Gathering data

to prove that we can implicitly crowdsource language-related datasets from language learners

Lionel Nicolas & Verena Lyding, Eurac Research, Bolzano, Italy

1st WG3/WG5 Meeting, 24th October 2018, Leiden, Netherlands

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

European Network for Combining Language Learning
with Crowdsourcing Techniques (enetCollect)

Summary

- Objectives of this presentation.
- Implicit crowdsourcing model combining language-related datasets with language learning exercises.
- Foreseen datasets:
  - tracked learners' answers to automatically-generated exercises,
  - tracked learners' answers to exercises with large sets of questions.
- Added value in providing datasets.
- 3 Questions to the audience.

Summary

- **Objectives of this presentation.**
- Implicit crowdsourcing model combining language-related datasets with language learning exercises.
- Foreseen datasets:
  - tracked learners’ answers to automatically-generated exercises,
  - tracked learners’ answers to exercises with large sets of questions.
- Added value in providing datasets.
- 3 Questions to the audience.

Objectives of this presentation

Overall objective

⇒ To start a data collection initiative that will enable several interesting lines of research within enetCollect.

⇒ Among others, it will allow to demonstrate the viability of an implicit crowdsourcing model.

⇒ You can participate in providing data for the collection or in helping to analyze the data.

Objectives of this presentation

This presentation IS NOT about:

- presenting (preliminary) results to prove that we can implicitly crowdsource language-related datasets from learners,
- presenting well-known / fully-demonstrated / widely-consensual facts about how we can implicitly crowdsource language-related datasets from learners.

Objectives of this presentation

This presentation is about:

- ✓ Explaining an implicit crowdsourcing theoretical model combining language-related datasets and language learning exercises.
- ✓ Starting to gather sets of learner data in the form of answers to exercises that a could serve multiple purposes of three WGs (WG2, WG4 & WG5).
- ✓ Explaining what is the added value in providing such data for you and for enetCollect.

Summary

✓ Objectives of this presentation.

➢ **Implicit crowdsourcing model combining language-related datasets with language learning exercises.**

➢ Foreseen datasets:

  ➢ tracked learners’ answers to automatically-generated exercises,

  ➢ tracked learners’ answers to exercises with large sets of questions.

➢ Added value in providing datasets.

➢ 3 Questions to the audience.

Implicit crowdsourcing theoretical model
[General idea]

Some language learning exercises can be automatically generated from a language-related (e.g. NLP) dataset (e.g. POS lexica, treebanks, wordnets).

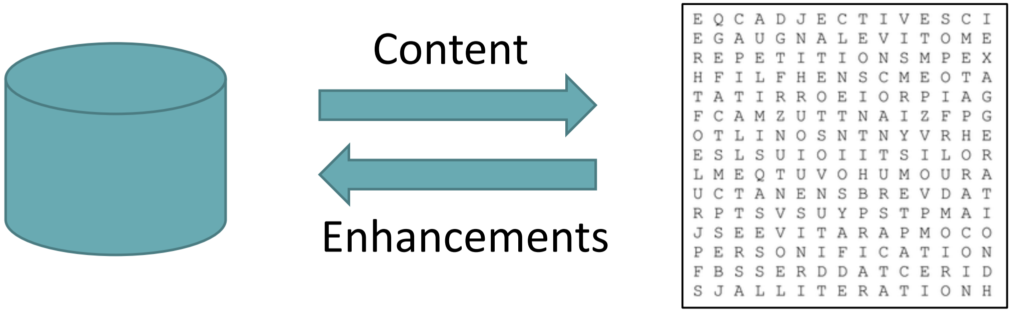
Theoretical model

IF

a dataset can be used to generate a language learning exercise,

THEN

the answers to such exercise can be used to enhance the dataset.

