Integrating Treebank Annotation and User Activity Data in Translation Research

Michael Carl, Henrik Høeg Müller, Bartolomé Mesa-Lao

{mc, hhm,bm}.isv@cbs.dk

Copenhagen Business School
Department of International Language Studies and Computational Linguistics

CBS, 07.03.2012

Overview – **CRITT centre**

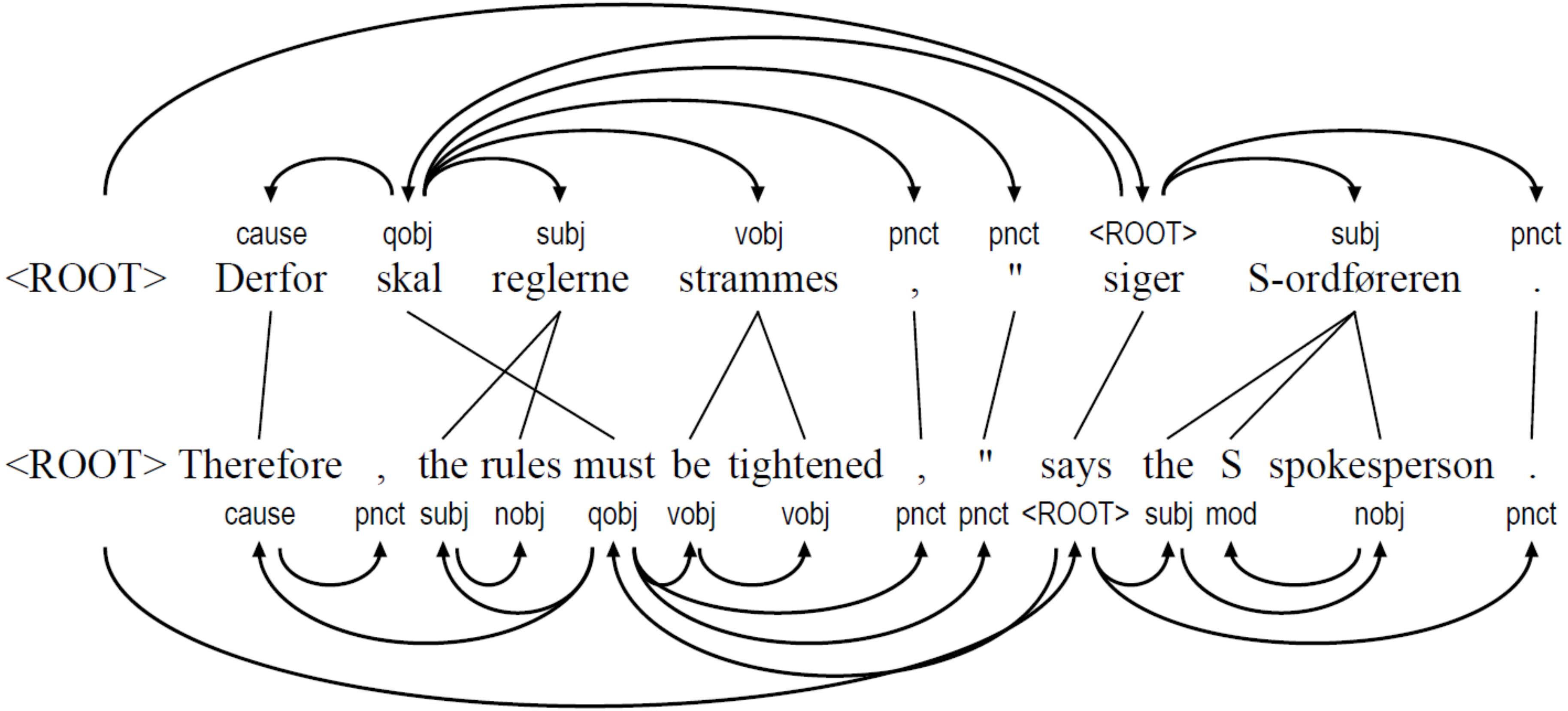
1. Focus on methodology for translation process research
2. Experiments with eyetracking reading
3. Keystroke background: Translog software (keylogging)
4. Integration of eyetracking and keylogging in Translog

