



**UNIVERSITÀ
DI TORINO**

Scuola di Dottorato in Digital Humanities
Academic years: 2018-2019, 2019-2020, 2020-2021
Academic disciplines: L-LIN/01 - INF/01



**UNIVERSITÀ
DEGLI STUDI
DI GENOVA**

PhD supervisors:
Prof. Cristina Bosco
Prof. Elisa Corino

ELISA DI NUOVO

VALICO-UD: ANNOTATING AN ITALIAN LEARNER CORPUS

Part 1

Introduction

Literature review

Overview

Part 2

VALICO-UD design

Error annotation

Linguistic annotation

Part 3

Parser evaluation

Extrinsic evaluation

Shared Tasks

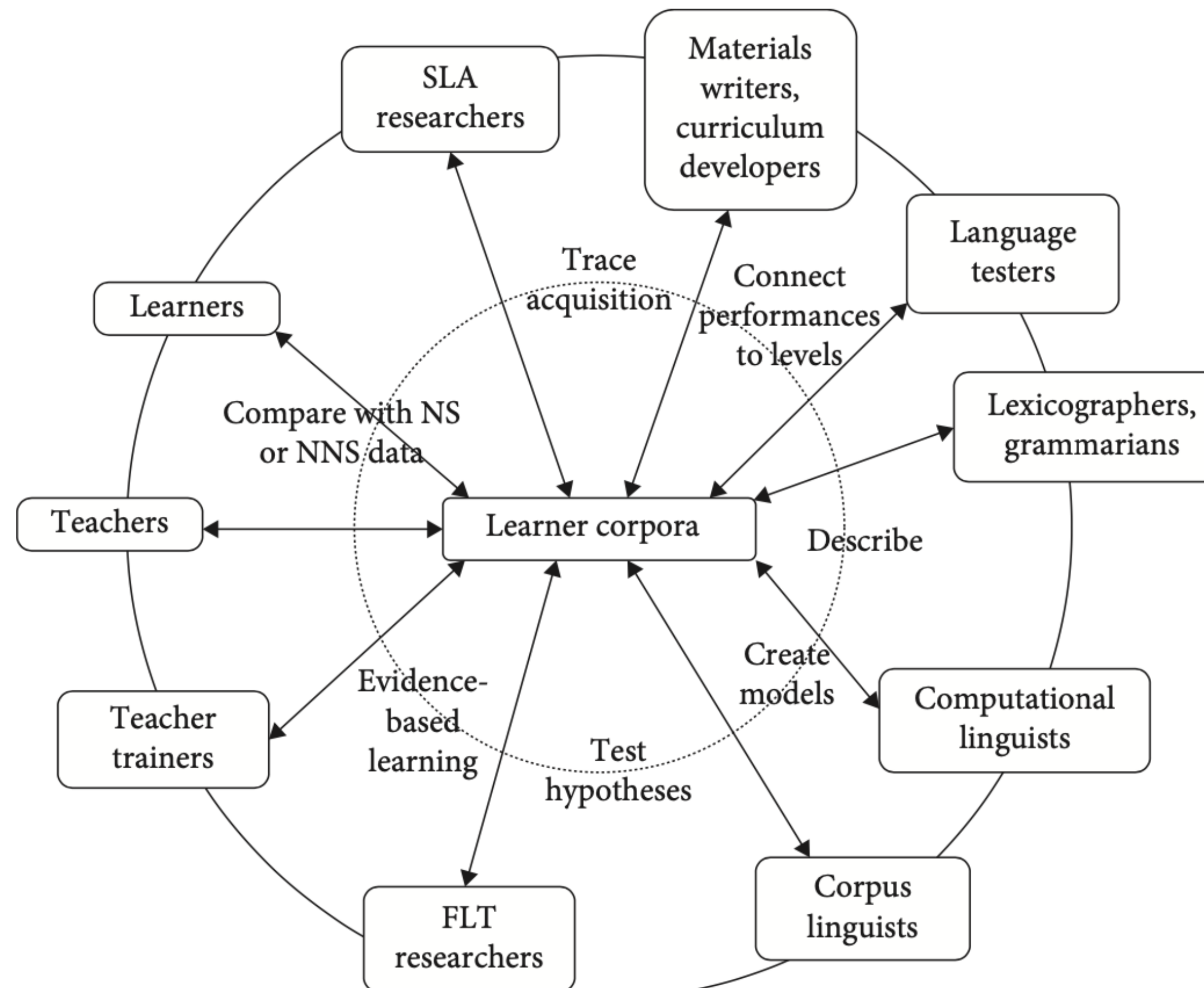
Part 4

Conclusions & Future work

LEARNER CORPUS USES AND USERS

Learner corpora are «*electronic collections of language data produced by L2 learners, that is, second or foreign-language learners.*»

(Granger, 2008)



For targeting NLP tasks:

- * Native Language Identification (NLI)
- * Grammatical Error Detection (GED)
- * Grammatical Error Correction (GEC)
- * Tagger and Parser development

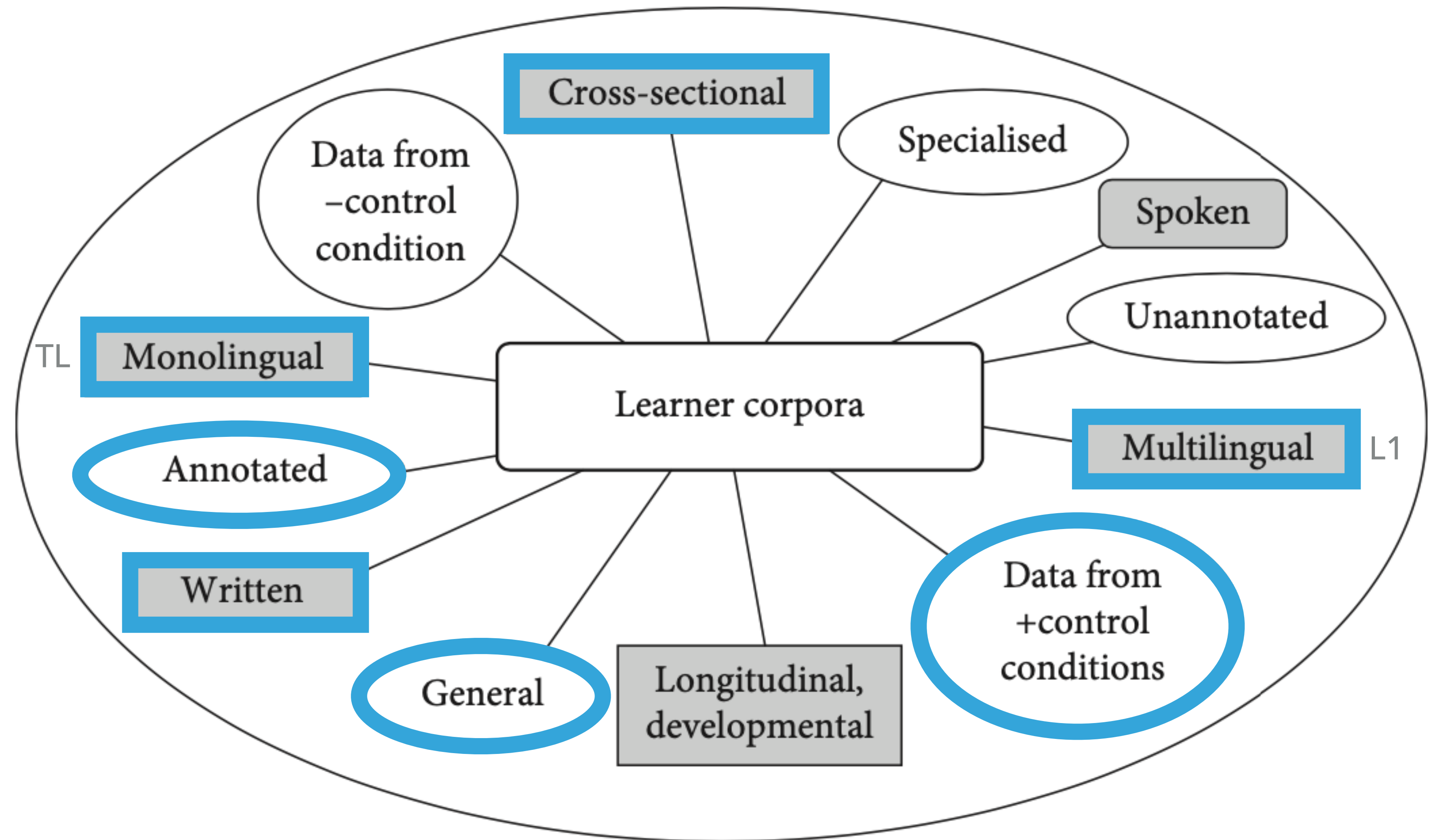
LEARNER CORPUS RESEARCH

Learner Corpus	Approximate Size	L2 Level	Target Language
CLC*	29,000,000 w	A1–C2	English
ICLEv3*	5,500,000 w	B2–C2	English
TOEFL11	4,000,000 t	A1–C2	English
NUCLE*	1,200,000 t	N/A	English
VALICO	380,000 t	A1–C2	Italian
MERLIN*	340,000 t	A1–C2	Czech / German / Italian
Falkov2*	280,000 t	B2–C2	German
PIL2	120,000 t	N/A	Italian
ESL* treebank	97,000 t	B2	English
CFL treebank	7,000 t	N/A	Chinese
SALLE treebank	600 t	A1–C2	English

LC marked with * are error-annotated.

VALICO DESIGN

- * Dimension: the biggest Italian learner corpus publicly available
- * Proficiency: various (expressed in year of study of Italian)
- * Availability: available for free
- * Academic



ERROR-ANNOTATED LC

Each annotation project has its own tag set (Díaz-Negrillo, Fernández-Domínguez, 2006)

Surface strategy taxonomy



Linguistic category taxonomy

CLC

Up to two letters (Nicholls, 2003:572):

- * general type of error
- * word class of the required word

Students have the possibility <#FV><#RT>to|of</#RT>leave|leaving</#FV> [...]

ICLE

More letters (Error Tagging Manual v. 1.1, 1996:4-7):

- * error category
- * sub-categorisation
- * following letters for more specification

Students have the (XNCO) possibility to leave \$possibility of leaving\$ [...]