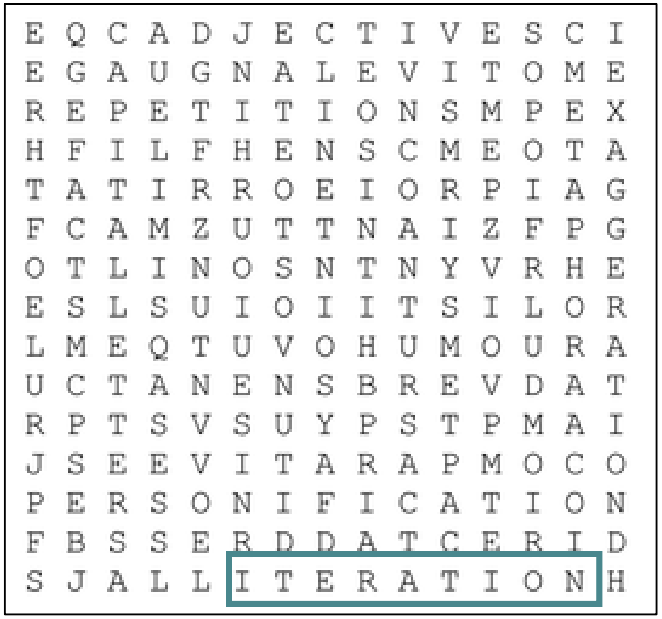


```
E Q C A D J E C T I V E S C I
E G A U G N A L E V I T O M E
R E P E T I T I O N S M P E X
H F I L F H E N S C M E O T A
T A T I R R O E I O R P I A G
F C A M Z U T T N A I Z F P G
O T L I N O S N T N Y V R H E
E S L S U I O I I T S I L O R
L M E Q T U V O H U M O U R A
U C T A N E N S B R E V D A T
R P T S V S U Y P S T P M A I
J S E E V I T A R A P M O C O
P E R S O N I F I C A T I O N
F B S S E R D D A T C E R I D
S J A L L I T E R A T I O N H
```

Implicit crowdsourcing theoretical model

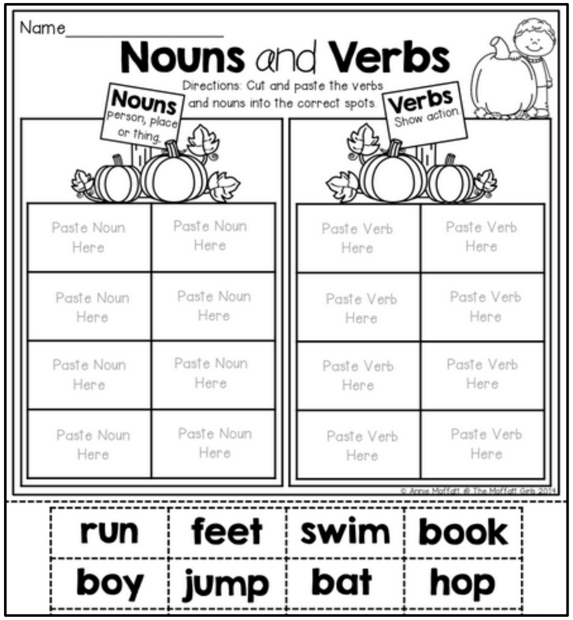
[Examples of exercises]

"Word search" exercises



```
E Q C A D J E C T I V E S C I
E G A U G N A L E V I T O M E
R E P E T I T I O N S M P E X
H F I L F H E N S C M E O T A
T A T I R R O E I O R P I A G
F C A M Z U T T N A I Z F P G
O T L I N O S N T N Y V R H E
E S L S U I O I I T S I L O R
L M E Q T U V O H U M O U R A
U C T A N E N S B R E V D A T
R P T S V S U Y P S T P M A I
J S E E V I T A R A P M O C O
P E R S O N I F I C A T I O N
F B S S E R D D A T C E R I D
S J A L L I T E R A T I O N H
```

"Classify words" exercises



Name____________

Nouns and Verbs

Directions: Cut and paste the verbs and nouns into the correct spots.

Nouns person, place or thing.

Verbs Show action.

