Natural Language Processing for Computer-Assisted Language Learning & Research on Language Acquisition

https://spraakbanken.gu.se/eng/icall/joint6thNLP4CALL-2ndNLP4LA2017

GÖTEBORGS UNIVERSITET

INSTITUTIONEN FÖR SVENSKA SPRÅKET

Språk-BANKEN

Ildikó Pilán
University of Gothenburg

Elena Volodina
University of Gothenburg

Lars Borin
University of Gothenburg

Stockholms universitet

Institutionen för lingvistik

13 submissions

36 PC members
(=reviewers)

3 reviews
7 accepted papers

Papers & authors, 2017

	Country	Submitted	Accepted
	Sweden	7 (15)	3 (7)
	Iceland	1 (6)	0 (6)
	Germany	3 (9)	2 (6)
	Switzerland	2 (2)	2 (2)
	Finland	2 (4)	2 (4)
	Estonia	1 (1)	1 (1)
	**Total (papers)**	**13**	**7**

We seem nice!

Workshop year	Submitted	Accepted	Acceptance rate
2012	12	8	67%
2013	8	4	50%
2014	13	10	77%
2015	9	6	67%
2016	14	10	71,5%
2017	13	7	54%

We seem nice!

Workshop year	Submitted	Accepted	Acceptance rate
2012	12	8	67%
2013	8	4	50%
2014	13	10	77%
2015	9	6	67%
2016	14	10	71,5%
2017	13	7	54%

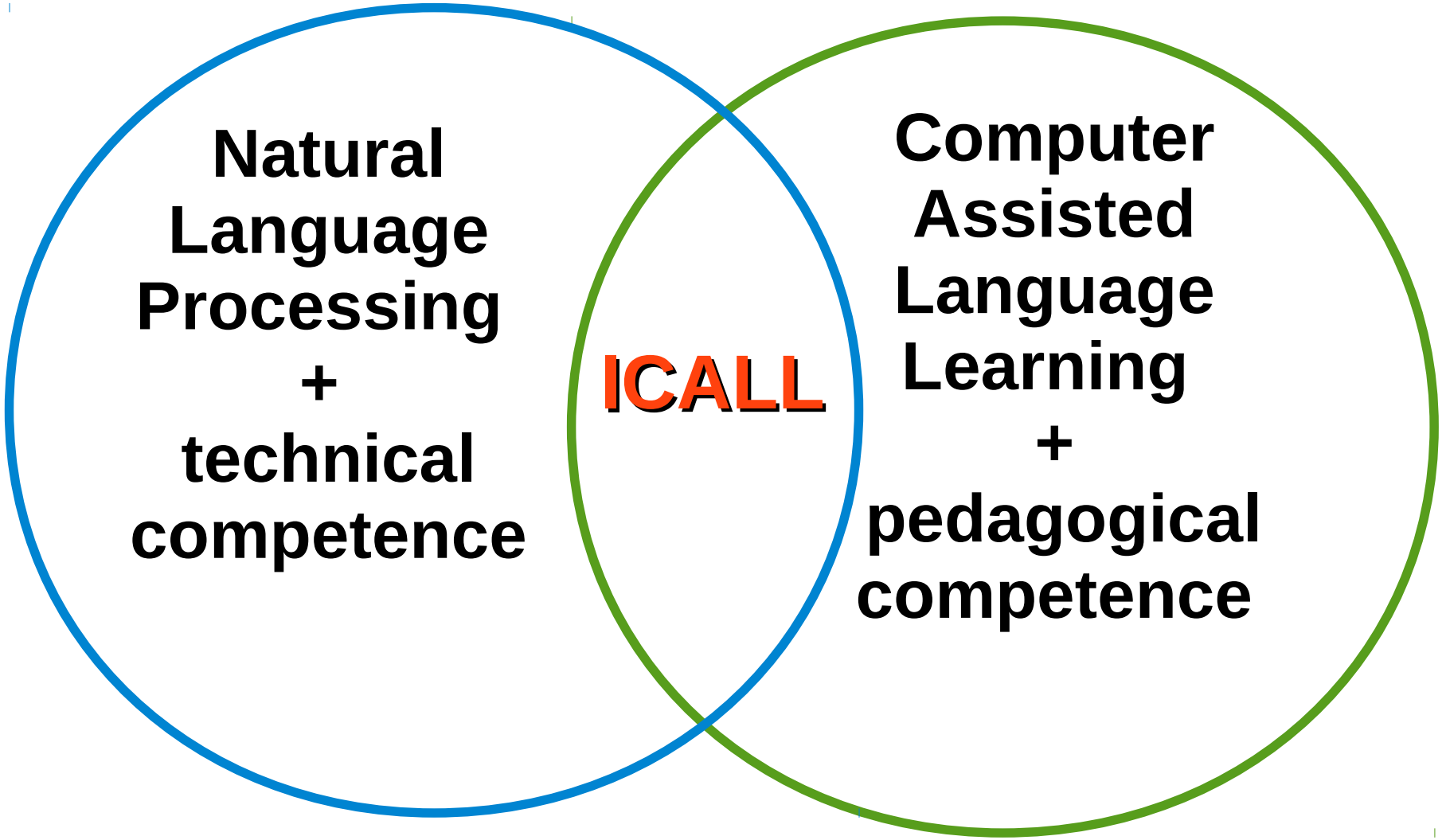
… but we are strict – it's your papers that are GOOD!