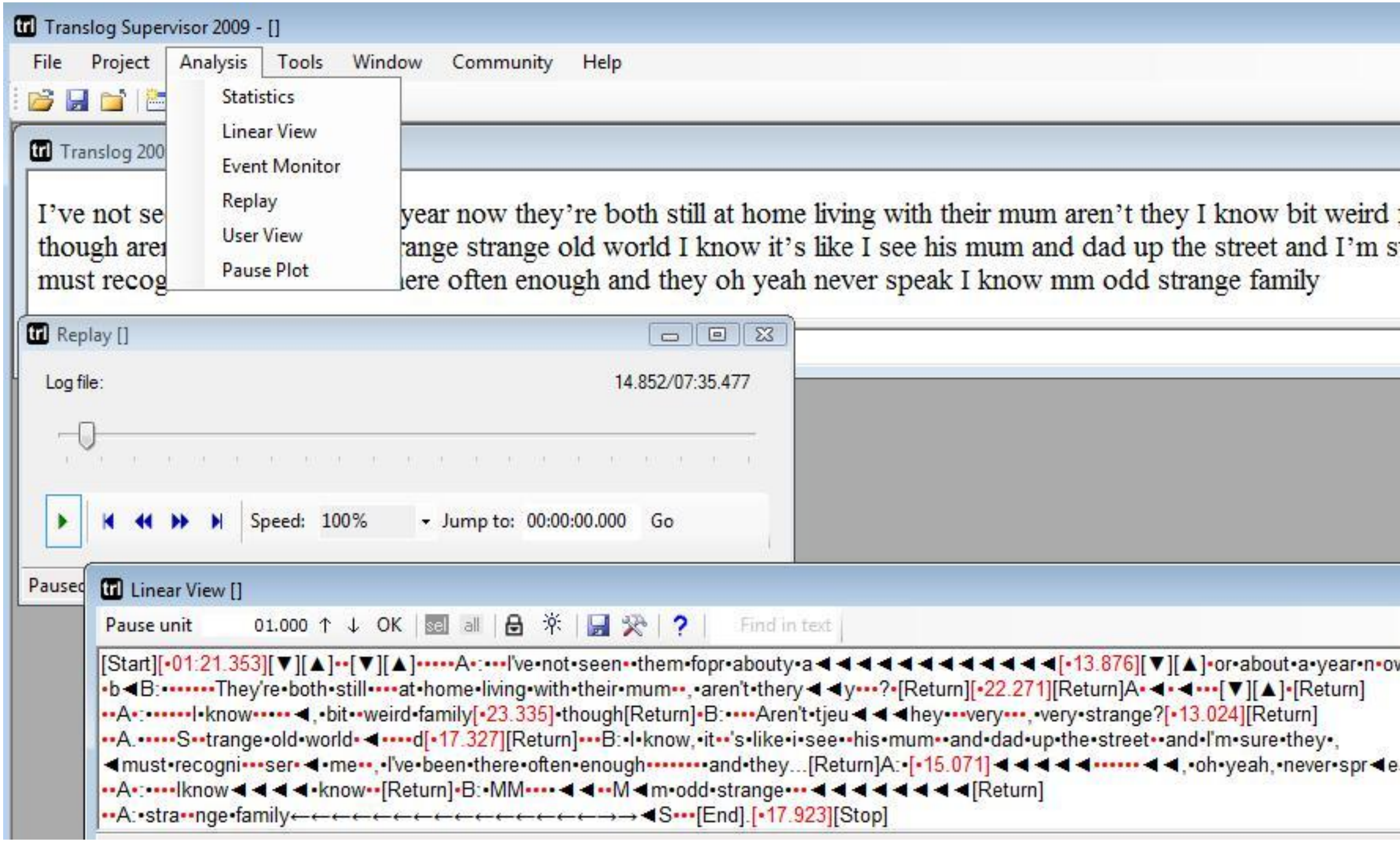
- ❑ Data representation and analysis (User Activity Data)
- ❑ Constant development of Translog-II

