Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean

Doktoregaia: Bertol Arrieta Kortajarena
Zuzendariak: Iñaki Alegria Loinaz
Arantza Diaz de Ilarraza Sanchez

Lengoaia eta Sistema Informatikoak/Lenguajes y Sistemas Informáticos
Euskal Herriko Unibertsitatea/Universidad del País Vasco

2010eko uztailaren 27a

Komak eta kateak

Nork jan du aita?

Komak eta kateak

Nork jan du, aita?

Komak eta kateak

(Nork) (jan du), (aita)?

Komak eta perpausak

Haserretu egin zen emaztea beste batekin ikusi zuenean.

Komak eta perpausak

Haserretu egin zen, emaztea beste batekin ikusi zuenean.
Haserretu egin zen emaztea, beste batekin ikusi zuenean.

Komak eta perpausak

(Haserretu egin zen, (emaztea beste batekin ikusi zuenean).)
(Haserretu egin zen emaztea, (beste batekin ikusi zuenean).)

Tesi honen helburu nagusiak

1. Euskarako kate- eta perpaus-identifikatzaile automatikoak sortzea

Tesi honen helburu nagusiak

1. Euskarako kate- eta perpaus-identifikatzaile automatikoak sortzea
2. Euskarako koma-zuzentzaile automatikoa garatzea
   - Euskarako komaren erabilera formalizatzea
   - Komaren erabilerak, euskarakoa eta ingelesekoa, konparatzea
   - Kate- eta perpaus-identifikatzaile automatikoen informazioa baliatzea
   - Informazio linguistikoa lortzeko dauzkagun tresnekiko mendekotasuna aztertzea
   - Ebaluazio kualitatiboa egitea

Bestelako helburuak

1. Ikasketa automatikoko teknikak aztertzea
   - Azaleko sintaxian
   - Erroreen detekzioan

Bestelako helburuak

1. Ikasketa automatikoko teknikak aztertzea
   - Azaleko sintaxian
   - Erroreen detekzioan
2. Erroreen detekziorako oinarrizko baliabideak garatzea
   - Erroreak etiketatuta dituzten corpusak
   - Euskarako erroreen sailkapena

Aurkezpenaren eskema

Aurkezpenaren eskema

Aurkezpenaren eskema

Aurkezpenaren eskema

Aurkezpenaren eskema

Ikasketa automatikoaren funtsa

Ikasketa automatikoa:

Ebatzi beharreko problema sailkapen-ataza batean bilakatzean datza.

Ikasketa automatikoaren funtsa

Ikasketa automatikoa:

Ebatzi beharreko problema sailkapen-ataza batean bilakatzean datza.

Ikasketa gainbegiratua:

- Ikasi nahi den atazarako corpus etiketatua behar da
- Ikasi nahi den kontzeptuari buruzko adibideekin ikasten du makinak $\Rightarrow$ adibide berrien portaera iragartzen du
- Bi dimentsioko matrizea: lerroetan, adibideak; zutabeetan, ezaugarriak
- Azken ezaugarria $\Rightarrow$ ikasi beharreko kontzeptua
- HPan, bereziki sintaxian, hitz bakoitza adibide bat da, eta inguruko hitzen informazioa gehitzeko $\Rightarrow$ leihoa

Ikasketa automatikoaren funtsa

Arropa	arropa	IZE	-	B-NP	(S(S*
mota	mota	IZE	-	I-NP	*
guztiak	guzti	DET	ABS	I-NP	*
gustatzen	gustatu	ADI	-	B-VP	*
zaizkidan	izan	ADL	-	I-VP	*
arren	arren	LOT	-	O	*
,	,	PUNT_KOMA	-	O	*S)
nahiago	nahi	IZE	ABS	B-VP	*
ditut	ukan	ADT	-	I-VP	*
soinekoak	soineko	IZE	ABS	B-NP	*
.	.	PUNT_PUNT	-	O	*S)

Taula: Perpausen ikasketarako matrize baten adibidea

Ikasketa automatikoaren funtsa

Arropa	arropa	IZE	-	B-NP	(S(S*
mota	mota	IZE	-	I-NP	*
guztiak	guzti	DET	ABS	I-NP	*
gustatzen	gustatu	ADI	-	B-VP	*
zaizkidan	izan	ADL	-	I-VP	*
arren	arren	LOT	-	O	*
,	,	PUNT_KOMA	-	O	*S)
nahiago	nahi	IZE	ABS	B-VP	*
ditut	ukan	ADT	-	I-VP	*
soinekoak	soineko	IZE	ABS	B-NP	*
.	.	PUNT_PUNT	-	O	*S)

Taula: Perpausen ikasketarako matrize baten adibidea

Ikasketa automatikoaren funtsa

Arropa	arropa	IZE	-	B-NP	**(S(S***
mota	mota	IZE	-	I-NP	*
guztiak	guzti	DET	ABS	I-NP	*
gustatzen	gustatu	ADI	-	B-VP	*
zaizkidan	izan	ADL	-	I-VP	*
arren	arren	LOT	-	O	*
,	,	PUNT_KOMA	-	O	***S)**
nahiago	nahi	IZE	ABS	B-VP	*
ditut	ukan	ADT	-	I-VP	*
soinekoak	soineko	IZE	ABS	B-NP	*
.	.	PUNT_PUNT	-	O	***S)**

Taula: Perpausen ikasketarako matrize baten adibidea

Ikasketa automatikoaren funtsa

Arropa	arropa	IZE	-	B-NP	**(S(S***
mota	mota	IZE	-	I-NP	*
guztiak	guzti	DET	ABS	I-NP	*
gustatzen	gustatu	ADI	-	B-VP	*
zaizkidan	izan	ADL	-	I-VP	*
arren	arren	LOT	-	O	*
,	,	**PUNT_KOMA**	-	**O**	***S)**
nahiago	nahi	IZE	ABS	B-VP	*
ditut	ukan	ADT	-	I-VP	*
soinekoak	soineko	IZE	ABS	B-NP	*
.	.	PUNT_PUNT	-	O	***S)**

Taula: Perpausen ikasketarako matrize baten adibidea

Ikasketa automatikoaren funtsa

Arropa	arropa	IZE	-	B-NP	**(S(S***
mota	mota	IZE	-	I-NP	*
guztiak	guzti	DET	ABS	I-NP	*
gustatzen	gustatu	ADI	-	B-VP	*
zaizkidan	**izan**	**ADL**	-	**I-VP**	*
arren	**arren**	**LOT**	-	**O**	*
,	**,**	**PUNT_KOMA**	-	**O**	***S)**
nahiago	**nahi**	**IZE**	**ABS**	**B-VP**	*
ditut	**ukan**	**ADT**	-	**I-VP**	*
soinekoak	soineko	IZE	ABS	B-NP	*
.	.	PUNT_PUNT	-	O	***S)**

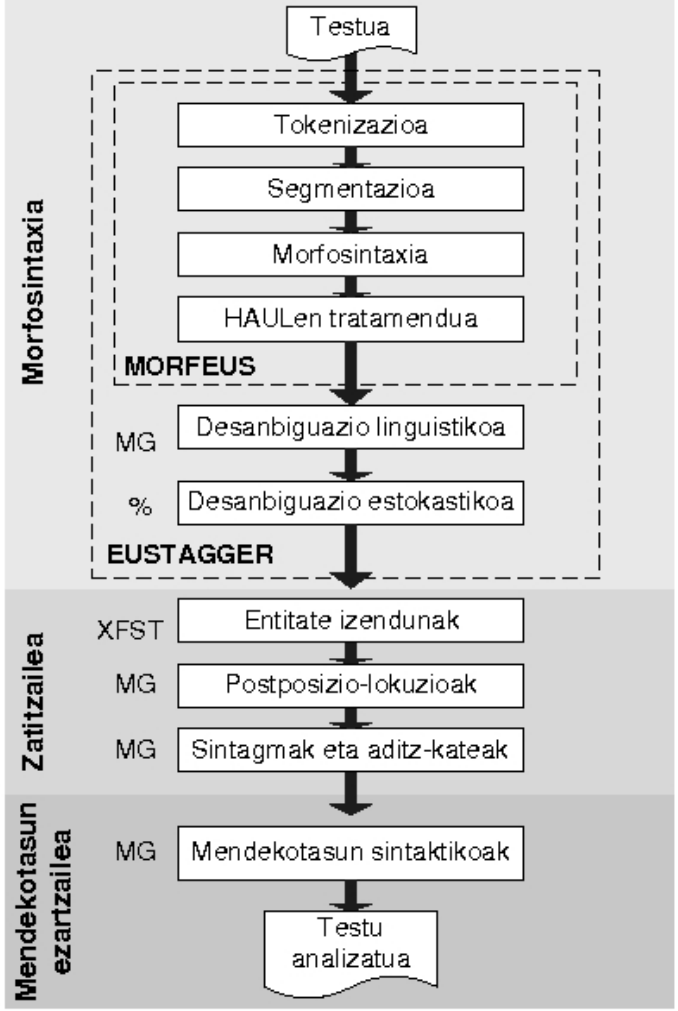
Taula: Perpausen ikasketarako matrize baten adibidea

Oinarrizko ikasketa-algoritmoak

	Estatistikoa	IA	Sinbolikoa	Azpisinbolikoa
Naive Bayes	√			√
Erabaki-zuhaitzak		√	√	
Pertzeptroiak		√		√
SVM		√		√

Taula: Tesi-lan honetan erabilitako oinarrizko ikasketa-eskemen sailkapena

Sintaxiaren tratamendu automatikoa IXA taldean



Euskarako erroreen detekzioa IXA taldean

Datak, postposizio-lokuzioak eta komunztadura (Oronoz, 2009)

*__Zentral nuklearrak__ *zakar erradiaktiboa* __eratzen dute__.

Euskarako erroreen detekzioa IXA taldean

Datak, postposizio-lokuzioak eta komunztadura (Oronoz, 2009)

*__Zentral nuklearrak__ *zakar erradiaktiboa* __eratzen dute__.

Determinatzaileak (Uria, 2009)

**Amak esan dit* __pinguino bezala__ *nabilela.*

Euskarako erroreen detekzioa IXA taldean

Datak, postposizio-lokuzioak eta komunztadura (Oronoz, 2009)

*__Zentral nuklearrak__ *zakar erradiaktiboa* __eratzen dute__.

Determinatzaileak (Uria, 2009)

**Amak esan dit* __pinguino bezala__ *nabilela.*

Ikasketa automatiko bidez:

- Proba txiki bat determinatzaile- eta komunztadura-akatsak detektatzeko: corpus txikia, errore gutxi
- Corpus egokia behar
- Komen ikasketa: corpusa etiketatzeko beharrik ez

Aurkezpenaren eskema

Hitz multzoak: definizio formal bat (Carreras, 2005)

$$\mathcal{P} = \{(s, e)_k | 1 <= s <= e, k \in \mathcal{K}\}$$

Hitz multzoak: definizio formal bat (Carreras, 2005)

$$\mathcal{P} = \{(s, e)_k | 1 <= s <= e, k \in \mathcal{K}\}$$

$p_1 = (s_1, e_1)$ eta $p_2 = (s_2, e_2)$ hitz multzoak **gainjartzen** dira ($p_1 \sim p_2$), baldin:

$$s_1 < s_2 <= e_1 < e_2 \text{ edo } s_2 < s_1 <= e_2 < e_1$$

Grafikoki: $(s_1 \quad (s_2 \quad)e_1 \quad)e_2$

Hitz multzoak: definizio formal bat (Carreras, 2005)

$$\mathcal{P} = \{(s, e)_k | 1 <= s <= e, k \in \mathcal{K}\}$$

$p_1 = (s_1, e_1)$ eta $p_2 = (s_2, e_2)$ hitz multzoak **gainjartzen** dira ($p_1 \sim p_2$), baldin:

$$s_1 < s_2 <= e_1 < e_2 \text{ edo } s_2 < s_1 <= e_2 < e_1$$

Grafikoki: $(s_1 \quad (s_2 \quad)e_1 \quad)e_2$

p_2 *hitz multzo*ak p_1 *hitz multzo*a **bere barnean hartzen** du ($p_1 \preceq p_2$), baldin:

$$s_2 <= s_1 <= e_1 <= e_2$$

Grafikoki: $(s_2 \quad (s_1 \quad)e_1 \quad)e_2$

Kateak eta perpausak hitz multzo gisa

Kateen soluzio-espazioa (Abney, 1991)

$$\mathcal{Y} = \{y \subseteq \mathcal{P} | \forall p_1, p_2 \in y (p_1 \not\asymp p_2 \wedge p_1 \not\trianglelefteq p_2)\}$$

(Lehendakariak) (bere erabakia) (berretsi zuen).

Kateak eta perpausak hitz multzo gisa

Kateen soluzio-espazioa (Abney, 1991)

$$\mathcal{Y} = \{y \subseteq \mathcal{P} | \forall p_1, p_2 \in y (p_1 \not\asymp p_2 \wedge p_1 \not\preceq p_2)\}$$

(Lehendakariak) (bere erabakia) (berretsi zuen).

Perpausen soluzio-espazioa:

$$\mathcal{Y} = \{y \subseteq \mathcal{P} | \forall p_1, p_2 \in y (p_1 \not\asymp p_2)\}$$

(Zubietarrak kezkatuta daude, (planak euren herria desitxuratuko duelako).)

Kateak: zenbait adibide

Ingelesez: lana banatzearen onurak

(The deficit) (will narrow) (to) (only 1.8 billion) (in) (September).
(The deficit) (will narrow) (to only 1.8 billion) (in September).

Kateak: zenbait adibide

Ingelesez: lana banatzearen onurak

(The deficit) (will narrow) (to) (only 1.8 billion) (in) (September).
(The deficit) (will narrow) (to only 1.8 billion) (in September).

Euskarako kateak: barne-hartzea?

(Ez dut (horretan) sakondu nahi).
(Ez dut) (horretan) (sakondu nahi).

Perpausak: zenbait adibide

Perpausen barnean perpausak:

((Hiriko agintari militarrek aditzera eman dutenez,) suziri batek eragin zuen leherketa.)
((Urteko helburu nagusia (bakea lortzea) izango dela) adierazi du.)

