

# **Crowdsourcing Second Language learner data: experiences and prospects**

*Elena Volodina*  
University of Gothenburg, Språkbanken, Sweden

# The Rise of Crowdsourcing

Remember *outsourcing*? Sending jobs to India and China is so 2003. The new pool of *cheap labor*: everyday people using their spare cycles to create content, solve problems and even do corporate R & D.

*Jeff Howe, 2006, Wired magazine*

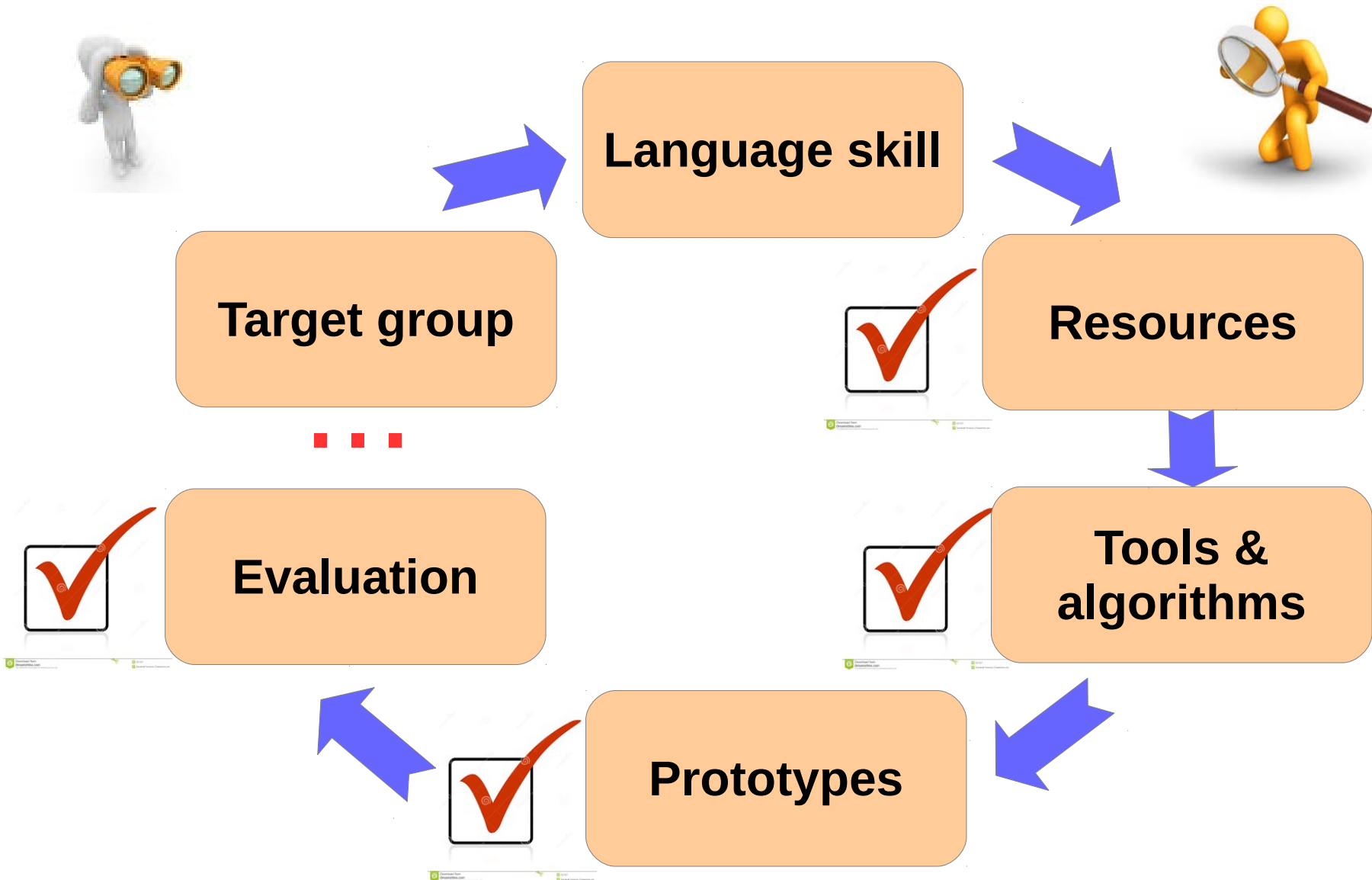
# WELCOME TO THE AGE OF THE CROWD

Use “plugged-in enthusiasts”,  
“take advantage of the networked world”,  
“discover ways to tap the latent talent of the  
crowd”