Overview – **Copenhagen Dependency Treebank**

1. NLP resource: parallel treebanks for Danish, English, German, Italian, and Spanish with 60,000 words in each language
2. Annotation levels:
   - MORPHOLOGY
   - SYNTAX
   - SEMANTICS
   - DISCOURSE
   - ANAPHORA
3. Unified account of morphology, syntax and discourse
4. Excellent basis for automatic parsers and MT-systems

Copenhagen Dependency Treebank


cause
qobj
subj
vobj
pnct
pnct
<ROOT>
subj
pnct
<ROOT> Derfor skal reglerne strammes , " siger S-ordføreren .
<ROOT> Therefore , the rules must be tightened , " says the S spokesperson .
cause
pnct
subj
nobj
qobj
vobj
vobj
pnct
pnct
<ROOT>
subj
mod
nobj
pnct



I've not se year now they're both still at home living with their mum aren't they I know bit weird
though aren ange strange old world I know it's like I see his mum and dad up the street and I'm s
must recog ere often enough and they oh yeah never speak I know mm odd strange family

```
[Start][•01:21.353][▼][▲]••[▼][▲]•••••A•:•••I've•not•seen••them•fopr•abouty•a◀◀◀◀◀◀◀◀◀◀◀◀[•13.876][▼][▲]•or•about•a•year•n•ow
•b◀B:•••••••They're•both•still••••at•home•living•with•their•mum••,•aren't•thery◀◀y•••?•[Return][•22.271][Return]A•◀•◀•••[▼][▲]•[Return]
••A•:••••••I•know•••••◀,•bit••weird•family[•23.335]•though[Return]•B:••••Aren't•tjeu◀◀◀hey•••very•••,•very•strange?[•13.024][Return]
••A.•••••S••trange•old•world•◀•••••d[•17.327][Return]•••B:•I•know,•it••'s•like•i•see••his•mum••and•dad•up•the•street••and•I'm•sure•they•,
◀must•recogni•••ser•◀•me••,•I've•been•there•often•enough••••••••and•they...[Return]A:•[•15.071]◀◀◀◀◀••••••◀◀,•oh•yeah,•never•spr◀e
••A•:••••Iknow◀◀◀◀•know••[Return]•B:•MM•••••◀◀••M◀m•odd•strange•••◀◀◀◀◀◀◀◀[Return]
••A:•stra••nge•family←←←←←←←←←←←←←←←→→◀S•••[End].[•17.923][Stop]
```