POS-ANNOTATED LC

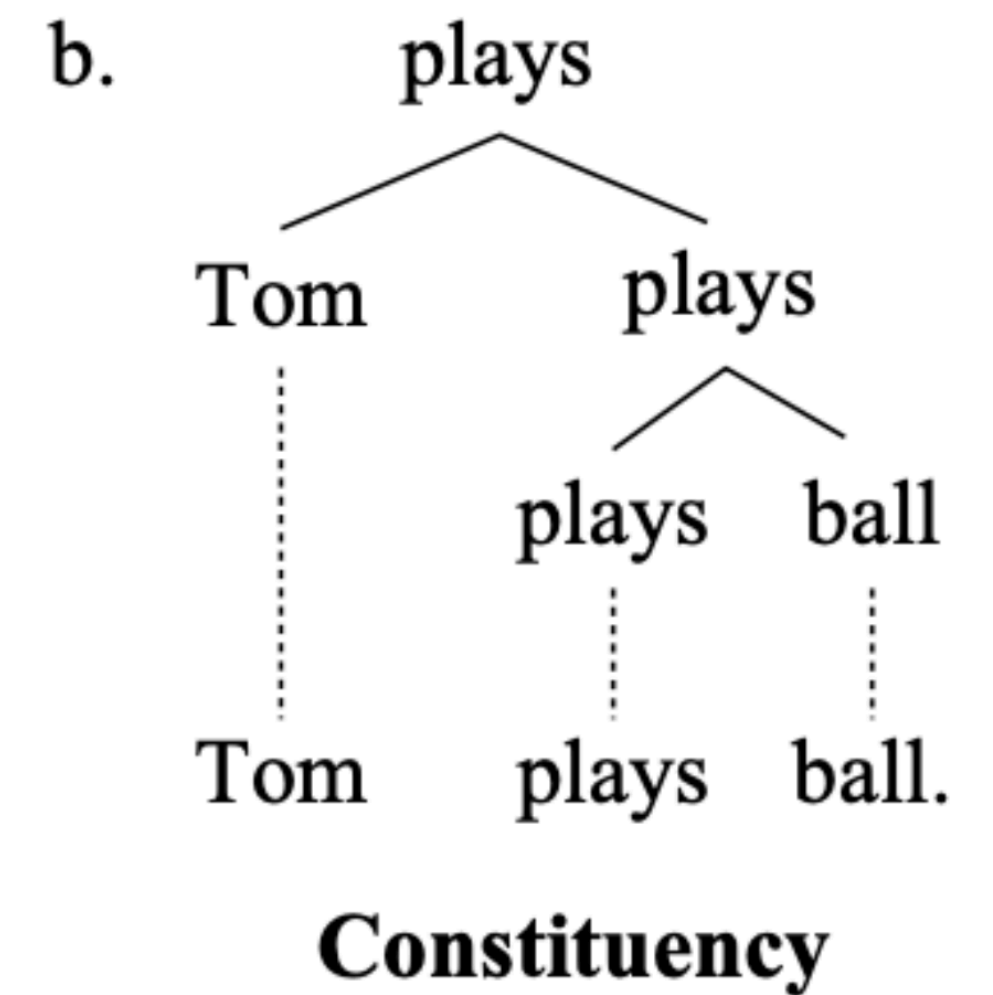
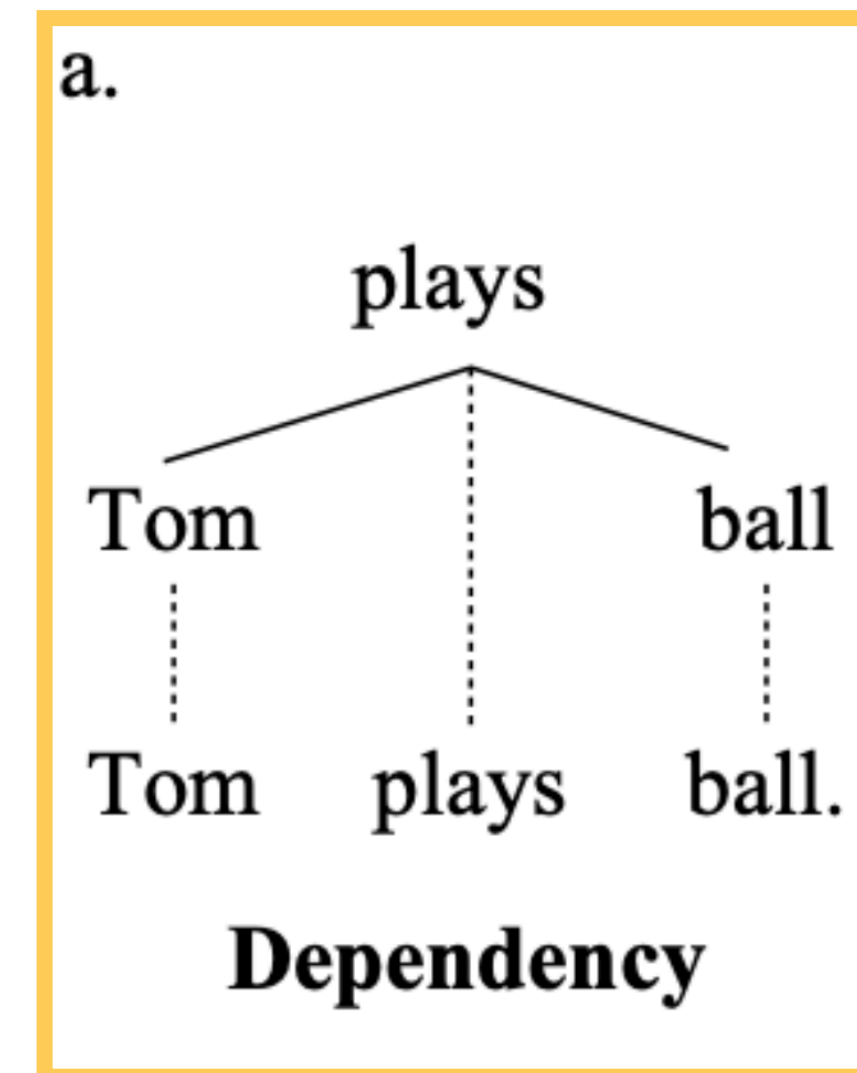
- * Second most commonly annotated feature on LC
- * It should include **lemmatisation** and other **morphological features**

- * Each LC uses a different PoS tag set

- * **Automatic** PoS annotation using pre-trained taggers (i.e. domain transfer), e.g. ICLE and MERLIN
- * **Manual** PoS annotation using:
 - * Target language PoS tags
 - * **Ad hoc** tags:
 - * More than one PoS tag per token (e.g. PIL2, NOCE and SALLE)
 - * One PoS tag per token guided by **literal** or **distributional** annotation (ESL and CFL)

SYNTACTICALLY ANNOTATED LC

- * Complex task performed on a few LC (SALLE, FALKO, MERLIN vs ESL, CFL)
- * **Automatic** annotation using pre-trained parsers (i.e. domain transfer)
- * **Manual** annotation using:
 - * Target language rules
 - * **Ad hoc** annotation:
 - * More than one syntactic tree per sentence (e.g. SALLE)
 - * One tree per sentence (ESL and CFL)
- * ESL and CFL annotated in Universal Dependencies



Osborne (2014:605)



UNIVERSAL DEPENDENCIES (UD) PROJECT

«developing cross-linguistically **consistent treebank annotation** for many languages, with the goal of facilitating **multilingual parser development**, **cross-lingual learning**, and **parsing research** from a language typology perspective»

(<https://universaldependencies.org>)

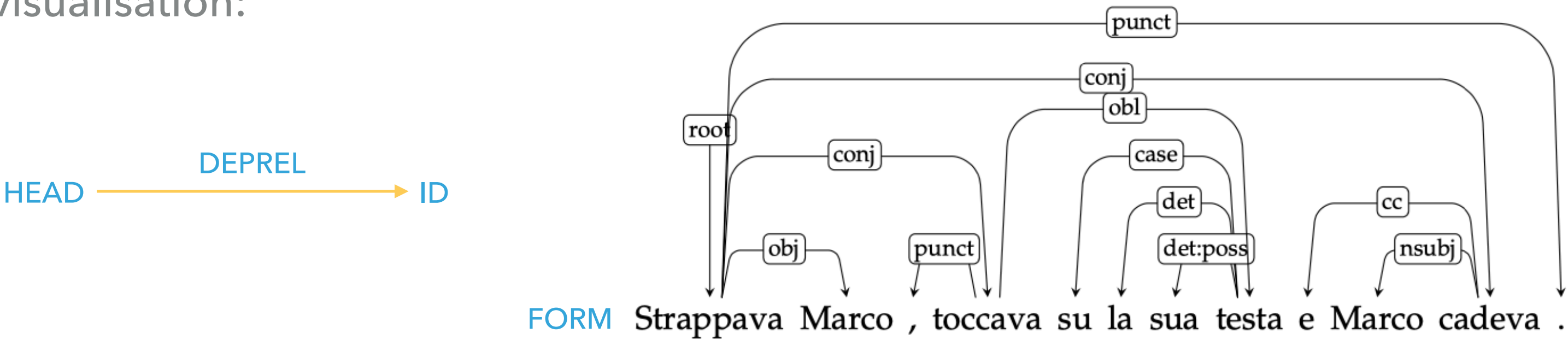
CoNLL-U encoding:

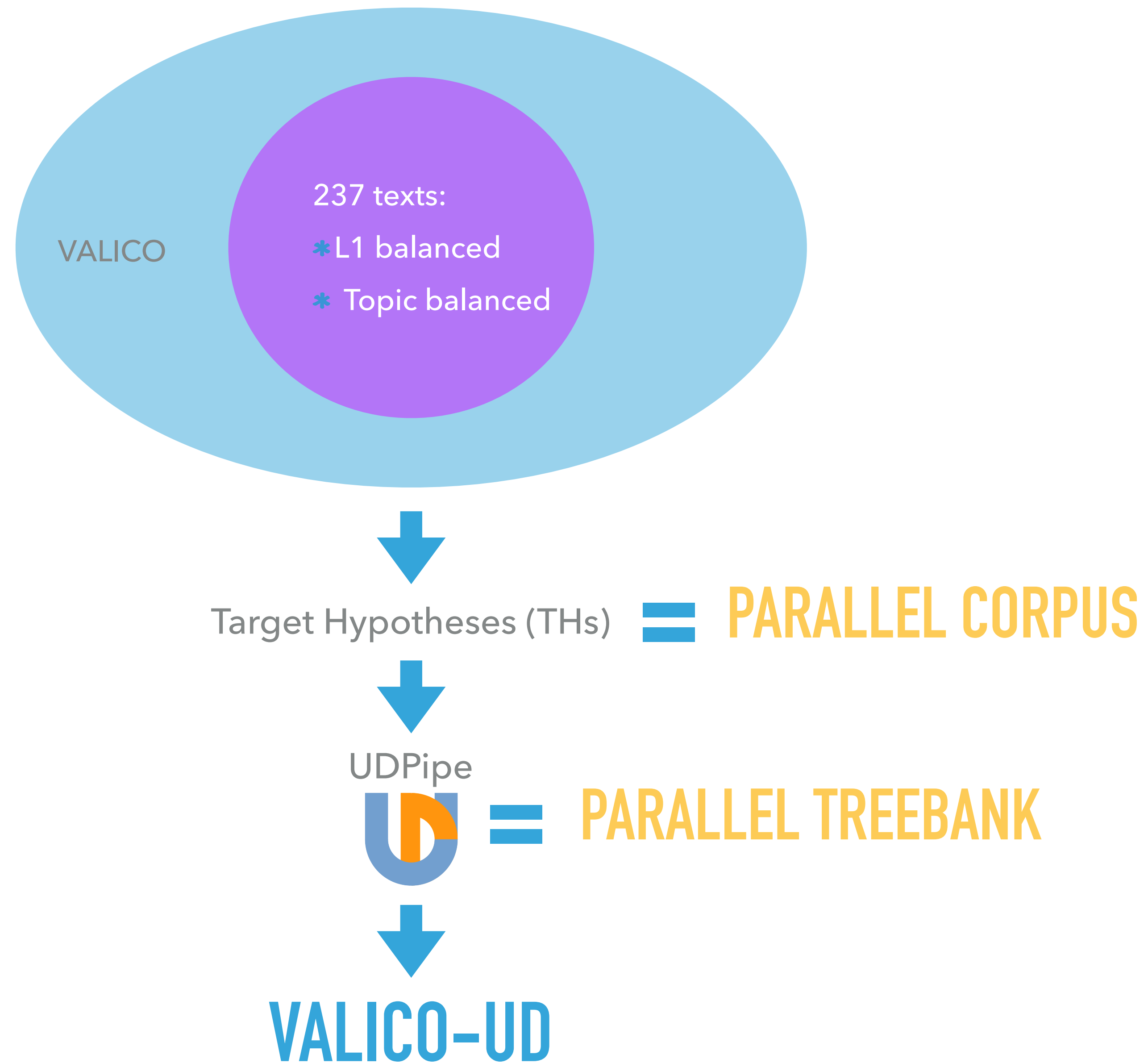
sent_id = 5-06_fr-3

text = Strappava **Marco**, toccava **sulla** sua testa e Marco **cadeva**.