Bente Ailin Svendsen
University of Oslo, Norway
Torsten Zesch
University of Duisburg-Essen, Germany

Estonia *2*
Finland *5*
Germany *4*
Norway *2*
Sweden *21*
Switzerland *2*

38 registrations
+Drop-in

NLP + CALL = ICALL



By combining CALL and LA

- We have extended the previous workshop concept in two dimensions:
  - L2 + L1 acquisition
  - NLP + corpus linguistics, psychology, cognitive science, and phonetics

The benefit

- shared ideas, tools, and methods
- broadening the community
- providing an environment for new and exciting collaborations

Time	Program point.
	Morning session 1. Chair: Kristina Nilsson Björkenstam
9.00 - 9.30	Opening. *Elena Volodina*
9.30 - 10.00	Allison Adams and Sara Stymne. *Learning with Learner Corpora: using the TLE for Native Language Identification.* [pdf]
10.00 - 10.30	Coffee break
	Morning session 2. Chair: Lars Borin
10.30 - 11.00	Xiaobin Chen and Detmar Meurers. *Developmental Benchmark of Syntactic Complexity.* [pdf]
11.00 - 11.30	Johannes Graën and Gerold Schneider. *Crossing the Border Twice: Reimporting Prepositions to Alleviate L1-Specific Transfer Errors.* [pdf]
11.30 - 12.20	Invited talk by **Torsten Zesch**: *Automatically ____ gap-fill exercise items.*
12.30 - 13.30	Lunch
	Afternoon session 1. Chair: Elena Volodina
13.30 - 14.00	Anisia Katinskaia, Roman Yangarber and Javad Nouri. *Tools for Language Learning and Supporting Endangered Languages.* [pdf]
14.00 - 14.30	Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy and Detmar Meurers. *Developing A Web-based Workbook for English Supporting the Interaction of Students and Teachers.* [pdf]
14.30 - 15.00	Sara Stymne, Eva Pettersson, Beáta Megyesi and Anne Palmér. *Annotating Errors in Student Texts: First Experiences and Experiments.* [pdf]
15.00 - 15.30	Coffee break

9.30 First talk

~ 50 minutes till coffee

~20 min intro

Common pitfalls
when developing ICALL applications

Pitfalls ...

… are based on

- our own experience
- our own observations
- comments from non-ICALLers
- comments from reviewers

ICALL tools for Second language (L2) learning

Language skill

Data & resources

Tools & algorithms

Prototypes

Evaluation

Target group

. . .

Challenges & lessons

- Data
- NLP researchers vs teachers → two different worlds?
- End-user applications → prototypes vs maintenance

Data

Two types of data

- Produced **BY** L2 learners
  - → essays
  - → exercise logs
  - → errors
  - → interviews

- Produced by experts **FOR** L2 learners
  - → reading comprehension texts
  - → exercises
  - → recordings of listening excerpt

Challenges:
L2 learner-produced data

- Electronic L2 essays/logs are very difficult to collect
  - → NOT available online
  - → Need learner permits
  - → Need learner variables (gender, age, native language, etc)
  - → Sensitive in nature
  - → Those who have it – don't want or CAN'T share

L2 essay collection

- Collaboration with parents (children)
  - → if you find any willing to sign permits
  - → ethical committee clearance
- Collaboration with teachers (children + adults)
  - → if you find any teachers willing to collaborate
  - → if you manage to convince students to sign permits'
- If you succeed:
  - → decipher names in their hand-writing
  - → digitize, anonymize, store, annotate for metadata

L2 essay pre-processing

SweLL workflow

Essay collection *ongoing* → **CEFR-grading** *teachers* → **Digitization** → **Manual registration** *editor in Lärka* → **Automatic annotation** *Korp pipeline*

Learner variables
Collected through permits*

Student variables:
- Age/birthyear
- Gender
- Mother tongue(s)
- Residence time in Sweden
- Education level

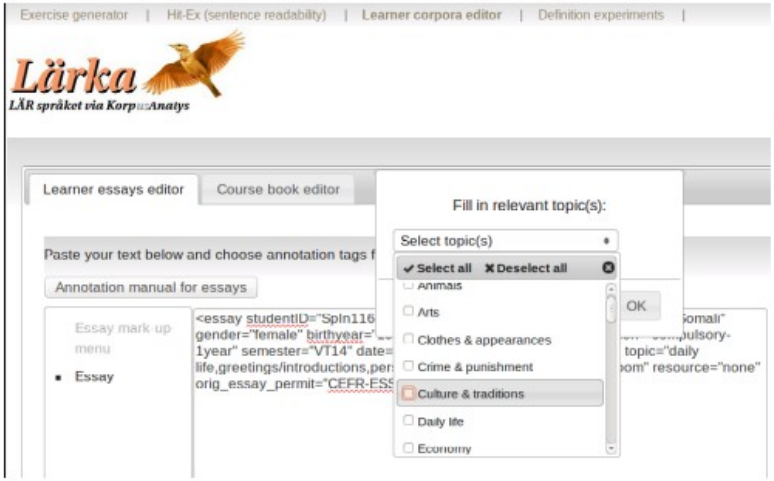
Essay variables
- Assigned CEFR level
- Essay setting (exam/home)
- Use of extra materials
- Academic term and date
- (Title, topic, genre, grade)