Translog-II

Logging behaviour with Translog-II

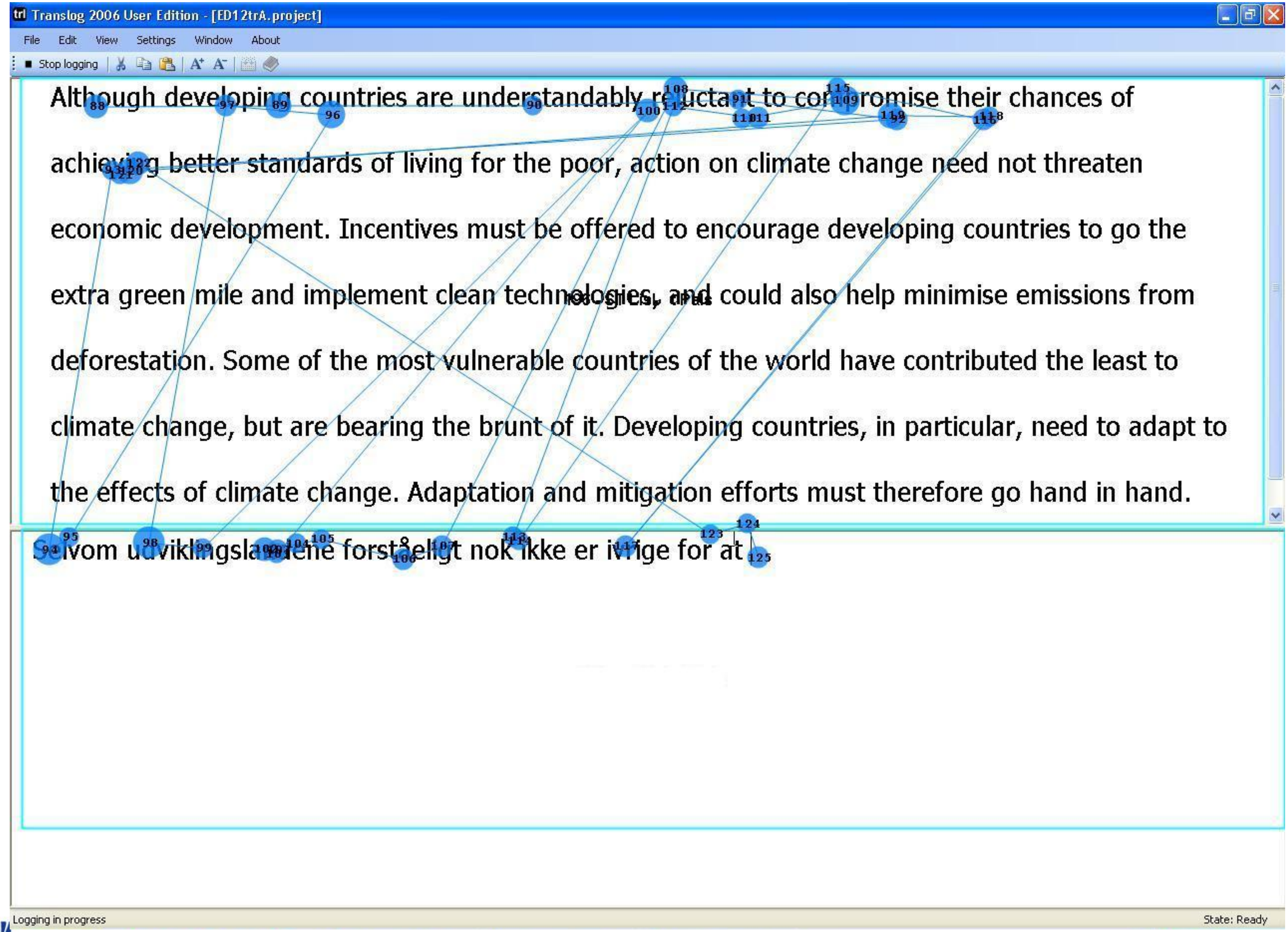
- *Experimental* approach - laboratory view of translation
- Data derived from *naturalistic*, but not fully natural translation events
- Translation is happening, but we can only record data occurring on the outside of the black box
- We get a very detailed, *microscopic* view of the translation process, but we do not get an inside view
- What goes on ’inside’, the cognitive activity, can still only be inferred

A bottom-up process: building a taxonomy of micro-behaviour (UAD) and recording dynamic interaction

- A bottom-up taxonomy where all the data is ’physical’ (eye and finger movements in time)
- Method: tracing what user activity (UAD) went before a certain action and what UAD followed it
(the ’history’ of a process)
- Aim: by tracking the way in which different processes succeed one another or interact dynamically and seeing what linguistic material they operate on, we aim to build a dynamic model of translation

Some basic assumptions

- ❑ What we record (the ’outside’) correlates somehow with the ’inside’
  - E.g. what the eyes fixate is what the mind is currently processing (’the eye-mind hypothesis’)
  - (eye movements are a window on the mind, but not necessarily a very clean and fully transparent window)
- ❑ Low-level ’outside’ data co-vary predictably (probabilistically) with the linguistic material which is being processed
  - E.g. longer pausing (less keystroke activity) before infrequent (less familiar) lexical items
- ❑ Higher-level, e.g. syntactic, processing phenomena can be identified from low-level data (’bottom-up’)
  - E.g. high probability of regressive eye movements to antecedents of anaphors



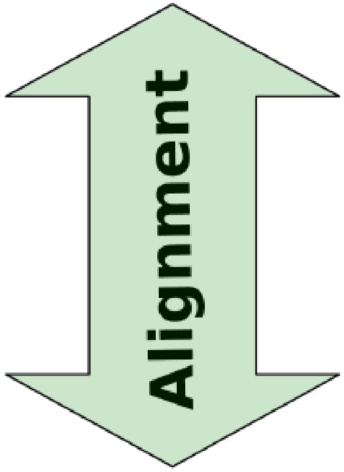
Although developing countries are understandably reluctant to compromise their chances of achieving better standards of living for the poor, action on climate change need not threaten economic development. Incentives must be offered to encourage developing countries to go the extra green mile and implement clean technologies, and could also help minimise emissions from deforestation. Some of the most vulnerable countries of the world have contributed the least to climate change, but are bearing the brunt of it. Developing countries, in particular, need to adapt to the effects of climate change. Adaptation and mitigation efforts must therefore go hand in hand.

