Using learner corpus tools in SLA research: the morpheme order studies revisited

Ana Díaz-Negrillo

anadiaznegrillo@ugr.es

Cristóbal Lozano

cristoballozano@ugr.es
http://wdb.ugr.es/~cristoballozano

Universidad de Granada

Corpus Linguistics Conference
Lancaster, UK
22th – 26th July, 2013

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Ref. FFI2012-30755

Revisiting the morpheme order studies (MOS) (1)

➢ The MOS (70s-80s) have been crucial in our understanding of IL in the SLA of English

- A remarkably consistent sequence independently of …
  - the learners' mother tongue (**L1**), **age** and learning **environment**
  - the testing method and the measuring instrument

1 progressive –ing

2 contractible copula –'s

3 plural –s

4 articles a(n)/the

5 contractible auxiliary (be) –'s

6 irregular past

7 regular past –ed

8 3rd person singular –s

9 possessive –'s

- Similar sequencing in child **L1 English**
- Different theoretical **explanations**: nativism (natural order), perceptual saliency, grammatical factors, etc.
- For overviews: Hawkins & Lozano 2006; Kwon 2005; Goldschneider & DeKeyser 2001

Revisiting the morpheme order studies (MOS) (2)

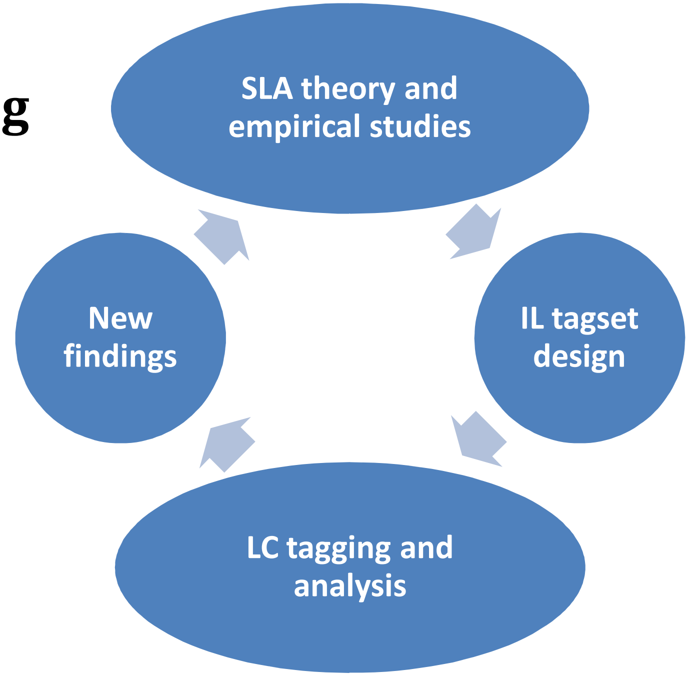
- Why are the MOS relevant for SLA and LCR?
- The MOS is a recently revived and controversial topic in SLA research (Goldschneider & DeKeyser 2001; Kwon 2005; Luk & Shirai 2009; Tono 2000)

> “The order that learners follow constitutes one of the most **important ‘facts’** that any theory of L2 acquisition must account for”
>
> (Ellis & Barkhuizen 2005: 91-92)

> “The study of learners’ use of morphemes through obligatory occasion analysis **still has much to offer SLA.** The descriptive information it provides serves as a basis for testing the validity of **different explanations of the order** of acquisition.”
>
> (Ellis & Barkhuizen 2005: 79)

Objectives

- Present an approach in LCR considering SLA as a point of departure
  - replicating **MOS**
    - replication in SLA is a necessary condition to (dis)confirm previous findings and to eliminate possible biases in the research method (Porte 2012)
  - using a different methodology
    - **learner corpora and corpus tagging**

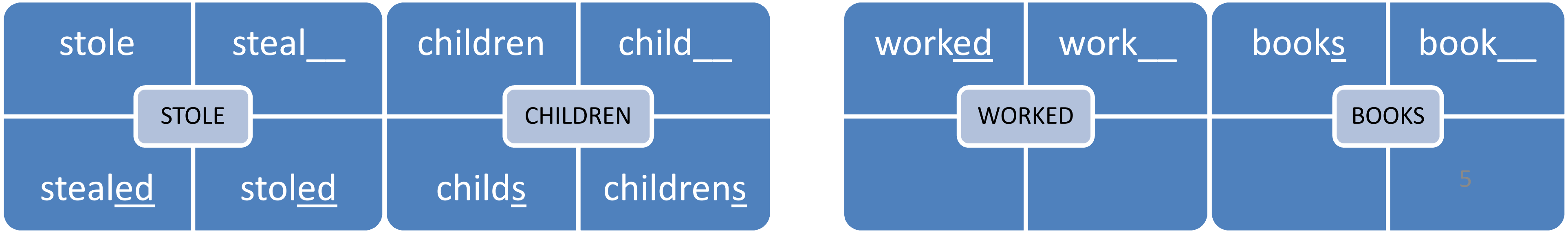


- Promote dialogue and synergies between LCR and SLA research (Tono 2003, Myles 2007)

"Learner corpus researchers should exchange ideas with SLA researchers in a more structured and systematic way. Many **corpus-based researchers** do not know enough about the theoretical background of SLA research to communicate with them effectively, while **SLA researchers** typically know little about what corpora can do for them. By improving the communication lines, we will be able to learn from each other." (Tono 2003: 806)

Methodological limitations of previous research (1)