Paste Noun Here	Paste Noun Here	Paste Verb Here	Paste Verb Here
Paste Noun Here	Paste Noun Here	Paste Verb Here	Paste Verb Here
Paste Noun Here	Paste Noun Here	Paste Verb Here	Paste Verb Here
Paste Noun Here	Paste Noun Here	Paste Verb Here	Paste Verb Here

© Annie Moffatt @ The Moffatt Girls 2014

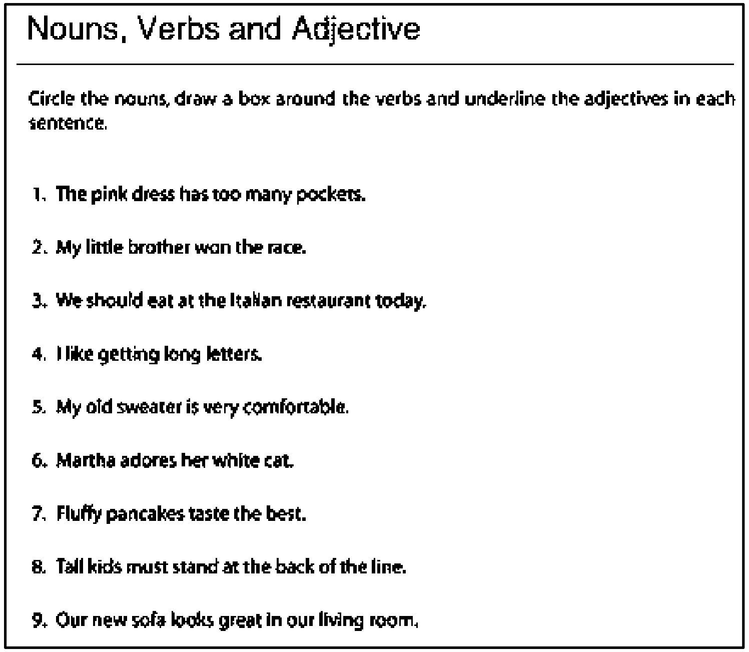
run	feet	swim	book
boy	jump	bat	hop

- Questions can be generated from a POS lexicon.
- Answers can be used to correct or extend a POS lexicon.

Implicit crowdsourcing theoretical model

[Examples of exercises]

"Circle the word" exercises

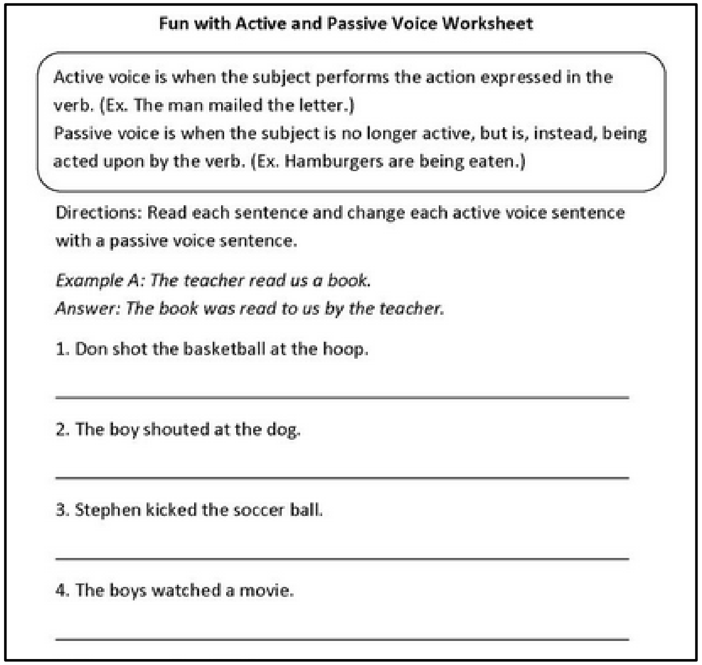


Nouns, Verbs and Adjective

Circle the nouns, draw a box around the verbs and underline the adjectives in each sentence.