Selvom udviklingslandene forståeligt nok ikke er ivrige for at

Copenhagen Business School
HANDELSHØJSKOLEN

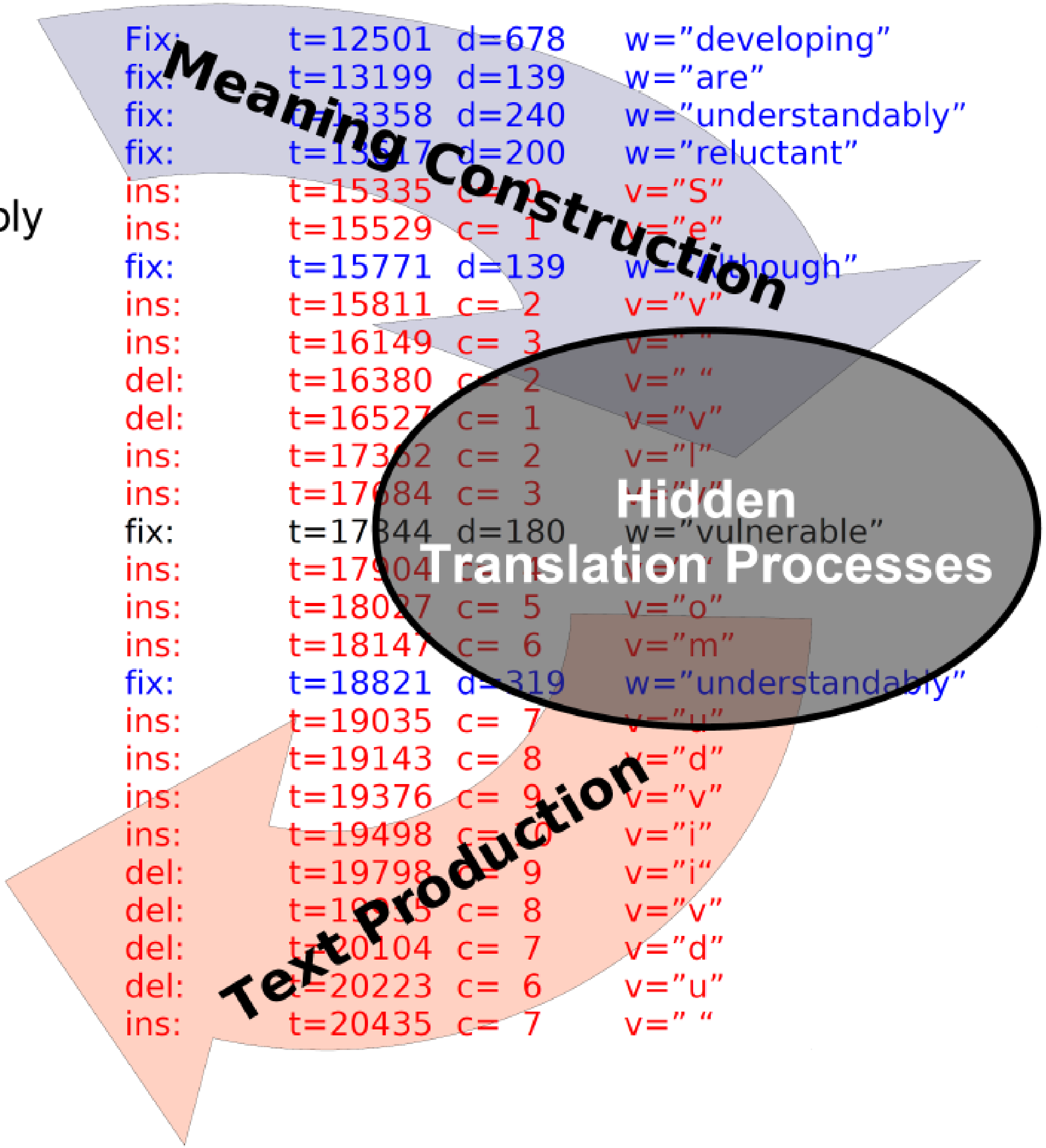
Source

Although developing countries are understandably reluctant to compromise ...