Perpausak: zenbait adibide

Perpausen barnean perpausak:

((Hiriko agintari militarrek aditzera eman dutenez,) suziri batek eragin zuen leherketa.)
((Urteko helburu nagusia (bakea lortzea) izango dela) adierazi du.)

Anbiguotasuna

(Haserretu egin zen (emaztea beste batekin ikusi zuenean).)
(Haserretu egin zen emaztea (beste batekin ikusi zuenean).)

Ingeleseko kateen identifikazioa

	Teknika	F_1
(Lee eta Wu, 2007)	*SVM*	94,22
(Zhang et al., 2002)	*Winnow* orokortua	94,13
(Shen eta Sarkar, 2005)	*Voted HMM*	94,01
(Kudo eta Matsumoto, 2001)	*SVM*	93,91
(Carreras et al., 2005)	*FR-perceptron*	93,74
(Molina eta Pla, 2002)	*Spec. HMM*	93,25
Oinarrizko neurria (baseline-a)	kategoriaren usuena	77,07

Taula: Ingeleseko kate-identifikatzaile onenak, *CoNLL 2000*ko baldintzetan

Ingeleseko kateen identifikazioa

	Teknika	F_1
(Lee eta Wu, 2007)	*SVM*	94,22
(Zhang et al., 2002)	*Winnow* orokortua	94,13
(Shen eta Sarkar, 2005)	*Voted HMM*	94,01
(Kudo eta Matsumoto, 2001)	*SVM*	93,91
(Carreras et al., 2005)	*FR-perceptron*	93,74
(Molina eta Pla, 2002)	*Spec. HMM*	93,25
Oinarrizko neurria (baseline-a)	**kategoriaren usuena**	**77,07**

Taula: Ingeleseko kate-identifikatzaile onenak, *CoNLL 2000*ko baldintzetan

Ingeleseko kateen identifikazioa

	Teknika	F_1
(Lee eta Wu, 2007)	*SVM*	94,22
(Zhang et al., 2002)	*Winnow* orokortua	94,13
(Shen eta Sarkar, 2005)	*Voted HMM*	94,01
(Kudo eta Matsumoto, 2001)	*SVM*	93,91
(Carreras et al., 2005)	**FR-perceptron**	**93,74**
(Molina eta Pla, 2002)	*Spec. HMM*	93,25
Oinarrizko neurria (baseline-a)	kategoriaren usuena	77,07

Taula: Ingeleseko kate-identifikatzaile onenak, *CoNLL 2000*ko baldintzetan

Ingeleseko perpausen identifikazioa

	Teknika	F_1
(Ram eta Devi, 2008)	*CRF + erregela ling.*	89,04
(Carreras et al., 2005)	*FR-Perceptron*	85,03
(Nguyen et al., 2009)	*joint-CRF*	84,66
(Carreras eta Màrquez, 2001)	*Boosting*	81,73
(Molina eta Pla, 2001)	*HMM*	70,68
(Tjong Kim Sang, 2001)	*MBL*	70,58
(Patrick eta Goyal, 2001)	*Erabaki-grafoak*	68,85
Oinarrizko neurria (baseline-a)	Esaldiak mugatu	47,71

Taula: Ingeleseko perpaus-identifikatzaile onenak, *CoNLL 2001*eko baldintzetan

Ingeleseko perpausen identifikazioa

	Teknika	F_1
(Ram eta Devi, 2008)	*CRF + erregela ling.*	89,04
(Carreras et al., 2005)	*FR-Perceptron*	85,03
(Nguyen et al., 2009)	*joint-CRF*	84,66
(Carreras eta Màrquez, 2001)	*Boosting*	81,73
(Molina eta Pla, 2001)	*HMM*	70,68
(Tjong Kim Sang, 2001)	*MBL*	70,58
(Patrick eta Goyal, 2001)	*Erabaki-grafoak*	68,85
Oinarrizko neurria (baseline-a)	**Esaldiak mugatu**	**47,71**

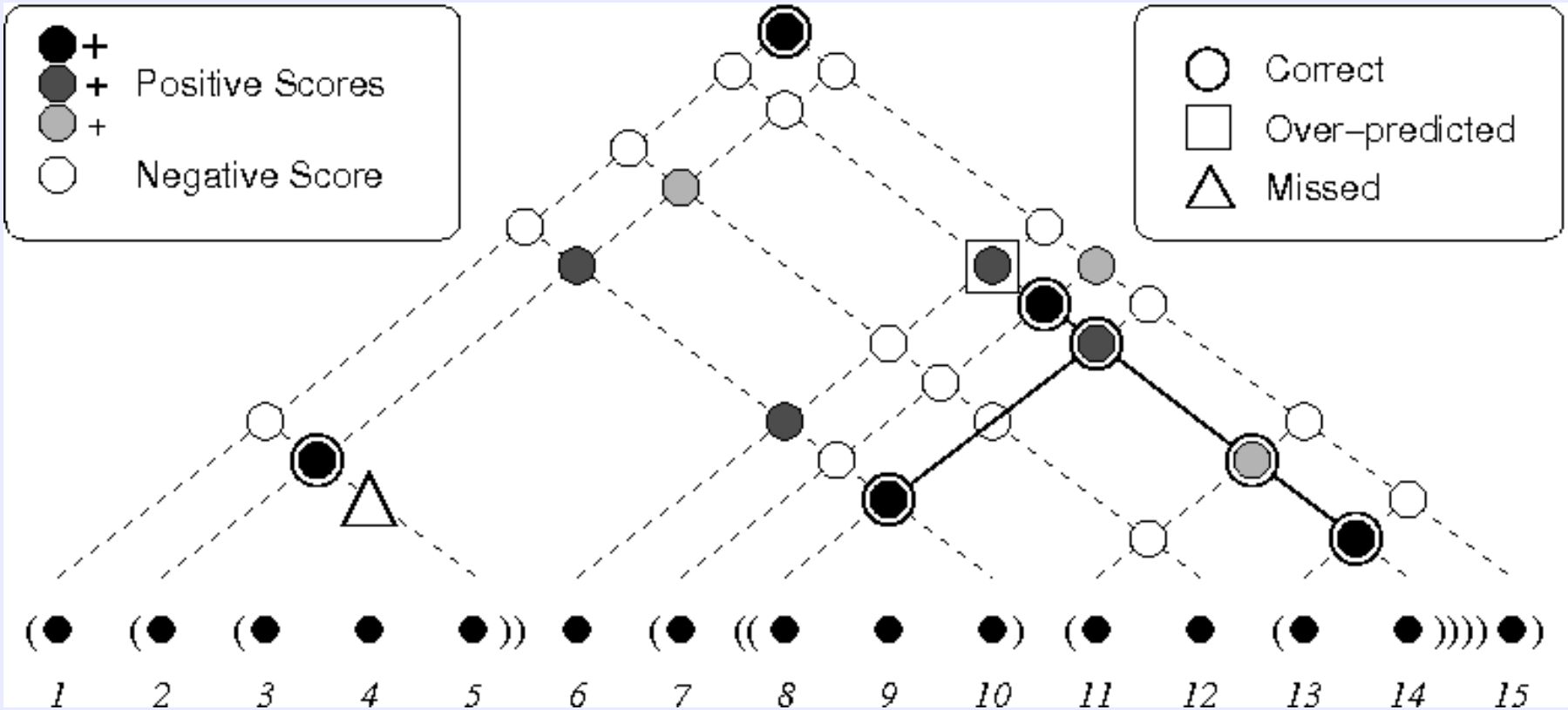
Taula: Ingeleseko perpaus-identifikatzaile onenak, *CoNLL 2001*eko baldintzetan

Ingeleseko perpausen identifikazioa

	Teknika	F_1
(Ram eta Devi, 2008)	*CRF + erregela ling.*	89,04
(Carreras et al., 2005)	**FR-Perceptron**	**85,03**
(Nguyen et al., 2009)	*joint-CRF*	84,66
(Carreras eta Màrquez, 2001)	*Boosting*	81,73
(Molina eta Pla, 2001)	*HMM*	70,68
(Tjong Kim Sang, 2001)	*MBL*	70,58
(Patrick eta Goyal, 2001)	*Erabaki-grafoak*	68,85
Oinarrizko neurria (baseline-a)	Esaldiak mugatu	47,71

Taula: Ingeleseko perpaus-identifikatzaile onenak, *CoNLL 2001*eko baldintzetan

FR-Perceptron algoritmoa: adibide bat (Carreras, 2005)



Esperimentuen prestaketa (I)

- Corpusa:

	Guztira	**Ikasketa-corpusa**	**Garapen-corpusa**	**Test-corpusa**
EPEC	150.128	104.956	22.548	22.624

Taula: Kateen eta perpausen identifikaziorako corpusa

- Ebaluaziorako neurriak:

  Doitasuna, estaldura eta F_1 neurria.

  $$F_1 = \frac{2 * Doitasuna * Estaldura}{(Doitasuna + Estaldura)}$$

Esperimentuen prestaketa (II): oinarrizko neurriak

		Doitasuna	**Estaldura**	**F_1 neurria**
Euskarako oinarrizko neurriak	**Kateak**	45,76	60,21	52,00
Ingeleseko oinarrizko neurriak	**Kateak**	72,58	82,14	77,07

Taula: **Kateen** identifikazioko oinarrizko neurriak: euskara vs ingelesa

Esperimentuen prestaketa (II): oinarrizko neurriak

		Doitasuna	Estaldura	F_1 neurria
Euskarako oinarrizko neurriak	**Kateak**	45,76	60,21	**52,00**
Ingeleseko oinarrizko neurriak	**Kateak**	72,58	82,14	**77,07**

Taula: **Kateen** identifikazioko oinarrizko neurriak: euskara vs ingelesa

Esperimentuen prestaketa (II): oinarrizko neurriak

		Doitasuna	Estaldura	F_1 neurria
Euskarako oinarrizko neurriak	**Sint.**	31,50	47,08	**37,74**
	ADK	77,65	80,63	**79,11**
	Kateak	45,76	60,21	52,00
Ingeleseko oinarrizko neurriak	**Kateak**	72,58	82,14	77,07

Taula: **Kateen** identifikazioko oinarrizko neurriak: euskara vs ingelesa

Esperimentuen prestaketa (II): oinarrizko neurriak

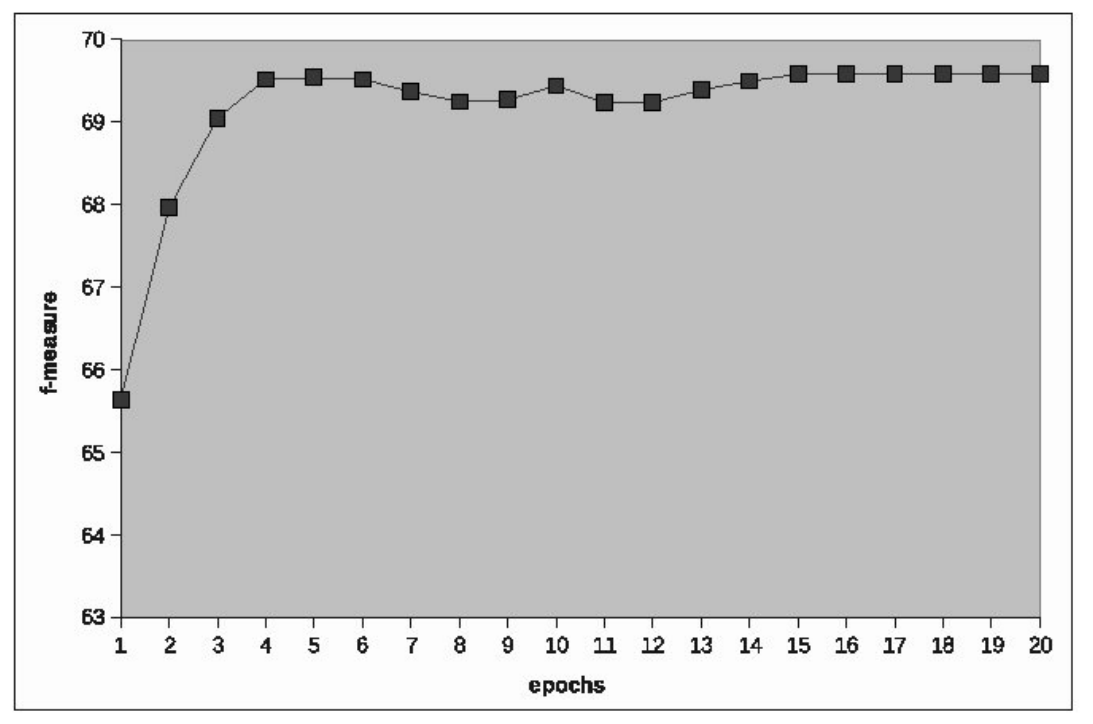
	Doitasuna	**Estaldura**	F_1 **neurria**
Euskarako oinarrizko neurriak	91,41	33,27	**48,79**
Ingeleseko oinarrizko neurriak	98,44	31,48	**47,71**