1. The pink dress has too many pockets.
2. My little brother won the race.
3. We should eat at the Italian restaurant today.
4. I like getting long letters.
5. My old sweater is very comfortable.
6. Martha adores her white cat.
7. Fluffy pancakes taste the best.
8. Tall kids must stand at the back of the line.
9. Our new sofa looks great in our living room.

"Passive / Active voice" exercise



Fun with Active and Passive Voice Worksheet

Active voice is when the subject performs the action expressed in the verb. (Ex. The man mailed the letter.)
Passive voice is when the subject is no longer active, but is, instead, being acted upon by the verb. (Ex. Hamburgers are being eaten.)

Directions: Read each sentence and change each active voice sentence with a passive voice sentence.

Example A: The teacher read us a book.
Answer: The book was read to us by the teacher.

1. Don shot the basketball at the hoop.

____

2. The boy shouted at the dog.

____

3. Stephen kicked the soccer ball.

____

4. The boys watched a movie.

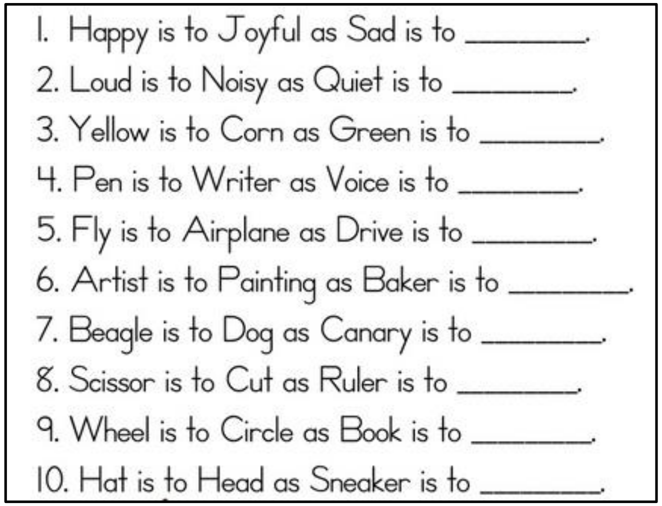
____

- Questions can be generated from a treebank.
- Answers can be used to correct or extend a treebank.

Implicit crowdsourcing theoretical model

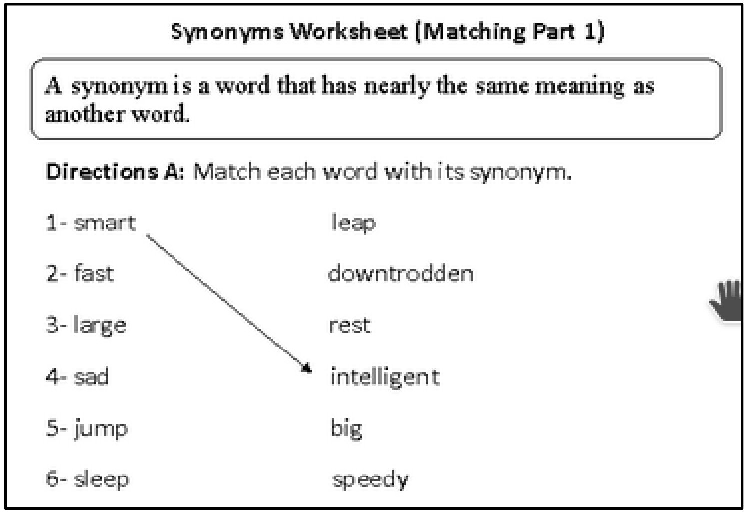
[Examples of exercises]

"Analogy" exercises



1. Happy is to Joyful as Sad is to ________.
2. Loud is to Noisy as Quiet is to ________.
3. Yellow is to Corn as Green is to ________.
4. Pen is to Writer as Voice is to ________.
5. Fly is to Airplane as Drive is to ________.
6. Artist is to Painting as Baker is to ________.
7. Beagle is to Dog as Canary is to ________.
8. Scissor is to Cut as Ruler is to ________.
9. Wheel is to Circle as Book is to ________.
10. Hat is to Head as Sneaker is to ________.