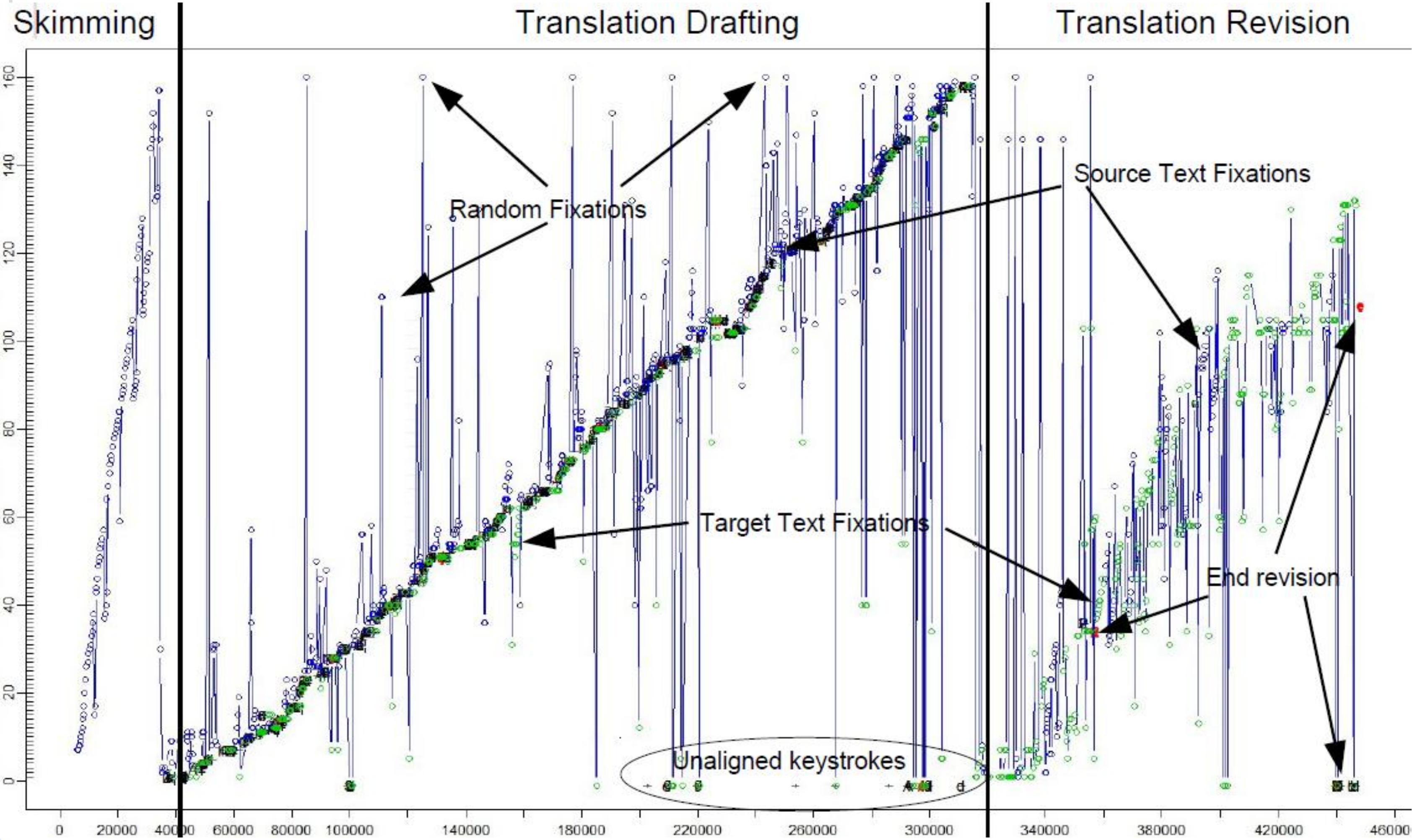
Selv om udviklingslandene forståeligt nok er modvillige mod at risikere deres chancer ...