Taula: **Perpausen** identifikazioko oinarrizko neurriak: euskara vs ingelesa

Kateen identifikazio automatikoa

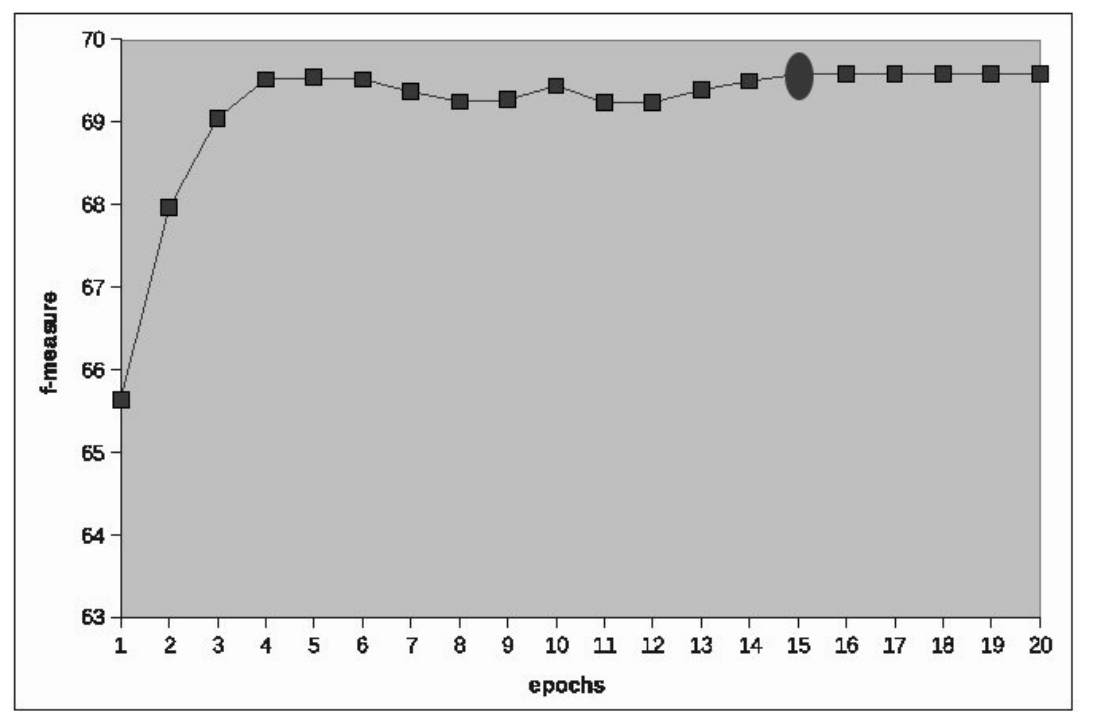
- Epoch zenbakiaren eragina
- Lehen probak oinarrizko ezaugarriekin
- Ezaugarri linguistikoak gehituz
- Kateen identifikazioa IA eta erregelak konbinatuz
- Ikasketa-corpusa handituz
- Emaitza idealak eskuz desanbiguatutako informazioarekin
- Azken emaitzak test-corpusean

Epoch zenbakiaren eragina



Irudia: *Epoch-zenbakiaren* eragina *FR-Perceptron* bidezko ikasketa automatikoan

Epoch zenbakiaren eragina



Irudia: *Epoch-zenbakiaren* eragina *FR-Perceptron* bidezko ikasketa automatikoan

Lehen probak oinarrizko ezaugarriekin

		Doitasuna	Estaldura	F_1 neurria
Oinarrizko neurriak (baseline-a)	**Kateak**	45,76	60,21	52,00
FR-Perceptron oinarrizko ezaugarriekin	**Kateak**	70,85	68,36	69,58

Taula: *FR-Perceptron* algoritmoan oinarritutako kate-identifikatzailearen emaitzak, *oinarrizko ezaugarriak* erabilita: **hitza eta kategoria**

Lehen probak oinarrizko ezaugarriekin

		Doitasuna	Estaldura	F_1 neurria
Oinarrizko neurriak (baseline-a)	**Kateak**	45,76	60,21	**52,00**
FR-Perceptron oinarrizko ezaugarriekin	**Kateak**	70,85	68,36	**69,58**

Taula: *FR-Perceptron* algoritmoan oinarritutako kate-identifikatzailearen emaitzak, *oinarrizko ezaugarriak* erabilita: **hitza eta kategoria**

Lehen probak oinarrizko ezaugarriekin

		Doitasuna	Estaldura	F_1 neurria
Oinarrizko neurriak (baseline-a)	**Sint.**	31,50	47,08	37,74
	ADK	77,65	80,63	79,11
	Kateak	45,76	60,21	52,00
FR-Perceptron oinarrizko ezaugarriekin	**Sint.**	60,54	55,49	57,90
	ADK	84,98	88,40	86,66
	Kateak	70,85	68,36	69,58

Taula: *FR-Perceptron* algoritmoan oinarritutako kate-identifikatzailearen emaitzak, *oinarrizko ezaugarriak* erabilita: **hitza eta kategoria**

Lehen probak oinarrizko ezaugarriekin

		Doitasuna	Estaldura	F_1 neurria
Oinarrizko neurriak (baseline-a)	**Sint.**	31,50	47,08	**37,74**
	ADK	77,65	80,63	**79,11**
	Kateak	45,76	60,21	52,00
FR-Perceptron oinarrizko ezaugarriekin	**Sint.**	60,54	55,49	**57,90**
	ADK	84,98	88,40	**86,66**
	Kateak	70,85	68,36	69,58

Taula: *FR-Perceptron* algoritmoan oinarritutako kate-identifikatzailearen emaitzak, *oinarrizko ezaugarriak* erabilita: **hitza eta kategoria**

Ezaugarri linguistikoak gehituz

	Doitasuna	**Estaldura**	**F_1 neurria**
oin	70,85	68,36	69,58
oin + ak	70,78	69,15	69,96
oin + dek	77,42	80,17	78,77
oin + l	71,53	68,91	70,20
oin + men	70,77	68,48	69,61
oin + ak + dek + l + men	77,67	79,70	78,67

Taula: Ezaugarri linguistikoak gehituz (ak: azpikategoria; dek: deklinabidea; l: lema; men: mendeko perpaus mota)

Ezaugarri linguistikoak gehituz

	Doitasuna	**Estaldura**	**F_1 neurria**
oin	70,85	68,36	69,58
oin + ak	70,78	69,15	69,96
oin + dek	77,42	80,17	**78,77**
oin + l	71,53	68,91	70,20
oin + men	70,77	68,48	69,61
oin + ak + dek + l + men	77,67	79,70	78,67

Taula: Ezaugarri linguistikoak gehituz (ak: azpikategoria; dek: deklinabidea; l: lema; men: mendeko perpaus mota)

Deklinabidearen garrantzia sintagmetan

	Sint. F_1 neurria	**ADK F_1 neurria**	**F_1 neurria**
Oin	57,90	86,66	69,58
Oin + dek	73,28	87,33	78,77

Taula: Deklinabidearen eraginaren konparazioa sintagmetan eta aditz-kateetan

Deklinabidearen garrantzia sintagmetan

	Sint. F_1 neurria	**ADK F_1 neurria**	**F_1 neurria**
Oin	57,90	**86,66**	69,58
Oin + dek	73,28	**87,33**	78,77

Taula: Deklinabidearen eraginaren konparazioa sintagmetan eta aditz-kateetan

Deklinabidearen garrantzia sintagmetan

	Sint. F_1 neurria	**ADK F_1 neurria**	**F_1 neurria**
Oin	**57,90**	86,66	69,58
Oin + dek	**73,28**	87,33	78,77

Taula: Deklinabidearen eraginaren konparazioa sintagmetan eta aditz-kateetan

Kateen identifikazioa IA eta erregelak konbinatuz

	Doitasuna	**Estaldura**	**F_1 neurria**
Erreg	50,06	52,98	51,48
FR-Perceptron oin	70,85	68,36	69,58
FR-Perceptron oin + erreg	75,51	77,61	76,54
FR-Perceptron oin + dek	77,42	80,17	78,77
FR-Perceptron oin + dek + erreg	77,68	80,80	79,21

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Kateen identifikazioa IA eta erregelak konbinatuz

	Doitasuna	**Estaldura**	**F_1 neurria**
Erreg	50,06	52,98	51,48
FR-Perceptron oin	70,85	68,36	**69,58**
FR-Perceptron oin + erreg	75,51	77,61	**76,54**
FR-Perceptron oin + dek	77,42	80,17	78,77
FR-Perceptron oin + dek + erreg	77,68	80,80	79,21

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Kateen identifikazioa IA eta erregelak konbinatuz

	Doitasuna	**Estaldura**	**F_1 neurria**
Erreg	50,06	52,98	51,48
FR-Perceptron oin	70,85	68,36	69,58
FR-Perceptron oin + erreg	75,51	77,61	76,54
FR-Perceptron oin + dek	77,42	80,17	**78,77**
FR-Perceptron oin + dek + erreg	77,68	80,80	**79,21**

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Ikasketa-corpusa handituz

	Doitasuna	**Estaldura**	**F_1 neurria**
Ikasketa-corpusaren % 25arekin	77,68	80,80	79,21
Ikasketa-corpusaren % 50arekin	79,46	83,06	81,22
Ikasketa-corpusaren % 100arekin	81,09	84,24	82,64

Taula: Ikasketa-corpusaren tamainaren arabera (% 100 = 104.956 token)

Ikasketa-corpusa handituz

	Doitasuna	**Estaldura**	**F_1 neurria**
Ikasketa-corpusaren % 25arekin	77,68	80,80	**79,21**
Ikasketa-corpusaren % 50arekin	79,46	83,06	**81,22**
Ikasketa-corpusaren % 100arekin	81,09	84,24	82,64

Taula: Ikasketa-corpusaren tamainaren arabera (% 100 = 104.956 token)

Ikasketa-corpusa handituz

	Doitasuna	**Estaldura**	**F_1 neurria**
Ikasketa-corpusaren % 25arekin	77,68	80,80	79,21
Ikasketa-corpusaren % 50arekin	79,46	83,06	**81,22**
Ikasketa-corpusaren % 100arekin	81,09	84,24	**82,64**

Taula: Ikasketa-corpusaren tamainaren arabera (% 100 = 104.956 token)

Emaitza idealak eskuz desanbiguatutako informazioarekin

Corpusa nola desanbiguatua	**Doit.**	**Est.**	**Sint.** F_1	**ADK** F_1	F_1
Automatikoki	81,09	84,24	78,00	89,84	82,64
Eskuz	89,61	91,46	87,85	94,52	90,52

Taula: Erabilitako informazio linguistikoaren kalitatearen arabera

Emaitza idealak eskuz desanbiguatutako informazioarekin

Corpusa nola desanbiguatua	**Doit.**	**Est.**	**Sint.** F_1	**ADK** F_1	F_1
Automatikoki	81,09	84,24	78,00	89,84	**82,64**
Eskuz	89,61	91,46	87,85	94,52	**90,52**

Taula: Erabilitako informazio linguistikoaren kalitatearen arabera

Emaitza idealak eskuz desanbiguatutako informazioarekin

Corpusa nola desanbiguatua	**Doit.**	**Est.**	**Sint.** F_1	**ADK** F_1	F_1
Automatikoki	81,09	84,24	**78,00**	89,84	**82,64**
Eskuz	89,61	91,46	**87,85**	94,52	**90,52**

Taula: Erabilitako informazio linguistikoaren kalitatearen arabera

Emaitza idealak eskuz desanbiguatutako informazioarekin

Corpusa nola desanbiguatua	**Doit.**	**Est.**	**Sint.** F_1	**ADK** F_1	F_1
Automatikoki	81,09	84,24	**78,00**	89,84	**82,64**
Eskuz	89,61	91,46	**87,85**	94,52	**90,52**

Taula: Erabilitako informazio linguistikoaren kalitatearen arabera

Desanbiguazio maila	F_1
1. mailan	95,92
2. mailan (ak)	95,42
3. mailan (ak + dek)	**90,36**

Taula: *Eustagger*-en emaitzak desanbiguazio mailaren arabera (Ezeiza, 2002)

Azken emaitzak test-corpusean

Proba-corpusa	**Doit.**	**Est.**	F_1
Garapen-corpusa	81,09	84,24	**82,64**
Test-corpusa	81,35	85,07	**83,17**

Taula: Garapen-corpuseko emaitzak vs test-corpuseko emaitzak

Perpausen identifikazio automatikoa

- Lehen probak oinarrizko ezaugarriekin
- Ezaugarri linguistikoak gehituz
- Kateen identifikazioa IA eta erregelak konbinatuz
- Ikasketa-corpusa handituz
- Emaitza idealak eskuz desanbiguatutako informazioarekin
- Azken emaitzak test-corpusean

Lehen probak oinarrizko ezaugarriekin

	Doitasuna	**Estaldura**	**F_1 neurria**
Oinarrizko neurriak (baseline-a)	91,41	33,27	48,79
FR-Perc. oinarrizko ezaugarriekin	76,72	64,34	**69,99**

Taula: Oinarrizko ezaugarriekin: **hitz, kategoria eta kateei buruzko informazioa**

Ezaugarri linguistikoak gehituz

	Doitasuna	**Estaldura**	**F_1 neurria**
oin	76,72	64,34	69,99
oin + ak	76,52	66,49	71,15
oin + dek	76,41	66,10	70,89
oin + l	75,66	67,58	71,39
oin + men	76,39	65,68	70,63
oin + ak + dek + l + men	76,82	69,94	73,22

Taula: Ezaugarri linguistikoak gehituz (ak: azpikategoria; dek: deklinabidea; l: lema; men: mendeko perpaus mota)

Ezaugarri linguistikoak gehituz

	Doitasuna	**Estaldura**	**F_1 neurria**
oin	76,72	64,34	69,99
oin + ak	76,52	66,49	71,15
oin + dek	76,41	66,10	70,89
oin + l	75,66	67,58	71,39
oin + men	76,39	65,68	70,63
oin + ak + dek + l + men	76,82	69,94	**73,22**

Taula: Ezaugarri linguistikoak gehituz (ak: azpikategoria; dek: deklinabidea; l: lema; men: mendeko perpaus mota)

Perpausen identifikazioa IA eta erregelak konbinatuz

	Doitasuna	**Estaldura**	F_1
erreg	50,84	48,63	49,71
oin + ak + dek + l + men	76,82	69,94	73,22
oin + ak + dek + l + men + erreg	78,03	71,35	74,54

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Perpausen identifikazioa IA eta erregelak konbinatuz

	Doitasuna	**Estaldura**	F_1
erreg	50,84	48,63	49,71
oin + ak + dek + l + men	76,82	69,94	73,22
oin + ak + dek + l + men + erreg	78,03	71,35	**74,54**

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Ikasketa-corpusa handituz

	Doitasuna	**Estaldura**	F_1
Ikasketa-corpusaren % 25arekin	78,03	71,35	74,54
Ikasketa-corpusaren % 50arekin	78,70	74,21	76,39
Ikasketa-corpusaren % 100arekin	80,13	76,18	78,11

Taula: Corpusaren tamainaren arabera (% 100 = 104.956 token)

