhetorical relations
- Signaling the RRs
- Delivery phase

3 Results
- Segmentation
- Central unit
- Rhetorical relations
- Signaling the RRs

4 Delivery phase

5 Conclusions and future work

Problems and solutions for RS annotation

- Discourse annotation is complex (Hovy, 2010)
  - Solution in CL: corpus annotation
    - Consistent: enough to support machine learning
    - Descriptive: enough to work with NLP advanced applications

The corpus

- The Basque RST TreeBank (Iruskieta et al., 2013a):
  - Short texts, but with complex RS
  - Abstracts: structured texts (Swales, 1990; Ripple et al., 2011)
  - Different domains
  - Parallel texts (Iruskieta et al., 2014a; da Cunha and Iruskieta, 2010; Iruskieta and da Cunha, 2010b)

Domain	Sub-corpus	Texts	Sentences	Words
Medicine	GMB	20	198	3010
Terminology	TERM	20	253	5664
Science	ZTF	20	352	6892
Total		**60**	**803**	**15566**

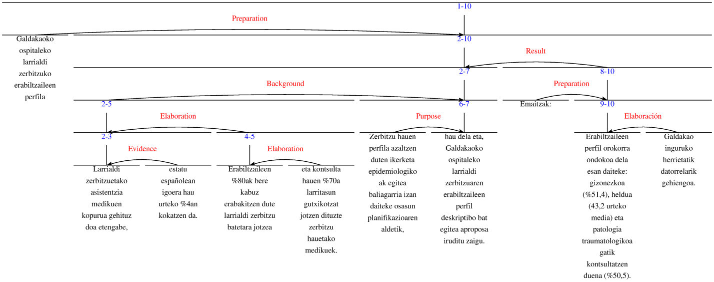
Description of the annotators and the super-annotator

- 4 linguists who had experience annotating texts at other language levels (morphologic, syntactic and semantic)
  - RST and RSTTool were introduced to 3 linguists
  - No previous training phase and no manual provided based on signals
    - To avoid circularity between RS and signaling
    - Because qualitative description was more important than reliability
    - Triple- (80%) and double-annotated (20%) corpus
- **Is there any way to gain reliability if previous training and manuals are avoided?**
  - A "super-annotator" (Hovy, 2010)
    - Experienced in RST
    - Criteria to harmonize annotations were established

Description of the annotators and the super-annotator

- 4 linguists who had experience annotating texts at other language levels (morphologic, syntactic and semantic)
    - RST and RSTTool were introduced to 3 linguists
    - No previous training phase and no manual provided based on signals
        - To avoid circularity between RS and signaling
        - Because qualitative description was more important than reliability
        - Triple- (80%) and double-annotated (20%) corpus
- **Is there any way to gain reliability if previous training and manuals are avoided?**
    - A “super-annotator” (Hovy, 2010)
        - Experienced in RST
        - Criteria to harmonize annotations were established

Different interpretations of GMB0401



1-10

Preparation

Galdakaoko ospitaleko larrialdi zerbitzuko erabiltzaileen perfila

2-10

Result

2-7

8-10

Background

Preparation

2-5

6-7

Emaitzak:

9-10

Elaboration

Purpose

Elaboración

2-3

4-5

Evidence

Elaboration

Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe,

estatu españolean igoera hau urteko %4an kokatzen da.

Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea

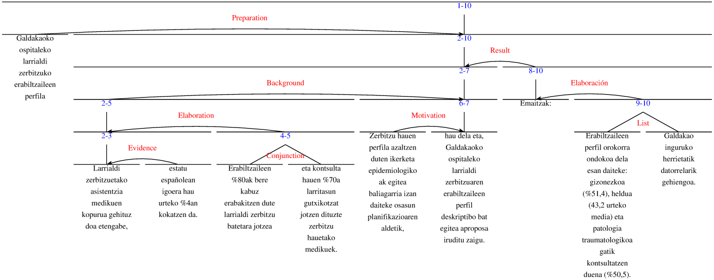
eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.

Zerbitzu hauen perfila azaltzen duten ikerketa epidemiologiko ak egitea baliagarria izan daiteke osasun planifikazioaren aldetik,

hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskriptibo bat egitea aproposa iruditu zaigu.

Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoa gatik kontsultatzen duena (%50,5).

Galdakao inguruko herrietatik datorrelarik gehiengoa.



1-10

Preparation

Galdakaoko ospitaleko larrialdi zerbitzuko erabiltzaileen perfila

2-10

Result

2-7

8-10

Background

Elaboración

2-5

6-7

Emaitzak:

9-10

Elaboration

Motivation

List

2-3

4-5

Evidence

Conjunction

Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe,

estatu españolean 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.

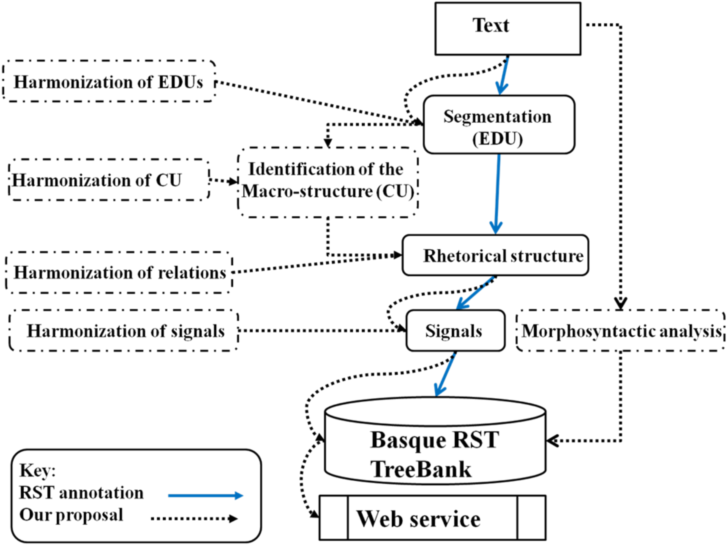
Zerbitzu hauen perfila azaltzen duten ikerketa epidemiologiko ak egitea baliagarria izan daiteke osasun planifikazioaren aldetik,

hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskriptibo bat egitea aproposa iruditu zaigu.

Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoa gatik kontsultatzen duena (%50,5).

Galdakao inguruko herrietatik datorrelarik gehiengoa.

Our annotation method



Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Segmentation guidelines

- Segmentation guidelines conflate RST and Basque clause combining (Tofiloski et al., 2009; Salaburu, 2012; Artiagoitia et al., 2003)

Clause type	EDU	Example
Perpaus independentea 'an independent sentence'	Yes	[*Whipple (EW) gaixotasunak hesteei eragiten die bereziki.*]$_1$ GMB0503
Perpaus nagusi koordinatua 'a main clause, part of sentence'	Yes	[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%).]$_2$
Aditz jokatudun adjuntu perpausa 'finite adjunct clauses'	Yes	[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'	Yes	[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'	Yes	[Dublin Hiriko Unibertsitateko atal bat da Fiontar,]$_1$ [zeinak Ekonomia, Informatika eta Enpresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.]$_2$ TERM23

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Harmonization of CU and its indicators

- Following van Dijk (1980a) texts ought to be coherent at
  - local level: (between words and) between clauses (or RRs)
  - global level: main topic (CU) with other thematic events (RRs)
- But the coherence of CU with other units (or RRs) is not considered in RST
  - in the annotation guidelines (Carlson et al., 2001)
  - in the evaluation method (Marcu, 2000a)
- CU annotation guidelines
  - *i*) Topic or thesis statement
  - *ii*) Purpose
  - *iii*) Method
  - *iv*) Results
  - *v*) Conclusions
- Description of some CU “indicators” (Paice, 1980)
  - Verb clustering with SUMO-category from MCR synset
  - Noun clustering with the WordNet synset

Outline

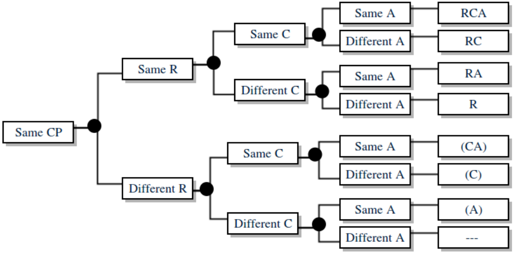
1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Our evaluation method: qualitative by Iruskieta et al. (2014a)