Jeff Howe

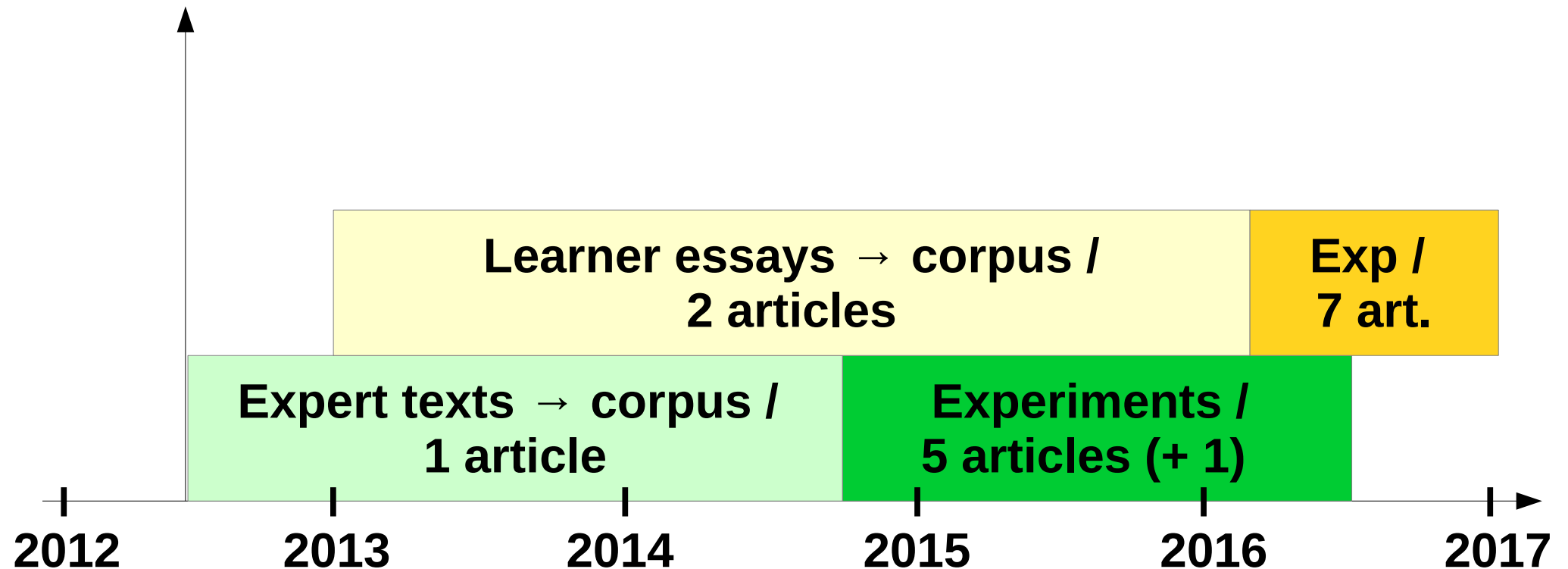


# ICALL tools for Second language (L2) learning



# Curious “time & effort” fact:

## Data vs experiments



# Lark Trills for Language Drills

Text-to-speech technology for language learners

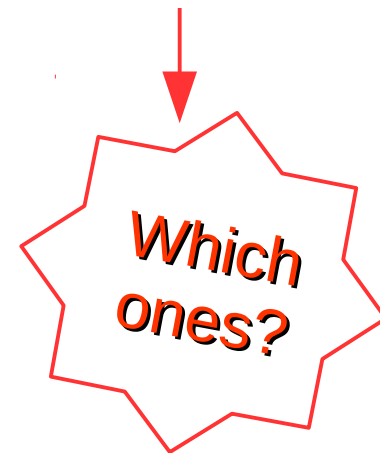
- Dictation and spelling exercise
- Focus on
  - evaluation of the quality of TTS
  - finding ways to give feedback on **spelling errors**

*Elena Volodina, Dijana Pijetlovic. 2015. Lark Trills for Language Drills: Text-to-speech technology for language learners. Proceedings of the 10th workshop on Building Educational Applications (BEA10), NAACL 2015, Denver, USA*

# Lark Trills for Language Drills

Text-to-speech technology for language learners

- Dictation and spelling exercise
- Focus on
  - evaluation of the quality of TTS
  - finding ways to give feedback on **spelling errors**



Fully automatic

self-study mode  test mode  timed test

word  inflected word  phrase  sentence  performance






### Result Tracker

Exercise name	Correct/Total	>>
Learners/spelling-word, self-study	1/2	

### Train spelling, word level

Type the word you hear

TIPS

Nr	Word	Correct answer	Links
2	<input type="text"/> <input type="button" value="submit"/>		  <input type="button" value="JSON"/>
1	<input type="text" value="matematik"/> <input type="button" value="submit"/>	 <b>matematik</b>	 