Ikasketa-corpusa handituz

	Doitasuna	**Estaldura**	F_1
Ikasketa-corpusaren % 25arekin	78,03	71,35	**74,54**
Ikasketa-corpusaren % 50arekin	78,70	74,21	**76,39**
Ikasketa-corpusaren % 100arekin	80,13	76,18	78,11

Taula: Corpusaren tamainaren arabera (% 100 = 104.956 token)

Ikasketa-corpusa handituz

	Doitasuna	**Estaldura**	F_1
Ikasketa-corpusaren % 25arekin	78,03	71,35	74,54
Ikasketa-corpusaren % 50arekin	78,70	74,21	**76,39**
Ikasketa-corpusaren % 100arekin	80,13	76,18	**78,11**

Taula: Corpusaren tamainaren arabera (% 100 = 104.956 token)

Emaitza idealak eskuz desanbiguatutako informazioarekin

Corpusa nola desanbiguatua	**Doit.**	**Est.**	F_1
Automatikoki	80,13	76,18	**78,11**
Eskuz	80,81	76,11	**78,39**

Taula: Informazio linguistikoaren kalitatearen arabera

Azken emaitzak test-corpusean

Proba-corpusa	**Doit.**	**Est.**	F_1
Garapen-corpusa	80,13	76,18	**78,11**
Test-corpusa	79,22	75,36	**77,24**

Taula: Garapen-corpuseko emaitzak vs test-corpuseko emaitzak

Aurkezpenaren eskema

Puntuazioa HPan

- Puntuazioaren garrantzia HPan
  - Komak esaldiaren sintaxian daukan garrantzia (Nunberg, 1990)
  - Puntuazioaren eragina analizatzaile sintaktiko batean (Briscoe eta Carroll, 1995)
  - *Workshop on punctuation in computational linguistics (1996)*

Puntuazioa HPan

- Puntuazioaren garrantzia HPan
  - Komak esaldiaren sintaxian daukan garrantzia (Nunberg, 1990)
  - Puntuazioaren eragina analizatzaile sintaktiko batean (Briscoe eta Carroll, 1995)
  - *Workshop on punctuation in computational linguistics (1996)*
- Lan esanguratsuenak
  - Komaren funtzioen sailkapena (Bayraktar et al., 1998)
  - Danierako koma-zuzentzailea (Hardt, 2001)
  - Komen rol sintaktikoen esleipena (Delden eta Gomez, 2002)
  - Komen berreskurapena ahotsaren ezagutzarako (Shieber eta Tao, 2003)
  - Puntuazioaren eta kasuaren berreskurapena (Baldwin eta Joseph, 2009)
  - Txekierako puntuazio-akatsak detektatzeko (Jakubicek eta Horak, 2010)

Puntuazioa HPan

- Puntuazioaren garrantzia HPan
  - Komak esaldiaren sintaxian daukan garrantzia (Nunberg, 1990)
  - Puntuazioaren eragina analizatzaile sintaktiko batean (Briscoe eta Carroll, 1995)
  - *Workshop on punctuation in computational linguistics (1996)*
- Lan esanguratsuenak
  - Komaren funtzioen sailkapena (Bayraktar et al., 1998)
  - Danierako koma-zuzentzailea (Hardt, 2001)
  - Komen rol sintaktikoen esleipena (Delden eta Gomez, 2002)
  - Komen berreskurapena ahotsaren ezagutzarako (Shieber eta Tao, 2003)
  - Puntuazioaren eta kasuaren berreskurapena (Baldwin eta Joseph, 2009)
  - Txekierako puntuazio-akatsak detektatzeko (Jakubicek eta Horak, 2010)

Puntuazio-zuzentzaile bat garatzeko arazoak

- Puntuazio-arauak hizkuntza askotan ez daude guztiz zehaztuta
- Puntuazio-marka fidagarriak (puntua, galdera-marka eta harridura-marka) vs ez-fidagarriak
- Komaren kasu berezia

Komaren erabilera: azterketa linguistikoa

- Aditz nagusiaren aurreko mintzagaiaren ondoren koma jartzea gomendatzen da (1), mintzagaia subjektua ez denean betiere (2). Aitzitik, behar-beharrezkoa da batzuetan, anbiguotasuna ekiditeko (3). Mintzagaia edo galdegaia (bietako bat, gutxienez) mendeko perpausa bada, mintzagaiaren ondoren koma jarri behar da derrigor (4,5) (Garzia, 1997).
  1. **Azkenean**, *zurekin joango naiz.*
  2. **Aita** *atzo iritsi zen.*
  3. **Batzuetan**, *irabazteko gogor jokatzea beharrezkoa da.*
  4. **Euria ari zuenez**, *etxean geratzea erabaki genuen.*
  5. **Gaur iritsiko zela esan zuen arren**, *atzo iritsi zen azkenean.*

Komaren erabilera: azterketa linguistikoa

- Aditz nagusiaren aurreko mintzagaiaren ondoren koma jartzea gomendatzen da (1), mintzagaia subjektua ez denean betiere (2). Aitzitik, behar-beharrezkoa da batzuetan, anbiguotasuna ekiditeko (3). **Mintzagaia edo galdegaia (bietako bat, gutxienez) mendeko perpausa bada, mintzagaiaren ondoren koma jarri behar da derrigor (4,5)** (Garzia, 1997).

1. **Azkenean**, *zurekin joango naiz.*
2. **Aita** *atzo iritsi zen.*
3. **Batzuetan**, *irabazteko gogor jokatzea beharrezkoa da.*
4. **Euria ari zuenez**, *etxean geratzea erabaki genuen.*
5. **Gaur iritsiko zela esan zuen arren**, *atzo iritsi zen azkenean.*

Komaren erabileraren konparaketa: euskara eta ingelesa

Euskarakoak	Ingelesekoak
Esaldi koordinatuak lotzean, juntagailuaren aurretik koma	Enumerazio batzuetan koma+juntagailua egitura
Enumerazioetan, osagaien artean koma	Enumerazioetan, osagaien artean koma
Deikiak komaz markatu	-
Estilo zuzeneko esaldiak komaz bereizi	Estilo zuzeneko esaldiak komaz bereizi
Aposizio ez-murriztaileak koma artean	Aposizio ez-murriztaileak koma artean
Tartekiak koma artean	Tartekiak koma artean
Lokailuak eta diskurtso-antolatzaileak koma artean	Izenen ondoko modifikatzaile ez-murriztaileak eta tartekiak koma artean
Aditz nagusiaren aurreko mintzagaiaren ondoren, koma	Esaldiaren hasieran, sarbide gisa doazen sintagmak edo perpausak komaz mugatuta
Aditz nagusiaren ondorengo perpaus zirkunstantzialen aurretik, koma	Esaldi bukaerako elementu osagarriak komaz bereizi
Aditza isildua dagoenean, komaz markatu	-

Taula: Koma jartzeko arauen konparazioa: euskara vs ingelesa

Komen zuzenketa hizkuntzaren ezagutzan oinarrituta

- Komak zuzentzeko 19 CG erregela:

```
MAP (&OKER_KOMA_FALTA_1_1)
    TARGET EDOZEIN_KAT
    IF (1 BAINA + JNT);
```

Komen zuzenketa hizkuntzaren ezagutzan oinarrituta

- Komak zuzentzeko 19 CG erregela:

```
MAP (&OKER_KOMA_FALTA_1_1)
    TARGET EDOZEIN_KAT
    IF (1 BAINA + JNT);
```

- Adibidez:
  "*Erabakia ez da bete beharrekoa baina ondorio politiko eta juridikoak izango ditu.*"

Komen zuzenketa hizkuntzaren ezagutzan oinarrituta

- Komak zuzentzeko 19 CG erregela:

```
MAP (&OKER_KOMA_FALTA_1_1)
    TARGET EDOZEIN_KAT
    IF (1 BAINA + JNT);
```

- Adibidez:
  "*Erabakia ez da bete beharrekoa, baina ondorio politiko eta juridikoak izango ditu.*"

Komen zuzenketa hizkuntzaren ezagutzan oinarrituta

- Komak zuzentzeko 19 CG erregela:

```
MAP (&OKER_KOMA_FALTA_1_1)
    TARGET EDOZEIN_KAT
    IF (1 BAINA + JNT);
```

- Adibidez:
  "*Erabakia ez da bete beharrekoa, baina ondorio politiko eta juridikoak izango ditu.*"
- Emaitzak:

	0			**1**		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Erregeletan oinarrituta	93,1	96,7	94,9	56,9	27,2	36,8

Taula: CG erregelekin lortutako komen berreskuratze-emaitzak. 0 klasea: atzetik komarik ez daramaten tokenak; 1 klasea: atzetik koma daramaten tokenak.

Komen zuzenketa hizkuntzaren ezagutzan oinarrituta

- Komak zuzentzeko 19 CG erregela:

```
MAP (&OKER_KOMA_FALTA_1_1)
    TARGET EDOZEIN_KAT
    IF (1 BAINA + JNT);
```

- Adibidez:
  "*Erabakia ez da bete beharrekoa, baina ondorio politiko eta juridikoak izango ditu.*"
- Emaitzak:

	0			**1**		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Erregeletan oinarrituta	93,1	96,7	94,9	56,9	27,2	**36,8**

Taula: CG erregelekin lortutako komen berreskuratze-emaitzak. 0 klasea: atzetik komarik ez daramaten tokenak; 1 klasea: atzetik koma daramaten tokenak.

Komen zuzenketa hizkuntzaren ezagutzan oinarrituta

- Komak zuzentzeko 19 CG erregela:

```
MAP (&OKER_KOMA_FALTA_1_1)
    TARGET EDOZEIN_KAT
    IF (1 BAINA + JNT);
```

- Adibidez:
  "*Erabakia ez da bete beharrekoa, baina ondorio politiko eta juridikoak izango ditu.*"
- Emaitzak:

	0			**1**		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Erregeletan oinarrituta	93,1	96,7	94,9	**56,9**	27,2	36,8

Taula: CG erregelekin lortutako komen berreskuratze-emaitzak. 0 klasea: atzetik komarik ez daramaten tokenak; 1 klasea: atzetik koma daramaten tokenak.

Komen zuzenketa ikasketa automatikoan oinarrituta

- Esperimentuen prestaketa
- Egindako saioak
- Komen zuzenketa erregelak eta ikasketa automatikoa konbinatuz
- Jatorrizko komen eragina saihesten
- Ebaluazio kualitatiboa
- Erroreen analisia

Esperimentuen prestaketa (I)

- Ikasketa-corpus gisa, komak ondo jarrita dituen edozein corpus erabil daiteke:

	Ikasketarako	**Garapenerako**	**Testerako**	**Guztira**
Tokenak	101.250	28.500	5.250	135.000

Taula: Komak ikasteko erabilitako *Euskaldunon Egunkaria*ko corpusa

Esperimentuen prestaketa (I)

- Ikasketa-corpus gisa, komak ondo jarrita dituen edozein corpus erabil daiteke:

	Ikasketarako	**Garapenerako**	**Testerako**	**Guztira**
Tokenak	101.250	28.500	5.250	135.000

Taula: Komak ikasteko erabilitako *Euskaldunon Egunkaria*ko corpusa

- Oinarrizko neurriak:

	0			**1**		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
baseline_%8	92,7	92,4	92,6	7,6	7,9	7,8
baseline_100	94,1	80,2	86,6	12,5	35,8	18,5
baseline_200	94,4	75,6	84,0	12,1	42,7	18,9
baseline_300	94,5	72,4	82,0	11,2	46,5	18,7
baseline_ikasketakoak	94,6	55,6	70,0	9,6	59,6	16,5

Esperimentuen prestaketa (I)

- Ikasketa-corpus gisa, komak ondo jarrita dituen edozein corpus erabil daiteke:

	Ikasketarako	**Garapenerako**	**Testerako**	**Guztira**
Tokenak	101.250	28.500	5.250	135.000

Taula: Komak ikasteko erabilitako *Euskaldunon Egunkaria*ko corpusa

- Oinarrizko neurriak:

	0			**1**		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
baseline_%8	92,7	92,4	**92,6**	7,6	7,9	7,8
baseline_100	94,1	80,2	86,6	12,5	35,8	18,5
baseline_200	94,4	75,6	84,0	12,1	42,7	18,9
baseline_300	94,5	72,4	82,0	11,2	46,5	18,7
baseline_ikasketakoak	94,6	55,6	70,0	9,6	59,6	16,5

Esperimentuen prestaketa (I)

- Ikasketa-corpus gisa, komak ondo jarrita dituen edozein corpus erabil daiteke:

	Ikasketarako	**Garapenerako**	**Testerako**	**Guztira**
Tokenak	101.250	28.500	5.250	135.000

Taula: Komak ikasteko erabilitako *Euskaldunon Egunkaria*ko corpusa

- Oinarrizko neurriak:

	0			**1**		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
baseline_%8	92,7	92,4	**92,6**	7,6	7,9	7,8
baseline_100	94,1	80,2	86,6	12,5	35,8	18,5
baseline_200	94,4	75,6	84,0	12,1	42,7	**18,9**
baseline_300	94,5	72,4	82,0	11,2	46,5	18,7
baseline_ikasketakoak	94,6	55,6	70,0	9,6	59,6	16,5

Esperimentuen prestaketa (II)

- Ikasi beharrekoa $\Rightarrow$ token bakoitzaren ondoren koma datorren ala ez. Horretarako:
  - Corpusetik jatorrizko komak kendu
  - Token bakoitzari azken atributu gisa gehitu: atzetik koma daraman ala ez

Esperimentuen prestaketa (III)

```
@RELATION komak_eu
@ATTRIBUTE hitza REAL
@ATTRIBUTE lema REAL
@ATTRIBUTE kat {-,ADB,ADI,ADJ,...}
...
@ATTRIBUTE puntu {0,1}
@ATTRIBUTE bi_puntu {0,1}
...
@ATTRIBUTE zenbat_ADK_ezk REAL
@ATTRIBUTE zenbat_ADK_esk REAL
...
@ATTRIBUTE koma {0,1}
```