- (Quasi)experimental methods have traditionally been used in the MOS:
  - **small L2 samples** under controlled conditions (except for Tono 2000; McEnery, Xiao & Tono 2006)
  - **native-oriented** approach (Ellis & Barkhuizen 2005: 92): unable to tell us about the forms that arise in learners' Interlanguage (IL)
    - Bley-Vroman's (1983) **Comparative Fallacy** (→see slides later)
  - **coarse-grained** in their analysis of learner productions since they do not fully explore all the subtypes of errors typical of learners' IL (**stealed, *stoled, *foots, *feets*, etc.)
    - We consider: ***U-shaped learning*** and the ***Dual Mechanism*** (→ see slide later)
    - We consider: ***Asymmetry in irreg. vs. reg. forms:***



stole
steal__
STOLE
stealed
stoled

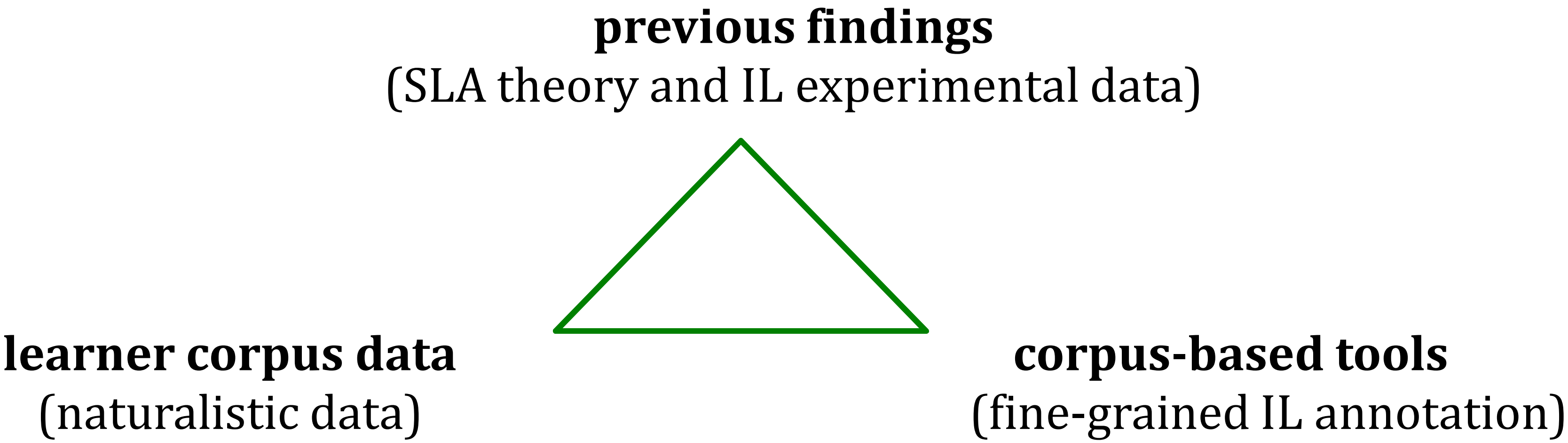
children
child__
CHILDREN
childs
childrens

worked
work__
WORKED

books
book__
BOOKS

Methodological limitations of previous research (2)

- Our approach aims to compensate limitations of MOS and LCR
  - by combining the methodological strengths of LCR and the theoretical explanatory power of SLA in MOS
  - for a fully-rounded picture of the acquisition of L2 English morphemes we need to triangulate:



previous findings
(SLA theory and IL experimental data)

learner corpus data
(naturalistic data)

corpus-based tools
(fine-grained IL annotation)

Our approach (1)

❶ **Learner corpus data**

- COREFL, CORpus of English as a Foreign Language
- Narrative written EFL texts, *Frog where are you?* - L1 Spanish
- Age: 12-17 (secondary school)
- Standardized proficiency level test: A1- C1 (*English Unlimited Placement Test*, CUP 2010)
- Size: approx. 100,000 words
- Ongoing (2012-)

❷ **Corpus-related methodology combined with SLA**:

- It moves away from bottom-up / corpus-driven / hypothesis-finding ← descriptive accounts of learner performance in LCR
- It takes a **top-down / corpus-based / hypothesis-testing approach** (cf. Myles 2005, 2007)

Our approach (2)

❸ **Corpus techniques combined with SLA: IL Annotation (ILA)**

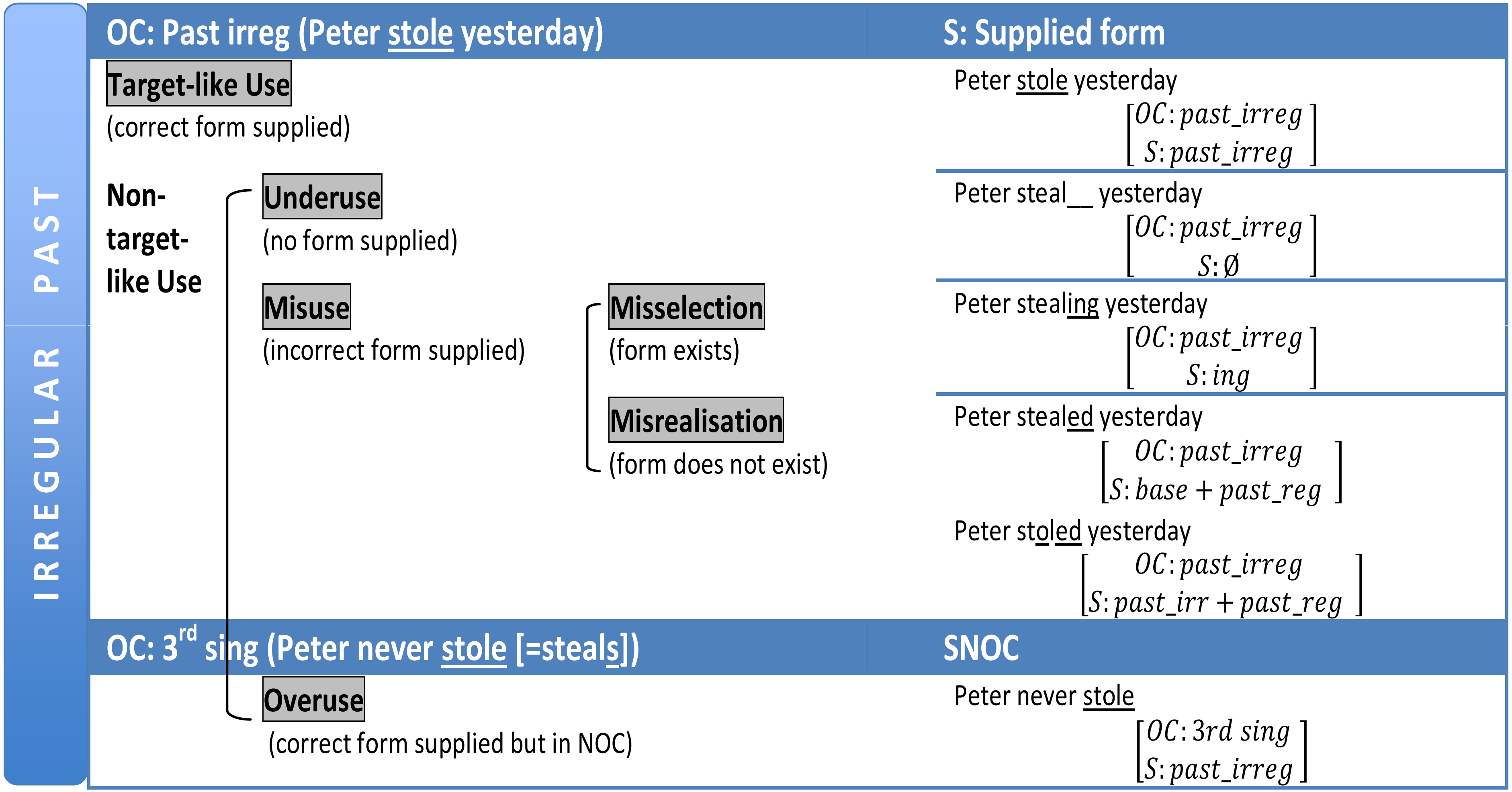
It moves away from the coarse-grained, all-purpose tagging of learners' errors. (cf., for example, Dagneaux et al. 1996; see Díaz-Negrillo & Fernández-Domínguez 2006 for an overview of error tagsets)