▶ Saldo morphology: *matematik*

▶ Wikipedia: *matematik*

▶ Wiktionary: *matematik*

▼ Monica: listen to pronunciation



matematik



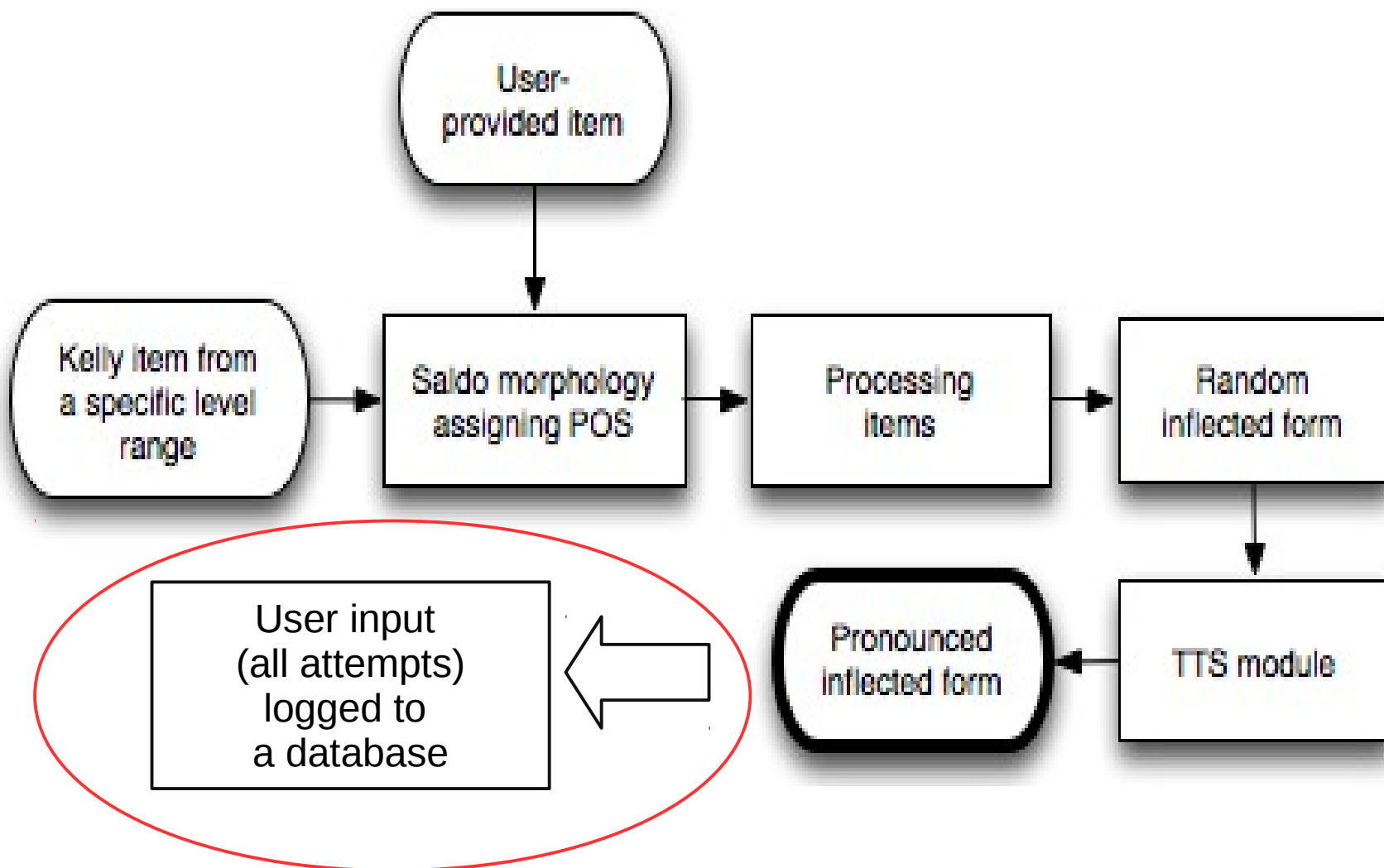




# Pipeline



for word & (inflected word) levels



# SPEED

## SPELLing Error Database

- For each correct item (base form + word class) we store:
  - session ID (no personal data, such as L1)
  - incorrect spelling(s)