Assessors
Minimum of two trained assessors

Inter-annotator agreement
- A degree to which several annotators agree about assigning attributes
- Reported for SW1203 subcorpus
- Krippendorff's alpha for pairwise agreement = 0.80
- **0.80 = good annotation quality** (Artstein & Poesio 2008)

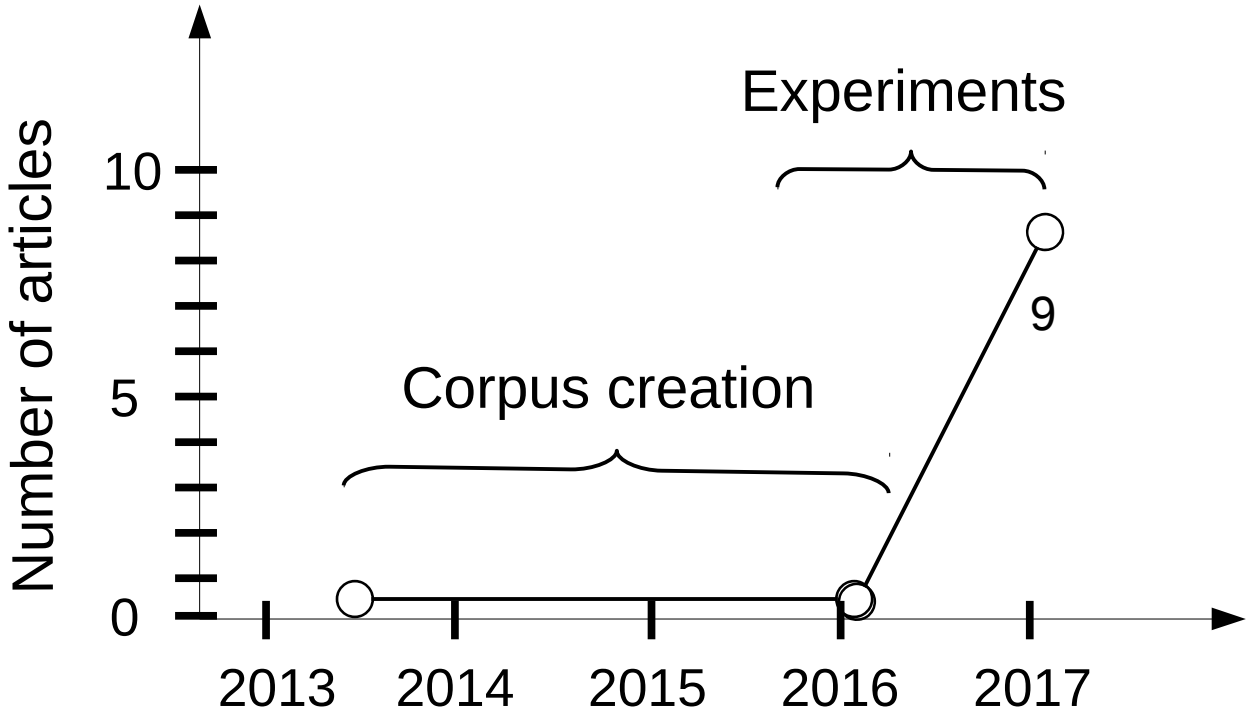
SweLL digitization principles

1. **Do not revel author identity**
   * revealing names → replace with *NN*
   * addresses → replace with *NN-street*
2. **Do not correct errors**
   * if several interpretations possible → make *positive assumption*, i.e. that the learner made no mistake
3. **Preserve illegible handwriting**
   * each illegible letter → replace with @
   * stricken text → leave out



* *http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/tillstand eng-24042013 v03.pdf*

Curios “time & effort” fact: Data vs experiments



Essay corpus, SweLL-corpus, creation and SweLL-based publications

Lesson 1

- Do not understimate the time it takes to collect and prepare data

Time-effect ratio consequences

- Researchers skip compiling own data
  - → use what is available
  - → in the end often targeting English

SweLL corpus

Sub-corpus	A1	A2	B1	B2	C1	Un-known	Total
Tisus	-	-	-	27	78	-	105
Sw1203	-	-	33	45	11	1	90
SpIn	16	83	42	2	-	1	144
Total	16	83	75	74	89	2	**339**

Språk-
BANKEN

GÖTEBORGS
UNIVERSITET

Permission

to use learner essays for research purposes

I, (name, surname) _____

grant my permission to the University of Gothenburg, Språkbanken, to use my essays for research purposes (check one alternative),

- ☐ only for restricted use by approved user groups and protected by password
- ☐ for unrestricted use provided my identity will remain anonymous

Personal information:

Gender: ☐ Woman ☐ Man Age_____

Mother tongue (one or more)_____

Residence time in Sweden: _____years _____ months

Education level: ☐ Elementary school nr of years _____

☐ High school nr of years _____

☐ Upper secondary school nr of years _____

☐ College/University nr of years _____

☐ Post-graduate studies nr of years _____

Place and date Signature:

Digitization & anonymization principles

SweLL digitization principles

1. **Do not revel author identity**
   * revealing names → replace with *NN*
   * addresses → replace with *NN-street*

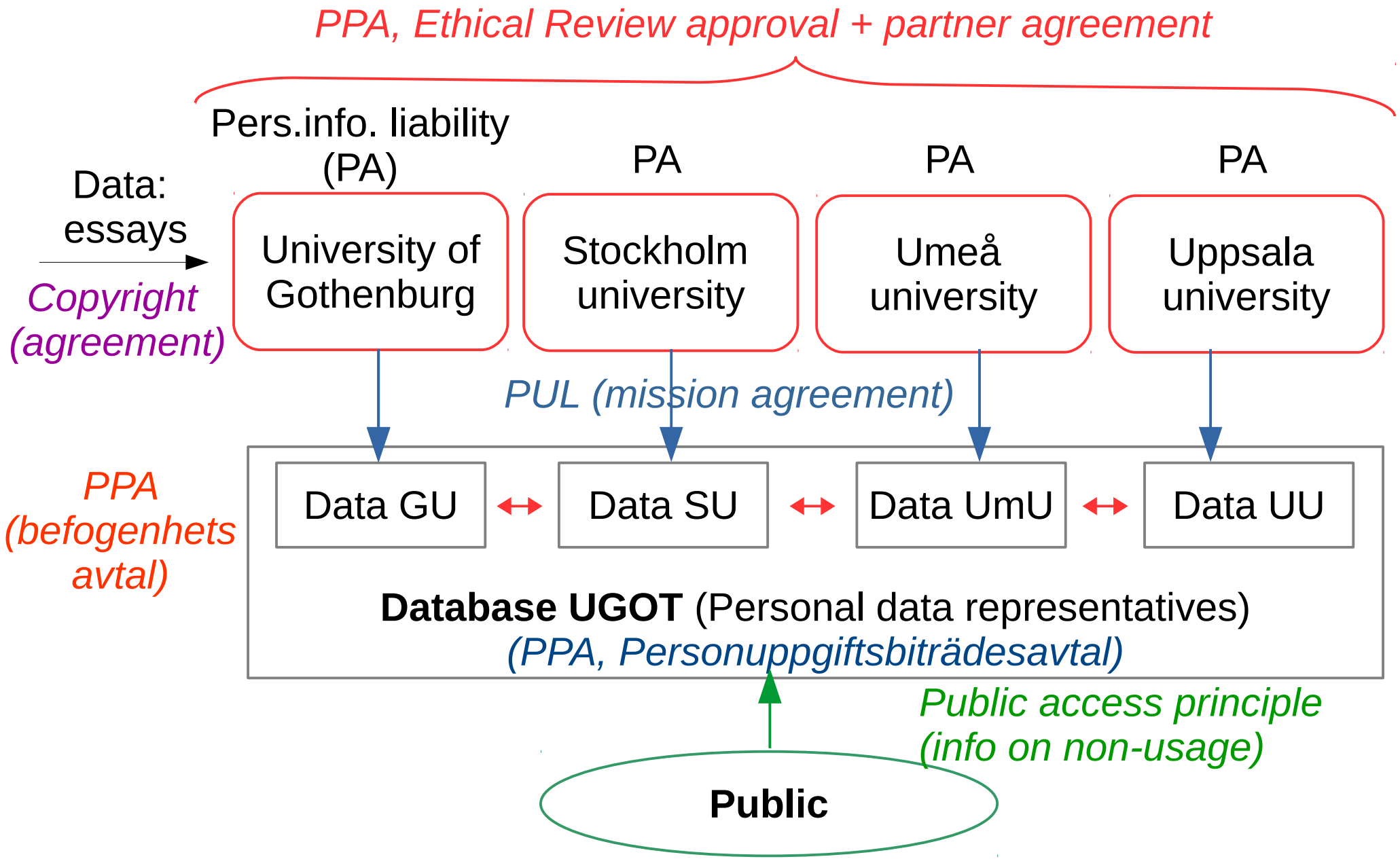
2. **Do not correct errors**
   * if several interpretations possible → make *positive assumption*, i.e. that the learner made no mistake

3. **Preserve illegible handwriting**
   * each illegible letter → replace with @
   * stricken text → leave out

Laws and regulation

that need to be taken into consideration

- Personal Privacy Act (European-wide)
- Ethical Review Act
- Freedom of writing (government) → public access
- Copyright


PPA, Ethical Review approval + partner agreement
Pers.info. liability (PA)
PA
PA
PA
Data: essays
Copyright (agreement)
University of Gothenburg
Stockholm university
Umeå university
Uppsala university
PUL (mission agreement)
PPA (befogenhets avtal)
Data GU
Data SU
Data UmU
Data UU
Database UGOT (Personal data representatives)
(PPA, Personuppgiftsbiträdesavtal)
Public access principle (info on non-usage)
Public

Lesson 2

- Take time to study legal regulations, not to “waste” the previously collected data

  - ➔ *There are loopholes, but not without information loss*

Different worlds?