☑ **purposed-oriented:** designed for the study of morpheme acquisition.

☑ **fine-grained:** it categorises learner performance in detailed categories based on previous IL theory and findings.

☑ it considers both **non-target like (NTLU)** and **target-like (TLU)** uses.

☑ the tagset considers **all subtypes of NTLU uses**, some of which have been overlooked in previous tagging systems ➔ **rich tagset**



IRREGULAR PAST

OC: Past irreg (Peter stole yesterday)

S: Supplied form

Target-like Use
(correct form supplied)

Peter stole yesterday
$\begin{bmatrix} OC: past\_irreg \\ S: past\_irreg \end{bmatrix}$

Non-target-like Use

Underuse
(no form supplied)

Peter steal__ yesterday
$\begin{bmatrix} OC: past\_irreg \\ S: \emptyset \end{bmatrix}$

Misuse
(incorrect form supplied)

Misselection
(form exists)

Peter stealing yesterday
$\begin{bmatrix} OC: past\_irreg \\ S: ing \end{bmatrix}$

Misrealisation
(form does not exist)

Peter stealed yesterday
$\begin{bmatrix} OC: past\_irreg \\ S: base + past\_reg \end{bmatrix}$

Peter stoled yesterday
$\begin{bmatrix} OC: past\_irreg \\ S: past\_irr + past\_reg \end{bmatrix}$

OC: 3rd sing (Peter never stole [=steals])

SNOC

Overuse
(correct form supplied but in NOC)

Peter never stole
$\begin{bmatrix} OC: 3rd\ sing \\ S: past\_irreg \end{bmatrix}$

☑ It considers a bi-contextual approach: both **obligatory contexts (OC)** and **non-obligatory contexts (SNOC)**

the boy and the dog falled into the river

falled: **OC irregular past** → misrealization (=misformation)

falled: **NOC regular past** → overuse (SNOC)

☑ it considers a **bi-layered approach:** the **native** and the **non-native (IL)** perspective so as to overcome the 'Comparative Fallacy' (Bley-Vroman 1983),

e.g. **OC: reg. past**

Work in progress

And not wanted (Target: "And he didn't want")

Native layer: Overuse (SNOC)

IL layer: TLU

they climbed up into a tree

Native layer: TLU

IL layer: TLU

❹ **IL Scoring (ILS)** (frequency-based)

Work in progress

$$\textbf{ILS} = \frac{N\ correct\ suppliance\ in\ obligatory\ contexts + \boxed{(N\ SMOC\ *0.5)}}{N\ obligatory\ contexts + \boxed{N\ suppliance\ in\ non-obligatory\ contexts}} = \frac{SOC + (SMOC * 0.5)}{OC + SNOC}$$

Our learner corpus analysis with ILT

Corpus: COREFL

- sample of approx. 5,000 words
- 44 texts
- A2 and B1 levels (years 1-3, secondary education)

Interlanguage Annotion (ILA)

UAM corpus tool

	Irregular past	Regular past
A2	94	80
B1	157	153
TOTAL TAGS	**251**	**233**

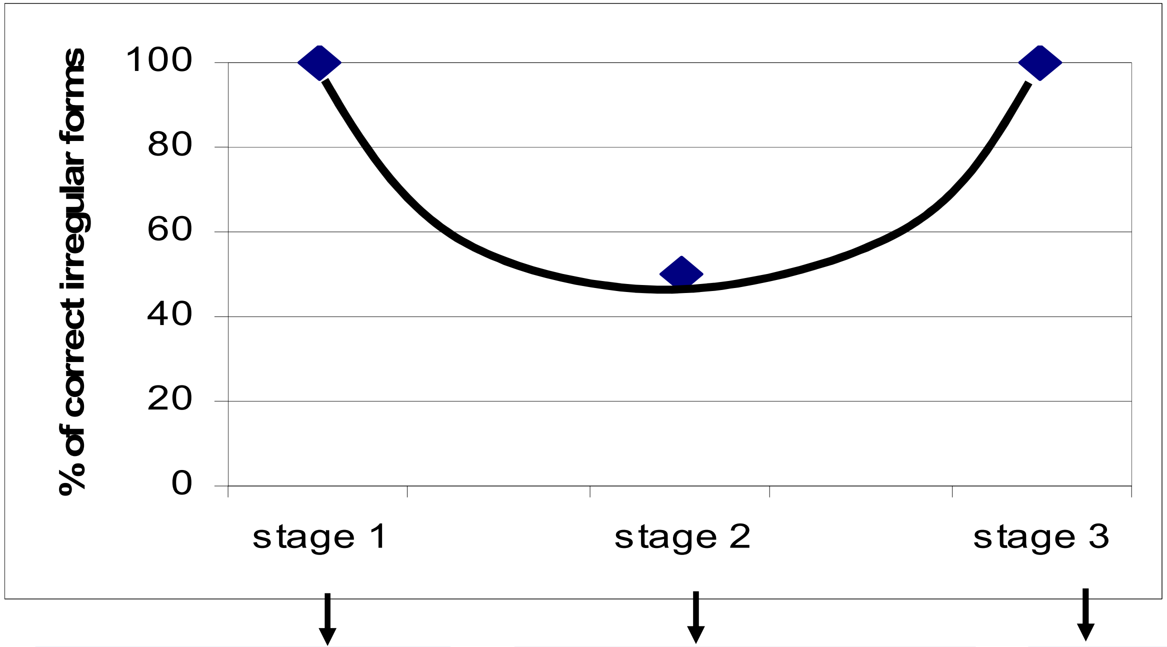
Interlanguage Scoring (ILS)

$$= \frac{SOC + (SMOC * 0.5)}{OC + SNOC}$$

Work in progress

A bit of experimental evidence on regular vs. irreg past before interpreting the corpus evidence…