# L2 spelling error database, SPEED

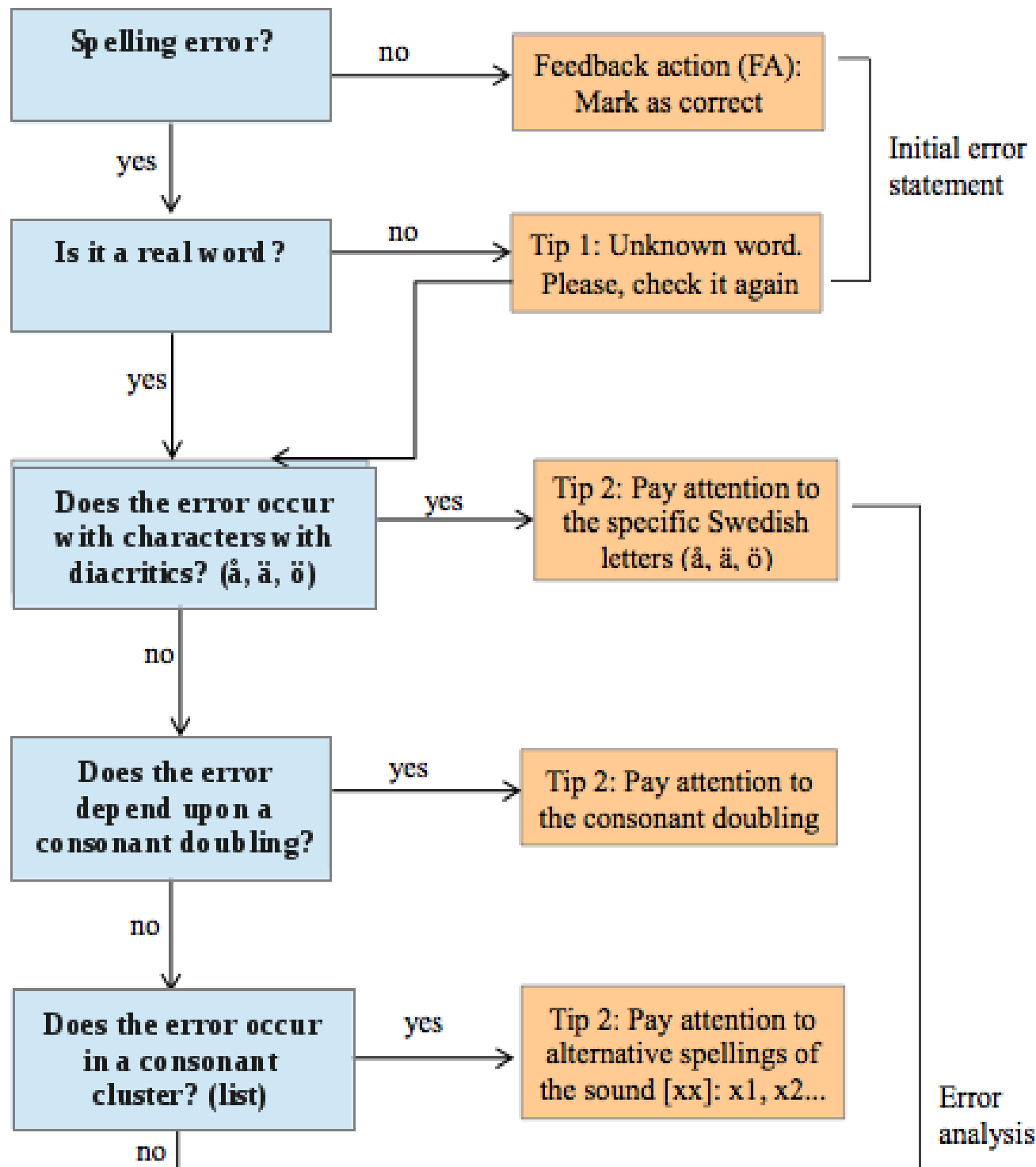
```
- <LexicalEntry uid="LexicalEntry-58d3459f-5acb-43f8-b60e-deb45a986c56">  
  <Sense id="speed--kelly-6950" uid="Sense-b1d45016-bdb5-4584-9ec5-11f780ecbf8a"/>  
  <word lang="swe" pos="AV" uid="word-4d9ed4cb-83b7-4293-a786-ff11f398e2d2">förträfflig</word>  
  <misspelling sessionID="2013-05-13-22-27-28" time="22:58:25" uid="misspelling-3ba31d83-f1ad-4c99-bc33-2c6a3a3c7849">förtrevlig</misspelling>  
- <modification uid="modification-4673c2b2-63a1-4c3c-b2a5-c6a80bc5dd20">  
  <feat att="updatedBy" val="laerka" uid="feat-ec50eb25-d870-4f53-b86c-ed620e3a332c"/>  
  <feat att="modificationDateTime" val="2013-05-13T22:58:26.01+02:00" uid="feat-4032acab-afa3-42c3-b6b0-3f904e345b76"/>  
  <feat att="modificationAccepted" val="pending" uid="feat-2e43c1b4-1106-4364-8df6-3991a4da6578"/>  
  <feat att="modificationComment" val="" uid="feat-22d76dd2-c26f-4ccd-b7a9-e6997082e0cd"/>  
</modification>  
</LexicalEntry>
```