L2 vs NLP researchers

or do we really work interdisciplinary?

- Terminology (e.g. corpora & annotation vs
  genre pedagogy & processability theory)
- Example: normalization

Two views

- Teachers

  → want control

- NLP researchers

  → want to automatize

Two views

- Teachers
  - → want control
  - → keep to fixed practices
  - → want 100% correctness
  - → sceptic about automatic solutions
- NLP researchers
  - → want to automatize
  - → want to revolutionize
  - → work within reasonable margins
  - → enthusiastic about automatic solutions


ARE YOU THE NEW ENGLISH TEACHER ?
YES I ARE!

Two views

- Teachers
  - → want control
  - → keep to fixed practices
  - → want 100% correctness
  - → sceptic about automatic solutions
- NLP researchers
  - → want to automatize
  - → want to revolutionize
  - → work within reasonable margins
  - → enthusiastic about automatic solutions

Meeting half-way?

- Educational Testing Service (ETS)
  - → from Criterion/e-rater to Language Muse (20 years)
  - → http://languagemuse.10clouds.com/
- WERTI/VIEW
  - → no automatic pre-selection of exercise text

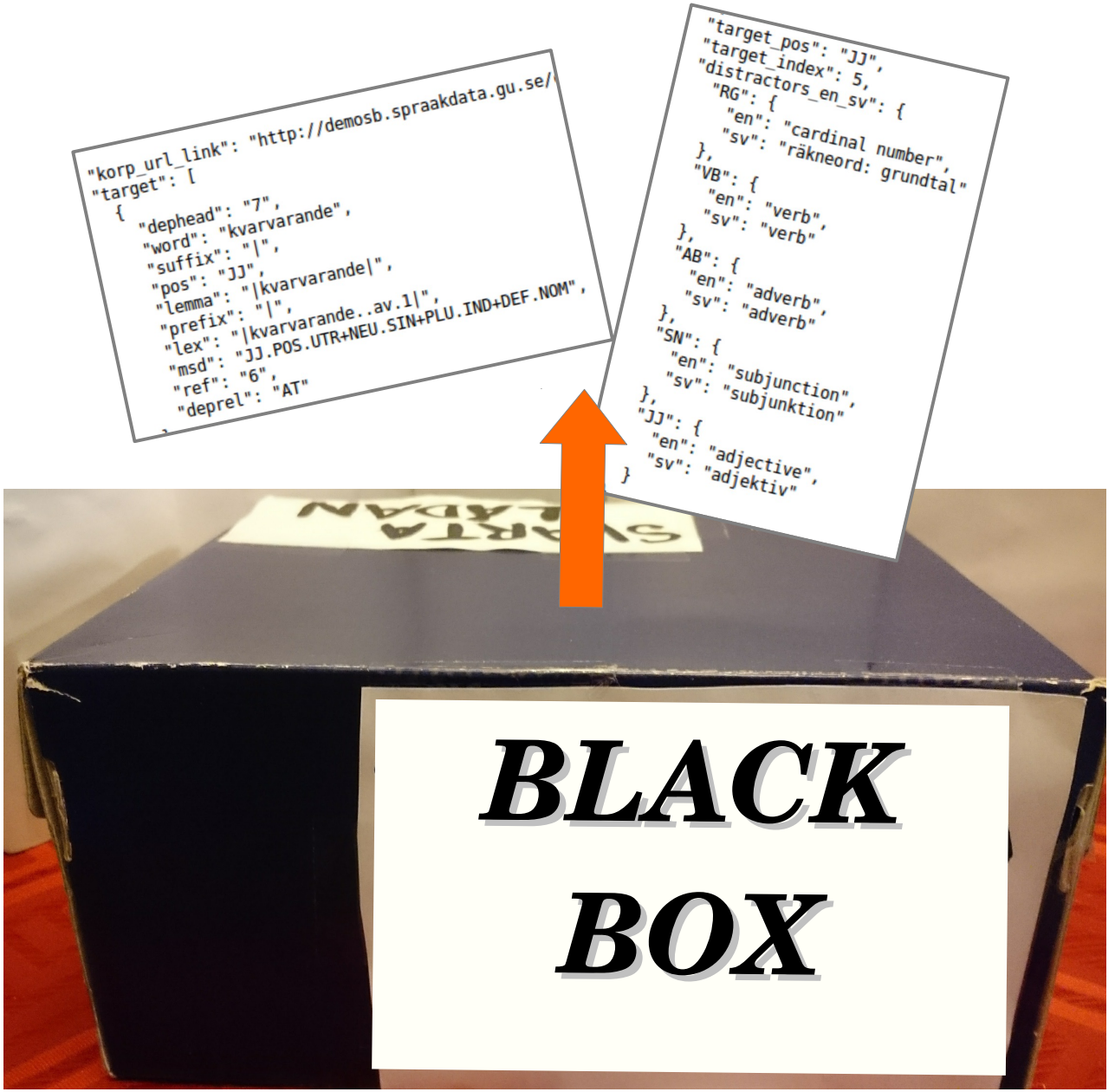
Lesson 3

- Take time to study what makes L2 researchers & L2 teachers “tick”, and vice versa

