A (systematic) evaluation of predictions of the ACT-R sentence processing model

Paul Mätzig

December 15, 2016

Motivation

What I'm going to talk about:

How to evaluate the (cf. Engelmann, Vasishth, Engbert, & Kliegl, 2013) EMMA/ACT-R model, and how to compare it to the model of Lewis and Vasishth (2005) (LV05)? More specifically, what are my ideas on that until now?

Why I am presenting:

This work in progress for my Master's thesis, so I want as much feedback as possible!

Outline

Learning from models

The LV05 and EMMA/ACT-R models

Testing models: Roberts and Pashler (2000)

Exploring the coverage of EMMA/ACT-R

Learning from models

Learning from models

Assumption: Models allow for *surrogative reasoning*, that is, by studying models, we can gain knowledge about the subject of the model (what problem in the real world the model represents).

This gives rise to a *model-based style of reasoning.*

But how is learning knowledge through models possible?

DDI account

Hughes (1997) proposes the 'DDI account' as a methodological framework to study model (and study with models).

It is a three step procedure:

1. Denotation
2. Demonstration
3. Interpretation

Denotation: Establish a representation relationship between the model and the real world problem.

Example: Define a set of assumptions and rules from which smaller, more specific 'submodels' can be derived.

DDI account

Demonstration: Investigate the model-internal behaviour and features to generate hypotheses (predictions) about the reality.

Applicable at two points in time: while building the model, and while manipulating the fully constructed model.

Both steps are important for computational modelling: On the one hand, to fix and implement the assumptions of one's theory, and on the other hand to simulate data from the finished model, showing its predictions.

DDI account

Interpretation: Transform knowledge about the model into propositions about the real world problem.

Problems that have to be dealt with at interpretation, especially with simulation results:

- discreteness of simulations (not the whole, continuous parameter space can be examined)
- a large amount of free parameters (overfitting)
- use of conceptually premature models and unfitting assumptions