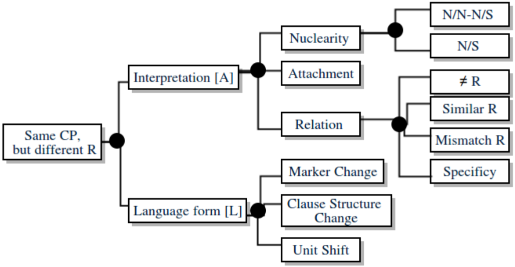
- Quantitative RS-tree evaluation method (Marcu, 2000a) by means of EDUs, spans, nuclearity and RRs
  - Shortcomings
    - Evaluated factors (nuclearity and RRs) are not independent (van der Vliet, 2010)
    - RRs are not (well) compared (Iruskieta et al., 2013b)
  - But well formalized (automated by Maziero and Pardo (2009))
- Appropriate qualitative measurement
  - Independent factors
  - Qualitative description of
    - agreement (RCA, RA, RC and R)
    - disagreement (annotators interpretations and language forms)
- Measurement of RS
  - with the same language: Basque-Basque (Iruskieta et al., 2013a)
  - in parallel texts: Basque-Spanish (da Cunha and Iruskieta, 2010) and Basque-English-Spanish (Iruskieta et al., 2014a)

Our evaluation method: decision trees

- Qualitative agreement



- Qualitative disagreement



RR harmonization guidelines: a proposal

Relation by relation	**Text by text**
Distinguish annotators, search consistency	**Top-down revision (CU and nuclearity)**
But cannot edit the RRs	**First, the RRs linked to CU**
Cannot decide nuclearity	**Then, incremental and modular**

- From confusion matrix(3)
  - Scale of informativeness: ELABORATION 47.21%
- Scale of informativeness is necessary (Kortmann, 1991)
  - Not all the relations needed (Mol, 2005) and some of them were adapted

RR	ANTITHESIS			
most	CONCESSION			
informative	CONTRAST			
↑	CONDITION			
↑	MEANS	ENABLEMENT	PURPOSE	MOTIVATION
↑	CAUSE	RESULT		
↑	SEQUENCE	SOLUTIONHOOD		
↑	JUSTIFY	INTERPRETATION	EVALUATION	EVIDENCE
↑	ELABORATION	BACKGROUND	RESTATEMENT	SUMMARY
↑	LIST	DISJUNCTION		
↑	CONJUNCTION			
RR	CIRCUMSTANCE			
least	PREPARATION			
informative	JOINT			

Outline

1 Introduction
- Aims
- RS-structure in Basque studies and in CLs

2 Methodology
- Preparation phase
- Segmentation
- Central unit
- Rhetorical relations
- Signaling the RRs
- Delivery phase

3 Results
- Segmentation
- Central unit
- Rhetorical relations
- Signaling the RRs

4 Delivery phase

5 Conclusions and future work

Signaling the RRs

- Portuguese (B): 100 texts 4100 RRs (Pardo and Nunes, 2004)
- Spanish: 84 texts 1275 RRs (da Cunha, 2013)
- English: 40 texts 1304 RRs (Taboada and Das, 2013)
- Basque: 60 texts 1292 RRs
  - Annotation tool: Rhetorical Data-Base (Pardo, 2005)
    - Relation by relation
    - Searches can be done to maintain consistency
  - **What is signaling?**
    a) DM annotation
    b) Annotation of the frequent forms (Taboada and Das, 2013)

Types	**Examples**	**Translations**
DMs	hala ere, horrenbestez, izan ere	however, thus, in fact
Morphological	-t(z)eko, -lako, -gatik, bait-	“purpose morpheme”, “cause morphemes”
Lexical	lortu, helburu, emaitza	to achieve, aim, result
Entity	“izen bereziak”	“proper names”
Semantic	handitu-txikitu, ugari-gutxi, irtenbide-arazo, patologia-gaixotasun, legamia-onddo	to increase-to decrease, abundant-few, solution-problem, pathology-disease, leaven-fungus
Syntactic	“Galde-perpausa eta erantzun perpausa”	“question sentence and reply”
Graphical-numerical	1. 2., a) b)	
Double signals	-t(z)eko asmoz, era honetan (...) lortu	with the intention of, to achieve (...) in this way