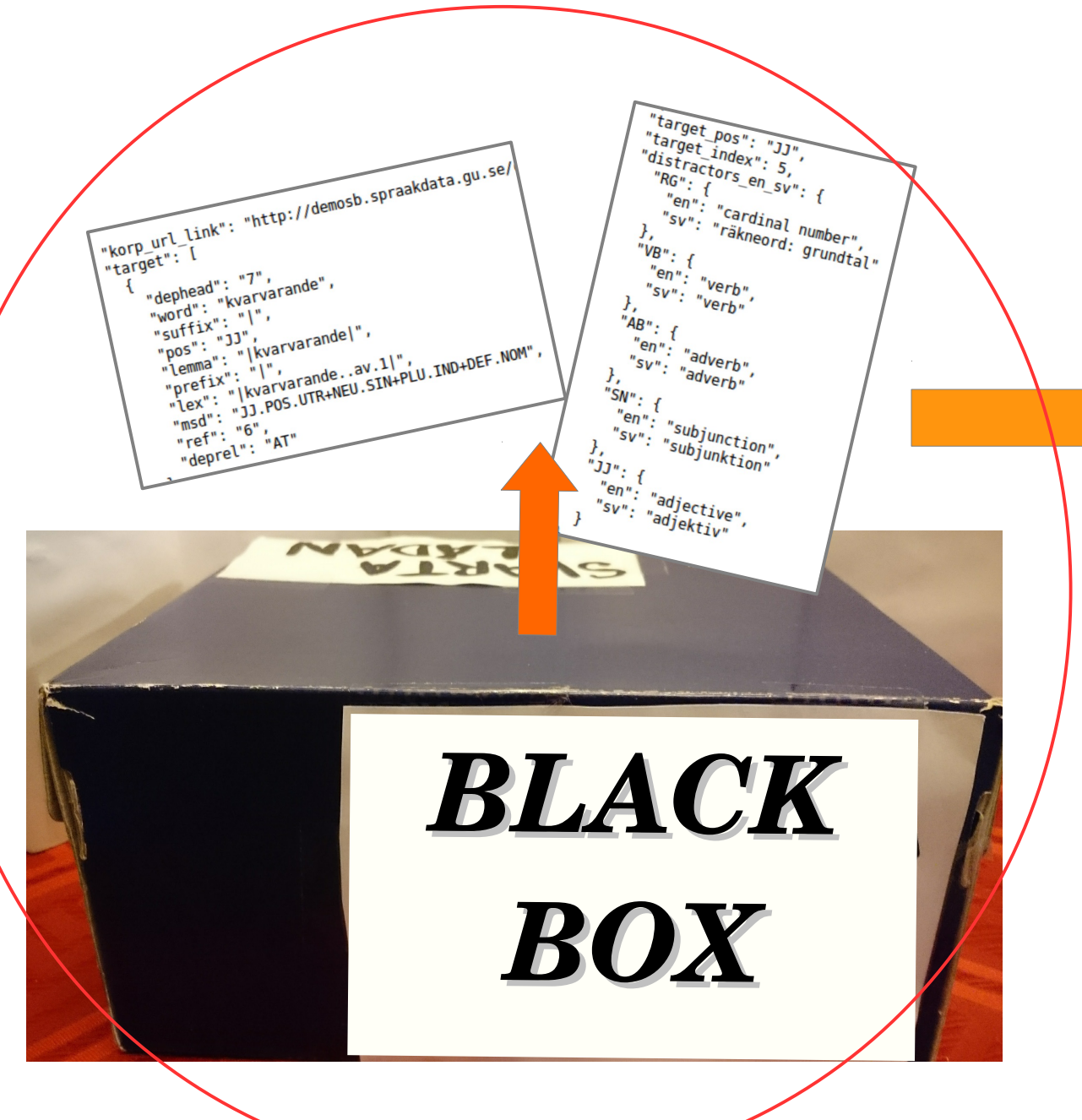
- Be ready to compromise → on both sides