- **U-shaped learning ➔ Dual Mechanism** for processing irregular vs. regular morphology (Pinker 1998).
- Observed in **L1** (Marcus et al. 1992, Pinker 1995) and **L2** (Zobl 1998, Birdsong & Flege 2001, Murphy 2004), *inter alia* --- but only L2 experimental evidence, **no corpus data.**

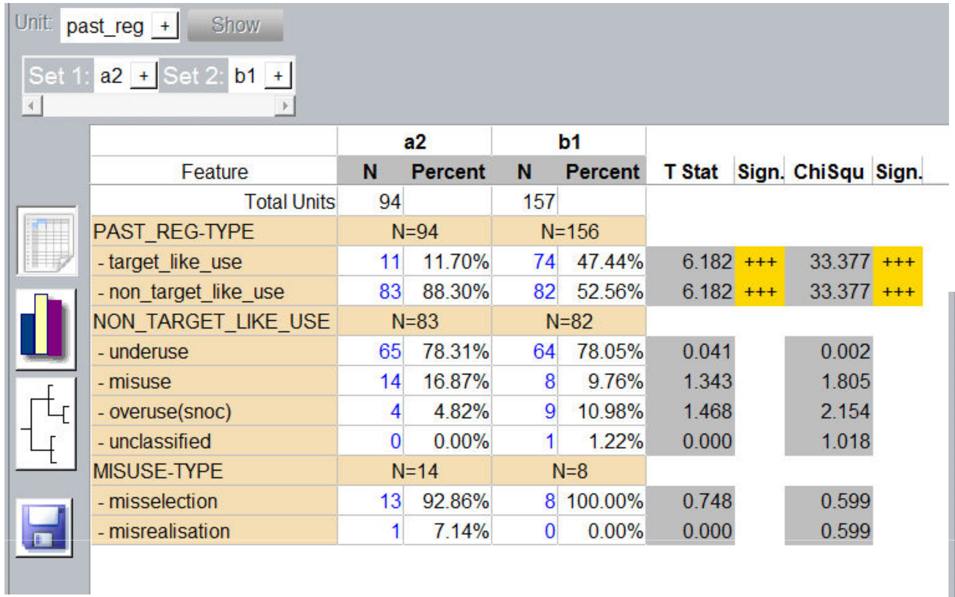


Form in associative memory:
BOUGHT
MICE

Acquisition of symbolic rule STEM+ "ed" / "s"
**BUYED*
**MOUSES*

Blocking of rule:
BOUGHT
MICE

Learner corpus analysis with ILA: results (1)

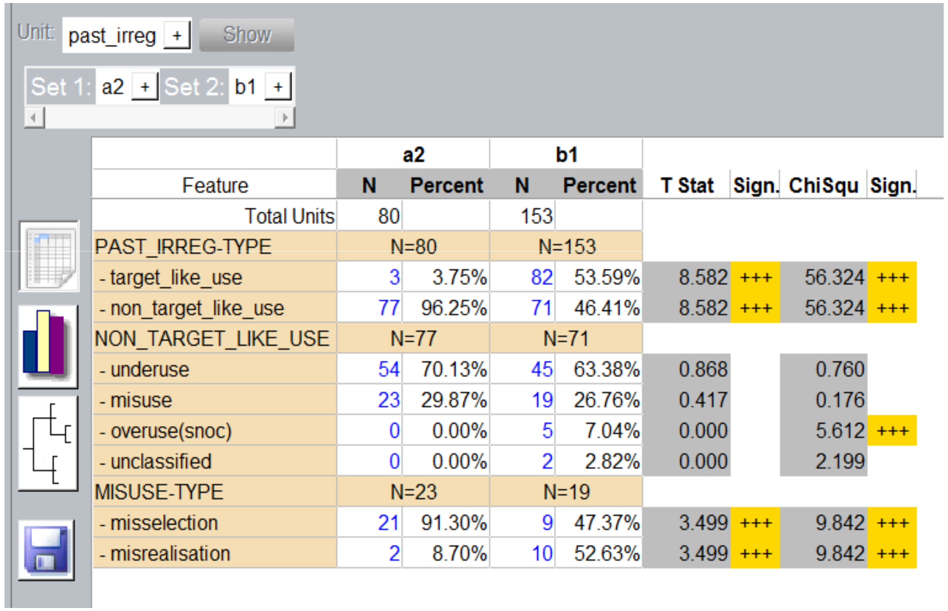


Unit: past_reg

Set 1: a2 Set 2: b1

	a2		b1					
Feature	N	Percent	N	Percent	T Stat	Sign.	ChiSqu	Sign.
Total Units	94		157					
PAST_REG-TYPE	N=94		N=156					
- target_like_use	11	11.70%	74	47.44%	6.182	+++	33.377	+++
- non_target_like_use	83	88.30%	82	52.56%	6.182	+++	33.377	+++
NON_TARGET_LIKE_USE	N=83		N=82					
- underuse	65	78.31%	64	78.05%	0.041		0.002	
- misuse	14	16.87%	8	9.76%	1.343		1.805	
- overuse(snoc)	4	4.82%	9	10.98%	1.468		2.154	
- unclassified	0	0.00%	1	1.22%	0.000		1.018	
MISUSE-TYPE	N=14		N=8					
- misselection	13	92.86%	8	100.00%	0.748		0.599	
- misrealisation	1	7.14%	0	0.00%	0.000		0.599	

