Second language acquisition modeling based on crowdsourced learner data

enetCollect - CA16105

Ildikó Pilán
ildiko.pilan@svenska.gu.se
University of Gothenburg, Sweden

Overview of the MCIF proposal

- **NLP** in language learning: potential for **individualized solutions**
- **Goal**: model **beginner** L2 **English** learners' **development** on a **large scale** in relation to the Common European Framework of Reference for Languages (**CEFR**) and **Reference Level Descriptions** (RLD)
- **Motivation**: make the learning process more efficient (shorten learning time, increase / maintain learner motivation)

- **Research questions**:
  - *1. Given previous errors of learners, how efficiently can we predict their future errors?*
  - *2. Are there more persistent error types across learners?*
  - *3. Are the error types observed in accordance with CEFR RLD?*
  - *4. Is there a difference in L2 development based on the geographical location of learners?*
- **Contributions**: 1. L2 acquisition model
  2. frequent error types related to CEFR and location

Methodology

- Both **quantitative** and **qualitative** methods (experimental method?)
- Data-driven approaches based on **machine learning**
- Data required: **longitudinal** learner **data** → Duolingo Shared Task
- **Universal dependencies**: facilitates a comparative cross-lingual analysis and use of results
- Develop (or re-use) **error-detection** and **categorization** methods
  - Data from 2014 Shared Task on Grammatical Error Correction
- Potential **comparison** to **additional corpora** (e.g. learner essays)

Funded by the Horizon 2020 Framework Programme of the European Union

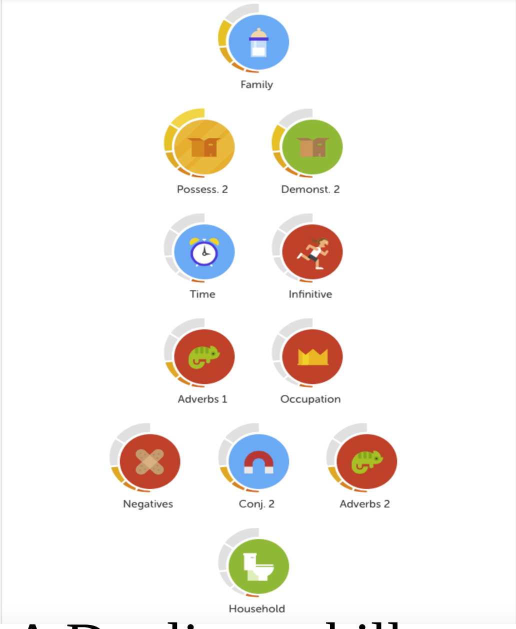
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Språk-BANKEN

GÖTEBORGS UNIVERSITET

Duolingo Shared Task

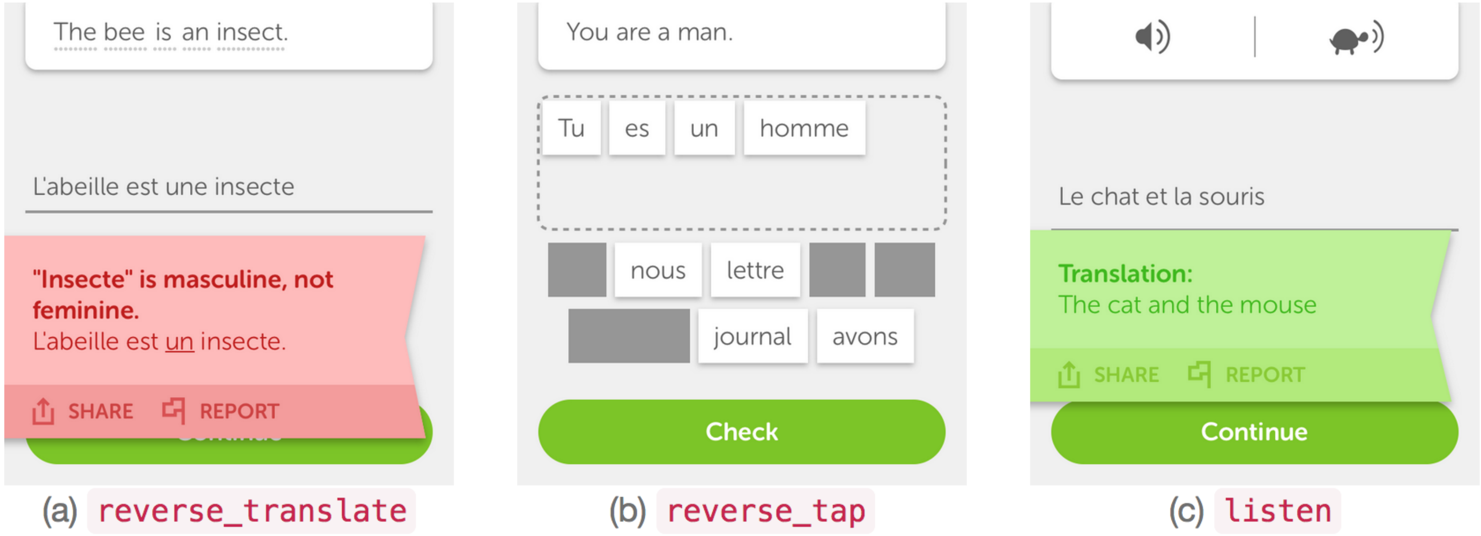
- Topic: Second Language Acquisition Modeling
  - concluding March – April 2018
  - **publicly available** dataset (Creative Commons)
- Duolingo:
  - a free online language learning platform
  - game-like language courses