Application life cycle

ICALL tools for L2 learning



ICALL tools for L2 learning



ICALL tools for Second language (L2) learning

Language skill

Data & resources

Tools & algorithms

Prototypes

Evaluation

Target group

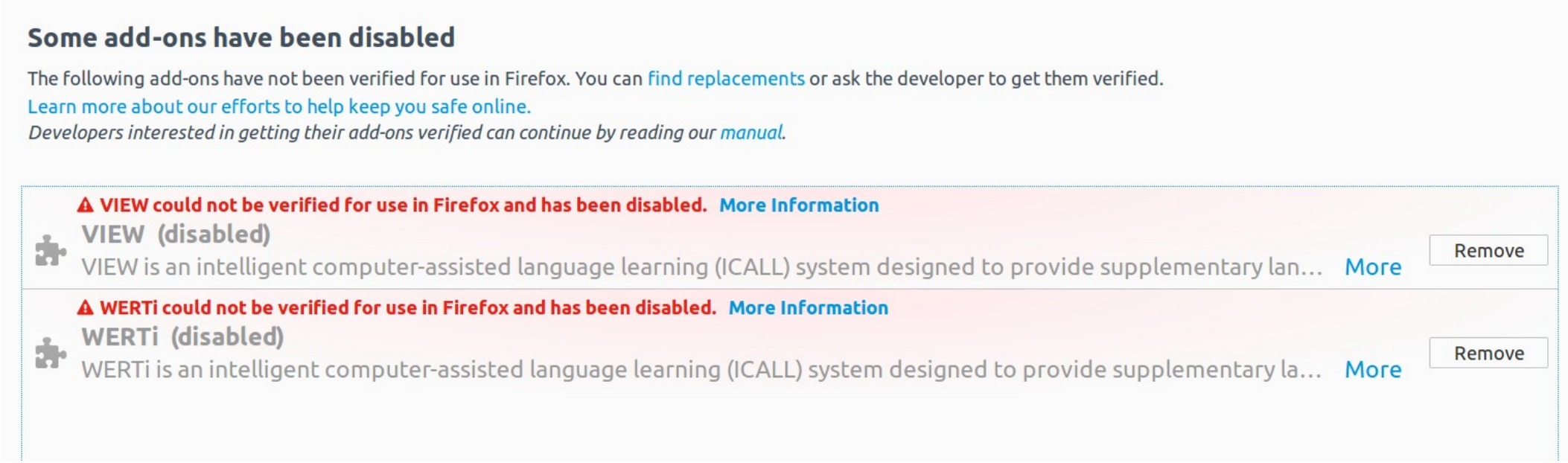
ICALL tools for L2 learning

Application-development and maintenance

versus

Prototype-development (and evaluation)

Recent example



Lesson 4

more of an insight

- (Most?) ICALL research remains within ICALLers' “comfortable zone”, i.e. on their desks; at the best goes into a prototype

- Researchers can at most develop prototypes as a “proof-of-concept”, but cannot maintain full-scale applications

- There is a need for a new type of funding, the one that would bring research findings to end-users

Lesson X...