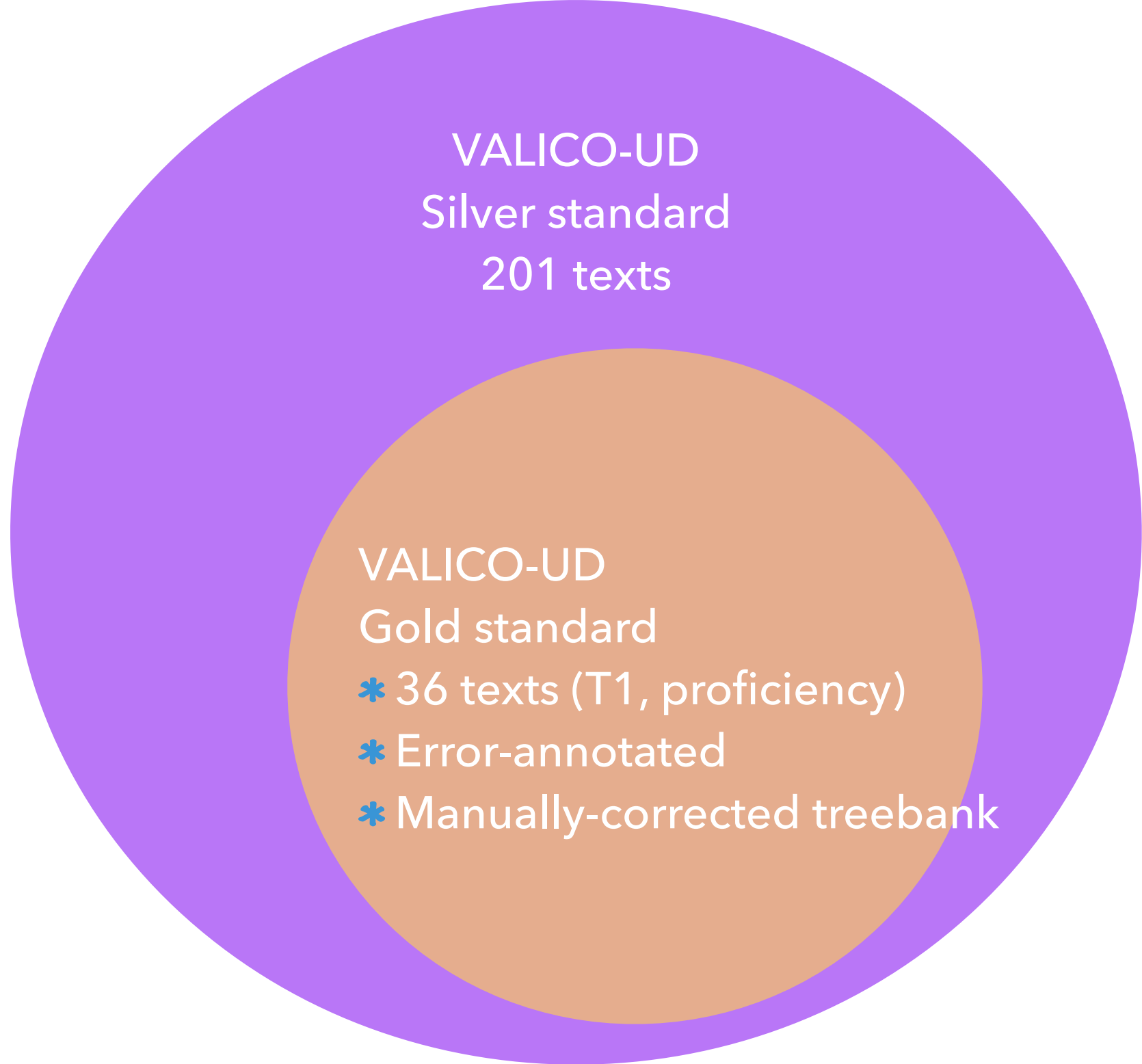
1	Strappava	strappare	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	0	root	-
2	Marco	Marco	PROPN	SP	-	1	obj	SpaceAfter=No
3	,	,	PUNCT	FF	-	4	punct	-
4	toccava	toccare	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	1	conj	-
5-6	sulla	-	-	-	-	-	-	-
5	su	su	ADP	E	-	8	case	-
6	la	la	DET	RD	Definite=Def Gender=Fem Number=Sing PronType=Art	8	det	-
7	sua	suo	DET	AP	Gender=Fem Number=Sing Poss=Yes PronType=Prs	8	det:poss	-
8	testa	testa	NOUN	S	Gender=Fem Number=Sing	4	obl	-
9	e	e	CCONJ	CC	-	11	cc	-
10	Marco	Marco	PROPN	SP	-	11	nsubj	-
11	cadeva	cadere	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	1	conj	SpaceAfter=No
12	.	.	PUNCT	FS	-	1	punct	SpacesAfter=\n
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	MISC

CoNLL-U visualisation:





L1	# Texts	# Sentences	# LS Tokens	# TH Tokens
	T1 T2	T1 T2		
German (DE)	58 29 29	622 280 342	8,729	8,838
English (EN)	60 42 18	662 474 188	9,834	10,029
Spanish (ES)	59 42 17	381 266 115	8,270	8,361
French (FR)	60 30 30	569 249 320	8,623	8,686
EN+FR+DE+ES	237	2,234	35,456	35,914



T1



Ieri al parco stavo leggendo il giornale quando che ho sentito urlare forte.
E dopo un po ho potuto vedere l'origine di questo: Un ragazzo di alta statura e muscoloso passava con la sua fidanzata sulla spalla.
Lei ancora stava urlando.
In questo momento ho capito che la fidanzata non voleva essere sulla spalla, però lei non ha potuto opporsi a lui.
Proprio per questo ho deciso di intervenire.
Si deve sapere che anche io sono di alta statura e mi alleno.
Così era semplice per me di sopraffare l'altro e liberare la ragazza.
Mi ho sentito come un'eroe.
Ma lei invece ha iniziato di urlare da nuova.
All'inizio non ho capito perché ma dopo un po ho potuto immaginare: lei era innamorata di lui e tutto era parte di un gioco tra di loro.

T2



La Stazione
Cera una volta una Stazione dove sono tanti persone, anche un cane.
Ce molto casino sono anche le Wurstel e un uomo e una donna si baciano.
Due donne e un uomo guardano sul grande Orologio.
Un Ladro ce anche, a che un grosso uomo con una cappa alla testa è fuma.
Una donna bere un caffè.
Sono tanti cofferi e un uomo qurda sulle due ce si baciano.
Anche una tasca e un pacceto.
Ce un Tavolo con i macharini e tanti treni.
Una donna aspetava a una sedia e una donna va con un uomo al treno.

TARGET HYPOTHESIS WRITING – MAIN GUIDING PRINCIPLES

- 1. **Formal criterion** (Andorno and Rastelli 2009, p.58)
The normalisation should be as far as possible near to what the learner wrote

LS	Ieri al parco uno studente era seduto sul banco .
TH1	Ieri al parco uno studente era seduto sulla panchina .
TH2	Ieri al parco uno studente era seduto sulla panca .
<i>Yesterday in the park a student was sitting on the bench.</i>	

- 2. **Reference resources**
All the forms considered as erroneous must be verified in:
VINCA (Corino and Marello 2017), CORIS (Rossini Fevretti et al. 2002), ISDT (Simi et al. 2014), ParTUT (Sanguinetti and Bosco, 2015), VIT (Alfieri and Tamburini, 2016), PoSTWITA (Sanguinetti et al. 2018), TWITTIRÒ (Cignarella et al. 2019), De Mauro’s Dictionary (De Mauro 2016), Italian reference grammar (Renzi et al. 2001)

LS	Mi ho sentito come un’eroe.
TH1	Mi sono sentito un eroe.
TH2	Mi sono sentito come un eroe.
<i>I felt like a hero.</i>	

ERROR TAG SET

1	2	3	Meaning
S	A C D J N O R T V X ⁷	–	Generic spelling error affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	S T	Capitalization or tokenization errors of a word whose class is identified by the letter in position 2.
	B	–	Spelling error affecting double consonants.
	E I P	M R U	Apostrophes, accents punctuation errors further specified as Missing, Replacement or Unnecessary.
D	A C D J N O R T V X	–	Derivation error affecting one of the word classes indicated by the letter in position 2.
F	A C D J N O R T V X	–	Form errors affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	G	Alternative form errors affecting gender.
	A C D J N O R T V X	L	Foreign forms that do not exist in Italian.
I	V	–	Conjugation errors.
	V	A B M T	Aspect, mood and tense, mood or tense errors.
	V	O	Gerund instead of relative clause.
	A D J N V	G N	Gender or number errors affecting one of the word classes indicated in position 2.
	A D V X	P	Person errors affecting one of the word classes indicated in position 2.
U	A C D J R T W	–	An unnecessary word, or more than one (W), whose word class in specified in position 2.
	J N R	G N	Unnecessary inflection of gender or number in the word class specified in position 2.
M	A C D N R T V X	–	A missing word whose word class is indicated by the letter in position 2.
R	A C D J N R T V X	–	Replace word with another whose word class is indicated by the letter in position 2. ⁸
	A C D J N O R T V X	L	Foreign words that do exist in Italian but with another meaning.
W	A C D J N R V W	–	Word order errors involving one of the word classes indicated in position 2.

ERROR TAG SET

1	2	3	Meaning
S	A C D J N O R T V X ⁷	–	Generic spelling error affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	S T	Capitalization or tokenization errors of a word whose class is identified by the letter in position 2.
	B	–	Spelling error affecting double consonants.
	E I P	M R U	Apostrophes, accents punctuation errors further specified as Missing, Replacement or Unnecessary.
D	A C D J N O R T V X	–	Derivation error affecting one of the word classes indicated by the letter in position 2.
F	A C D J N O R T V X	–	Form errors affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	G	Alternative form errors affecting gender.
	A C D J N O R T V X	L	Foreign forms that do not exist in Italian.
I	V	–	Conjugation errors.
	V	A B M T	Aspect, mood and tense, mood or tense errors.
	V	O	Gerund instead of relative clause.
	A D J N V	G N	Gender or number errors affecting one of the word classes indicated in position 2.
	A D V X	P	Person errors affecting one of the word classes indicated in position 2.
U	A C D J R T W	–	An unnecessary word, or more than one (W), whose word class in specified in position 2.
	J N R	G N	Unnecessary inflection of gender or number in the word class specified in position 2.
M	A C D N R T V X	–	A missing word whose word class is indicated by the letter in position 2.
R	A C D J N R T V X	–	Replace word with another whose word class is indicated by the letter in position 2. ⁸
	A C D J N O R T V X	L	Foreign words that do exist in Italian but with another meaning.
W	A C D J N R V W	–	Word order errors involving one of the word classes indicated in position 2.

ERROR TAG SET

1	2	3	Meaning
S	A C D J N O R T V X ⁷	–	Generic spelling error affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	S T	Capitalization or tokenization errors of a word whose class is identified by the letter in position 2.
	B	–	Spelling error affecting double consonants.
	E I P	M R U	Apostrophes, accents punctuation errors further specified as Missing, Replacement or Unnecessary.
D	A C D J N O R T V X	–	Derivation error affecting one of the word classes indicated by the letter in position 2.
F	A C D J N O R T V X	–	Form errors affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	G	Alternative form errors affecting gender.
	A C D J N O R T V X	L	Foreign forms that do not exist in Italian.
I	V	–	Conjugation errors.
	V	A B M T	Aspect, mood and tense, mood or tense errors.
	V	O	Gerund instead of relative clause.
	A D J N V	G N	Gender or number errors affecting one of the word classes indicated in position 2.
	A D V X	P	Person errors affecting one of the word classes indicated in position 2.
U	A C D J R T W	–	An unnecessary word, or more than one (W), whose word class in specified in position 2.
	J N R	G N	Unnecessary inflection of gender or number in the word class specified in position 2.
M	A C D N R T V X	–	A missing word whose word class is indicated by the letter in position 2.
R	A C D J N R T V X	–	Replace word with another whose word class is indicated by the letter in position 2. ⁸
	A C D J N O R T V X	L	Foreign words that do exist in Italian but with another meaning.
W	A C D J N R V W	–	Word order errors involving one of the word classes indicated in position 2.