A Duolingo skill tree

Data description

- English, Spanish and French L2 learners' data
- 3 different **tasks** ➡
- Collected during the first 30 days of platform use



The bee is an insect.

L'abeille est une insecte

"Insecte" is masculine, not feminine.
L'abeille est un insecte.

(a) reverse_translate

You are a man.

Tu es un homme

nous lettre journal avons

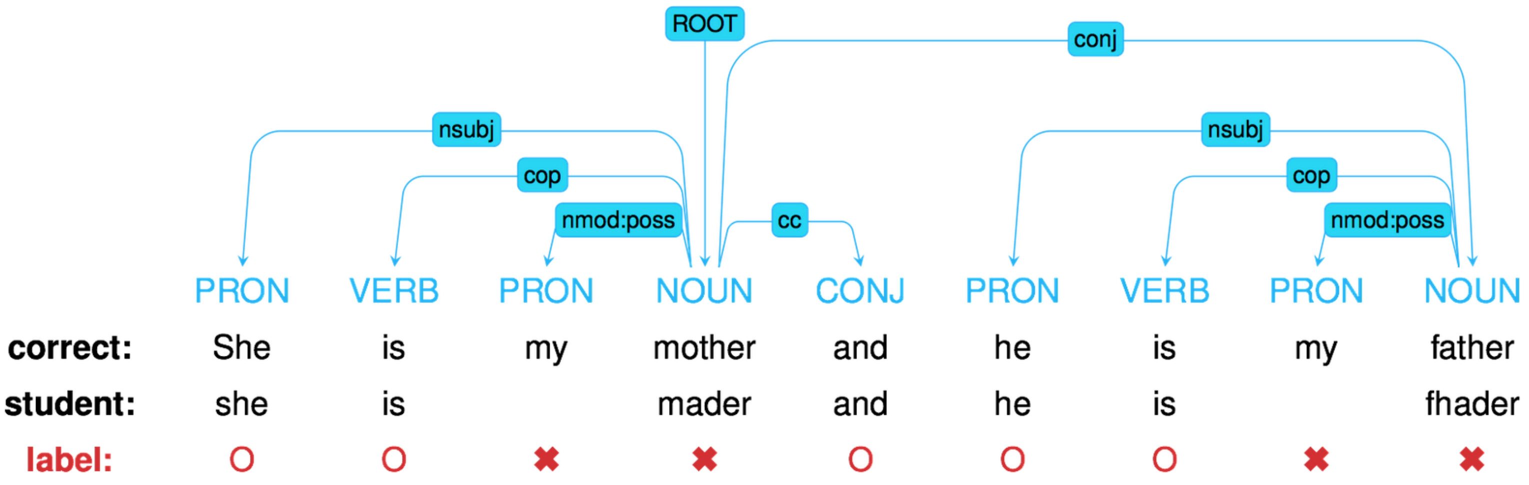
(b) reverse_tap

Le chat et la souris

Translation:
The cat and the mouse

(c) listen

- Available **annotations**:



	PRON	VERB	PRON	NOUN	CONJ	PRON	VERB	PRON	NOUN
correct:	She	is	my	mother	and	he	is	my	father
student:	she	is		mader	and	he	is		fhader
label:	O	O	✖	✖	O	O	O	✖	✖

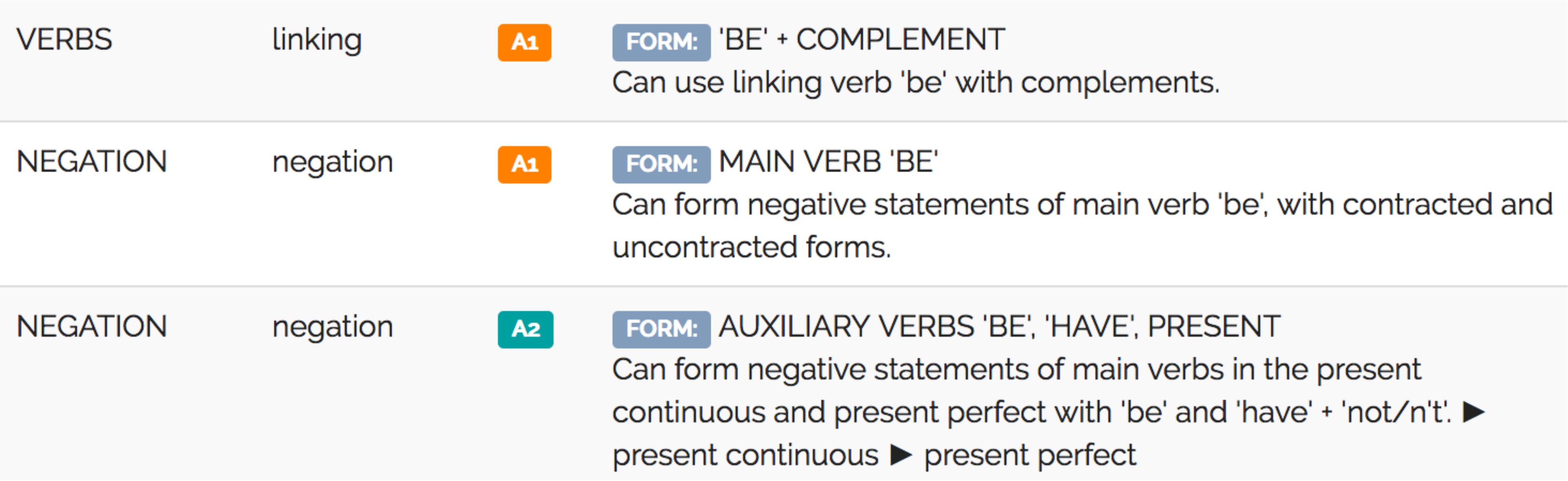
- **Additional** available **information**:
  - **Time**-related data: number of days **from** the **start** of platform use
    time **taken to submit** each answer
  - **Countries** of login
  - Another **language** known (English or Spanish)
- Distribution of tokens: 83% correct – 27% incorrect

Example of longitudinal Duolingo data

solution (bold = incorrect)	*days*	*format*
You are very welcome	2.695	reverse-translate
You are **very** welcome	2.703	reverse-translate
You are **very** welcome	2.739	reverse-translate
You are very welcome	5.725	reverse-translate
Thank you and **you are** welcome	9.205	listen
You are **welcome**	9.210	listen
Thank you **and** you **are** welcome	9.210	listen

(user ID:XEinXf5+)

CEFR RLDs: English Grammar Profile



VERBS	linking	A1	FORM: 'BE' + COMPLEMENT Can use linking verb 'be' with complements.
NEGATION	negation	A1	FORM: MAIN VERB 'BE' Can form negative statements of main verb 'be', with contracted and uncontracted forms.
NEGATION	negation	A2	FORM: AUXILIARY VERBS 'BE', 'HAVE', PRESENT Can form negative statements of main verbs in the present continuous and present perfect with 'be' and 'have' + 'not/n't'. ► present continuous ► present perfect

http://englishprofile.org/english-grammar-profile/egp-online

Additional dimensions to explore

- How does the amount and the spacing of repetitions affect the acquisition process?
- Are there more persistent grammatical error types across languages?

Limitations

- Availability of **actual** learner **answers** (not published yet, but promised to be released soon)
- **Enough data** to represent well **single error types**?
- **Not manually annotated**, but automatically corrected answers