Correct

Logged misspelling

# Error data

Error types	Nr,%	Example (correct → *incorrect*)
<b>Competence-based errors</b>	<b>55</b>	
Consonant doubling	28	stoppa → *stopa*
Diacritics (å, ä, ö)	23	högre → *hogre*
Phonetic errors (e.g. voiced vs voiceless)	25	relevans --> *relevanz*
Consonant clusters (phoneme-grapheme mappings, incl. cases of homonyms)	20	skön → *sjön*
Other (unclassifiable)	4	Israel → *visträv*
<b>Performance-based errors</b>	<b>17</b>	
Typos (neighbouring keys, addition, deletion, insertion, replacement)	17	förbättra → *förb'ttra*
<b>Across one word</b> (phrases & sentences)	<b>28</b>	se en bild → *sen bild*



# **SPEED**

## **SPELLing Error Database**

Advantages of collecting a corpus  
by applying this method:  
participants are quickly attracted,  
while cost, time and effort of  
collecting a corpus are reduced

**THIS is RESEARCH DATA!**

**And we need more of it!**

# What is infrastructure?



"'Infrastructure'? — You mean like rocks and sticks?"

# An electronic research infrastructure

- (free accessible) data in electronic format
- technical platform for exploring data, including tools and algorithms for data analysis, and visualization
- a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
- a network of experts in the relevant disciplines, incl. legal and ethical questions



# Partners

- University of Gothenburg: NLP, L2, language assessment  
*Elena Volodina, Julia Prentice, Monica Reichenberg*
- Stockholm university: NLP, L2  
*Mats Wirén, Gunlög Sundberg*
- Uppsala university: NLP  
*Beata Megyesi*
- Umeå university: L2/assessment  
*Lena Granstedt*



# Guess what?



- Riksbankens Jubileumsfond, infrastructure project IN16-0464:1



- 2017-2019

# **SweLL:** **electronic research infrastructure on** **Swedish learner language**

- SweLL – **S**wedish **L**earner **L**anguage
- Lärka-based L2 infrastructure
  - ... as a unit under Språkbanken's infrastructure
  - ... in the context of CLARIN

# Our focus is on...

- L2 essays (writing)
- **exercise logs (reading and listening comprehension, vocabulary and grammar training)**
- NO speech data – yet
- target group: adult learners

# L2 “alternative” data

- Logs – acc. to a defined research interest
- Steps:
  - Implement an activity for learners
  - Prepare database for storing (structured) data
  - Implement a way to browse logs, visualize statistics etc
  - If necessary – add extra annotation steps (manual, automatic)

# Pilot 1 on L2 “alternative” data

- Identifying most predictive features for a language proficiency level (for diagnostic purposes)
  - Multi-word expressions
  - Syntactic properties (e.g. word order)
  - Knowledge of word morphology (e.g. inflections)



*David Alfter*

# L2 “alternative” data (logs)



LÄR språket via **Korpus** Analys

## Exercise type evaluation

### Bundled gaps (variant 1)

Which word fits into these gaps? Each gap contains the same word. Write the word.

Hennes \_\_\_\_\_ var på hans lår , gned in värme i hans kalla ben .

En annan taxi tar \_\_\_\_\_ om skolbarnen .

Novelty hade flera trumf på \_\_\_\_\_ .

I första \_\_\_\_\_ har hon spelat dragspel och fiol .

### Evaluation

For which levels is this exercise type relevant?

A1  A2  B1  B2  C1

### Comments

# Pilot 2 on L2 “alternative” data

- Automatic assigning new words to a proficiency level
  - We predict the level automatically
  - Learners (of a known level) get the word in an exercise (or a series of exercises)
  - We see whether learners can cope with it



*David Alfter*



*Ildikó Pilán*



# L2 “alternative” data (logs)

<https://spraakbanken.gu.se/larkalabb/wordguess>



LÄR språket via **Korpus** Analys

## Word guess

Tries: 0/7

### Definition:

rör sig tyst och försiktigt för att inte bli upptäckt, tassar (pres ind aktiv)