ERROR TAG SET

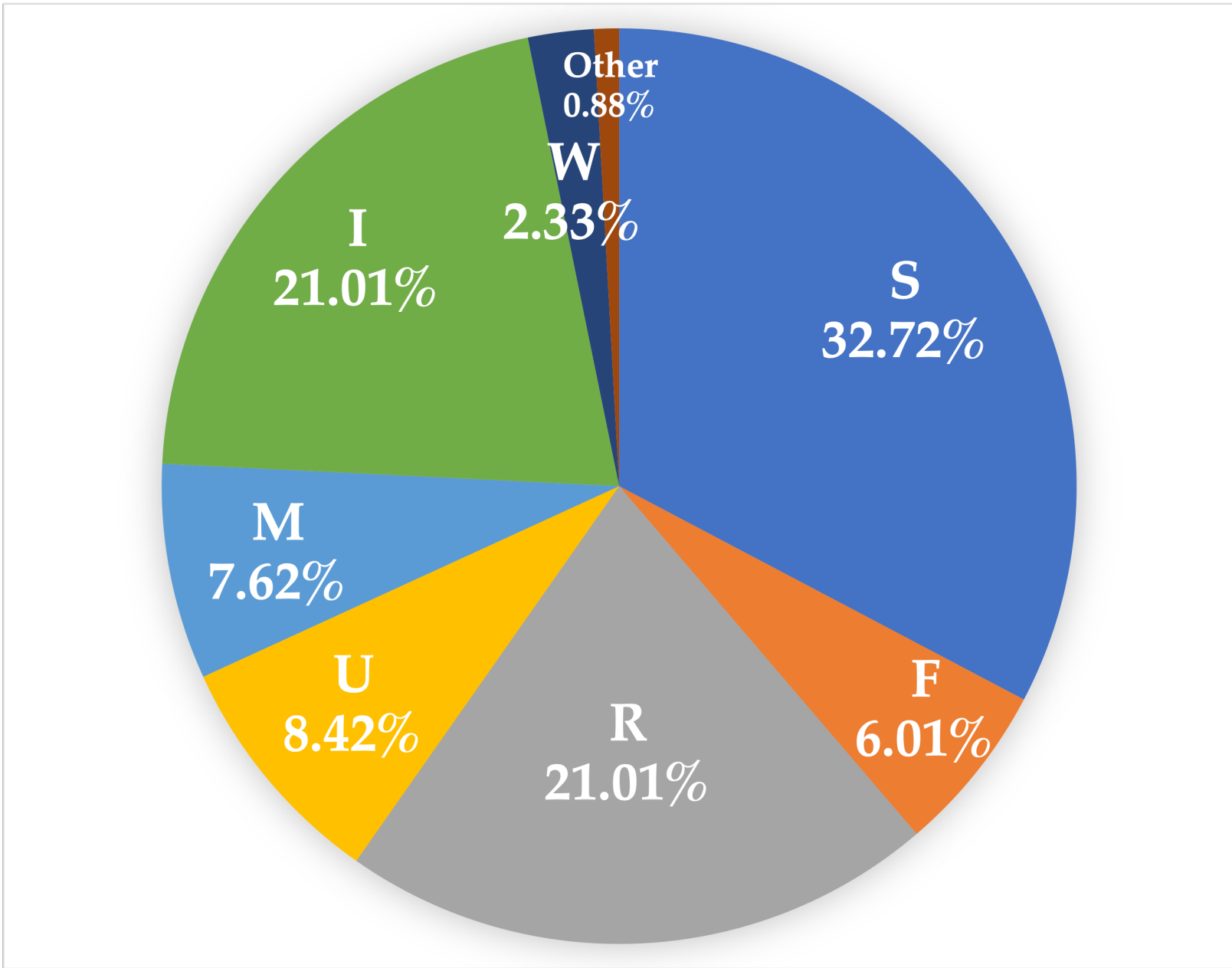
1	2	3	Meaning
S	A C D J N O R T V X ⁷	–	Generic spelling error affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	S T	Capitalization or tokenization errors of a word whose class is identified by the letter in position 2.
	B	–	Spelling error affecting double consonants.
	E I P	M R U	Apostrophes, accents punctuation errors further specified as Missing, Replacement or Unnecessary.
D	A C D J N O R T V X	–	Derivation error affecting one of the word classes indicated by the letter in position 2.
F	A C D J N O R T V X	–	Form errors affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	G	Alternative form errors affecting gender.
	A C D J N O R T V X	L	Foreign forms that do not exist in Italian.
I	V	–	Conjugation errors.
	V	A B M T	Aspect, mood and tense, mood or tense errors.
	V	O	Gerund instead of relative clause.
	A D J N V	G N	Gender or number errors affecting one of the word classes indicated in position 2.
	A D V X	P	Person errors affecting one of the word classes indicated in position 2.
U	A C D J R T W	–	An unnecessary word, or more than one (W), whose word class is specified in position 2.
	J N R	G N	Unnecessary inflection of gender or number in the word class specified in position 2.
M	A C D N R T V X	–	A missing word whose word class is indicated by the letter in position 2.
R	A C D J N R T V X	–	Replace word with another whose word class is indicated by the letter in position 2. ⁸
	A C D J N O R T V X	L	Foreign words that do exist in Italian but with another meaning.
W	A C D J N R V W	–	Word order errors involving one of the word classes indicated in position 2.

ERROR TAG SET

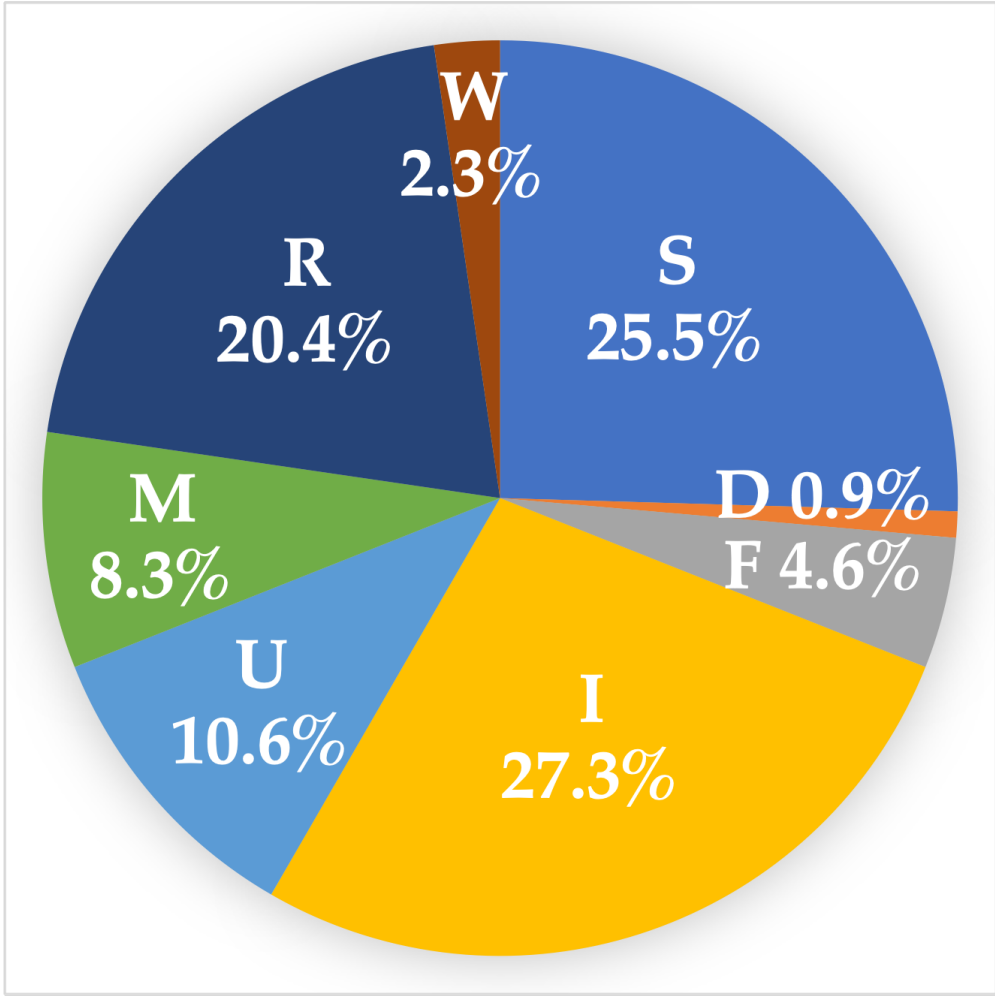
1	2	3	Meaning
S	A C D J N O R T V X ⁷	–	Generic spelling error affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	S T	Capitalization or tokenization errors of a word whose class is identified by the letter in position 2.
	B	–	Spelling error affecting double consonants.
	E I P	M R U	Apostrophes, accents punctuation errors further specified as Missing, Replacement or Unnecessary.
D	A C D J N O R T V X	–	Derivation error affecting one of the word classes indicated by the letter in position 2.
F	A C D J N O R T V X	–	Form errors affecting one of the word classes indicated by the letter in position 2.
	A C D J N O R T V X	G	Alternative form errors affecting gender.
	A C D J N O R T V X	L	Foreign forms that do not exist in Italian.
I	V	–	Conjugation errors.
	V	A B M T	Aspect, mood and tense, mood or tense errors.
	V	O	Gerund instead of relative clause.
	A D J N V	G N	Gender or number errors affecting one of the word classes indicated in position 2.
	A D V X	P	Person errors affecting one of the word classes indicated in position 2.
U	A C D J R T W	–	An unnecessary word, or more than one (W), whose word class in specified in position 2.
	J N R	G N	Unnecessary inflection of gender or number in the word class specified in position 2.
M	A C D N R T V X	–	A missing word whose word class is indicated by the letter in position 2.
R	A C D J N R T V X	–	Replace word with another whose word class is indicated by the letter in position 2. ⁸
	A C D J N O R T V X	L	Foreign words that do exist in Italian but with another meaning.
W	A C D J N R V W	–	Word order errors involving one of the word classes indicated in position 2.