Esperimentuen prestaketa (III)

```
@RELATION komak_eu
@ATTRIBUTE hitza REAL
@ATTRIBUTE lema REAL
@ATTRIBUTE kat {-,ADB,ADI,ADJ,...}
...
@ATTRIBUTE puntu {0,1}
@ATTRIBUTE bi_puntu {0,1}
...
@ATTRIBUTE zenbat_ADK_ezk REAL
@ATTRIBUTE zenbat_ADK_esk REAL
...
@ATTRIBUTE koma {0,1}

@DATA
380,235,ADB,ADOARR,-,-,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,4,0,1,1,18,0
2089,1054,LOT,LOK,-,-,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,4,0,1,2,17,1
638,1412,IZE,ARR,-,-,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,3,0,1,3,16,0
6459,669,ADJ,IZO,ABS,-,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,3,0,1,4,15,0
13,666,IZE,ARR,ERG,-,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,2,2,0,1,6,13,0
.....
```

Egindako saioak: leihoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
(-2,+5)	95,6	98,2	96,9	64,8	43,1	51,8
(-3,+5)	95,7	97,9	96,8	62,7	44,1	51,8
(-4,+5)	95,7	98,0	96,8	63,4	44,6	52,0
(-5,+5)	95,5	98,1	96,8	63,5	41,7	50,3
(-5,+4)	95,5	98,2	96,8	64,0	41,7	50,5
(-5,+3)	95,6	98,1	96,9	64,3	43,2	51,7
(-5,+2)	95,6	98,2	96,9	65,0	42,4	51,4
(-6,+2)	95,6	98,2	96,9	64,5	42,1	50,9
(-6,+3)	95,6	98,2	96,9	64,6	42,6	51,4
(-8,+2)	95,6	98,2	96,9	64,5	42,5	51,3
(-8,+3)	95,6	97,9	96,7	61,5	43,1	50,7
(-8,+8)	95,6	97,8	96,7	60,4	42,2	49,7

Taula: Leihoaren aukeraketa

Egindako saioak: leihoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
(-2,+5)	95,6	98,2	**96,9**	64,8	43,1	51,8
(-3,+5)	95,7	97,9	**96,8**	62,7	44,1	51,8
(-4,+5)	95,7	98,0	**96,8**	63,4	44,6	52,0
(-5,+5)	95,5	98,1	**96,8**	63,5	41,7	50,3
(-5,+4)	95,5	98,2	**96,8**	64,0	41,7	50,5
(-5,+3)	95,6	98,1	**96,9**	64,3	43,2	51,7
(-5,+2)	95,6	98,2	**96,9**	65,0	42,4	51,4
(-6,+2)	95,6	98,2	**96,9**	64,5	42,1	50,9
(-6,+3)	95,6	98,2	**96,9**	64,6	42,6	51,4
(-8,+2)	95,6	98,2	**96,9**	64,5	42,5	51,3
(-8,+3)	95,6	97,9	**96,7**	61,5	43,1	50,7
(-8,+8)	95,6	97,8	**96,7**	60,4	42,2	49,7

Taula: Leihoaren aukeraketa

Egindako saioak: leihoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
(-2,+5)	95,6	98,2	96,9	64,8	43,1	51,8
(-3,+5)	95,7	97,9	96,8	62,7	44,1	51,8
(-4,+5)	95,7	98,0	96,8	63,4	44,6	**52,0**
(-5,+5)	95,5	98,1	96,8	63,5	41,7	50,3
(-5,+4)	95,5	98,2	96,8	64,0	41,7	50,5
(-5,+3)	95,6	98,1	96,9	64,3	43,2	51,7
(-5,+2)	95,6	98,2	96,9	65,0	42,4	51,4
(-6,+2)	95,6	98,2	96,9	64,5	42,1	50,9
(-6,+3)	95,6	98,2	96,9	64,6	42,6	51,4
(-8,+2)	95,6	98,2	96,9	64,5	42,5	51,3
(-8,+3)	95,6	97,9	96,7	61,5	43,1	50,7
(-8,+8)	95,6	97,8	96,7	60,4	42,2	49,7

Taula: Leihoaren aukeraketa

Egindako saioak: leihoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
(-2,+5)	95,6	98,2	96,9	64,8	43,1	51,8
(-3,+5)	95,7	97,9	96,8	62,7	44,1	51,8
(-4,+5)	95,7	98,0	96,8	63,4	44,6	**52,0**
(-5,+5)	95,5	98,1	96,8	63,5	41,7	50,3
(-5,+4)	95,5	98,2	96,8	64,0	41,7	50,5
(-5,+3)	95,6	98,1	96,9	64,3	43,2	51,7
(-5,+2)	95,6	98,2	96,9	**65,0**	42,4	51,4
(-6,+2)	95,6	98,2	96,9	64,5	42,1	50,9
(-6,+3)	95,6	98,2	96,9	64,6	42,6	51,4
(-8,+2)	95,6	98,2	96,9	64,5	42,5	51,3
(-8,+3)	95,6	97,9	96,7	61,5	43,1	50,7
(-8,+8)	95,6	97,8	96,7	60,4	42,2	49,7

Taula: Leihoaren aukeraketa

Egindako saioak: leihoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
(-2,+5)	95,6	98,2	96,9	64,8	43,1	51,8
(-3,+5)	95,7	97,9	96,8	62,7	44,1	51,8
(-4,+5)	95,7	98,0	96,8	**63,4**	44,6	**52,0**
(-5,+5)	95,5	98,1	96,8	63,5	41,7	50,3
(-5,+4)	95,5	98,2	96,8	64,0	41,7	50,5
(-5,+3)	95,6	98,1	96,9	64,3	43,2	51,7
(-5,+2)	95,6	98,2	96,9	**65,0**	42,4	**51,4**
(-6,+2)	95,6	98,2	96,9	64,5	42,1	50,9
(-6,+3)	95,6	98,2	96,9	64,6	42,6	51,4
(-8,+2)	95,6	98,2	96,9	64,5	42,5	51,3
(-8,+3)	95,6	97,9	96,7	61,5	43,1	50,7
(-8,+8)	95,6	97,8	96,7	60,4	42,2	49,7

Taula: Leihoaren aukeraketa

Egindako saioak: algoritmoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
C4.5	95,6	98,2	96,9	65,2	42,4	51,4
Naive Bayes	94,8	95,6	95,2	37,6	33,5	35,5
SVM	93,6	99,4	96,5	67,2	14,3	23,6

Taula: Ikasketa-algoritmoaren aukeraketa

Egindako saioak: algoritmoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
C4.5	95,6	98,2	96,9	65,2	42,4	**51,4**
Naive Bayes	94,8	95,6	95,2	37,6	33,5	35,5
SVM	93,6	99,4	96,5	67,2	14,3	23,6

Taula: Ikasketa-algoritmoaren aukeraketa

Egindako saioak: algoritmoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
C4.5	95,6	98,2	96,9	65,2	42,4	**51,4**
Naive Bayes	94,8	95,6	95,2	37,6	33,5	35,5
SVM	93,6	99,4	96,5	**67,2**	14,3	23,6

Taula: Ikasketa-algoritmoaren aukeraketa

Egindako saioak: algoritmoaren aukeraketa

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
C4.5	95,6	98,2	96,9	**65,2**	42,4	**51,4**
Naive Bayes	94,8	95,6	95,2	37,6	33,5	35,5
SVM	93,6	99,4	96,5	**67,2**	14,3	**23,6**

Taula: Ikasketa-algoritmoaren aukeraketa

Egindako saioak: corpus motaren eragina

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Egunkaria	93,0	97,9	95,4	48,7	21,4	29,7
Filosofia-itzulpena	93,0	97,0	95,0	60,4	38,8	47,3
Literatura	92,7	97,5	95,0	50,4	25,0	33,4
Zientzia eta teknika	94,9	98,5	96,6	49,5	22,4	30,8

Taula: *Cross-validation* emaitzak corpus motaren arabera (25.000 tokeneko corpusak)

Egindako saioak: corpus motaren eragina

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Egunkaria	93,0	97,9	95,4	48,7	21,4	29,7
Filosofia-itzulpena	93,0	97,0	95,0	60,4	38,8	**47,3**
Literatura	92,7	97,5	95,0	50,4	25,0	33,4
Zientzia eta teknika	94,9	98,5	96,6	49,5	22,4	30,8

Taula: *Cross-validation* emaitzak corpus motaren arabera (25.000 tokeneko corpusak)

Egindako saioak: ingeleseko corpusarekin

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Euskara	95,6	98,2	96,9	65,2	42,4	51,4
Ingelesa	97,8	99,7	98,7	83,3	38,7	52,8

Taula: Hizkuntzaren araberako emaitzak: euskara vs ingelesa

Egindako saioak: ingeleseko corpusarekin

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Euskara	95,6	98,2	**96,9**	65,2	42,4	**51,4**
Ingelesa	97,8	99,7	**98,7**	83,3	38,7	**52,8**

Taula: Hizkuntzaren araberako emaitzak: euskara vs ingelesa

Egindako saioak: atributu berriekin

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Atributu berririk gabe	95,6	98,2	96,9	65,2	42,4	51,4
(1) 300 atributu berri	96,0	98,3	97,2	69,6	48,6	57,2
(2) 3 atributu berri	96,0	98,1	97,0	66,5	48,1	55,8

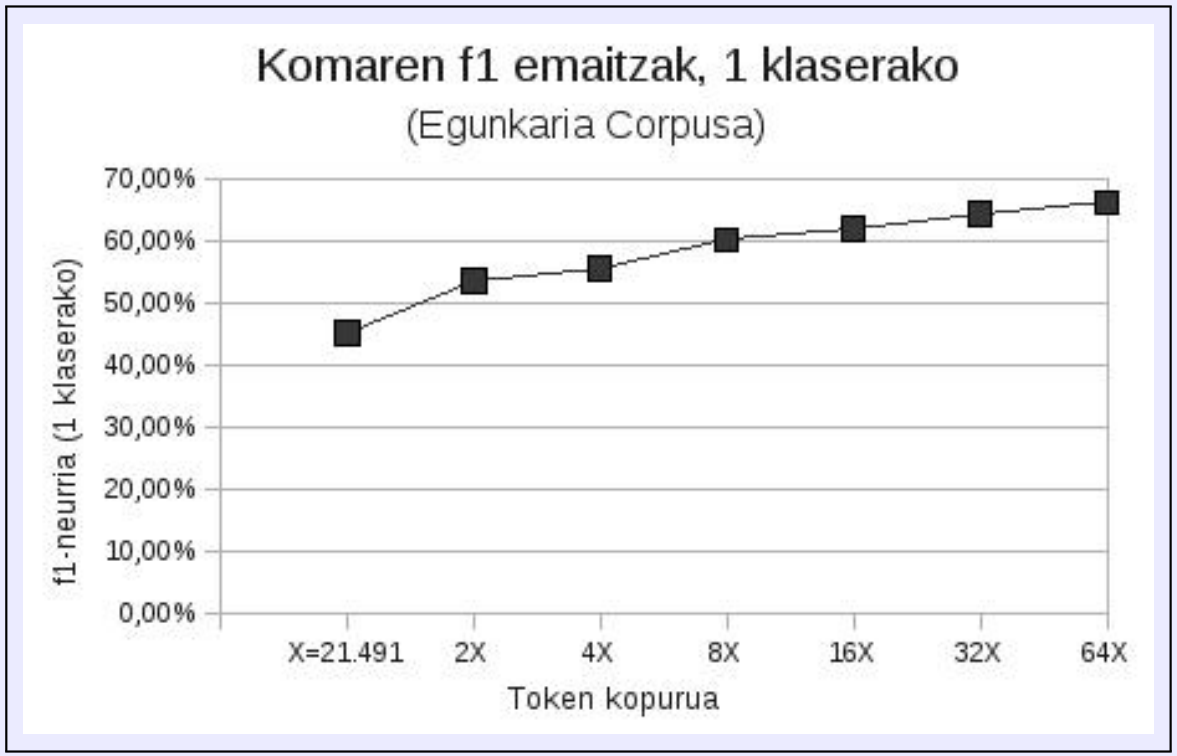
Taula: Komen aurretik maizen agertzen diren hitzak atributu gisa gehituz

Egindako saioak: atributu berriekin

	0			1		
	Doit.	**Est.**	F_1	**Doit.**	**Est.**	F_1
Atributu berririk gabe	95,6	98,2	96,9	65,2	42,4	51,4
(1) 300 atributu berri	96,0	98,3	97,2	**69,6**	**48,6**	**57,2**
(2) 3 atributu berri	96,0	98,1	97,0	66,5	48,1	55,8

Taula: Komen aurretik maizen agertzen diren hitzak atributu gisa gehituz

Egindako saioak: corpusaren tamainaren eragina



Irudia: Corpusaren tamainaren eragina (hamar zatiko *cross-validation* baliatuta)

Egindako saioak: kateen eta perpausen informazioa gehituz

Ikasketa-corpusa	**Test-corpusa**	**Desanb.**	F_1 *neurria*
Komaduna	*Komaduna*	autom.	83,17
Komagabea	*Komagabea*	autom.	82,24

Taula: Komaren eragina *FR-P* bidezko **kate-identifikatzailean**

Egindako saioak: kateen eta perpausen informazioa gehituz

Ikasketa-corpusa	**Test-corpusa**	**Desanb.**	F_1 *neurria*
Komaduna	*Komaduna*	autom.	83,17
Komagabea	*Komagabea*	autom.	82,24

Taula: Komaren eragina *FR-P* bidezko **kate-identifikatzailean**

Ikasketa-corpusa	**Test-corpusa**	**Desanb.**	F_1 *neurria*
Komaduna	*Komaduna*	autom.	77,24
Komagabea	*Komagabea*	autom.	73,66

Taula: Komaren eragina *FR-P* bidezko **perpaus-identifikatzailean**

Egindako saioak: kateen eta perpausen informazioa gehituz

	1		
	Doit.	**Est.**	F_1
Kate-info. eta perpaus-info. gabe	69,6	48,6	57,2
Kate-ident komagabearen info. gehituta	70,4	48,5	57,4
Kate- eta perpaus-ident komagabeen info. gehituta	76,6	55,7	64,5
Kate-ident komadunaren info. gehituta	73,0	50,7	59,8
Kate- eta perpaus-ident komadunen info. gehituta	78,4	59,8	67,9