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

Signaling the RRs

- Portuguese (B): 100 texts 4100 RRs (Pardo and Nunes, 2004)
- Spanish: 84 texts 1275 RRs (da Cunha, 2013)
- English: 40 texts 1304 RRs (Taboada and Das, 2013)
- Basque: 60 texts 1292 RRs
  - Annotation tool: Rhetorical Data-Base (Pardo, 2005)
    - Relation by relation
    - Searches can be done to maintain consistency
  - **What is signaling?**
    - *a)* DM annotation
    - *b)* Annotation of the frequent forms (Taboada and Das, 2013)

Types	**Examples**	**Translations**
DMs	hala ere, horrenbestez, izan ere	however, thus, in fact
Morphological	-t(z)eko, -lako, -gatik, bait-	"purpose morpheme", "cause morphemes"
Lexical	lortu, helburu, emaitza	to achieve, aim, result
Entity	"izen bereziak"	"proper names"
Semantic	handitu-txikitu, ugari-gutxi, irtenbide-arazo, patologia-gaixotasun, legamia-onddo	to increase-to decrease, abundant-few, solution-problem, pathology-disease, leaven-fungus
Syntactic	"Galde-perpausa eta erantzun perpausa"	"question sentence and reply"
Graphical-numerical	1. 2., a) b)	
Double signals	-t(z)eko asmoz, era honetan (...) lortu	with the intention of, to achieve (...) in this way

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

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Delivery phase (Iruskieta et al., 2013a)

- First rhetorical structure annotated corpus in Basque
- The Basque RST TreeBank's delivery phase (Ide and Pustejovsky, 2010)
- Innovations: a number of operations can be carried out with this annotated corpus

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Segmentation results

	Measure	State of the Art	Basque
Manual annotation	Kappa	> 0.8	0.6337
Segmenter	F-score	73% - 85%	57%

- Discourse parsers: EDUs (F_1)
  - Machine learning (French): 73% (Afantenos et al., 2010)
  - DiSeg, rule based (Spanish): 80% (da Cunha et al., 2010a)
- Preliminary results in Basque: end boundaries (F_1)
  - Transformed segmenter: 66.94% (Iruskieta et al., 2011a)
  - Constraint Grammar-based rules: 69.69%
  - Syntactic dependency based heuristics: 80.68%

Granularity and RR agreement

- Less agreement at intra-sentential than at sentential agreement (-13.74%), but more agreement in RRs ($+14.19\%$) and more robust (RCA $+9.5\%$) (Iruskieta et al., 2011b)
  - Parallelism: syntax-discourse (Marcu and Echihabi, 2002)
  - Some RRs 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)

Outline

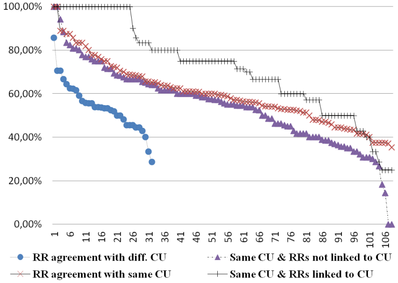
1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

CU annotation results (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%

- CU annotation by 2 non-professionals:
  - Extracted from RS-tree: 65% (GMB)
  - First CU: 85% (TERM and ZTF)

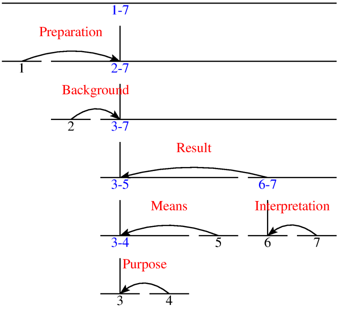
- When CU is the same, bigger agreement in RRs (+5.04%, T-test: 0.013)
- When RR is linked to CU, bigger agreement (+17.29% T-test: 0.001)



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

- Within the RRs linked to the CU, those with an IMRaD structure appear most frequently (unless ELABORATION)