ERROR DISTRIBUTION PER MACRO-CATEGORY

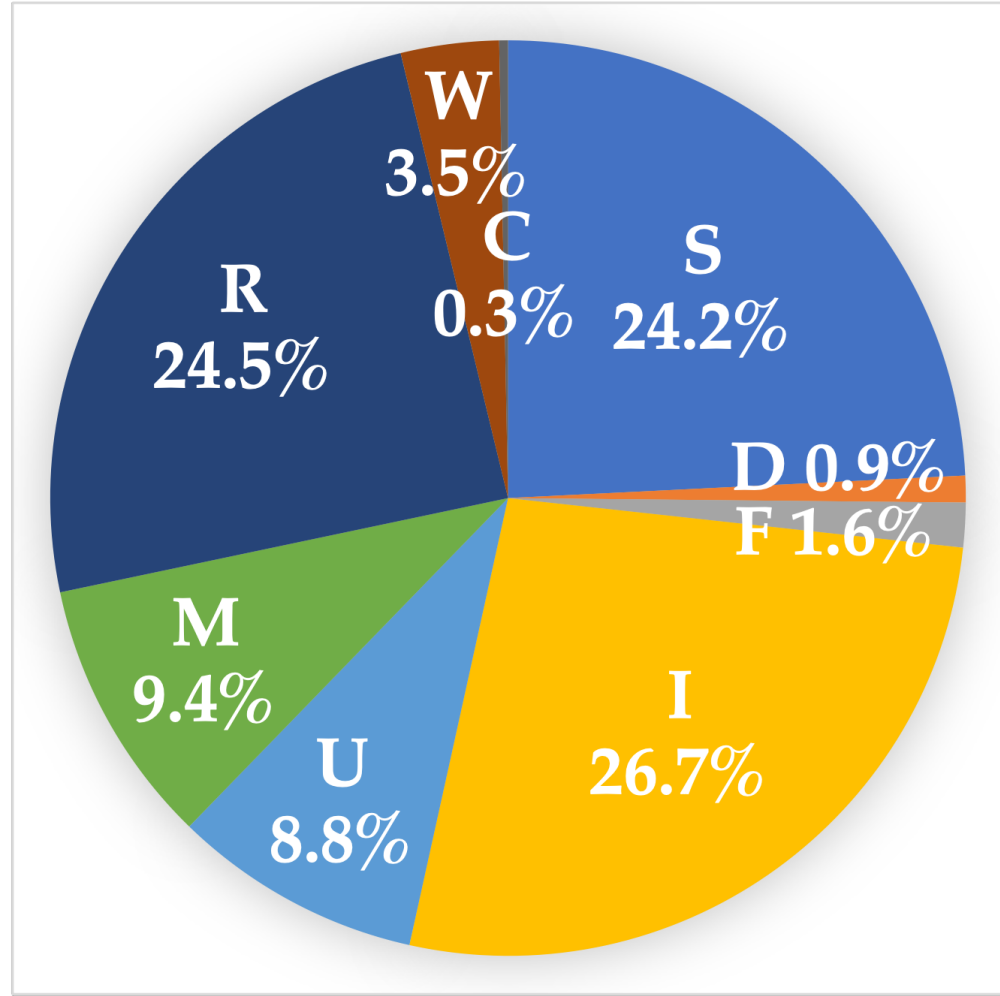
All L1



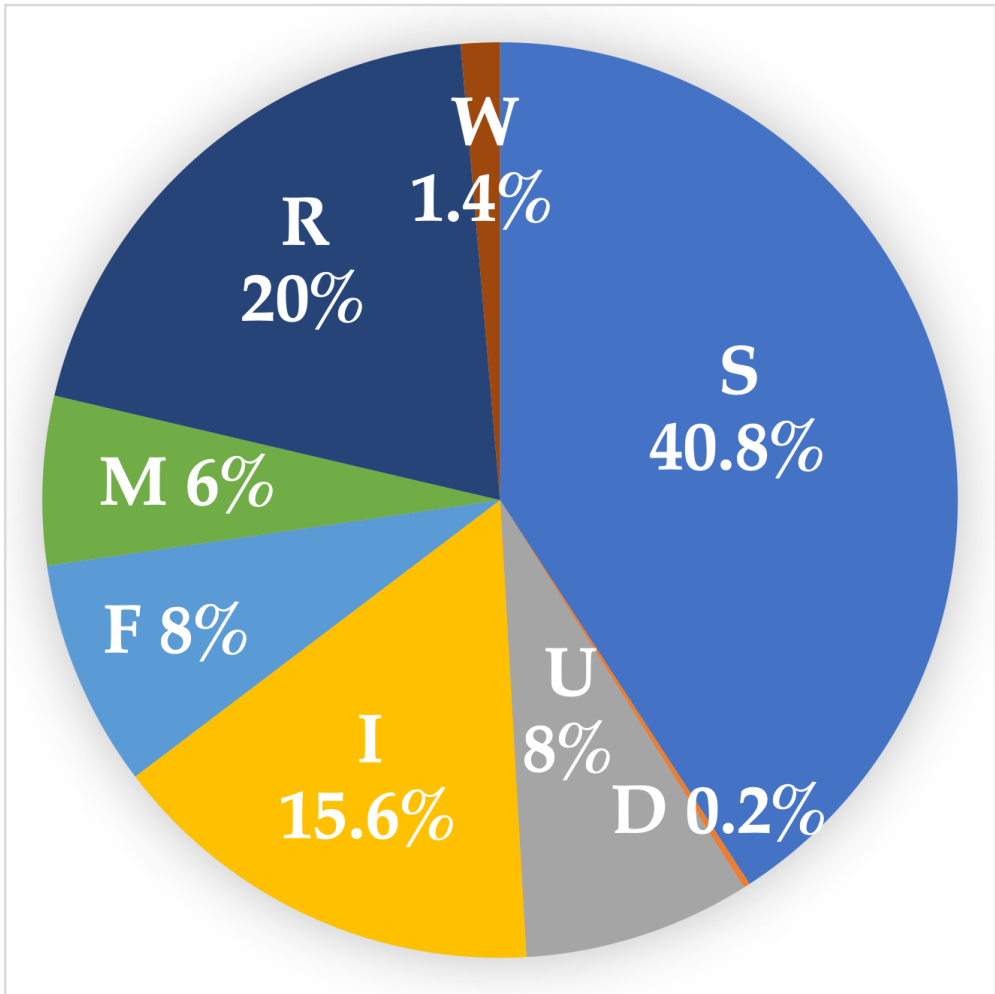
DE



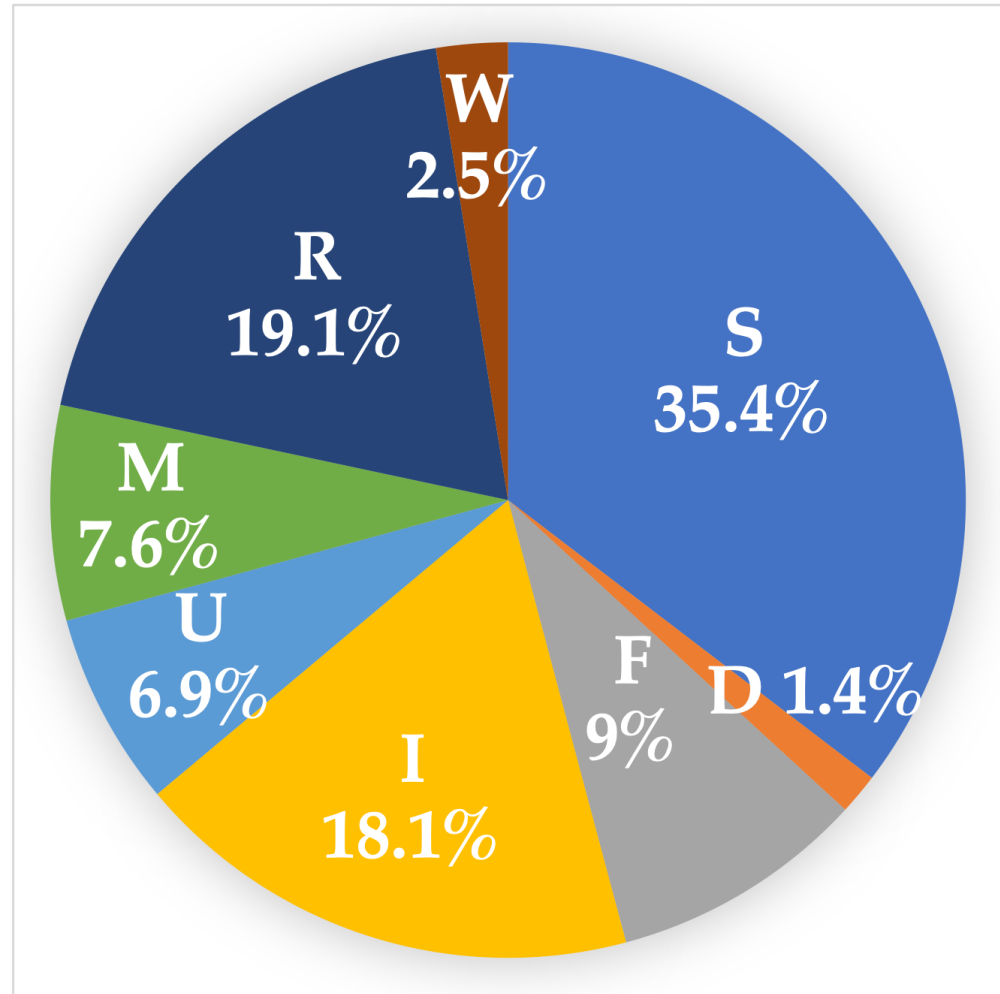
EN



ES



FR



ERROR DISTRIBUTION PER L1

Statistics		Population								
Test 1	one-way ANOVA	L1	DE		EN		ES		FR	
		# Texts	9		9		9		9	
Test 2	unpaired <i>t</i> test	YoS	1				3-4			
		L1	DE	EN	ES	FR	DE	EN	ES	FR
		# Texts	3	3	3	3	3-0	3-0	1-1	3-3
		Total	12				14			

Test 1 – one-way ANOVA

➡

p value 0.000085

➡

Extremely statistically significant

Test 1a – Post Hoc Tukey HSD

➡

ES group differs significantly from each of the other 3 L1s

➡

ES-DE $p = 0.00010$
ES-EN $p = 0.00171$
ES-FR $p = 0.01315$

Test 2 – unpaired *t* test

➡

Error distribution between initial and advanced learners

➡

No statistically significant difference

INTER-ANNOTATOR AGREEMENT

2 annotators: **Annotator 1** from Piedmont (north Italy) **Annotator 2** from Sicily (south Italy)

1. What should be considered as 'error'?

➡ Inter-Annotator Agreement on **error identification**

Cascade error

<i>seduto</i>	0	
<i>sul</i>	1	sulla
<i>banco</i>	1	panca

Error distribution: **23.33%**
Cohen's *K* = 0.82

VALICO-UD (picture-elicited) *K* = 0.82
ComiGS (picture-elicited) *K* = 0.79

CREG (reading comprehension answers) *K* = 0.68

NUCLE (essays) *K* = 0.39



2. How should this error be corrected?

➡ Inter-Annotator Agreement on **error correction**

<i>seduto</i>	0	
<i>sul</i>	1	sulla
<i>banco</i>	1	panca

Cohen's *K* = 0.69

3. How much the error tag set is effective?

➡ Inter-Annotator Agreement on **error tags**

120 different tags
(+ 28 cascade)

Cohen's *K* = 0.95

VALICO-UD: POS ANNOTATION

Two rules to annotate them all

- * Literal annotation – following Italian grammar rules we coherently describe what the learner wrote
 - * Distributional annotation – non-words or existing words in inappropriate cotext (except closed-class PoS)
- } SALLE, CFL, ESL

Examples

LS: [...] gente che soltanto **voglie** chiamare un pò l'attenzione.

Lemma: [...] volere [...]

PoS: [...] AUX [...]

[...] *people who just want to call some attention to themselves.*

LS: **Durante** un ragazzo è passato.

Lemma: [...] durante [...]

PoS: [...] ADP [...]

Meanwhile, a boy passed by.

VALICO-UD: MORPHOLOGICAL FEATURES

LEXICAL AND INFLECTIONAL FEATURE ENCODED

- * Depending on the annotated UPOS (e.g. FEATS of verb in Italian, gender only for partitive verbs)
- * Some UPOS cannot have FEATS annotated, i.e. proper names, foreign words, prepositions, conjunctions, punctuation, adverbs

IF FEATS ANNOTATION CAN BE OF INTEREST

- * In VALICO-UD we avoided the UD constrains adding FEATS info in the MISC column (i.e. foreign words, evaluative suffixes, MWEs)

sent_id = 34-07_en-3

text = [...] tale come camminare il canino al parco. [such as walking the dog in the park.]