"Synonymy" exercises



Synonyms Worksheet (Matching Part 1)

A synonym is a word that has nearly the same meaning as another word.

Directions A: Match each word with its synonym.

1- smart	leap
2- fast	downtrodden
3- large	rest
4- sad	intelligent
5- jump	big
6- sleep	speedy

- Questions can be generated from a Wordnet.
- Answers can be used to correct or extend a Wordnet.

Summary

✓ Objectives of this presentation.

✓ Implicit crowdsourcing model combining language-related datasets with language learning exercises.

➢ **Foreseen datasets:**

  ➢ **tracked learners’ answers to automatically-generated exercises,**

  ➢ **tracked learners’ answers to exercises with large sets of questions.**

➢ Added value in providing datasets.

➢ 3 Questions to the audience.

Foreseen datasets

- The datasets foreseen are simple => sets of tracked learner answers to language learning exercises, as collected by some LL platforms.

- Two similar types of datasets foreseen (see slides after).

- Only limited amount of information needed for each answer, for example:
  - learner id,
  - exercise id,
  - question id,
  - timestamp,
  - answer,
  - correctness of the answer.

(1) Tracked learners’ answers to automatically-generated exercises

- ➢ Automatically-generated exercises can have a large number of different questions for one exercise type.

- ➢ Collecting learner answers would allow to study if the set of learner answers overall confirm or contradict the correct answer (i.e. the data used to generate the question).

- ➢ If the learners answer contradict the correct solution more than they do for other questions, then we can assume that:

  (1) the question is more difficult => the learner answers are of less quality,

  (2) the solution is incorrect (i.e. and so is the data that generated it).

  ⇒This would prove that learners can help correcting the datasets and that the theoretical model is viable in such use-case.
  ⇒The studies would already allow to detect corrections.

(2) Tracked learners’ answers to exercises with large sets of questions

- An exercise type with a sufficiently large number of different questions could be considered as automatically-generated from a small dataset.

- The reasoning explained previously for automatically generated-exercises is still valid even though there are no real datasets here.

- The number of questions of the exercise should be large enough to run some statistics (e.g. more than 100).

Summary

- ✓ Objectives of this presentation.
- ✓ Implicit crowdsourcing model combining language-related datasets with language learning exercises.
- ✓ Foreseen datasets:
  - ✓ tracked learners’ answers to automatically-generated exercises,
  - ✓ tracked learners’ answers to exercises with large sets of questions.
- ➢ **Added value in providing datasets.**
- ➢ 3 Questions to the audience.

Added value in providing datasets

On a theoretical level, helping to demonstrate the model will help developing two noticeable win-win aspects for multiple stakeholders (you included):

- ➢ First win-win: every improvement to the dataset benefit the crowd of learners by improving the quantity and validity of the generated questions.
- ➢ Second win-win: the more crowdsourced manpower = the more support from language-related R&I actors for the language learning community.

On a practical level:

- ➢ You would concretively contribute to enetCollect,
- ➢ You would enable interesting works in other WGS (WG2, WG4 & WG5),
- ➢ Your datasets will be studied and you will receive feedback,
- ➢ You will foster your involvement in subsequent initiatives (e.g. publications, project proposals).

Summary

- ✓ Objectives of this presentation.
- ✓ Implicit crowdsourcing model combining language-related datasets with language learning exercises.
- ✓ Foreseen datasets:
  - ✓ tracked learners’ answers to automatically-generated exercises,
  - ✓ tracked learners’ answers to exercises with large sets of questions.
- ✓ Added value in providing datasets.
- ➢ **3 Questions to the audience.**

3 Questions to the audience

1) Who has understood (most of) this presentation?
=> If you haven’t, feel free to ask questions now.

2) Who knows a language learning platform that tracks or could track learner answers?

3) Who would also be interested in performing statistical studies on such sets of learner answers?