Regular past



Unit: past_irreg

Set 1: a2 Set 2: b1

	a2		b1					
Feature	N	Percent	N	Percent	T Stat	Sign.	ChiSqu	Sign.
Total Units	80		153					
PAST_IRREG-TYPE	N=80		N=153					
- target_like_use	3	3.75%	82	53.59%	8.582	+++	56.324	+++
- non_target_like_use	77	96.25%	71	46.41%	8.582	+++	56.324	+++
NON_TARGET_LIKE_USE	N=77		N=71					
- underuse	54	70.13%	45	63.38%	0.868		0.760	
- misuse	23	29.87%	19	26.76%	0.417		0.176	
- overuse(snoc)	0	0.00%	5	7.04%	0.000		5.612	+++
- unclassified	0	0.00%	2	2.82%	0.000		2.199	
MISUSE-TYPE	N=23		N=19					
- misselection	21	91.30%	9	47.37%	3.499	+++	9.842	+++
- misrealisation	2	8.70%	10	52.63%	3.499	+++	9.842	+++

Irregular past

Let's explore each of these in detail…

Learner corpus analysis with ILA: results (2)

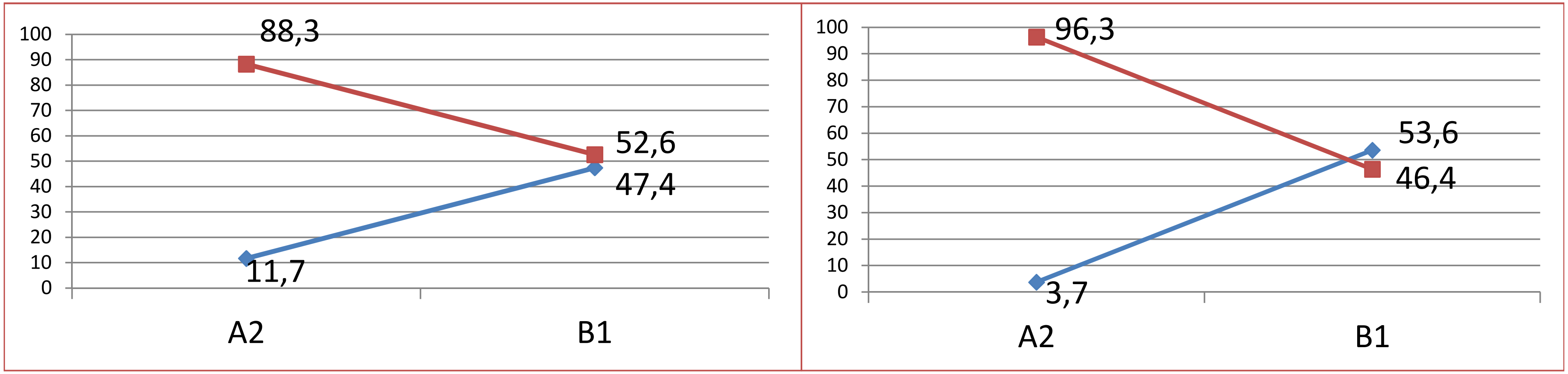
➢ TLU vs. NTLU

Regular past

- 🟦 *were sleeping the frog escaped from the vase*
- 🟥 *while the boy was sleeping, the frog scape*

Irregular past

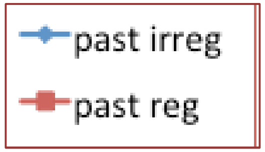
- 🟦 *Then they saw a deer*
- 🟥 *They leave the forest and moved the hand*



- **Development from A2 to B1**: significant and drastic decrease in NTLU for both regular and irregular past ($p<0.05$) ➔ L2ers start to acquire past tense from intermediate stages (B1 onwards).

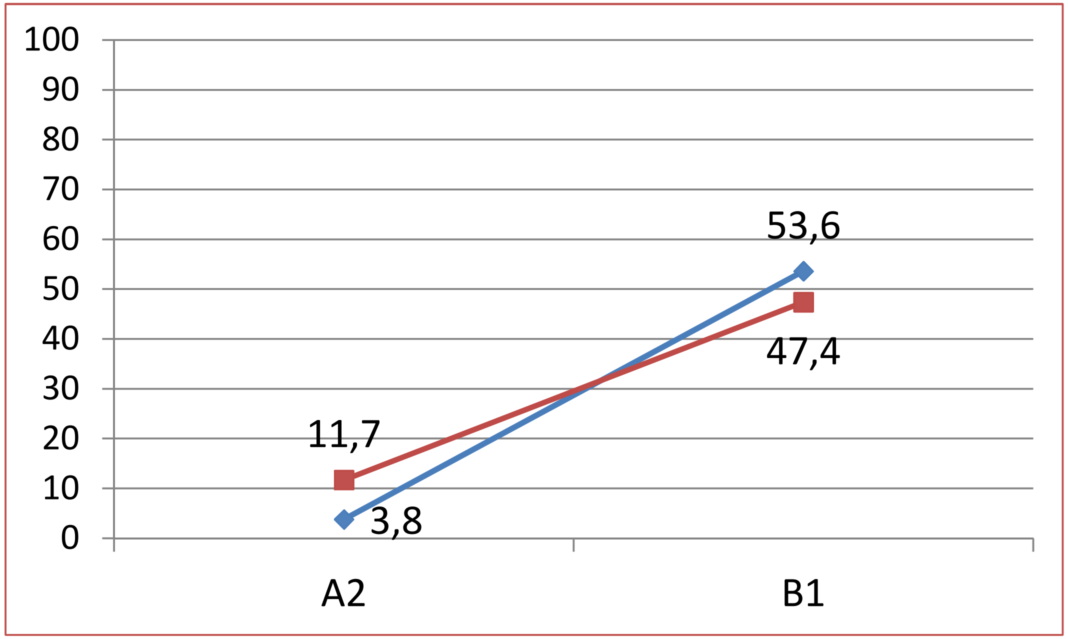
Learner corpus analysis with ILA: results (3)

➢ **TLU**



🟦 *Then they <u>saw</u> a deer*

🟥 *were sleeping the frog <u>escaped</u> from the vase*



- Inverted results for A2 and B1 groups
- It is only at **B1** (low intermediate) that **irregular** > **regular** past ($p<0.05$) ➔ irregular forms precede regular forms (in line with MOS)

Learner corpus analysis with ILA: results (4)