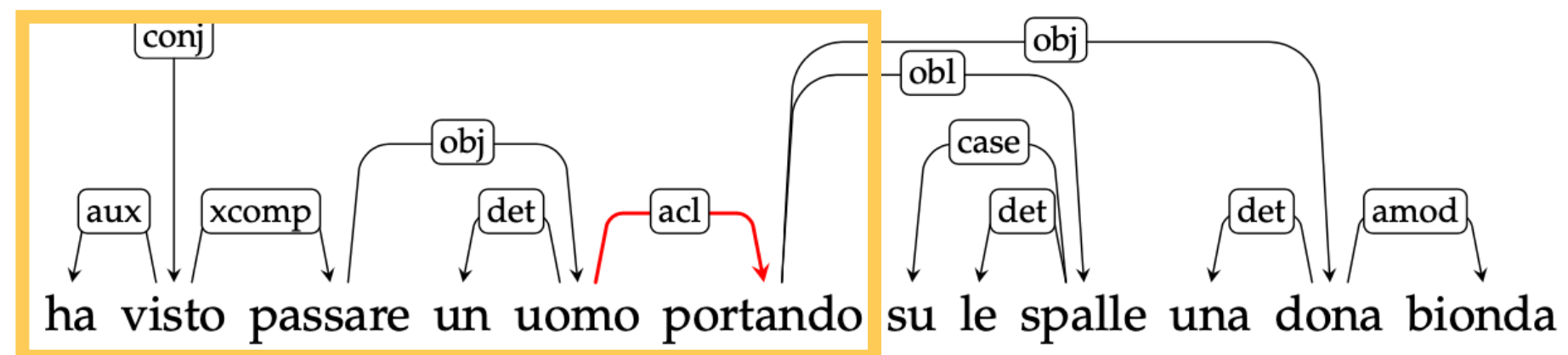
9	tale	tale	ADJ	A	[...]	11	mark	_
10	come	come	SCONJ	CS	_	9	fixed	_
11	camminare	camminare	VERB	V	[...]	4	acl	_
12	il	il	DET	RD	[...]	13	det	_
13	canino	cane	NOUN	S	[...]	11	obj	EvalMorph=Dim
14-15	al	_	_	_	_	_	_	_
14	a	a	ADP	E	_	16	case	_
15	il	il	DET	RD	[...]	16	det	_
16	parco	parco	NOUN	S	[...]	11	obl	SpaceAfter=No
17	.	.	PUNCT	FS	_	2	punct	SpacesAfter=\n

VALICO-UD: DEPENDENCY ANNOTATION

Two rules to annotate them all

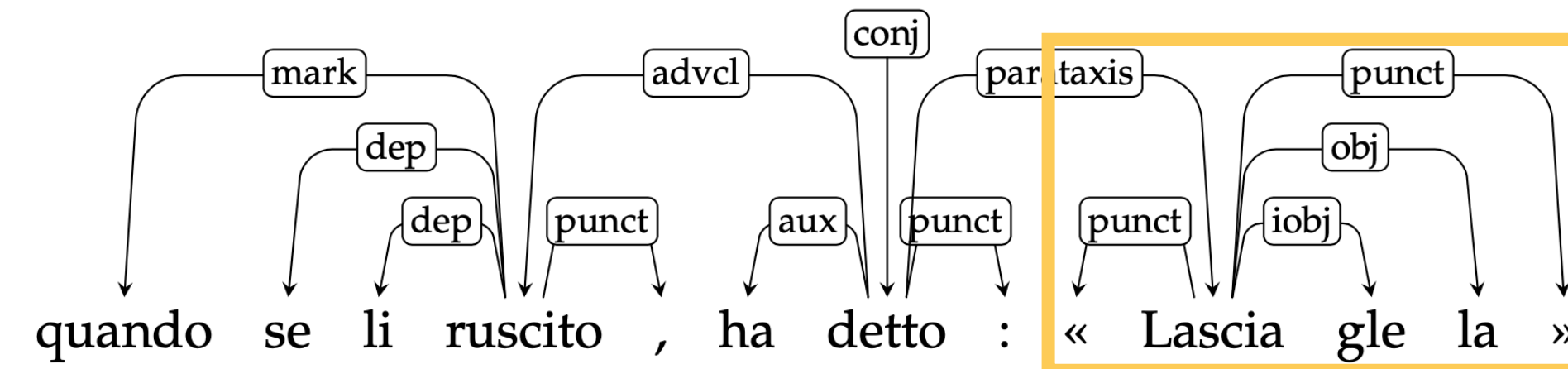
- * Literal annotation
 - * Distributional annotation
- } SALLE, CFL, ESL

Literal annotation



Yesterday in the park a man, wearing glasses, was sitting on a bench, reading the newspaper, when he saw a very strong man pass by, with a tattooed heart on his arm, CARRYING on his shoulders a blond woman who was screaming and seemed to be asking for help.

Distributional annotation



LS: Lui era un ragazzo buono e ardito: si è alzato, e li è seguito; quando **se li ruscito**, ha detto: «Lasciaglela»

He was a good and bold guy: he got up and followed them; when if he did, he said: «Leave it to her».

TH: Lui era un ragazzo buono e ardito: si è alzato e li ha seguiti; quando li ha raggiunti, ha detto: «Lasciala».

He was a good and bold guy: he got up and followed them; when he reached them, he said: «Leave her».

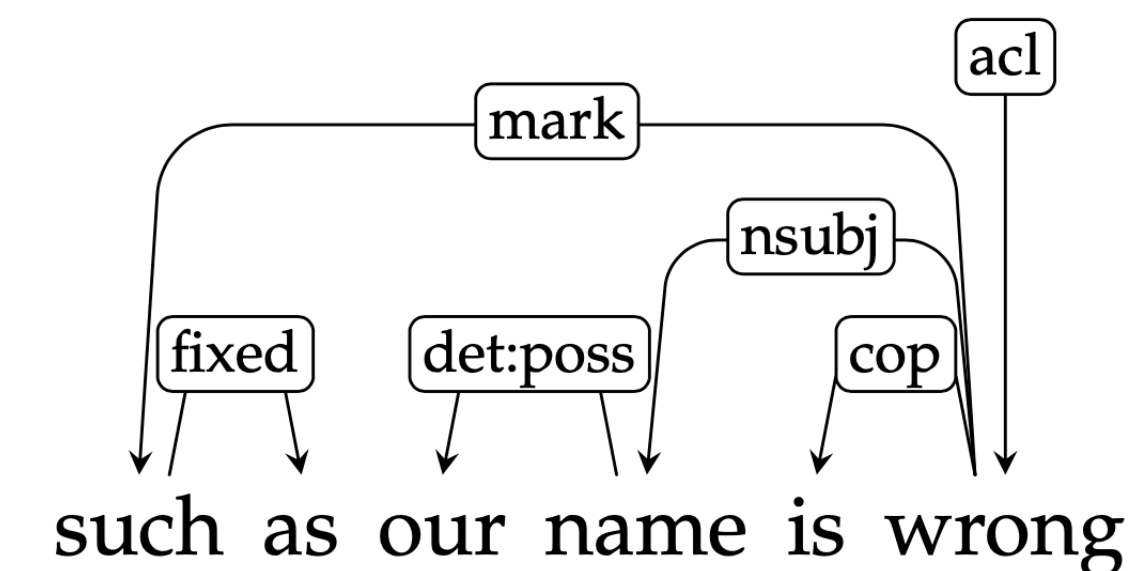
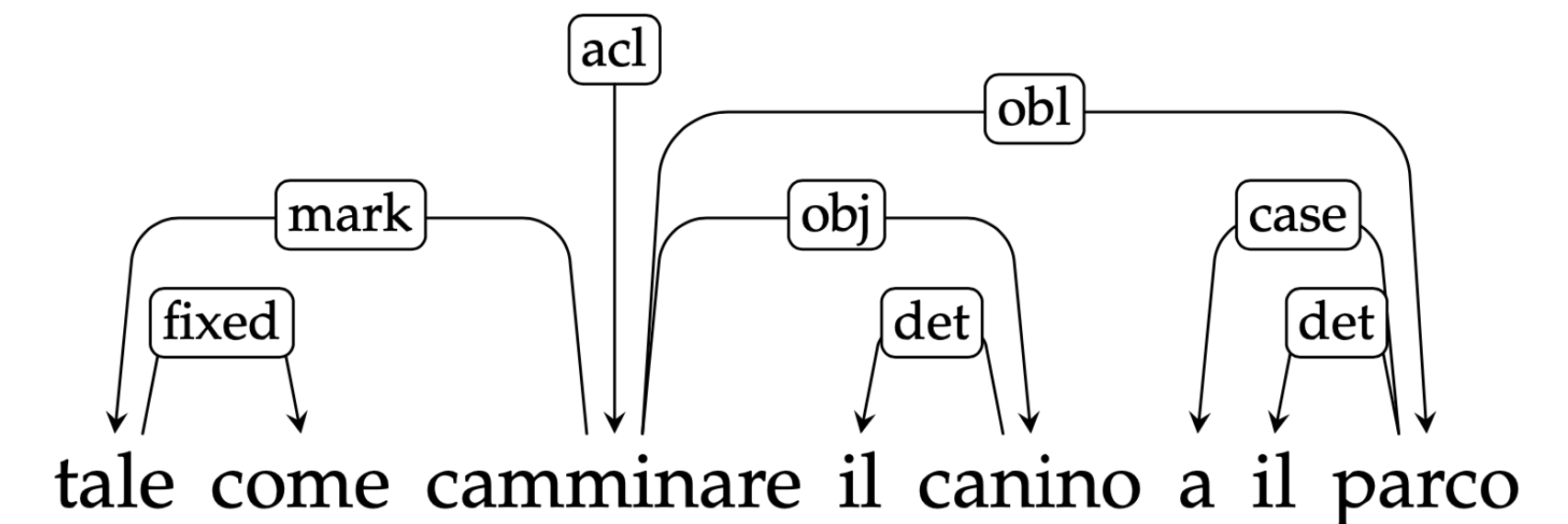
VALICO-UD: DEPENDENCY ANNOTATION

Contrastive interlanguage analysis

LS: Ho provato molte strategie per attrarre le ragazze tale come camminare il canino al parco.

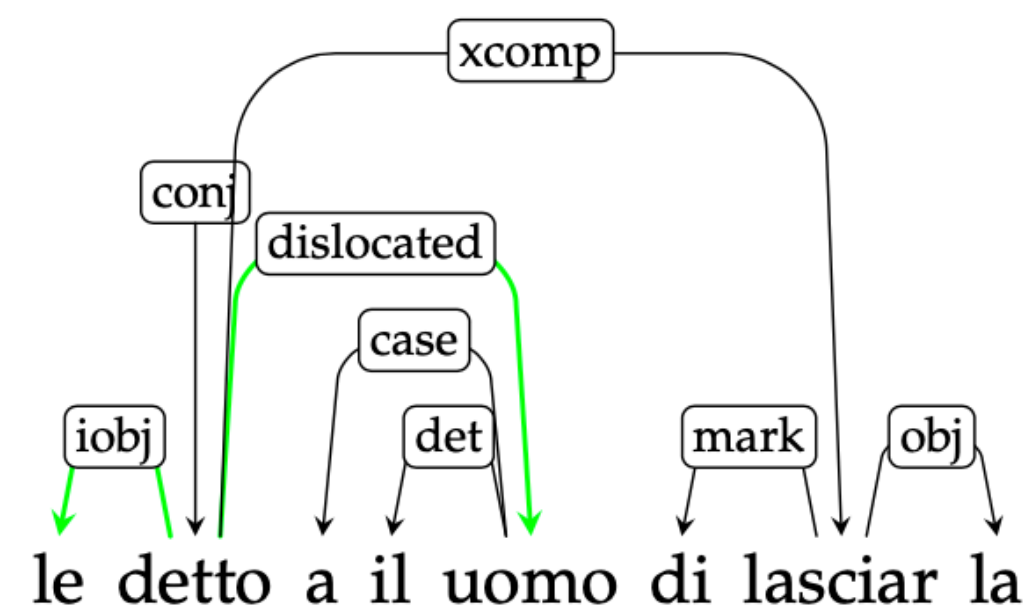
I have tried many strategies to attract girls such as walking the doggy in the park.

EWT:¹⁵ I looked at the UEComm Master and had some comments—such as our name is wrong, [...]