	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**



Verb and noun indicators of the CU and their strength

- Different verb group and indicator's strength in each domain

SUMO	GMB %	TERM %	ZTF %
Reasoning IPP	46.15	22.73	8.70
Comparing IPP	26.92		
Communication SI	3.85	45.45	4.35
Predicate	3.85	4.55	34.78

			Verbs strength		
Lemma	MCR	SUMO	GMB %	TERM %	ZTF %
aztertu	analyze$_1$	Reasoning IPP	58.82	25.00	2.78
aurkeztu	present$_2$	Communication SI	50.00	25.00	
izan, ukan		Predicate	0.47	4.29	2.95

- Some nouns' synsets are good indicator

Indicator	GMB	TERM	ZTF	Total	!	Noun strength	WN 3.0
ikerkuntza$_3$		1	3	19	28	67.86	research$_2$
ikerketa$_2$	1		6				
azterlan$_3$	6	1					
ikerlan$_3$			1				
lan$_3$	2	4	7	13	45	28.89	work$_2$
xede$_1$			2	12	39	30.77	goal$_1$
helburu$_2$		2	8				
komunikazio		10		10	15	66.67	paper$_5$
bide$_2$	2	6	1	9	15	60.00	means$_1$

A list of indicators for Basque (verbs and nouns)

- New indicators from our corpus in gray for an automatic detection of the CU in Basque
    - New indicators (Paice, 1980) in gray
    - Good indicators in blue (≥50.00%)
    - But analysis of other categories are needed to detect the CU

Verbs		Nouns	
BSQ	ENG$_{MCR}$	BSQ	ENG$_{MCR}$
aztertu	examine$_1$	abiapuntu$_1$	starting_point$_1$
analizatu	examine$_1$	arlo$_1$	subject_field$_1$
oinarritu	base$_1$	artikulu$_7$	article$_1$
baloratu	value$_2$	asmo$_2$	purpose$_1$
azaldu	recount$_1$	bide$_2$	means$_1$
aurkeztu	present$_2$	gai$_6$	topic$_1$
aipatu	present$_2$	ikerkuntza$_3$	research$_2$
berri eman	present$_2$	ikerketa$_2$	
jardun	present$_2$	azterlan$_3$	
plazaratu	present$_2$	ikerlan$_3$	
izan / ukan		arazo$_3$	problem$_2$
erabili	use$_1$	irtenbide$_2$	resolution$_4$
ikertu	investigate$_1$	komunikazio	papers$_5$
		hitzaldi$_2$	speech$_1$
		lan$_3$	work$_2$
		lan-ildo	——
		lerro$_{11}$	line$_8$
		ikerketa-lerro	
		proiektu$_2$	project$_2$
		ikerketa-proiektu	
		talde$_1$	group$_1$
		ikerketa-talde	
		xede$_1$	goal$_1$
		helburu$_2$	

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

RR annotation results

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%	

- The Basque RST TreeBank (Iruskieta et al., 2013a)
  - 0,568 κ or 61.47% F_1 (2 annotators, 60 texts: 1470 EDUs)
- The Dutch TreeBank (van der Vliet et al., 2011)
  - 0.57 κ (2 annotators, 4 texts)
- The RST TreeBank (Carlson et al., 2001)
  - from 0.5973 to 0.7921 κ (2 annotators, 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., 2010b)
  - 77.64% F_1 (2 trained annotators: 84 texts, 694 EDUs)

RR confusion matrix

		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

- To go back to RR's harmonization(4)

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	183
MEANS	ELABORATION	30	
LIST	CONJUNCTION	29	
ELABORATION	RESULT	27	
ELABORATION	LIST	26	
ELABORATION	CONJUNCTION	21	
INTERPRETATION	RESULT	13	69
PREPARATION	ELABORATION	12	
PURPOSE	ELABORATION	12	
JUSTIFY	CAUSE	11	
SEQUENCE	LIST	11	
MEANS	BACKGROUND	10	
SOLUTIONHOOD	BACKGROUND	9	60
ELABORATION	INTERPRETATION	9	
ELABORATION	JOINT	8	
CONJUNCTION	RESULT	8	
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 annotators: 0.7% (SOLUTIONHOOD-BACKGROUND)

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

CAUSE subgroup signaling agreement

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

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

  - Annotating signals are more ambiguous than DMs
    - DMs' disagreement 15.27%
    - Other signals' disagreement 68.13%