Target

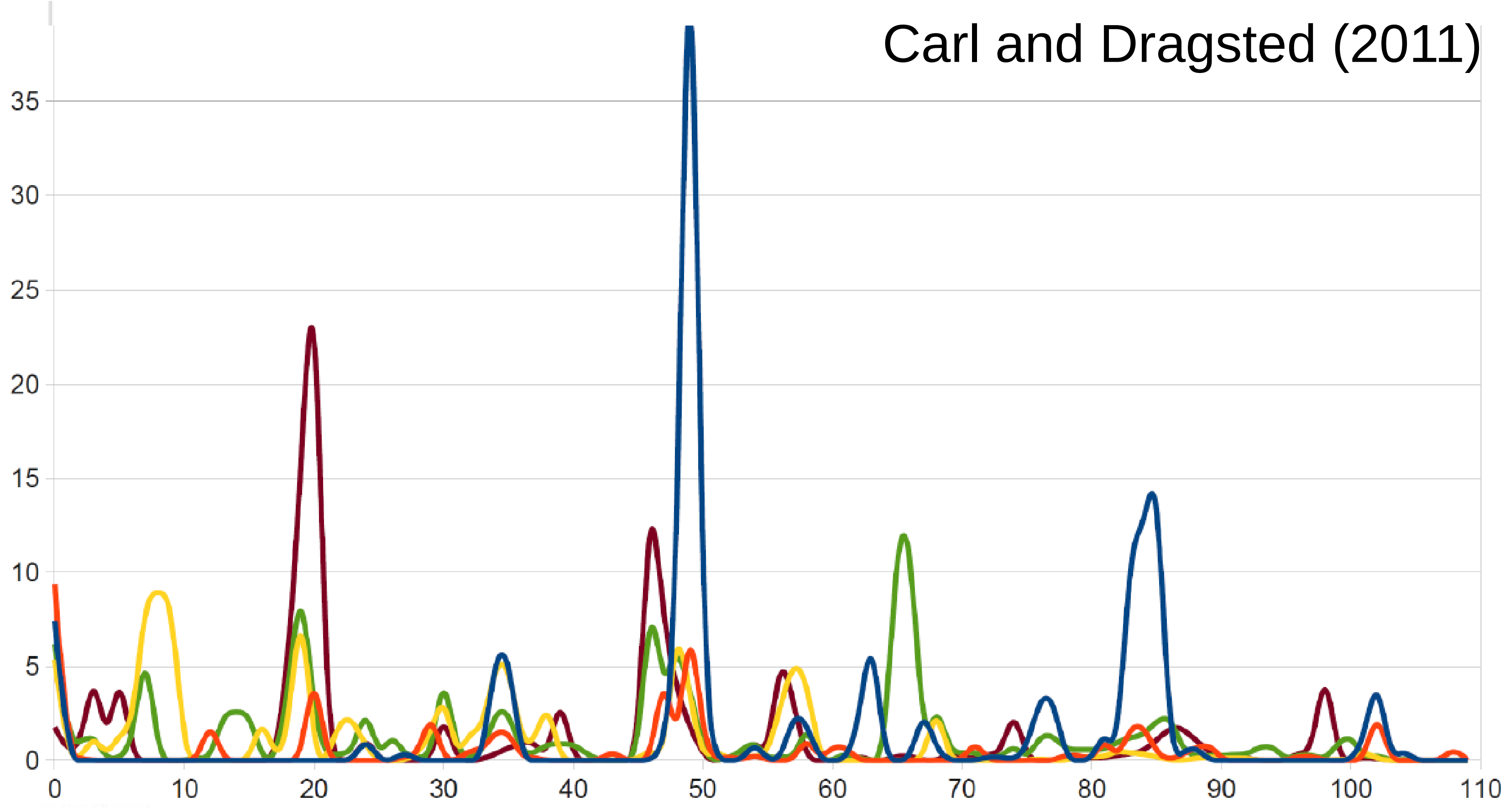


```
Fix:    t=12501  d=678    w="developing"
fix:    t=13199  d=139    w="are"
fix:    t=13358  d=240    w="understandably"
fix:    t=13917  d=200    w="reluctant"
ins:    t=15335  c= 0     v="S"
ins:    t=15529  c= 1     v="e"
fix:    t=15771  d=139    w="although"
ins:    t=15811  c= 2     v="v"
ins:    t=16149  c= 3     v=" "
del:    t=16380  c= 2     v=" "
del:    t=16527  c= 1     v="v"
ins:    t=17362  c= 2     v="l"
ins:    t=17684  c= 3     v="v"
fix:    t=17844  d=180    w="vulnerable"
ins:    t=17904  c= 4     v=" "
ins:    t=18027  c= 5     v="o"
ins:    t=18147  c= 6     v="m"
fix:    t=18821  d=319    w="understandably"
ins:    t=19035  c= 7     v="u"
ins:    t=19143  c= 8     v="d"
ins:    t=19376  c= 9     v="v"
ins:    t=19498  c=10     v="i"
del:    t=19798  c= 9     v="i"
del:    t=19935  c= 8     v="v"
del:    t=20104  c= 7     v="d"
del:    t=20223  c= 6     v="u"
ins:    t=20435  c= 7     v=" "
```