Taula: Komen zuzenketa, kate- eta perpaus-identifikatzaile komadunen edo komagabeen informazioa gehituta

Egindako saioak: kateen eta perpausen informazioa gehituz

	1		
	Doit.	**Est.**	F_1
Kate-info. eta perpaus-info. gabe	69,6	48,6	57,2
Kate-ident komagabearen info. gehituta	70,4	48,5	**57,4**
Kate- eta perpaus-ident komagabeen info. gehituta	76,6	55,7	64,5
Kate-ident komadunaren info. gehituta	73,0	50,7	59,8
Kate- eta perpaus-ident komadunen info. gehituta	78,4	59,8	67,9

Taula: Komen zuzenketa, kate- eta perpaus-identifikatzaile komadunen edo komagabeen informazioa gehituta

Egindako saioak: kateen eta perpausen informazioa gehituz

	1		
	Doit.	**Est.**	F_1
Kate-info. eta perpaus-info. gabe	69,6	48,6	57,2
Kate-ident komagabearen info. gehituta	70,4	48,5	57,4
Kate- eta perpaus-ident komagabeen info. gehituta	76,6	55,7	**64,5**
Kate-ident komadunaren info. gehituta	73,0	50,7	59,8
Kate- eta perpaus-ident komadunen info. gehituta	78,4	59,8	67,9

Taula: Komen zuzenketa, kate- eta perpaus-identifikatzaile komadunen edo komagabeen informazioa gehituta

Egindako saioak: kateen eta perpausen informazioa gehituz

	1		
	Doit.	**Est.**	F_1
Kate-info. eta perpaus-info. gabe	69,6	48,6	57,2
Kate-ident komagabearen info. gehituta	70,4	48,5	57,4
Kate- eta perpaus-ident komagabeen info. gehituta	76,6	55,7	64,5
Kate-ident komadunaren info. gehituta	73,0	50,7	**59,8**
Kate- eta perpaus-ident komadunen info. gehituta	78,4	59,8	67,9

Taula: Komen zuzenketa, kate- eta perpaus-identifikatzaile komadunen edo komagabeen informazioa gehituta

Egindako saioak: kateen eta perpausen informazioa gehituz

	1		
	Doit.	**Est.**	F_1
Kate-info. eta perpaus-info. gabe	69,6	48,6	57,2
Kate-ident komagabearen info. gehituta	70,4	48,5	57,4
Kate- eta perpaus-ident komagabeen info. gehituta	76,6	55,7	64,5
Kate-ident komadunaren info. gehituta	73,0	50,7	59,8
Kate- eta perpaus-ident komadunen info. gehituta	78,4	59,8	**67,9**

Taula: Komen zuzenketa, kate- eta perpaus-identifikatzaile komadunen edo komagabeen informazioa gehituta

Komen zuzenketa erregelak eta ikasketa automatikoa konbinatuz

	1		
	Doit.	**Est.**	F_1
CG erregelak	56,9	27,2	36,8
Ikasketa automatikoa KPI-komagabearekin	76,6	55,7	64,5
CG erregelak + ikask. autom. KPI-komagabearekin	77,8	55,0	64,4
Ikask. autom. KPI-komadunarekin	78,4	59,8	67,9
CG erregelak + ikask. autom. KPI-komadunarekin	79,0	61,4	69,1

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Komen zuzenketa erregelak eta ikasketa automatikoa konbinatuz

	1		
	Doit.	**Est.**	F_1
CG erregelak	56,9	27,2	36,8
Ikasketa automatikoa KPI-komagabearekin	**76,6**	55,7	**64,5**
CG erregelak + ikask. autom. KPI-komagabearekin	**77,8**	55,0	**64,4**
Ikask. autom. KPI-komadunarekin	78,4	59,8	67,9
CG erregelak + ikask. autom. KPI-komadunarekin	79,0	61,4	69,1

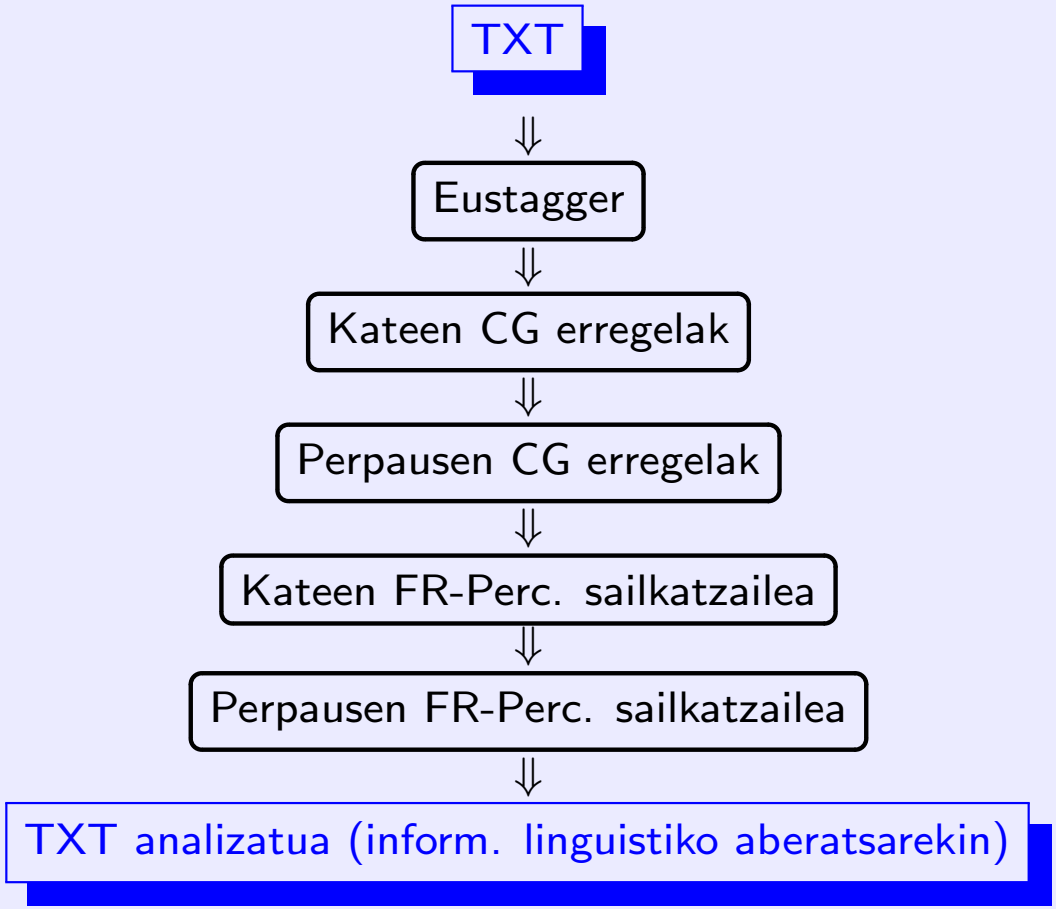
Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

Komen zuzenketa erregelak eta ikasketa automatikoa konbinatuz

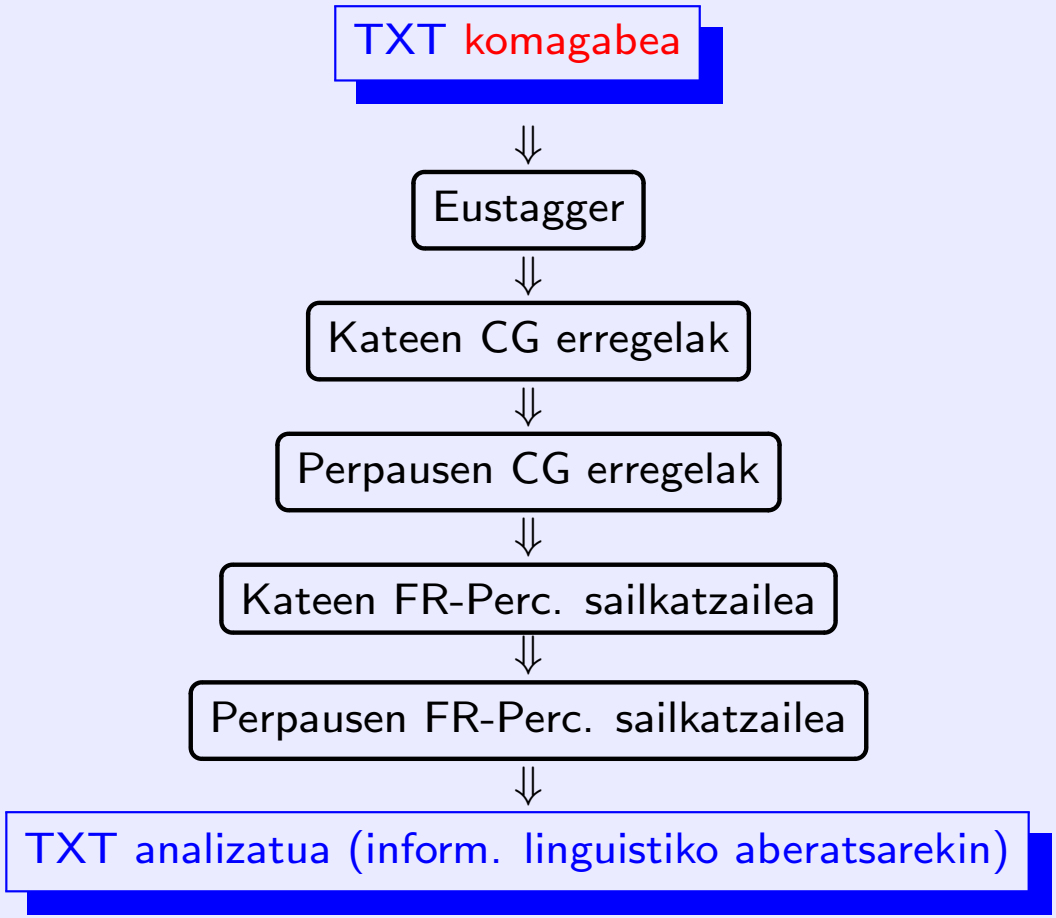
	1		
	Doit.	**Est.**	F_1
CG erregelak	56,9	27,2	36,8
Ikasketa automatikoa KPI-komagabearekin	76,6	55,7	64,5
CG erregelak + ikask. autom. KPI-komagabearekin	77,8	55,0	64,4
Ikask. autom. KPI-komadunarekin	78,4	59,8	**67,9**
CG erregelak + ikask. autom. KPI-komadunarekin	79,0	61,4	**69,1**