CAUSE subgroup signals ($\geq$ 3 occurrences)

- CAUSE signals:
  - *bait-*, *-gatik*, *-nez*, *-ela eta* (CAUSE morphemes)
  - *-ren ondorioz* 'as consequence', *arrazoia* 'reason'
  - *izan ere* 'in fact'
- RESULT signals:
  - *ondorioz* 'consequence' and *emaitza* 'result'
  - *lortu* 'to achieve', *ekarri* 'to bring'
  - *eta* 'and', *beraz* 'thus'
  - *-tuz* '-ing'
- PURPOSE signals:
  - *-t(z)eko*, *-t(z)eko asmoz* 'with the intention of', *-t(z)eko helburuarekin* 'with the aim of'
  - *helburu* 'aim', *asmo* 'intention'
  - "Subjuntiboko aditzak" (subjunctive verbs)

Results of the RRs and their signals

Rhetorical relations			Signal%		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

RRs and signals: interpretation of the results

- Signaling tendencies:
  - Low signaling ($\leq 25\%$):
    - PREPARATION, BACKGROUND
  - Middle signaling ($\geq 25\%$ eta $\leq 75\%$):
    - EVIDENCE, RESTATEMENT, SUMMARY, ELABORATION, LIST, SEQUENCE
  - High signaling ($\geq 75\%$):
    - ENABLEMENT, MOTIVATION, JUSTIFY, ANTITHESIS, CONCESSION, MEANS, CIRCUMSTANCE, CONDITION, SOLUTIONHOOD, UNCONDITIONAL, INTERPRETATION, EVALUATION, CAUSE, RESULT, PURPOSE, CONTRAST, CONJUNCTION, DISJUNCTION
- The 4 most annotated RRs 48.44%, only signaled at 29.20%
  - ELABORATION, LIST, PREPARATION, BACKGROUND
    - General RRs (not very informative)
- The signaling of other 22 RRs has a high frequency 86.28%

Signaling and RR 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

- Detection of some RRs based on non-ambiguous signals

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Importance of the delivery phase

- Delivery phase is of paramount importance (Hovy, 2010), to provide place for interesting studies
  - But often forgotten
  - Not in the RST Spanish Treebank (da Cunha et al., 2011b)
    - Extract RRs from the corpus (to analyze the RRs patterns)
- Is there any place for improvements?
  1. The SEARCH section 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.

Importance of the delivery phase

- Delivery phase is of paramount importance (Hovy, 2010), to provide place for interesting studies
  - But often forgotten
  - Not in the RST Spanish Treebank (da Cunha et al., 2011b)
    - Extract RRs from the corpus (to analyze the RRs patterns)
- **Is there any place for improvements?**
  1. The SEARCH section 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 *taldeek* 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.

EDUs and CUs in RS-trees: *SEGMENTS* section

- **Extra advanced functionalities:**
  2. CU and RRs linked to CU
  3. Annotator's info

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

RELATIONS section

- **Extra advanced functionalities:**
  4. Specific RRs and the search of their signals

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

SIGNALS section

- **Extra advanced functionalities:**
  5. To search in which RR is the specific signal

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”,	

Outline

1. Introduction
   - Aims
   - RS-structure in Basque studies and in CLs
2. Methodology
   - Preparation phase
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
   - Delivery phase
3. Results
   - Segmentation
   - Central unit
   - Rhetorical relations
   - Signaling the RRs
4. Delivery phase
5. Conclusions and future work

Goal 1: describe the RS of a Basque corpus

- The Basque RST TreeBank (1,315 RR, 783 signals)
- Adjunct verb clause-based segmentation (81.14% F_1 in pairs)
  - A prototype of intra-sentential discourse segmentation (57.81% F_1)
- CU decision tree (61.42% F_1 in threes) and indicators (verbs and nouns) for CU detection
- RR (61.81% F_1 in pairs) and signal (76.82% F_1 in pairs) description
  - 22 RRs (51.16%) signaled with a high frequency 86.28%
  - Signals in DU_2 (67.94%) and in satellite unit (91.36%)
  - IMRaD structure was observed in RRs frequency linked to CU
  - Consistent cause subgroup harmonization for their detection