➢ **NTLU**

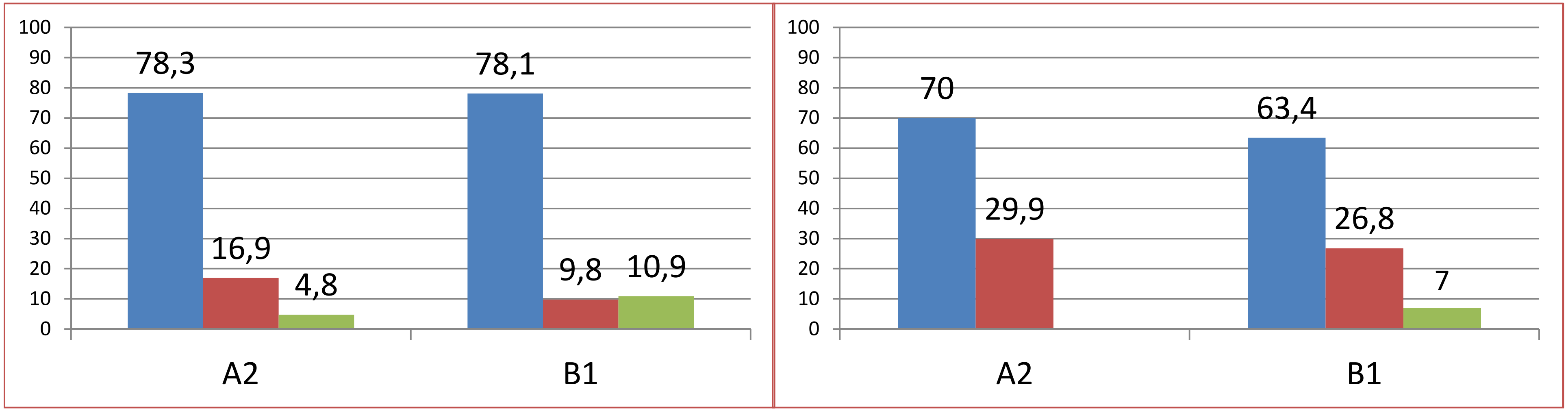
Regular past

🟦 while the boy was sleeping, the frog scape
🟥 He searchs for all over the river
🟩 a deer catched the boy.

🟦 Underuse
🟥 Misuse
🟩 Overuse(SNOC)

Irregular past

🟦 the boy go to sleep because was latter
🟥 a deer catched the boy
🟩 He don’t found the frog.

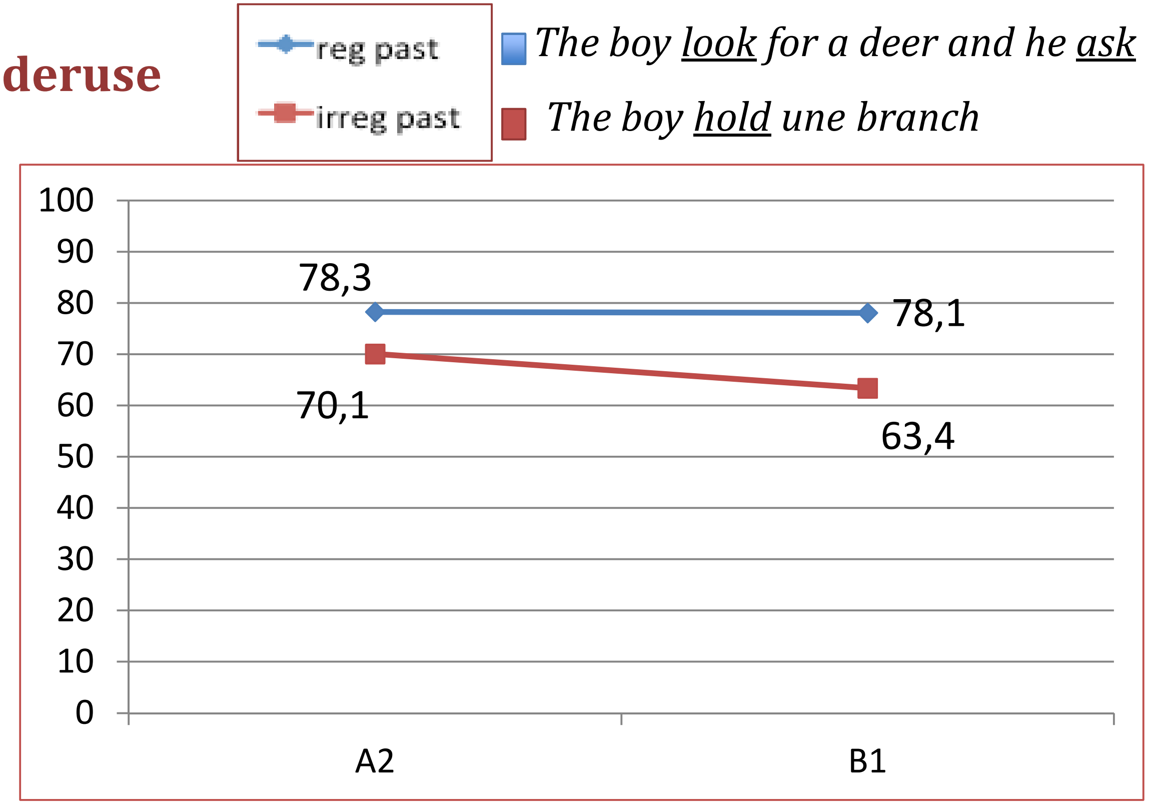


- **Underuse** is by far the most frequent error at all levels and with both morphemes ➔ learners have not fully acquired yet the inflected forms (-ed) and the irregular forms.
- **Misuse**: irregular>regular at both levels ➔ to be discussed in detail later
- **Overuse** is the least frequent tag in all levels and in both morphemes

Let’s explore each of these in detail…

Learner corpus analysis with ILA: results (5)