Taula: Erregelen informazioa gehitzea, pilaratzearen bidez

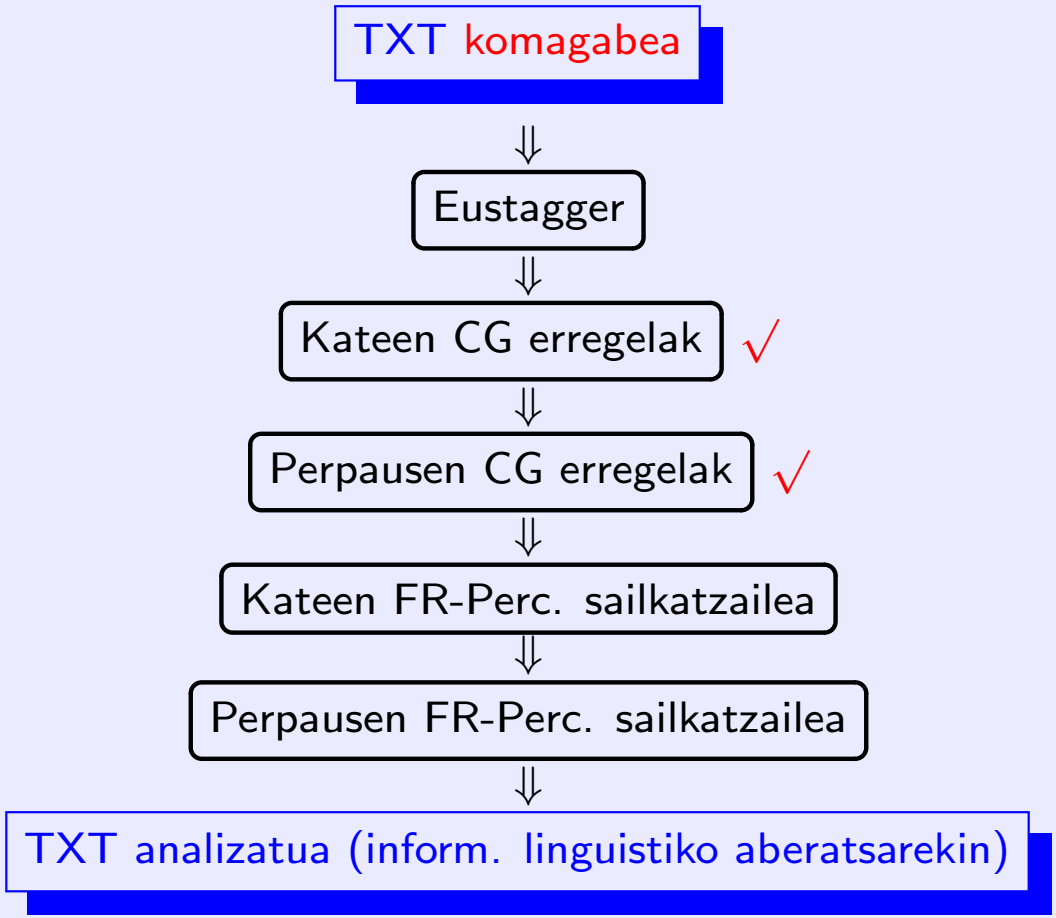
Jatorrizko komen eragina saihesten



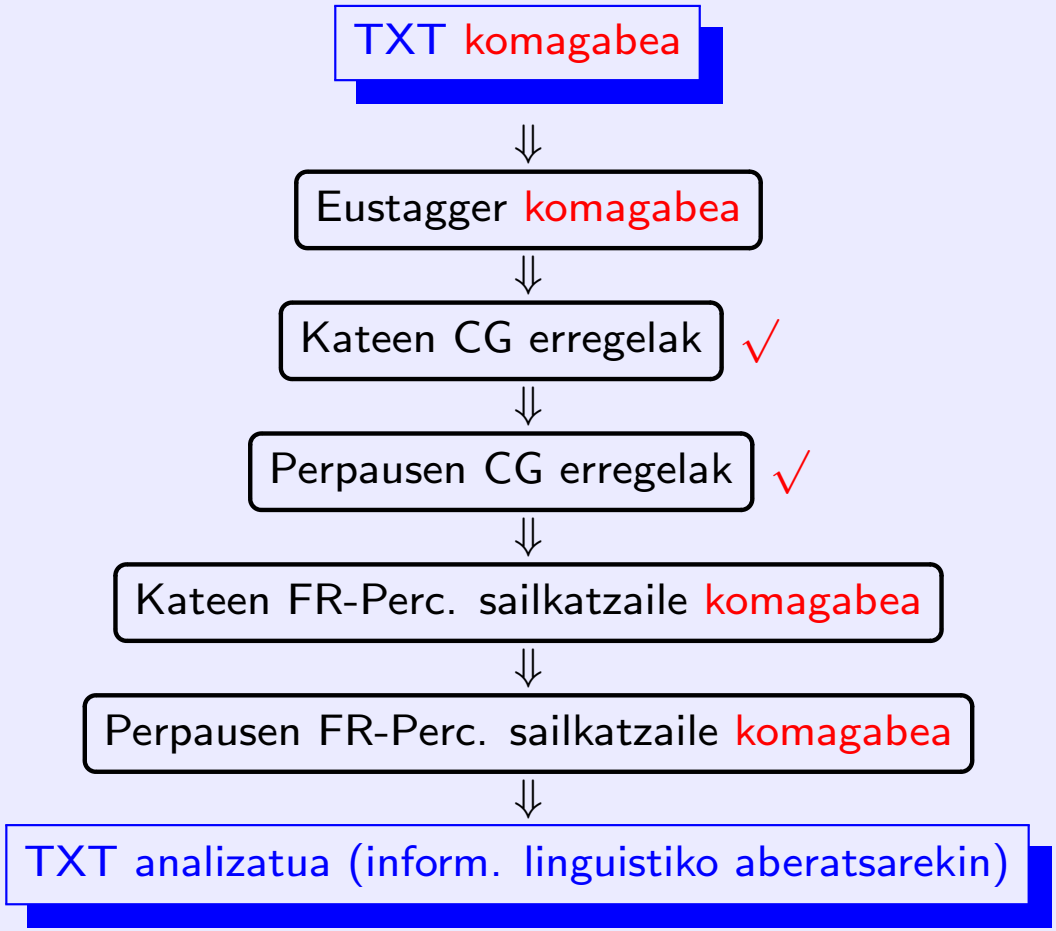
Jatorrizko komen eragina saihesten



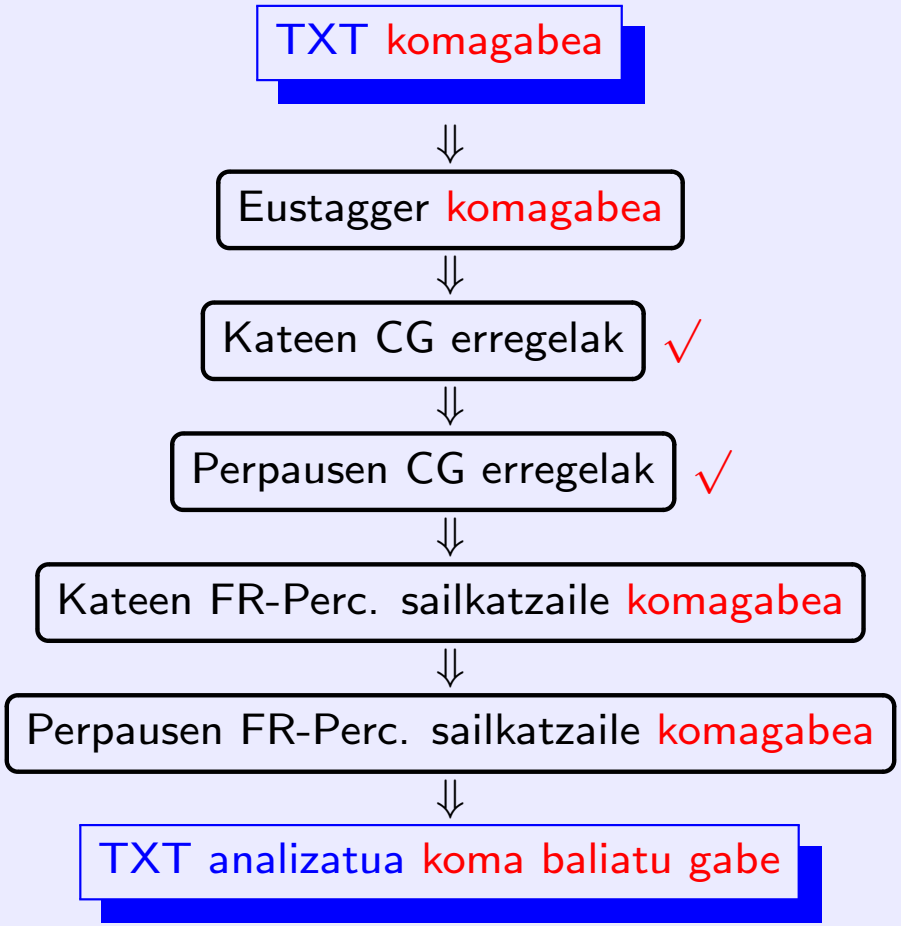
Jatorrizko komen eragina saihesten



Jatorrizko komen eragina saihesten



Jatorrizko komen eragina saihesten



Jatorrizko komen eragina saihesten

	1		
	Doit.	**Est.**	F_1
Corpus komaduna + Eustagger komaduna	77,8	55,0	64,4
Corpus komagabea + Eustagger komagabea	69,3	33,3	45,0

Taula: Eustagger komagabeak edo komadunak emandako informazio linguistikoaren arabera

Jatorrizko komen eragina saihesten

	1		
	Doit.	**Est.**	F_1
Corpus komaduna + Eustagger komaduna	77,8	55,0	**64,4**
Corpus komagabea + Eustagger komagabea	69,3	33,3	**45,0**

Taula: Eustagger komagabeak edo komadunak emandako informazio linguistikoaren arabera

Jatorrizko komen eragina saihesten

	1		
	Doit.	**Est.**	F_1
Corpus komaduna + Eustagger komaduna	77,8	55,0	**64,4**
Corpus komagabea + Eustagger komagabea	**69,3**	33,3	**45,0**

Taula: Eustagger komagabeak edo komadunak emandako informazio linguistikoaren arabera

Ebaluazio kualitatiboa

	1		
	Doit.	**Est.**	F_1
Ikask. autom. KPI-komagabe + CG erregelak	77,6	52,7	62,8
Hizkuntzalari1	79,1	85,9	82,3
Hizkuntzalari2	76,1	76,4	76,3

Taula: Test-corpuseko emaitzak vs hizkuntzalarien etiketatzearen emaitzak

Ebaluazio kualitatiboa

	1		
	Doit.	**Est.**	F_1
Ikask. autom. KPI-komagabe + CG erregelak	77,6	52,7	**62,8**
Hizkuntzalari1	79,1	85,9	82,3
Hizkuntzalari2	76,1	76,4	**76,3**

Taula: Test-corpuseko emaitzak vs hizkuntzalarien etiketatzearen emaitzak

Ebaluazio kualitatiboa

	1		
	Doit.	**Est.**	F_1
Ikask. autom. KPI-komagabe + CG erregelak	**77,6**	52,7	62,8
Hizkuntzalari1	79,1	85,9	82,3
Hizkuntzalari2	**76,1**	76,4	76,3

Taula: Test-corpuseko emaitzak vs hizkuntzalarien etiketatzearen emaitzak

Ebaluazio kualitatiboa

	1		
	Doit.	**Est.**	F_1
Ikask. autom. KPI-komagabe + CG erregelak	77,6	52,7	62,8
Hizkuntzalari1	79,1	85,9	82,3
Hizkuntzalari2	76,1	76,4	**76,3**

Taula: Test-corpuseko emaitzak vs hizkuntzalarien etiketatzearen emaitzak

	1		
	Doit.	**Est.**	F_1
Hizkuntzalari1, hizkuntzalari2-rekiko	73,8	79,7	**76,6**

Taula: Hizkuntzalarien arteko adostasuna

Ebaluazio kualitatiboa

	1		
	Doit.	**Est.**	F_1
Ikask. autom. KPI-komagabearekin + CG erregelak	77,6	52,7	62,8
Ebaluazio kualitatiboa (hiru erref.)	**83,01**	58,46	**68,61**

Taula: Sailkatzailearen bateragarritasuna test-corpusarekiko eta hiru erreferentziekiko

Erroreen analisia

1. *Gurean igandeko egunkariek aste osoa ematen dute komunetik&SOBRAN bueltaka eta jiraka&FALTAN orain toalleroan&FALTAN orain erradiadorean.*
2. *Pasa ziren egiazko ospakizunak eta itxurazkoak&FALTAN etorri ziren lehen adierazpenak eta hasierako azterketak&KOMA argazkiak eta gezur ezkutatuak.*
3. *Batzuen ustez&KOMA inora ez doana&KOMA ezer egiten ez duena&KOMA eta beste batzuentzat&FALTAN gehiegi egiten duena&KOMA urrutiegi eta arinegi doana.*
4. *UEUk&FALTAN EIREk&FALTAN Euskal Adarrak&KOMA Barrutiak eta Uniekimenak "eztabaidaren erdigunean" jarri nahi dute aldarrikapena.*

Aurkezpenaren eskema

Ekarpen garrantzitsuenak

- Euskarako kate- eta perpaus-identifikatzailearen garapena

Ekarpen garrantzitsuenak

- Euskarako kate- eta perpaus-identifikatzailearen garapena
  - Analisi sintaktiko osoa bideratzeko
  - HPko hainbat arlotarako: informazioaren erauzketa, laburpenen sorkuntza, itzulpen automatikoa...
  - Koma-zuzentzailerako

Ekarpen garrantzitsuenak

- Euskarako kate- eta perpaus-identifikatzailearen garapena
  - Analisi sintaktiko osoa bideratzeko
  - HPko hainbat arlotarako: informazioaren erauzketa, laburpenen sorkuntza, itzulpen automatikoa...
  - Koma-zuzentzailerako
- Euskarako koma-zuzentzaile automatikoaren garapena

Ekarpen garrantzitsuenak

- Euskarako kate- eta perpaus-identifikatzailearen garapena
  - Analisi sintaktiko osoa bideratzeko
  - HPko hainbat arlotarako: informazioaren erauzketa, laburpenen sorkuntza, itzulpen automatikoa...
  - Koma-zuzentzailerako
- Euskarako koma-zuzentzaile automatikoaren garapena
  - Estilo- eta gramatika-zuzentzailean txertatzeko
  - Analizatzaile eta desanbiguatzaile sintaktiko automatikoak hobetzeko
  - Ahotsaren ezagutza-sistemetan integratzeko

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	51,48
Baseline-a	-	Autom.	52,00
FR-P oin	*% 25*	Autom.	69,58
FR-P oin + dek	*% 25*	Autom.	78,77
FR-P oin + dek + Erreg	*% 25*	Autom.	79,21
FR-P oin + dek + Erreg	*% 100*	Autom.	82,64
FR-P oin + dek + Erreg	*% 100*	Eskuz	90,52

Taula: **Garapen-corpusean** egindako ebaluazioa

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	**51,48**
Baseline-a	-	Autom.	**52,00**
FR-P oin	*% 25*	Autom.	**69,58**
FR-P oin + dek	*% 25*	Autom.	78,77
FR-P oin + dek + Erreg	*% 25*	Autom.	79,21
FR-P oin + dek + Erreg	*% 100*	Autom.	82,64
FR-P oin + dek + Erreg	*% 100*	Eskuz	90,52

Taula: **Garapen-corpusean** egindako ebaluazioa

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	51,48
Baseline-a	-	Autom.	52,00
FR-P oin	*% 25*	Autom.	**69,58**
FR-P oin + dek	*% 25*	Autom.	**78,77**
FR-P oin + dek + Erreg	*% 25*	Autom.	79,21
FR-P oin + dek + Erreg	*% 100*	Autom.	82,64
FR-P oin + dek + Erreg	*% 100*	Eskuz	90,52

Taula: **Garapen-corpusean** egindako ebaluazioa

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	51,48
Baseline-a	-	Autom.	52,00
FR-P oin	*% 25*	Autom.	69,58
FR-P oin + dek	*% 25*	Autom.	**78,77**
FR-P oin + dek + Erreg	*% 25*	Autom.	**79,21**
FR-P oin + dek + Erreg	*% 100*	Autom.	82,64
FR-P oin + dek + Erreg	*% 100*	Eskuz	90,52

Taula: **Garapen-corpusean** egindako ebaluazioa

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	51,48
Baseline-a	-	Autom.	52,00
FR-P oin	*% 25*	Autom.	69,58
FR-P oin + dek	*% 25*	Autom.	78,77
FR-P oin + dek + Erreg	*% 25*	Autom.	**79,21**
FR-P oin + dek + Erreg	*% 100*	Autom.	**82,64**
FR-P oin + dek + Erreg	*% 100*	Eskuz	90,52

Taula: **Garapen-corpusean** egindako ebaluazioa

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	51,48
Baseline-a	-	Autom.	52,00
FR-P oin	*% 25*	Autom.	69,58
FR-P oin + dek	*% 25*	Autom.	78,77
FR-P oin + dek + Erreg	*% 25*	Autom.	79,21
FR-P oin + dek + Erreg	*% 100*	Autom.	**82,64**
FR-P oin + dek + Erreg	*% 100*	Eskuz	**90,52**

Taula: **Garapen-corpusean** egindako ebaluazioa

Kate-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	51,48
Baseline-a	-	Autom.	52,00
FR-P oin	*% 25*	Autom.	69,58
FR-P oin + dek	*% 25*	Autom.	78,77
FR-P oin + dek + Erreg	*% 25*	Autom.	79,21
FR-P oin + dek + Erreg	*% 100*	Autom.	**82,64**
FR-P oin + dek + Erreg	*% 100*	Eskuz	90,52