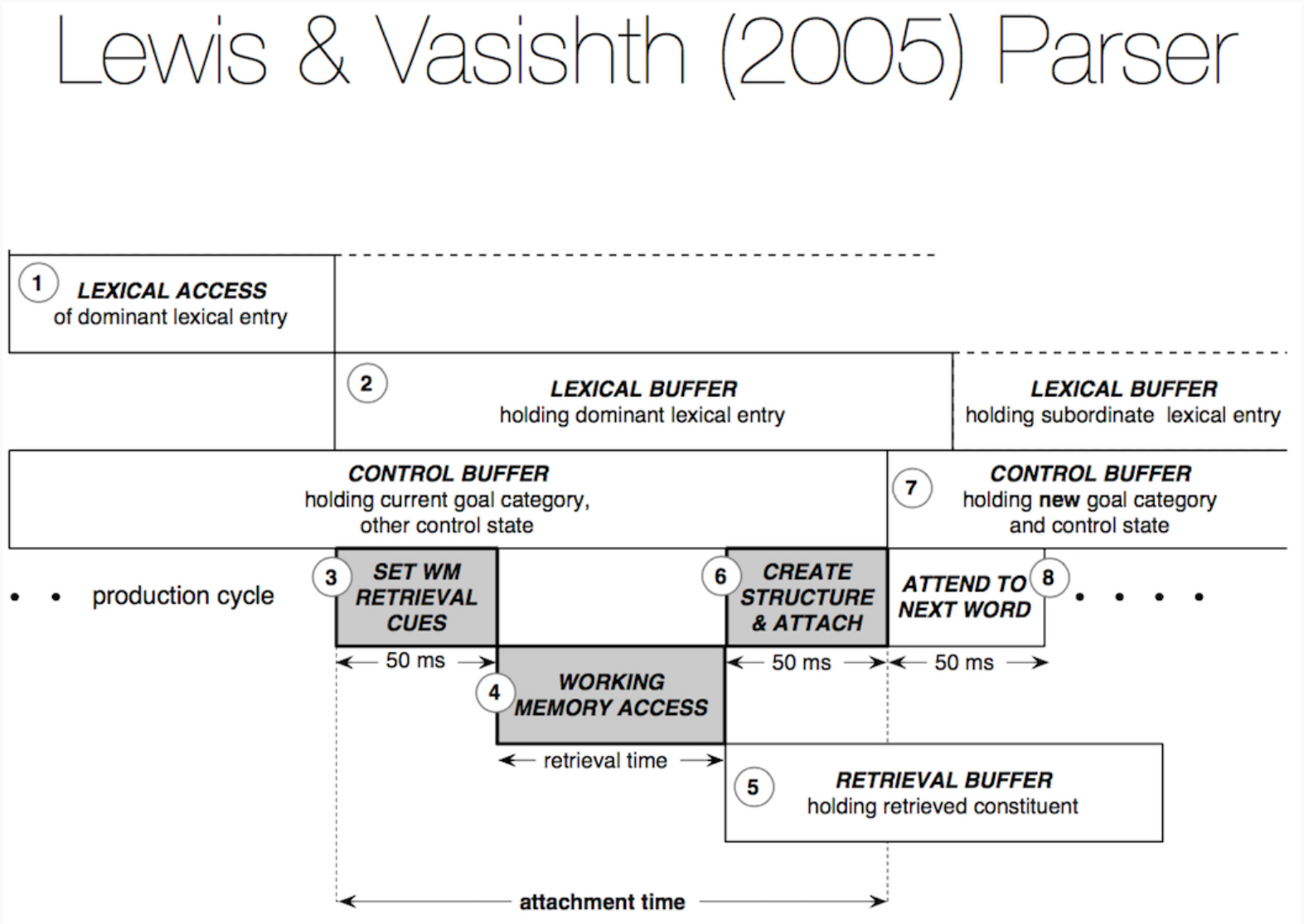
The LV05 and EMMA/ACT-R models

The LV05 model

Some core assumptions and features of the model:

- implemented in the ACT-R cognitive architecture
  - separation of production and declarative memory
  - limited working memory focus
  - activation-based mechanism
- left-corner parser
- cue-based retrieval

The LV05 model



(from Engelmann, 2016 (PhD Thesis))

EMMA/ACT-R (Engelmann et al., 2013)

Crucial novelty: Integration of EMMA (**?**) into the LV05 model

Features EMMA (very generally):

- calculation of encoding time of a word based on frequency of occurence and eccentricity from the current viewing location
- basic 'algorithm' of interaction between EMMA and ACT-R:
  1. find the nearest word to the right,
  2. shit attention and start encoding,
  3. start memory retrieval (in ACT-R)
  4. when sentence is finished, stop reading.

EMMA/ACT-R (Engelmann et al., 2013)

New mechanism: timeout regressions (i.e., buying time for unfinished integration of a word)!

What else is different to the LV05 model?

The dependent measure: reaction/reading times in LV05, but eye movement measures in EMMA/ACT-R!

Testing models: Roberts and Pashler (2000)

Testing models: Roberts and Pashler (2000)

Testing models by only fitting them to empirical data is not enough.

Roberts and Pashler (2000) argue that a good fit . . .

- gives no full information on what the model predicts (how it constrains the possible outcomes),
- doesn't say anything about the variability of the data in the space of possible outcomes, and it
- ignores the a priori probability that the theory will fit (perhaps it could fit anything).

Testing models: Roberts and Pashler (2000)

What do Roberts and Pashler (2000) suggest instead?

1. Determine the predictions of the model.
2. Show the variability of the data.
3. Show plausible results that the theory cannot fit.

How could these steps be applied to the model comparison at hand?

Exploring the coverage of EMMA/ACT-R

Questions for EMMA/ACT-R evaluation

- Does the model cover the same empirical data as its LV05 predecessor?
- How variable are the data that we are trying to model?
- What are plausible results that the model does not predict?

Simulations in LV05

The empirical phenomena that the LV05 model was fit to were:

1. SRs vs. ORs (Exp. 1 of Grodner & Gibson, 2005)
2. Estimating time for working memory retrieval
3. effect of interpolated material on MV and EmbV (Exp. 2 of Grodner & Gibson, 2005)
4. Length and interference effects (Van Dyke & Lewis, 2003)
5. storage load effects (Grodner, Gibson & Tunstall, 2002)

Predictions: one possible procedure

1. Determine a phenomenon to test the model's predictions on.
2. In *n* iterations with the 'default' parameter setting acquired in Engelmann et al. (2013), determine a baseline for the phenomenon at hand.
3. Choose (increasingly many?) parameters to vary, and perform a *grid search* of simulations through all *plausible* parameter value combinations.
4. Visualise the change in dependent variables, depending on the varied parameter(s).

Example: SR/OR sentence processing in aphasics (Caplan, 2015)

Plausible results that the model cannot fit

This is the point where Roberts and Pashler (2000) argue that predictions of a model should be surprising.

Reasoning behind that (again): if none of the model's predictions is surprising, and there are no plausible results that it cannot predict, the model is useless.

References

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*(3), 452–474.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science*, 325–336.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? *Psychological Review*, *107*(2), 358–367.