➢ **NTLU 1: Underuse**



🟦 *The boy look for a deer and he ask*

🟥 *The boy hold une branch*

Regular past

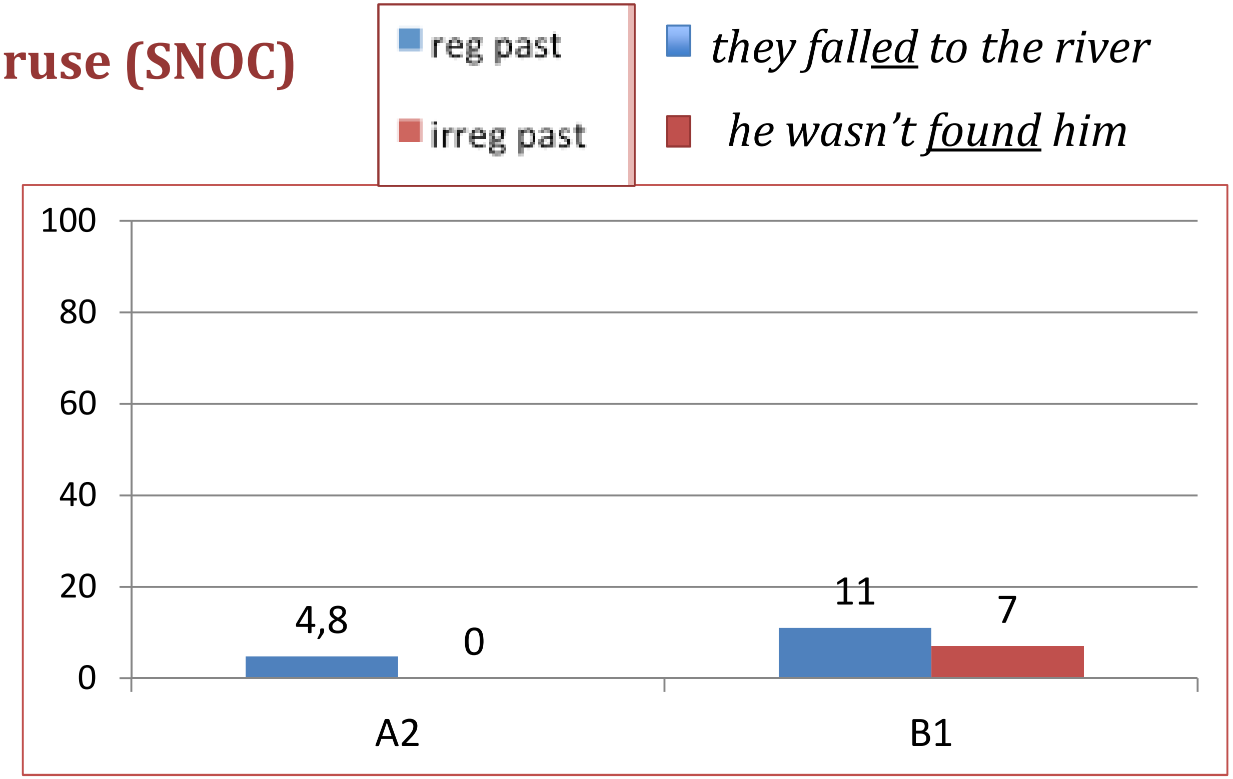
- production of **Ø morphology** is **stable** across levels; more **likely** to appear with **regular verbs** (irregular morph. is listed in associative memory in the mental lexicon).
- Not even the inflection for 3rd ps. sing.. This is more frequent in A2 learners.

Irregular past

- a NTLU **decrease** from A2 (70%) to B1 (63%) **signals TLU of irregulars** (recall: irreg>reg in intermediates).
- some **frequent irreg verbs** are **inflected** (*saw*, *went* vs. *hold*, *fall*) → high frequency prevents overregularizations according to Blocking Principle in '**Dual mechanism**' (Marcus et al 1992).

Learner corpus analysis with ILA: results (6)

➢ **NTLU 2: Overuse (SNOC)**



🟦 *they falled to the river*

🟥 *he wasn't found him*

Regular past

- An **increase in overuse** of **-ed** morpheme in irregular past contexts (4.8% at A2 and 11% at B1) reflects overregularisation at **intermediate** (B1) stages, as predicted by 'Dual Mechanism' model.

Irregular past

- All examples involve negative constructions (results to be taken cautiously).
- Double marking strategy??? **[PAST] → wasn't + irreg_past**

Learner corpus analysis with ILA: results (7)

➢ NTLU 3: Misuse (misslelection vs. misrealization)

Regular Past

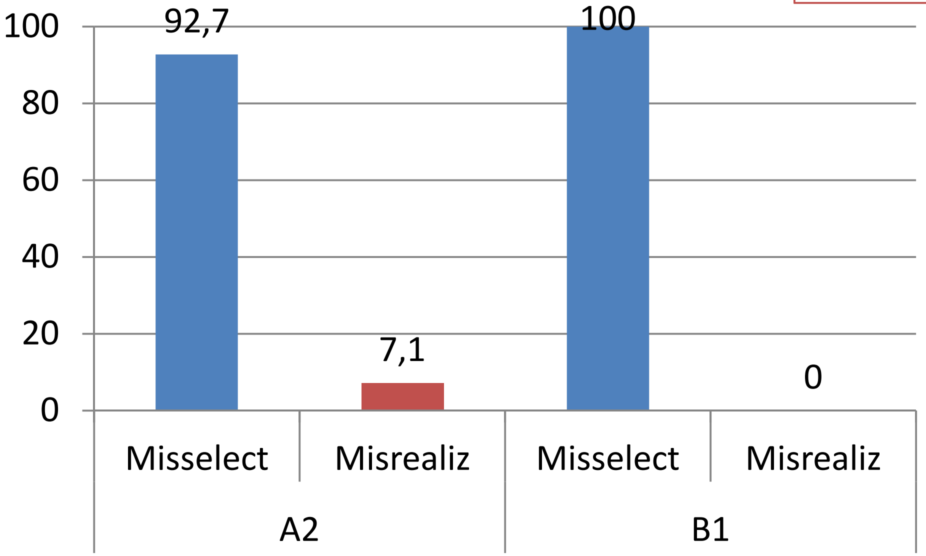
- 🟦 because tries to run away
- 🟥 The boy and the dog worrieds