Thank you all for your attention

Lionel Nicolas & Verena Lyding

chair.enetcollect@eurac.edu

Type your motivational phrase here and impress the audience.

This is a Title of This Slide

If you must include text on the same slide as an information graphic, keep it minimal so as not to overwhelm your audience.

Title and a Two-Column Text

Type your highly motivational phrase here and impress the audience.

Title, Picture and a Text

- This is a short description of the picture.
- You don’t have to use bullets always. You don’t have to use bullets always. You don’t have to use bullets always.

This is a Title

- You don’t have to use bullets always. Be clear and short.
- Here goes a bullet
- http://enetcollect.eurac.edu
- This is also a bullet
    - Try to avoid sub-bullets as much as possible

“If you wish, type your highly motivational phrase here and impress the audience.”

References and Resources

- ARHAR HOLDT, Špela, Iztok KOSEM, and Polona GANTAR, 2017: Corpus-based resources for L1 teaching: The case of Slovene. A. Marcus-Quinn (ed.): Handbook on Digital Learning for K-12 Schools. Springer International Publishing. 91–113.
- ČIBEJ, Jaka, FIŠER, Darja, KOSEM, Iztok. The role of crowdsourcing in lexicography. I. Kosem (ed,). Electronic lexicography in the 21st century : linking lexical data in the digital age : proceedings of eLex 2015 Conference, Herstmonceux Castle, United Kingdom. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing. 70-83.
- KILGARRIFF, Adam, Pavel RYCHLY, Pavel SMRZ in David TUGWELL. 2004. The Sketch Engine. G. Williams in S. Vessier (eds.): Proceedings of the Eleventh EURALEX International Congress, Lorient, France. Universite de Bretagne-sud. 105–116.
- KOSEM, Iztok, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. Analiza jezikovnih težav učencev: korpusni pristop. Trojina, zavod za uporabno slovenistiko.
- KOSEM, Iztok, HUSAK, Miloš, MCCARTHY, Diana, 2011. GDEX for Slovene. I. Kosem et al. (ed.). Electronic lexicography in the 21st century: new applications for new users: Proceedings of eLex 2011, Bled, Slovenia. Ljubljana: Trojina, Institute for Applied Slovene Studies. 150-159.
- KREK, Simon, 2012. Slovenski jezik v digitalni dobi = The Slovene language in the digital age, (White paper). Heidelberg [etc.]: Springer.
- LEECH, Geoffrey, 1997: Teaching and language corpora: A convergence. A. Wichmann, S. Fliegelstone, T. McEnery and G. Knowles (eds.): Teaching and language corpora. London: Longmann. 1-23.
- PILÁN, Ildikó, Elena VOLODINA, Lars BORIN. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. To appear in the Special issue of Traitement Automatique des Langues (TAL) journal, Special issue on NLP for learning and teaching. [pre-print]
- ROZMAN, Tadeja, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. Empirični pogled na pouk slovenskega jezika. Trojina, zavod za uporabno slovenistiko.
- ARHAR HOLDT, Špela, Gaja ČERV, Polona GANTAR, Iztok KOSEM, Karmen KOSEM, Irena KRAPŠ VODOPIVEC, Simon KREK, Sara MOŽE, Tadeja ROZMAN, Ana Marija SOBOČAN, Mojca STRITAR KUČUK and Ana ZWITTER VITEZ, 2013. Pedagoški slovnični portal. [Ljubljana]: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://slovnica.slovenščina.eu/
- ROZMAN, Tadeja, Mojca STRITAR KUČUK, Iztok KOSEM, Simon KREK, Irena KRAPŠ VODOPIVEC, Špela ARHAR HOLDT and Marko STABEJ, 2012. Šolar. [Ljubljana]: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://www.korpus-solar.net/
- Communication in Slovene: http://eng.slovenscina.eu/
- Šolar 2.0: http://solar.trojina.si/
- Pybossa: http://pybossa.com/

Thank you.

Name
name@mail.com