VALICO-UD: DEPENDENCY ANNOTATION

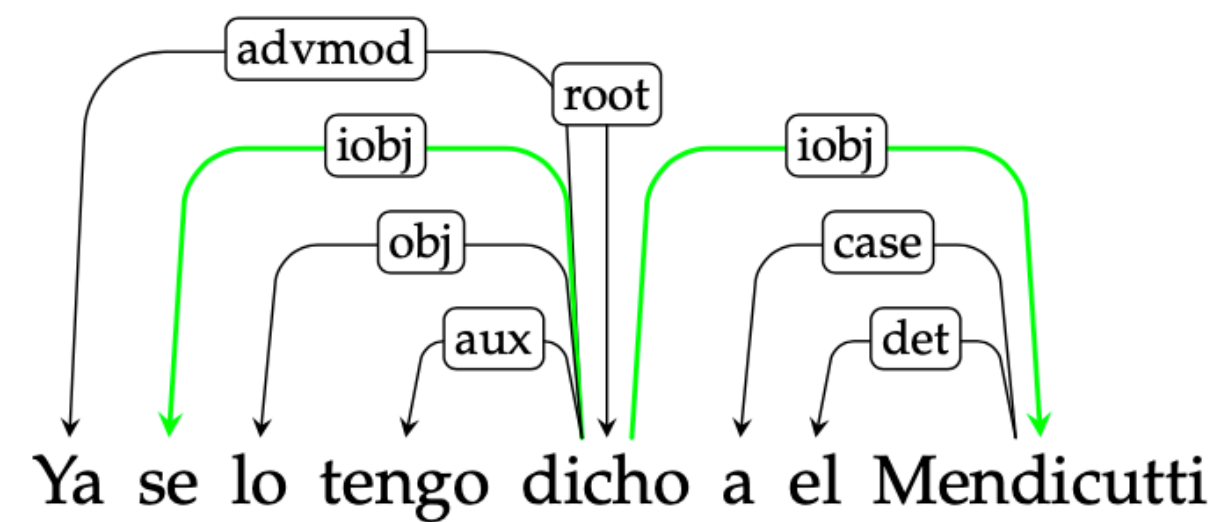
Clitic doubling (Pericchi et al. 2017) – Spanish



LS: MI he alzato e **le** detto **al** **uomo** de lasciarla perchè le faceva danno.

TH: Mi sono alzato e **gli** ho detto di lasciarla perché le faceva male.

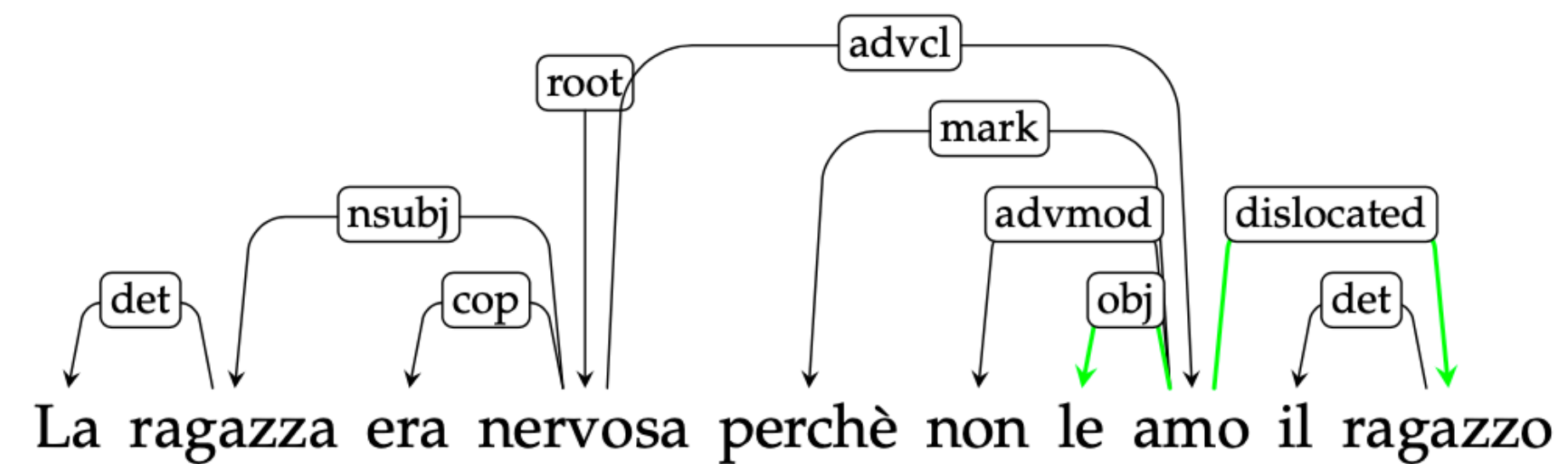
*I got up and told **him** to leave her because he was hurting her.*



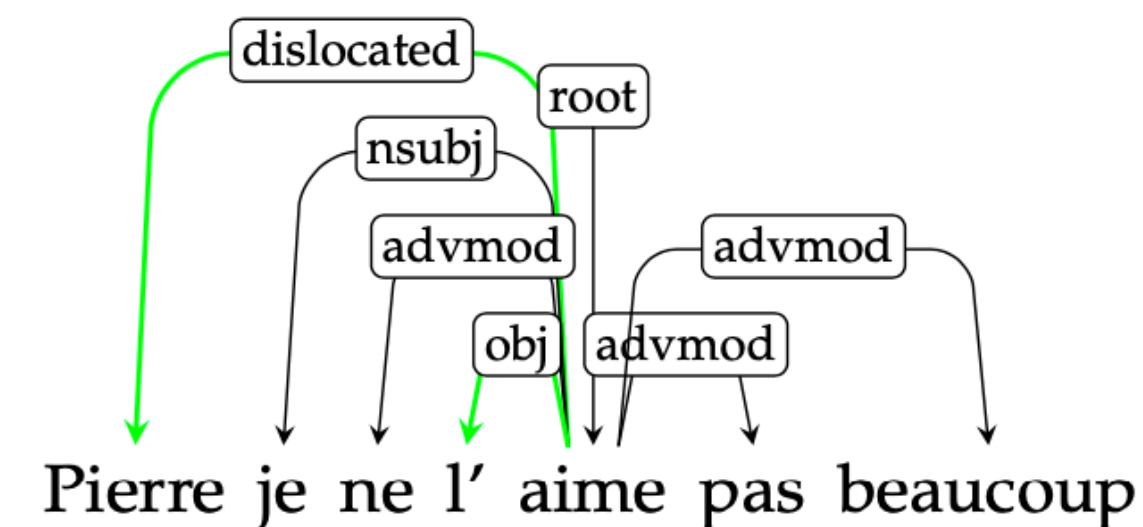
sent_id = train-s953 Ya se lo tengo dicho al Mendicutti [...]

I have already told this to Mendicutti

Clitic doubling / dislocation – French



*The girl was nervous because she did not love **him**, the **boy***



***Pierre** I don't like **him** very much*

VALICO-UD: PARALLEL READING OF (SUB)TREES

CASE STUDY: Comparing LSs and THs NOUN PHRASES

Python script available in GitHub (https://github.com/ElisaDiNuovo/VALICO-UD_NP_parallelReading)

It loops all the children of the noun linked via det, adj, nmod, acl:relcl and allows retrieving:

- * Unnecessary inflection of gender in gender invariable adjectives (e.g. *fragila* vs *fragile*)
- * Wrong lexical gender with correct internal agreement (e.g. *le uccelle* vs *gli uccelli*)
- * Different noun phrase structures (e.g. *l'uomo che aveva la rabbia* vs *l'uomo che era arrabbiato*)

RESULTS:

- * 39 agreement errors found, 34 concerning gender, 5 number
- * Gender- agreement found 21 times in noun-det pairs, 13 in noun-det-adj triplets

Results per L1:

- * Number agreement: ES 0.8%; FR 2.9%
- * Gender agreement: DE and EN 14.3%; FR 13.6%; ES 7.5%.

INCREMENTAL PARSER EVALUATION

Do interlanguage features affect UDPipe model performance?

- * Model trained on ISDT & PoSTWITA
- * TH gold standard to measure out-of-domain performance loss
- * LS gold standard to measure interlanguage influence on performance loss

	TH (UPOS LAS)	LS (UPOS LAS)
Rough data	96.18 83.45	94.20 78.51
Gold segmentation	96.19 84.69	94.22 79.61
Gold tokenisation	97.01 86.05	95.39 81.54
Gold tagging	Given 89.39	Given 87.96

INCREMENTAL PARSER EVALUATION

Do interlanguage features affect UDPipe model performance?

- * Model trained on ISDT & PoSTWITA
- * TH gold standard to measure out-of-domain performance loss
- * LS gold standard to measure interlanguage influence on performance loss

	TH (UPOS LAS)		LS (UPOS LAS)	
Rough data	96.18 83.45		94.20 78.51	
Gold segmentation	96.19 84.69		94.22 79.61	
Gold tokenisation	97.01	86.05	95.39	81.54
Gold tagging	Given	89.39	Given	87.96

1. CORRELATING LEARNERS' PROFICIENCY AND MT METRICS

Distance and similarity metrics comparing LSs and THs for distinguishing initial and advanced learners

VALICO-UD automatically-annotated section (silver standard)

Learners' L1	# texts per year of study					
	1	2	3	4	>4	?
DE	8	2	10	11	14	4
EN	7	21	3	13	3	4
ES	22	3	0	2	0	23
FR	7	13	5	4	20	2

A. TER (Translation Error Rate) (Snover et al. 2006): string distance

- * Variant of WER (Word Error Rate)—i.e. metrics that compute the Levenshtein distance (Levenshtein 1966) between the words of the system output (i.e. the LS) and the words of the reference translation (i.e. the TH), divided by the length of the reference translation (i.e. the TH)—overcoming the word-ordering limitation of WER.
- * HTER, Human-mediated Translation Error Rate—reference translations created by humans for calculating the minimum number of edits required to transform the system output into a fluent sentence with the same meaning as the references.

B. UDAPI (Popel et al. 2017): tree similarity

- * Compares LS and TH trees in pairs, calculating F1.

1A. CORRELATING LEARNERS' PROFICIENCY AND STRING DISTANCE (TER)

Distance and similarity metrics comparing LSs and THs for distinguishing initial and advanced learners

VALICO-UD automatically-annotated section (silver standard)

TER COM: string distance

TER value 0.375

LS: Ieri al parco è successo qualcosa stana.

TH: Ieri al parco è successo qualcosa di strano.

Yesterday in the park something strange happened.

Initial learners (average): **TER 0.29** (SD = 0.10, min = 0.11, max = 0.50)

Advanced learners (average): **TER 0.20** (SD = 0.10, min = 0.04, max = 0.39)

Unpaired *t* test: **$p < 0.0001$**

There is an extremely statistically significant difference between initial and advanced learners when described using the **TER string distance metric** between LSs and THs