Goal 2: to establish an annotation method

- A new annotation phase (global coherence before local)
- A new method to gain reliability avoiding circularity (harmonizing RRs)
- A qualitative evaluation method for RS-trees
- A robust and innovative delivery phase to different theoretical studies (consistency, patterns, ambiguity)

Goal 3: validate annotation method and analyze disagreement cases

- Incremental annotation: intra-sentential segmentation was 13.74% lower than sentential but 14.19% higher for RRs.
- Macro-structure (CU) before, micro-structure (RRs)
  - Higher CU agreement (from 10% to 30%), even though the probability was smaller
  - Higher RR agreement when same CU is annotated (6.17%, t-test: $p < 0.013$)
  - Higher RR agrement when RRs are linked to CU (11.52%, t-test: $p < 0.001$)
- Signaling RRs to avoid implicit RRs, is more ambiguous than marking with DMs
  - Then, it is necessary to put more attention in signal evaluation
- Relevant disagreements in RR confusion matrix:
  - Most widely user RR (ELABORATION) in 47.21%
  - Not well understood RRs: EVIDENCE, INTERPRETATION, ANTITHESIS, EVALUATION and SUMMARY
  - Not used RR: SOLUTIONHOOD

Future work

- To measure the adequacy of the **segmentation** criteria and of the **RRs** harmonization criteria
- To extent the **corpus** to other genres and domains
  - The Reference Corpus for the Processing of Basque (EPEC) (Aduriz et al., 2006) manually-annotated at different language levels
  - From the abstracts to their full articles: summarization (da Cunha, 2008)
- To apply in **advanced applications**
  - Discourse segmenter
  - Detection of CU via indicators: summarization
  - IMRaD structure detection: assessment of written abstracts
  - Qualitative evaluation of RS-trees
  - Detection of the cause subgroup: discourse structure analysis
  - Re-annotate the **signals** of some RRs with more annotators, to gain reliability and detect RRs

Publications

Papers	Topic
Iruskieta (2012)	Explanation of RST
Iruskieta et al. (2011a)	Automatic segmentation
Iruskieta et al. (2014b)	Central unit
Iruskieta et al. (2013b)	The drawbacks of quantitative evaluation
Iruskieta et al. (2011b)	Relation and segmentation levels
da Cunha and Iruskieta (2010)	Qualitative evaluation of relations
Iruskieta et al. (2014a)	Qualitative evaluation of relations
Iruskieta et al. (2009)	DM for signals
Iruskieta and da Cunha (2010b)	DM for signals (Spanish and Basque)
Iruskieta and da Cunha (2010a)	Using DMs and RRs to discriminate domains (medicine and terminology)
Iruskieta et al. (2008)	Study of DM and its ambiguity
Garcia and Iruskieta (2013)	DMs of reformulation
Iruskieta et al. (2013a)	The RST Basque *TreeBank*

Annotated corpus:
http://ixa2.si.ehu.es/diskurtsoa/

Thesis-report in Basque:
http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena.pdf

Abbreviated translation of the thesis-report in English:
http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena_itzulita.pdf

Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean

A description of pragmatics rhetorical structure and its evaluation in computational linguistics

Candidate: Mikel Iruskieta Quintian

Advisors: Arantza Díaz de Ilarraza Sánchez
Mikel Lersundi Ayestaran

University of the Basque Country (UPV/EHU)

February 26, 2014

References I

Aduriz, I., Aranzabe, M. J., Arriola, J., Atutxa, A., Diaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and Urizar, R. (2006). Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15.

Afantenos, S. D., Denis, P., Muller, P., and Danlos, L. (2010). Learning recursive segments for discourse parsing. In *Seventh conference on International Language Resources and Evaluation*, pages 3578–3584, Paris, France.

Aierbe, A. (2008). Birformulazio-estrategiak eta komunikagarritasuna administrazioko testuetan. Technical report.

Alberdi, X. and Landa, J. (2013). EUDIMA corpusetik adibideak erauzteko lan-tresna: diskurtso unitate fraseologikoak aztertzeko lanabesa. *ASJU*, 45(2).

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.

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.

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.

References II

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.