Visualization - Progression graphs



Predictability of translators' behaviour

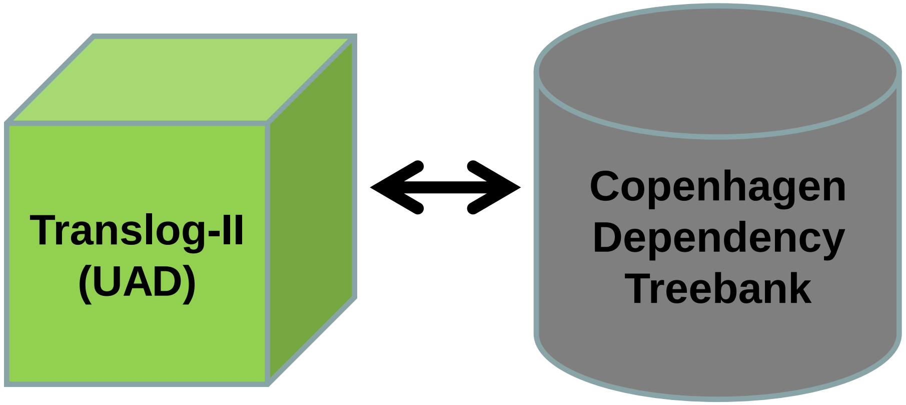


Carl and Dragsted (2011)

Future perspectives

- Using Translog-II:
  - Analysis of the typing process (pauses & typing)
  - Analysis of eye movements in the reading process
  - Analysis of coordination of reading and typing

ENRICHING THE COPENHAGEN DEPENDENCY TREEBANK WITH **USER ACTIVITY DATA** (UAD)



Future perspectives

- High expectations on the combination of UAD and CDT data to facilitate inquiries into isomorphism between:
  - PAUSES & GAZE LOCATION/DURATION
  - LINGUISTIC LAYERS IN THE BROADEST SENSE
- New insights about source text decoding, memory retrieval and encoding of textual segments

Conclusions

- By mapping UAD to a parallel treebank with multilevel linguistic annotation, the possibilities of systematically analyzing correlations between **gaze fixations** / **keystrokes** and underlying **linguistic structure** of the texts are promising to uncover translation processes.

- We assume that the integration of these two NLP resources would allow us to correlate patterns of UAD with patterns of morphological, syntactic or discourse structure.

Integrating Treebank Annotation and User Activity Data in Translation Research

Michael Carl, Henrik Høeg Müller, Bartolomé Mesa-Lao

{mc, hhm,bm}.isv@cbs.dk

Copenhagen Business School
Department of International Language Studies and Computational Linguistics

CBS, 07.03.2012