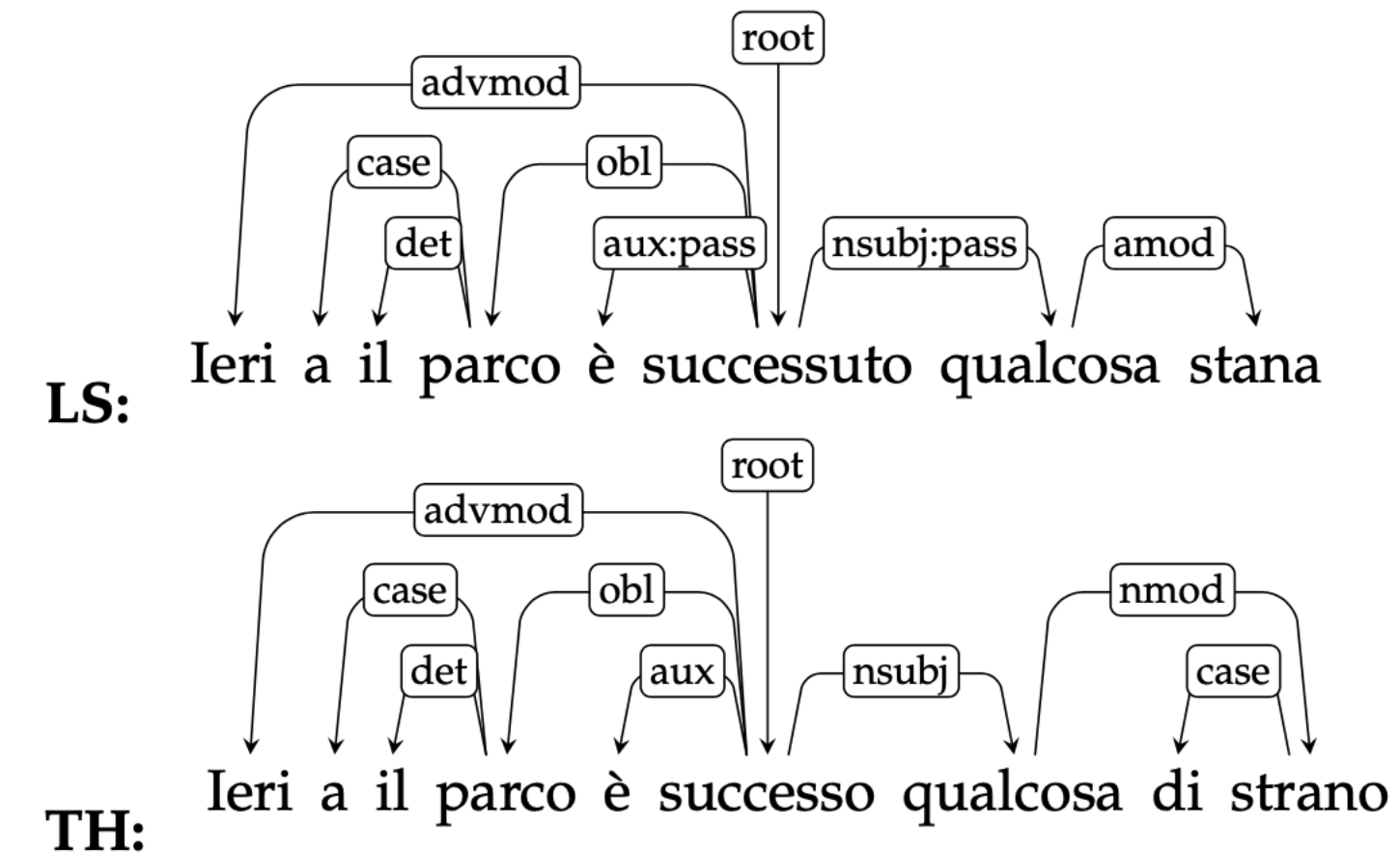
1B. CORRELATING LEARNERS' PROFICIENCY AND TREE SIMILARITY (F1)

Distance and similarity metrics comparing LSs and THs for distinguishing initial and advanced learners

VALICO-UD automatically-annotated section (silver standard)

UDAPI F1: tree similarity

F1 value 73.68%



Initial learners (average): **F1 0.80** (SD = 0.07, min = 65%, max = 93%)

Advanced learners (average): **F1 0.86** (SD = 0.07, min = 69%, max = 98%)

Unpaired *t* test: **$p < 0.0001$**

There is an extremely statistically significant difference between initial and advanced learners when described using the **F1 tree similarity metric** between LSs and THs

NLI

Performed using a linear multi-class SVM

Data Description:

237 texts:

- * 60 texts → EN
- * 59 texts → ES
- * 58 texts → DE
- * 60 texts → FR

Train/test splitting:

- * randomly selected 80%/20% (test: 48 texts, random_state=44)

Features:

- * TF-IDF(word ngram range=(1, 3))

Accuracy: 69%

HUMAN NLI

Performed by 25 humans (TT)

Data Description:

Test set: 48 texts used for NLI

	# Text/L1	TT/L1
EN	11	100%
FR	14	72%
DE	10	28%
ES	13	56%

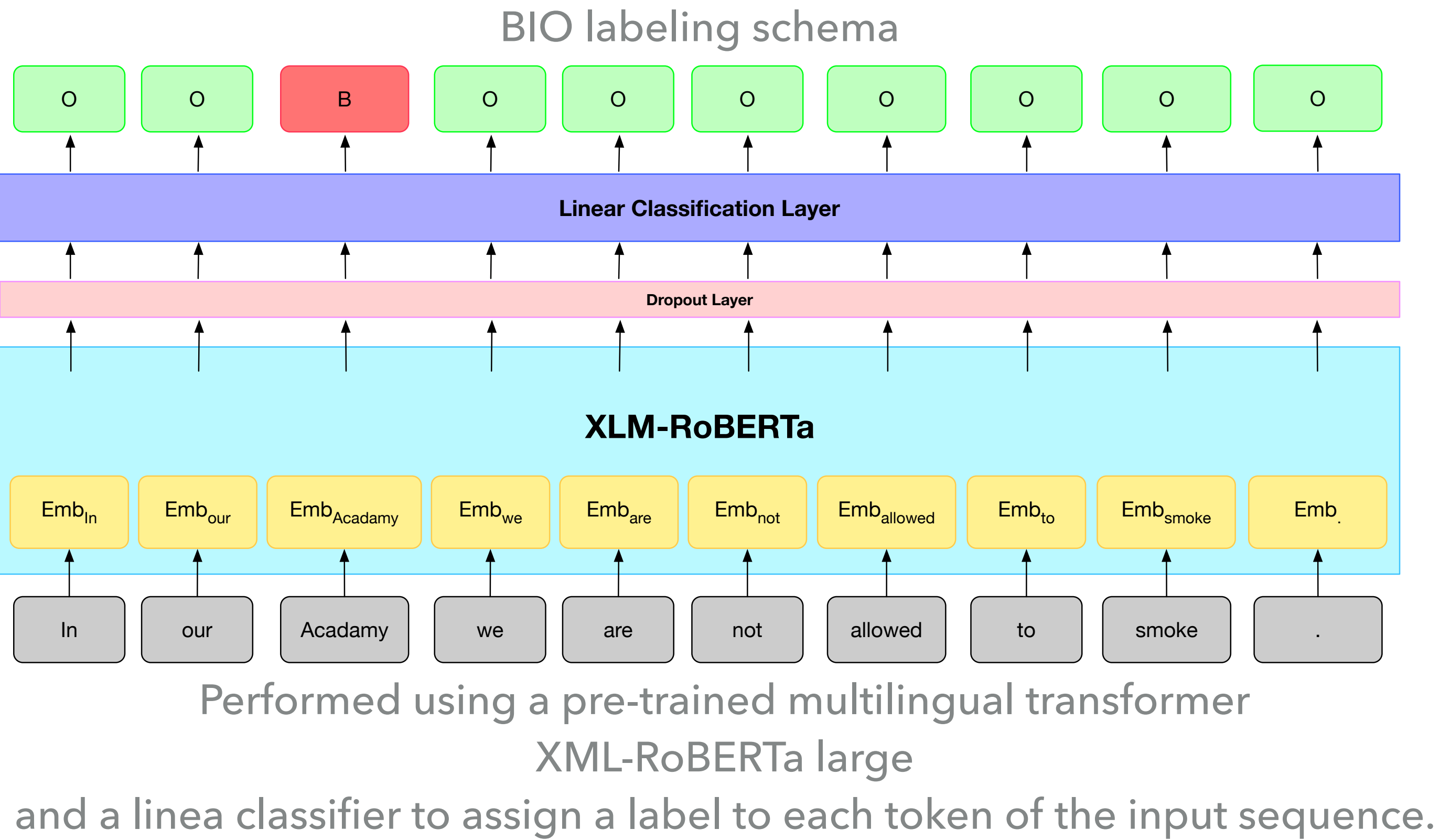
Features:

Knowledge of the L1s involved, teaching experience

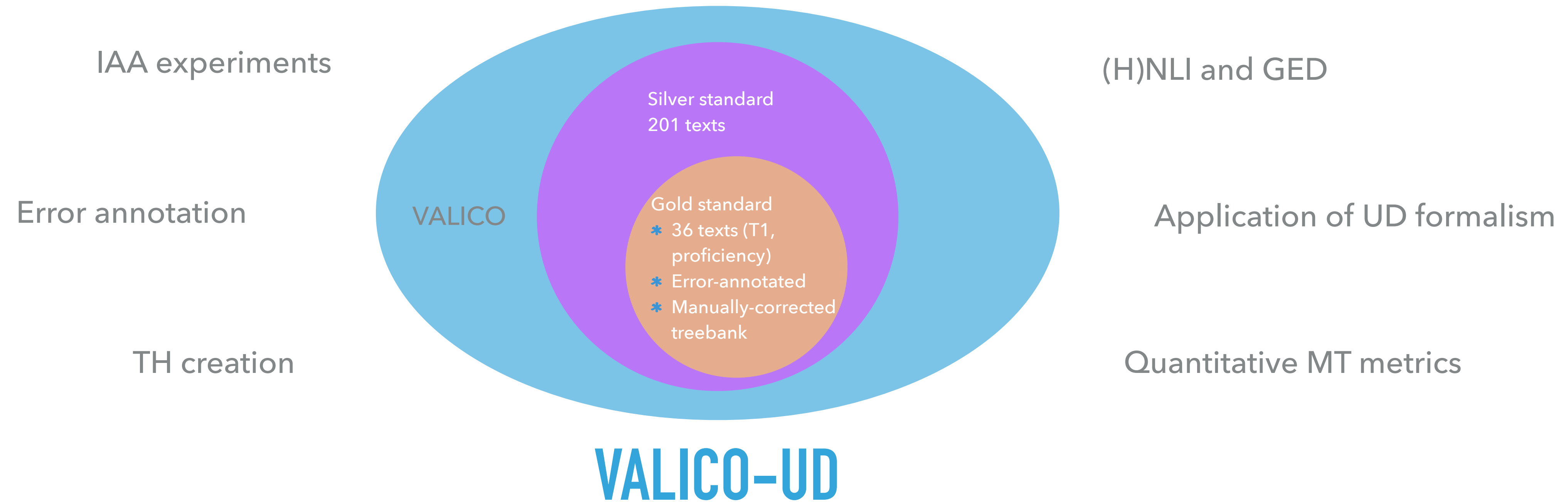
Accuracy: 54.2% (best TT)

MULTIGED – MULTILINGUAL GRAMMATICAL ERROR DETECTION

Source corpus	Language	Split	# Tokens	# Unique words
GECCC	Czech	train	333,995	37,228
		dev	32,071	8,145
		test	35,075	8,764
FCE	English	train	465,038	13,972
		dev	35,463	3,569
		test	42,545	3,800
REALEC	English	train	–	–
		dev	88,698	6,208
		test	90,391	6,300
Falko-MERLIN	German	train	306,847	20,561
		dev	39,627	5,606
		test	36,763	5,478
MERLIN	Italian	train	82,040	6,957
		dev	9,326	2,041
		test	10,300	2,176
SweLL-gold	Swedish	train	115,547	10,791
		dev	15,713	3,225
		test	14,666	3,141



System	Czech			English - FCE			English - REALEC			System	German			Italian			Swedish		
	P	R	F0.5	P	R	F0.5	P	R	F0.5		P	R	F0.5	P	R	F0.5	P	R	F0.5
ELICoDE _{ALL}	82.01	51.79	73.44	71.67	50.74	66.21	43.69	40.74	43.07	ELICoDE _{ALL}	84.78	73.75	82.32	86.67	67.96	82.15	81.80	66.34	78.16
ELICoDE _{MLT}	83.06	50.72	73.66	73.85	50.08	67.45	44.36	42.29	43.93	ELICoDE _{MLT}	83.47	72.52	81.02	85.30	69.64	81.63	82.24	65.94	78.36
ELICoDE _{MLTALL}	82.79	49.56	73.01	75.01	48.94	67.79	45.34	40.29	44.23	ELICoDE _{MLTALL}	84.80	71.09	81.65	85.71	65.95	80.87	83.34	64.37	78.70



Correlating error annotation and parser performance loss

Enlarge the gold standard:

- * Error annotation (for profiling learner proficiency through errors, or for NLI, GEC & GEI)
- * Gold tagging (e.g. improve the automatic parsing)
- * Have at least 2 annotators (200 sentences – treebank and 796 sentences – error annotation)



Thank you. Questions?

elisa.di-nuovo@ec.europa.eu