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. (2008). Hacia un modelo lingüístico de resumen automático de artículos médicos en español. Doktore-tesia, IULA, Universitat Pompeu Fabra.

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. (2010a). Discourse segmentation for Spanish based on shallow parsing. In *9th Mexican international conference on Advances in artificial intelligence: Part I*, pages 13–23, Pachuca, Mexico. Springer-Verlag.

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

da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011a). 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.

da Cunha, I., Torres-Moreno, J.-M., Sierra, G., Cabrera-Diego, L.-A., and Castro-Rolón, B.-G. (2011b). The RST Spanish Treebank On-line Interface. In *International Conference Recent Advances in NLP*, Bulgaria.

Esnal, P. (2008). *Testu-antolatzaileen erabilera estrategikoa*, volume 51. Euskaltzaindia, Bilbo.

Euskaltzaindia (1990). *Euskal gramatika. Lehen urratsak III (Lokailuak)*. Euskaltzaindia, Bilbo.

References III

Euskaltzaindia (1994). *Euskal gramatika: lehen urratsak (EGLU) VI (juntagailuak)*. Euskaltzaindia, Bilbo.

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

Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A., and Webber, B. L. (2003). D-ltag system: Discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Logic, Language and Information*, 12(3):261–279.

Garcia, J. and Iruskieta, M. (2013). *Birformulatzaile zuzentzaileak testu idatzietan*. Eridenen du zerzaz kontenta. Sailkideen omenaldia Henrike Knörr irakasleari (1947-2008). EHU, Bilbo.

García, I. M. (2010). *Estrategias textuales y discursivas en el aprendizaje de la exposición oral de dos materias distintas*, pages 155–162. Modos y formas de la comunicación humana. AESLA, Vigo.

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).

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.

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.

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

Ibarra, O. (2013). Sobre estrategias discursivas del lenguaje de los jóvenes vascoparlantes: aspectos pragmáticos y discursivos (conectores, marcadores). *ASJU*, pages 395–411.

Ide, N. and Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. In *2nd Int. Conf. Global Interoperability Lang. Res*, Hong Kong.

References IV

Iruskieta, M. (2012). Pragmatika. http://www.ehu.es/seg/hizk/1/6.

Iruskieta, M., Aranzabe, M. J., Diaz de Ilarraza, A., 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.

Iruskieta, M. and da Cunha, I. (2010a). El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera y español. *Calidoscópio*, 8(3):181–202.

Iruskieta, M. and da Cunha, I. (2010b). 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. (2014a). A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. Submitted to Language Resources and Evaluation.

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2008). Análisis de los marcadores del discurso para el euskera: denominación, clases, relaciones semánticas y tipos de ambigüedad. In *Proceedings of 26th AESLA International Conference*, pages 1271–1282.

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2009). Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso. In *Proceedings of 27th AESLA International Conference*, pages 963–971.

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2011a). Bases para la implementación de un segmentador discursivo para el euskera. In *8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*.

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2011b). 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.

References V

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2014b). Deecting the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. Submitted to Coling.

Kortmann, B. (1991). *Free adjuncts and absolutes in English: Problems of control and interpretation*. Psychology Press, New York.

Larringan, L. (1995). Testu-antolatzaileak bi testu motatan: testu informatiboa eta argudiapenezkoa. Doktore-tesia, Euskal Herriko Unibertsitatea, Gasteiz.

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.

Mol, S. (2005). Causality in a cross-linguistic perspective: So, therefore, and thus versus så, derfor, and således. Technical Report 27.

References VI

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

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5-6):601–638.

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.

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.

References VII

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). Disambiguating rhetorical structure. *Research on Language and Computation*, 6(3):311–332.

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

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592.

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.

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.

Urrutia, A. (2008). Legeen eta administrazioaren hizkera, testu-antolatzaileen ikuspegitik. *Euskera: Euskaltzaindiaren lan eta agiriak*, 53(2):525–546.

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

References VIII

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.

Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean

A description of pragmatics rhetorical structure and its evaluation in computational linguistics

Candidate: Mikel Iruskieta Quintian

Advisors: Arantza Díaz de Ilarraza Sánchez
Mikel Lersundi Ayestaran

University of the Basque Country (UPV/EHU)

February 26, 2014