...to be continued

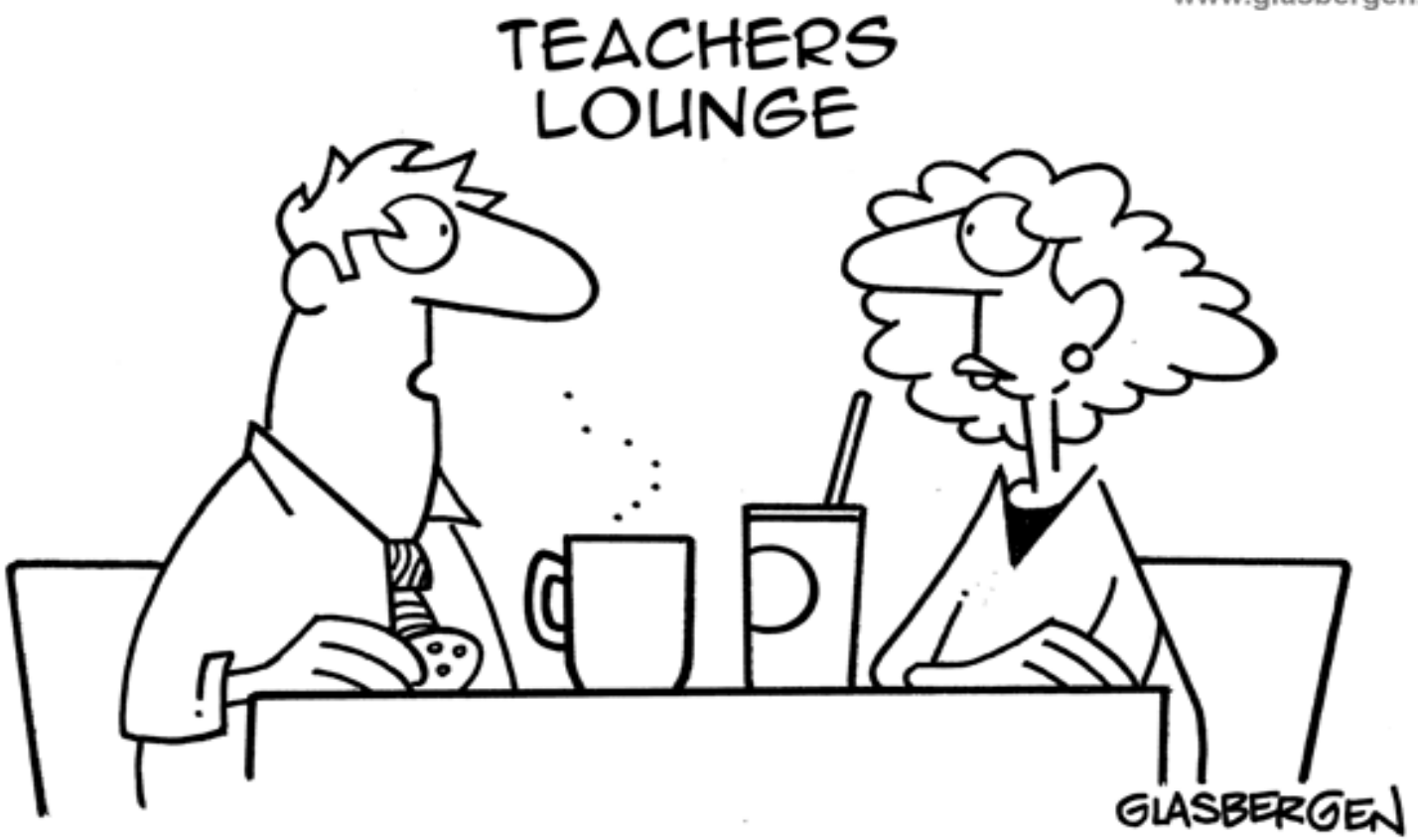
- …
- …

We need to

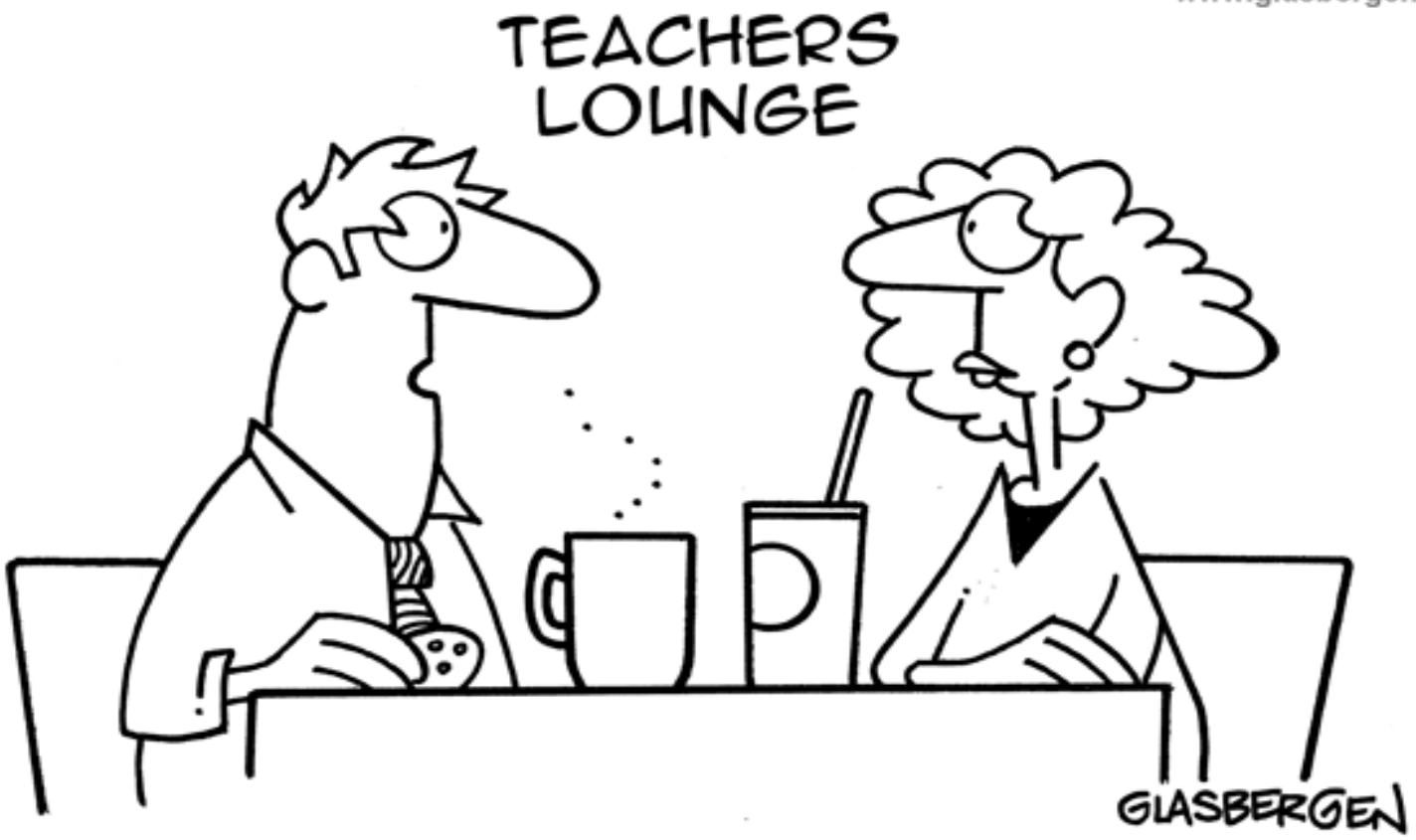
- re-examine our practices
- take these issues to discussion
- make newcomers aware of the pitfalls

We need to

get back to these issues now and then



"The kids don't listen, so I have to repeat myself. I'm always repeating myself. You know, always saying the same thing more than once. I say it once, and then they make me say it again..."



"The kids don't listen, so I have to repeat myself. I'm always repeating myself. You know, always saying the same thing more than once. I say it once, and then they make me say it again..."