Score: 0

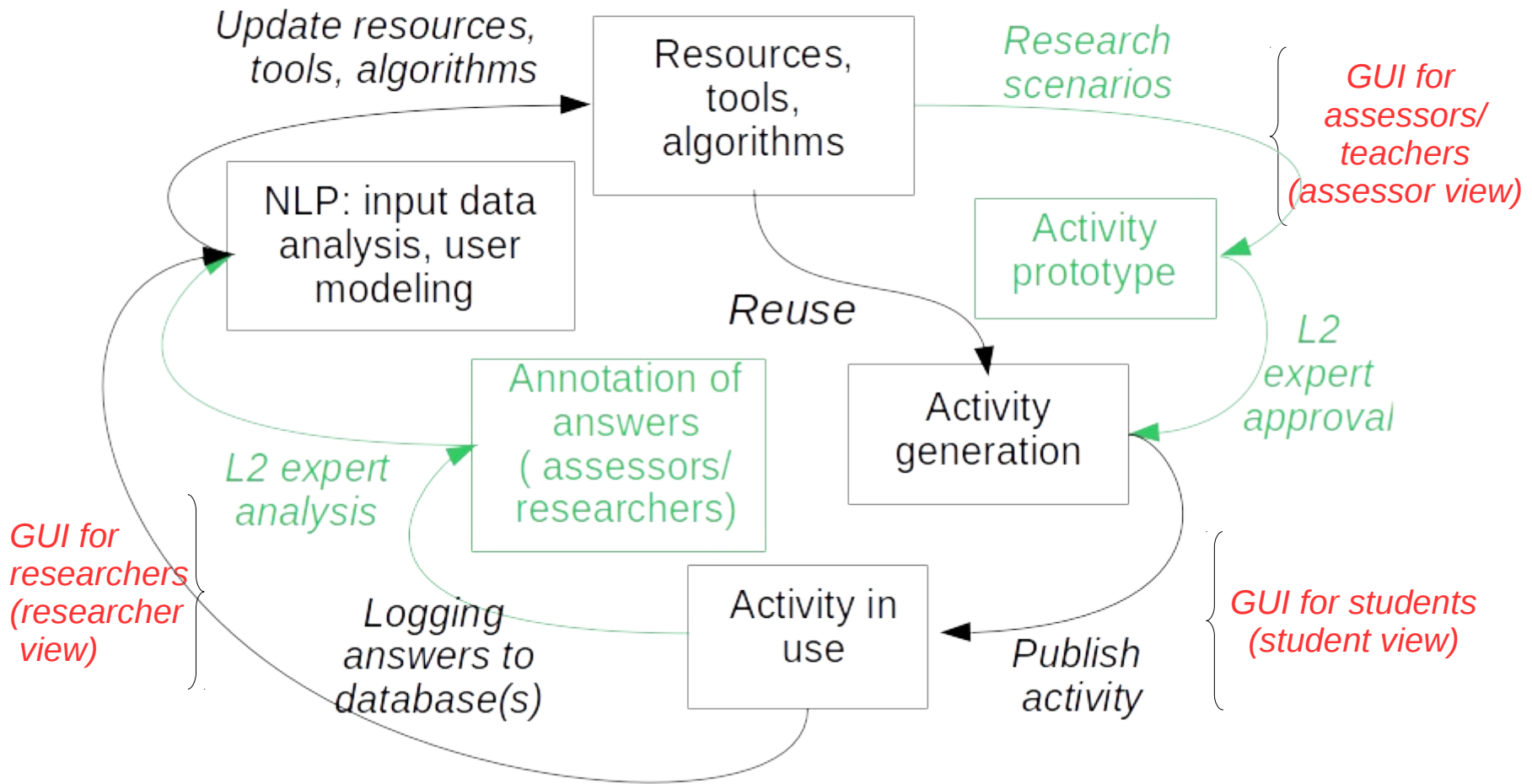
### Hjälp:

slip; steal, sneak, creep

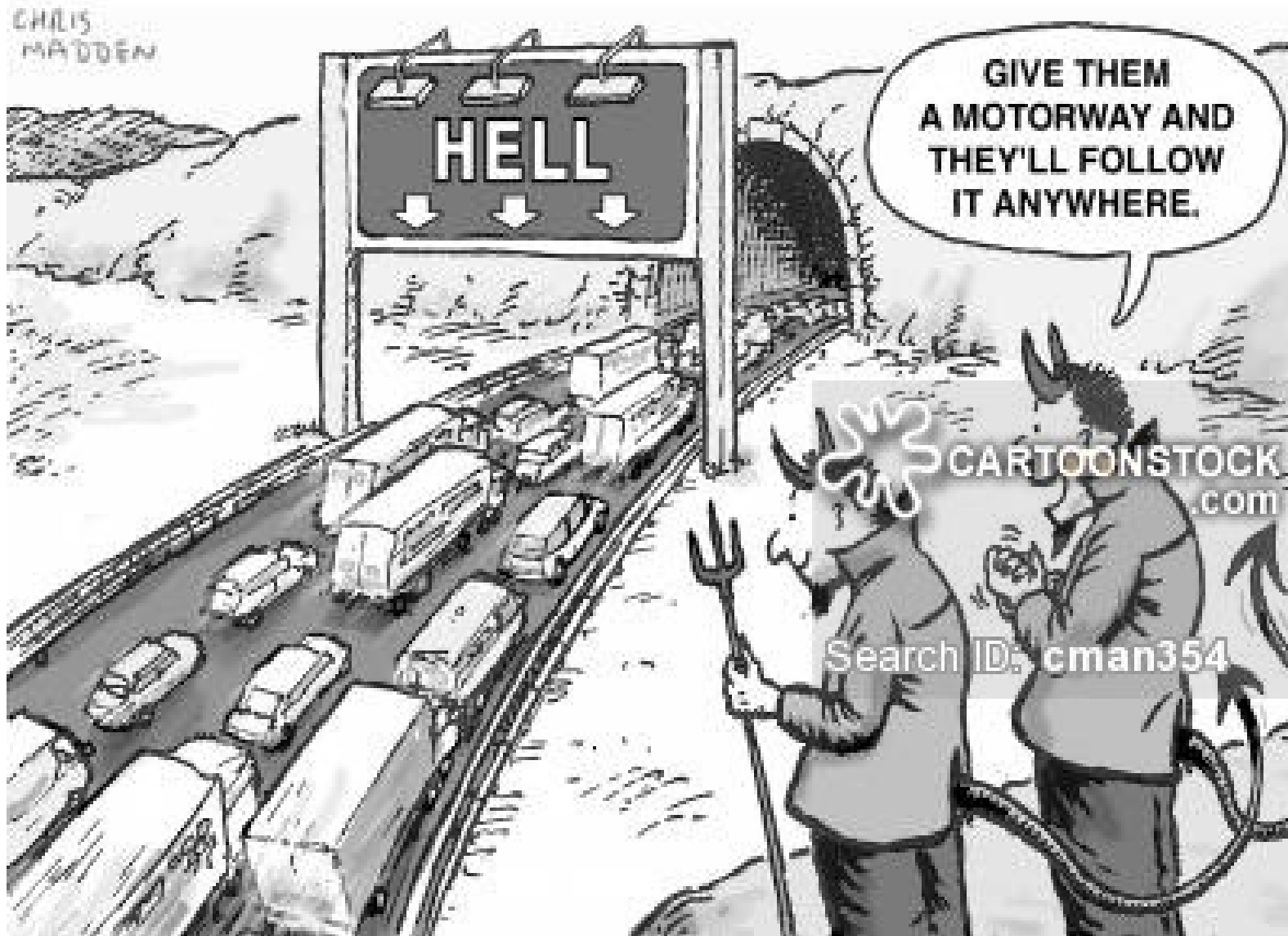


# The ultimate goal

## L2 infrastructure activity development cycle



# Where will this lead?





**Thank you!**

**Questions?**