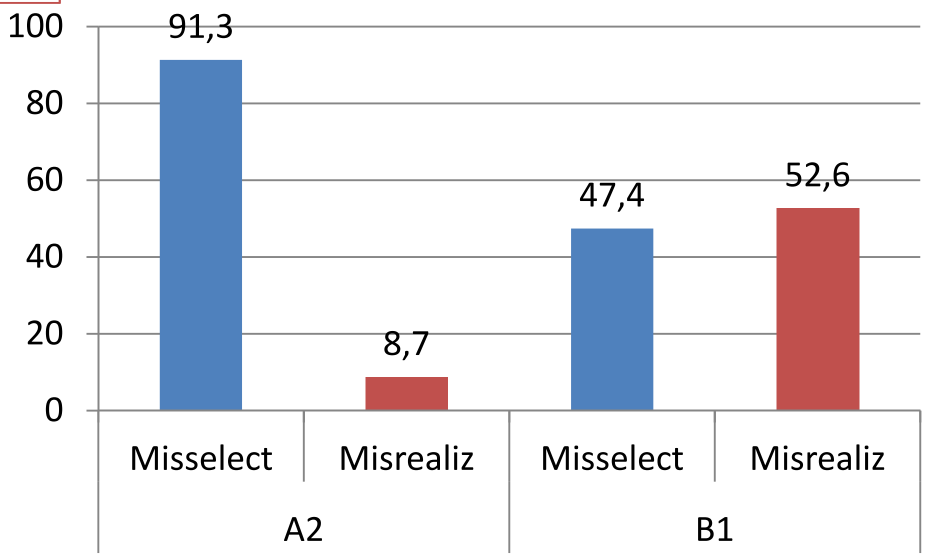
🟦 Misselect

🟥 Misrealiz

Irregular Past

- 🟦 then he falls
- 🟥 they broked their legs



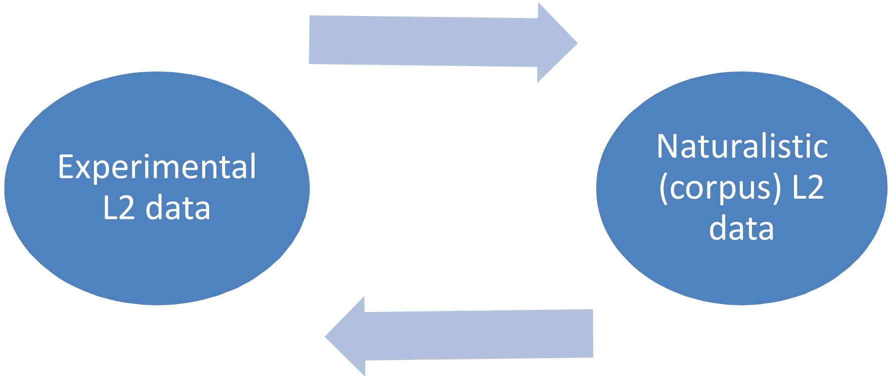


Imbalance regular vs irregular past:

- **Regular past:**
  - **misselection > misrealiz at all proficiency levels .**
  - **93% + errors are misselect. of 3rd sing –s: *escapes* (=*escaped*) etc.**
  - **only 7% errors are misrealiz (agreement added to past tense): worrieds (=worried)**
- **Irregular past: Proficiency effect**
  - **A2 (beginners): misselection > misrealiz: again 3rd singular –s: *falls* (=*fell*) etc.**
  - **B1 (low interm.): misselection ≤ misrealiz: *falled* (=*fell*) etc. ➔overregularization clearly starts at intermediate stages (Dual Mechanism)**

Conclusion

- This study has illustrated a different approach in LCR which
  - sets off from SLA theory...
  - uses learner corpus research methods…
  - proposes ILA (Interlanguage Annotation)
- Future work
  - **annotation** of the corpus for the rest of the morphemes
  - further exploration of the **bi-layered approach**
  - further **specification of the annotation** categories based on SLA findings:
    - tense-aspect categories: telicity, accomplishments, states, etc.
    - interface with other aspects: negation, passivization, etc.
  - triangulation of **corpus data** with **experimental data**



References

Birdsong, D., & Flege, J. E. (2001). Regular-irregular dissociations in L2 acquisition of English morphology. In *BUCLD 25: Proceedings of the 25th Annual Boston University Conference on Language Development* (pp. 123–132). Boston, MA: Cascadilla Press.

Brown, R. (1973). *A First Language.* Cambridge, MA: Harvard University Press.

Cambridge University Press (2010). *English Unlimited Placement Test*. Cambridge: CUP.

Díaz Negrillo, A. 2009. *EARS: A User's Manual*. Munchen: Licon.

Dulay, H.C. & Burt, M.K. (1980). On acquisition orders., In S. Felix (Ed.), *Second Language Development: Trends and Issues*. Tübingen: Narr.

Ellis, R., & Barkhuizen, G. P. (2005). *Analyzing Learner Language*. Oxford University Press.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, 51(1), 1–50.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 259–275). Berlin: Mouton de Gruyter.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 13–32). Amsterdam: John Benjamins.

Granger, S. (2012). How to use Foreign and Second Language Learner Corpora. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 5–29). Oxford: Wiley-Blackwell.

Hawkins, R. & Lozano, C. (2006). Second Language Acquisition of Phonology, Morphology and Syntax. In: K. Brown (ed). *The Encyclopedia of Language and Linguistics* (2nd Edition) (pp. 67-74). Oxford: Elsevier.

Kwon, E.-Y. (2005). The "Natural Order" of morpheme acquisition: A historical survey and discussion of three putative determinants. *Columbia University Working Papers in TESL & Applied Linguistics*, 5(1), 1–21.

References (2)

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order or grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural -s, articles, and possessive ’s. *Language Learning*, 59(4), 721–754.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, R. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 54(4 (serial no. 228)).

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book* (Unit C3). London: Routledge.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4), 373–391.

Murphy, V. A. (2004). Dissociable systems in second language inflectional morphology. *Studies in Second Language Acquisition*, 26(3), 433–459.

Myles, F. (2007). Using electronic corpora in SLA research. In D. Ayoun (Ed.), *Handbook of French Applied Linguistics* (pp. 377–400). Amsterdam: John Benjamins.

Pica, T. (1984). Methods of morpheme quantification: their effect on the interpretation of second language data. *Studies in Second Language Acquisition,* 6, 69-78.

Pinker, S. (1995). Invitation to Cognitive Science. Volume 1: Language. In M. Gleitman & M. Liberman (Eds.), Why the child holded the baby rabbits: A case study in language acquisition (2nd ed.). Cambridge, MA: MIT Press.

Pinker, S. (1998). Words and rules. *Lingua*, 106, 219–242.

Porte, G. (Ed.). (2012). *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.

Tono, Y. (2000). A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: papers from the Third International Conference on Teaching and Language Corpora* (pp. 123–132). Peter Lang.

Tono, Y. (2003). Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the 2003 Corpus Linguistics Conference* (pp. 800–809). UCREL, Lancaster University: UCREL Technical Paper number 16.