Taula: **Garapen-corpusean** egindako ebaluazioa

Hizkuntza	Teknika	Desanbiguatua	F_1
Euskara	*FR-P* oin + dek + Erreg	Autom.	**83,17**
Ingelesa	*FR-P*	Autom.	93,74

Taula: **Test-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	**Desanbiguatua**	F_1
Erreg	-	Autom.	49,71
Baseline-a	-	Autom.	48,79
FR-P oin	*% 25*	Autom.	69,99
FR-P oin+ak+dek+l+m	*% 25*	Autom.	73,22
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	74,54
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	78,11
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	78,39

Taula: **Garapen-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	**49,71**
Baseline-a	-	Autom.	**48,79**
FR-P oin	*% 25*	Autom.	**69,99**
FR-P oin+ak+dek+l+m	*% 25*	Autom.	73,22
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	74,54
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	78,11
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	78,39

Taula: **Garapen-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	49,71
Baseline-a	-	Autom.	48,79
FR-P oin	*% 25*	Autom.	**69,99**
FR-P oin+ak+dek+l+m	*% 25*	Autom.	**73,22**
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	74,54
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	78,11
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	78,39

Taula: **Garapen-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	49,71
Baseline-a	-	Autom.	48,79
FR-P oin	*% 25*	Autom.	69,99
FR-P oin+ak+dek+l+m	*% 25*	Autom.	**73,22**
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	**74,54**
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	78,11
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	78,39

Taula: **Garapen-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	49,71
Baseline-a	-	Autom.	48,79
FR-P oin	*% 25*	Autom.	69,99
FR-P oin+ak+dek+l+m	*% 25*	Autom.	73,22
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	**74,54**
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	**78,11**
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	78,39

Taula: **Garapen-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	49,71
Baseline-a	-	Autom.	48,79
FR-P oin	*% 25*	Autom.	69,99
FR-P oin+ak+dek+l+m	*% 25*	Autom.	73,22
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	74,54
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	**78,11**
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	**78,39**

Taula: **Garapen-corpusean** egindako ebaluazioa

Perpaus-identifikatzailea

	Ik. corpusa	Desanbiguatua	F_1
Erreg	-	Autom.	49,71
Baseline-a	-	Autom.	48,79
FR-P oin	*% 25*	Autom.	69,99
FR-P oin+ak+dek+l+m	*% 25*	Autom.	73,22
FR-P oin+ak+dek+l+m+Erreg	*% 25*	Autom.	74,54
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Autom.	**78,11**
FR-P oin+ak+dek+l+m+Erreg	*% 100*	Eskuz	78,39

Taula: **Garapen-corpusean** egindako ebaluazioa

Hizkuntza	Teknika	Desanbiguatua	F_1
Euskara	*FR-P* oin+ak+dek+l+m+e+Erreg	Autom.	**77,24**
Ingelesa	*FR-P* oin	Autom.	84,36

Taula: **Test-corpusean** egindako ebaluazioa

Euskarako koma-zuzentzailea

- Komari buruzko azterketa teorikoa
- Euskara vs ingelesa
- Kate- eta perpaus-identifikatzaileak baliatu ditugu, koma-zuzentzailea hobetzeko
- Ebaluazio kualitatiboa egin dugu: 3 erreferentzia
- Eustagger-en koma erabiltzearen eragina aztertu dugu

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Baseline-200	12,1	42,7	18,9
Err	56,9	27,2	36,8
C4.5, -5+2	65,2	42,4	51,4
C4.5, -5+2, 300 atrib.	69,6	48,6	57,2
C4.5, -5+2, 300 atrib., KPI komagabearekin	76,6	55,7	64,5
Err + C4.5, -5+2, 300 atrib., KPI komagabearekin	77,8	55,0	64,4
C4.5, -5+2, 300 atrib., KPI komadunarekin	78,4	59,8	67,9
Err + C4.5, -5+2, 300 atrib., KPI komadunarekin	79,0	61,4	69,1
Err + C4.5, -5+2, 300 atrib., KPI eta Eustagger komagabeekin	69,3	33,3	45,0

Taula: Koma-zuzentzailearen emaitzen laburpena **garapen-corpusean**

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Baseline-200	12,1	42,7	**18,9**
Err	56,9	27,2	**36,8**
C4.5, -5+2	65,2	42,4	**51,4**
C4.5, -5+2, 300 atrib.	69,6	48,6	57,2
C4.5, -5+2, 300 atrib., KPI komagabearekin	76,6	55,7	64,5
Err + C4.5, -5+2, 300 atrib., KPI komagabearekin	77,8	55,0	64,4
C4.5, -5+2, 300 atrib., KPI komadunarekin	78,4	59,8	67,9
Err + C4.5, -5+2, 300 atrib., KPI komadunarekin	79,0	61,4	69,1
Err + C4.5, -5+2, 300 atrib., KPI eta Eustagger komagabeekin	69,3	33,3	45,0

Taula: Koma-zuzentzailearen emaitzen laburpena **garapen-corpusean**

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Baseline-200	12,1	42,7	18,9
Err	56,9	27,2	36,8
C4.5, -5+2	65,2	42,4	**51,4**
C4.5, -5+2, 300 atrib.	69,6	48,6	**57,2**
C4.5, -5+2, 300 atrib., KPI komagabearekin	76,6	55,7	64,5
Err + C4.5, -5+2, 300 atrib., KPI komagabearekin	77,8	55,0	64,4
C4.5, -5+2, 300 atrib., KPI komadunarekin	78,4	59,8	67,9
Err + C4.5, -5+2, 300 atrib., KPI komadunarekin	79,0	61,4	69,1
Err + C4.5, -5+2, 300 atrib., KPI eta Eustagger komagabeekin	69,3	33,3	45,0

Taula: Koma-zuzentzailearen emaitzen laburpena **garapen-corpusean**

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Baseline-200	12,1	42,7	18,9
Err	56,9	27,2	36,8
C4.5, -5+2	65,2	42,4	51,4
C4.5, -5+2, 300 atrib.	69,6	48,6	**57,2**
C4.5, -5+2, 300 atrib., KPI komagabearekin	76,6	55,7	**64,5**
Err + C4.5, -5+2, 300 atrib., KPI komagabearekin	**77,8**	55,0	64,4
C4.5, -5+2, 300 atrib., KPI komadunarekin	78,4	59,8	67,9
Err + C4.5, -5+2, 300 atrib., KPI komadunarekin	79,0	61,4	69,1
Err + C4.5, -5+2, 300 atrib., KPI eta Eustagger komagabeekin	69,3	33,3	45,0

Taula: Koma-zuzentzailearen emaitzen laburpena **garapen-corpusean**

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Baseline-200	12,1	42,7	18,9
Err	56,9	27,2	36,8
C4.5, -5+2	65,2	42,4	51,4
C4.5, -5+2, 300 atrib.	69,6	48,6	**57,2**
C4.5, -5+2, 300 atrib., KPI komagabearekin	76,6	55,7	64,5
Err + C4.5, -5+2, 300 atrib., KPI komagabearekin	77,8	55,0	64,4
C4.5, -5+2, 300 atrib., KPI komadunarekin	78,4	59,8	**67,9**
Err + C4.5, -5+2, 300 atrib., KPI komadunarekin	79,0	61,4	**69,1**
Err + C4.5, -5+2, 300 atrib., KPI eta Eustagger komagabeekin	69,3	33,3	45,0

Taula: Koma-zuzentzailearen emaitzen laburpena **garapen-corpusean**

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Baseline-200	12,1	42,7	18,9
Err	56,9	27,2	36,8
C4.5, -5+2	65,2	42,4	51,4
C4.5, -5+2, 300 atrib.	69,6	48,6	57,2
C4.5, -5+2, 300 atrib., KPI komagabearekin	76,6	55,7	64,5
Err + C4.5, -5+2, 300 atrib., KPI komagabearekin	77,8	55,0	**64,4**
C4.5, -5+2, 300 atrib., KPI komadunarekin	78,4	59,8	67,9
Err + C4.5, -5+2, 300 atrib., KPI komadunarekin	79,0	61,4	**69,1**
Err + C4.5, -5+2, 300 atrib., KPI eta Eustagger komagabeekin	69,3	33,3	**45,0**

Taula: Koma-zuzentzailearen emaitzen laburpena **garapen-corpusean**

Euskarako koma-zuzentzailea

	1		
	Doit.	**Est.**	F_1
Hizkuntz.2: goi-muga	76,1	76,4	76,3
Ikask. autom. KPI komagabearekin + CG erregelak	77,6	52,7	62,8
Ebaluazio kualitatiboa	**83,01**	58,46	**68,61**

Taula: Koma zuzentzailearen emaitzak test-corpusean

Ondorioak

- Ikasketa automatikoko teknikak (FR-Perceptron), baliagarriak
- Kateak identifikatzea, perpausak baino errazago
- Euskarako kateak eta perpausak, ingelesekoak baino zailago
- Komaren erabilera antzekoa da euskaraz eta ingelesez
- Kate- eta perpaus-identifikatzaileen informazioa esanguratsua koma-zuzentzailearentzat
- Corpusean eta hizkuntza-ezagutzan oinarritutako teknikak uztartzeak emaitzak hobetzen ditu
- Koma-zuzentzailearen ebaluaziorako ez da nahikoa aukera bakarra ontzat ematea

Etorkizuneko lanak

- Kate- eta perpaus-identifikatzaileak hobetzea
  - Ikasketa-corpusa handituz
  - Euskarako analizatzaile eta desanbiguatzailea hobetuz

Etorkizuneko lanak

- Kate- eta perpaus-identifikatzaileak hobetzea
  - Ikasketa-corpusa handituz
  - Euskarako analizatzaile eta desanbiguatzailea hobetuz
- Koma-zuzentzailea hobetzea
  - Uneko tokenaren aurreko komak kontuan hartuz
  - Emaitza-klaseen desorekaren arazoa konponduz
  - Koma okerrekin lortutako informazio linguistikoarekin gertatzen dena aztertuz

Etorkizuneko lanak

- Kate- eta perpaus-identifikatzaileak hobetzea
  - Ikasketa-corpusa handituz
  - Euskarako analizatzaile eta desanbiguatzailea hobetuz
- Koma-zuzentzailea hobetzea
  - Uneko tokenaren aurreko komak kontuan hartuz
  - Emaitza-klaseen desorekaren arazoa konponduz
  - Koma okerrekin lortutako informazio linguistikoarekin gertatzen dena aztertuz
- FR-Perceptron algoritmoa hobetzea
- Bestelako erroreen detekzioa lantzea ikasketa automatikoko tekniken bidez
- Analisi-katean eta XUXENg-n integratzea

Tesiari hertsiki lotutako argitalpenak (I)

- Alegria I., Arrieta B., Carreras X., Díaz de Ilarraza A., Uria L. **Chunk and Clause Identification for Basque by Filtering and Ranking with Perceptrons.** *Revista del procesamiento del lenguaje natural, nº 41, pags: 5-12;* 2008.
- Aldabe I., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Niebla I., Oronoz M., Uria L. **Basque error corpora: a framework to classify and store it.** *In the Proceedings of the 4th Corpus Linguistic Conference*. Birmingham. UK. 2007.
- Alegria I., Arrieta B., Díaz de Ilarraza A., Izagirre E., Maritxalar M. **Using Machine Learning Techniques to Build a Comma Checker for Basque.** *Proceedings of Coling-ACL.* Sydney. Australia. 2006.
- Aduriz I., Arrieta B., Arriola J.M., Díaz de Ilarraza A., Izagirre E., Ondarra A. **Muga Gramatikaren Optimizazioa.** *EHU/LSI/TR 26-2005*. Donostia. Euskal Herria. 2005
- Aldabe I., Arrieta B., Díaz de Ilarraza A., Maritxalar M., Oronoz M., Uria L. **Propuesta de una clasificación general y dinámica para la definición de errores.** *Revista de Psicodidáctica, EHU. Vol 10, Nº 2, p. 47-60.* 2005.

Tesiari hertsiki lotutako argitalpenak (II)

- Ansa O., Arregi X., Arrieta B., Díaz de Ilarraza A., Ezeiza N., Fernandez I., Garmendia A., Gojenola K., Laskurain B., Martínez E., Oronoz M., Otegi A., Sarasola K., Uria L. **Integrating NLP Tools for Basque in Text Editors.** *Workshop on International Proofing Tools and Language Technologies.* University of Patras. Greece. 2004.
- Aldezabal I., Aranzabe M., Arrieta B., Maritxalar M., Oronoz M. **Toward a punctuation checker for Basque.** *ATALA workshop. Le role de la typographie et de la ponctuation dans le traitement automatique des langues.* Paris. France. 2003.
- Arrieta B., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Oronoz M. **A database system for storing second language learner corpora.** *Learner corpora workshop. Corpus linguistics 2003. Volume 16, Part 1. p.: 33-41;* Lancaster, UK. 2003.
- Aduriz I., Aldezabal I., Aranzabe M., Arrieta B., Arriola J., Atutxa A., Díaz de Ilarraza A., Gojenola K., Oronoz M., Sarasola K., Urizar R. **The design of a digital resource to store the knowledge of linguistic errors.** *DRH2002 (Digital Resources for the Humanities), pp 76-78*. Edinburgh. Scotland. 2002.

"Euskaran bezala bizitzan, gauza ustez txiki baina guztiz beharrezkoak baztertzeko joera dugu eta nahiago ditugu proiektu pinpirinak. Baina elementalena —eta neketsuena— koma baten garrantzia eta gisakoak azpimarratzea da."

A. Lertxundi (Berria; 2010/07/15)

Eskerrik asko!

Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean

Doktoregaia: Bertol Arrieta Kortajarena
Zuzendariak: Iñaki Alegria Loinaz
Arantza Diaz de Ilarraza Sanchez

Lengoaia eta Sistema Informatikoak/Lenguajes y Sistemas Informáticos
Euskal Herriko Unibertsitatea/Universidad del País Vasco

2010eko uztailaren 27a