UNIVERSITY OF Southampton

Towards the Domain Agnostic Generation of Natural Language Explanations from Provenance Graphs for Casual Users

Darren P. Richardson* and Luc Moreau

6th International Provenance and Annotation Workshop
McLean, Virginia, USA — 6th-9th June 2016

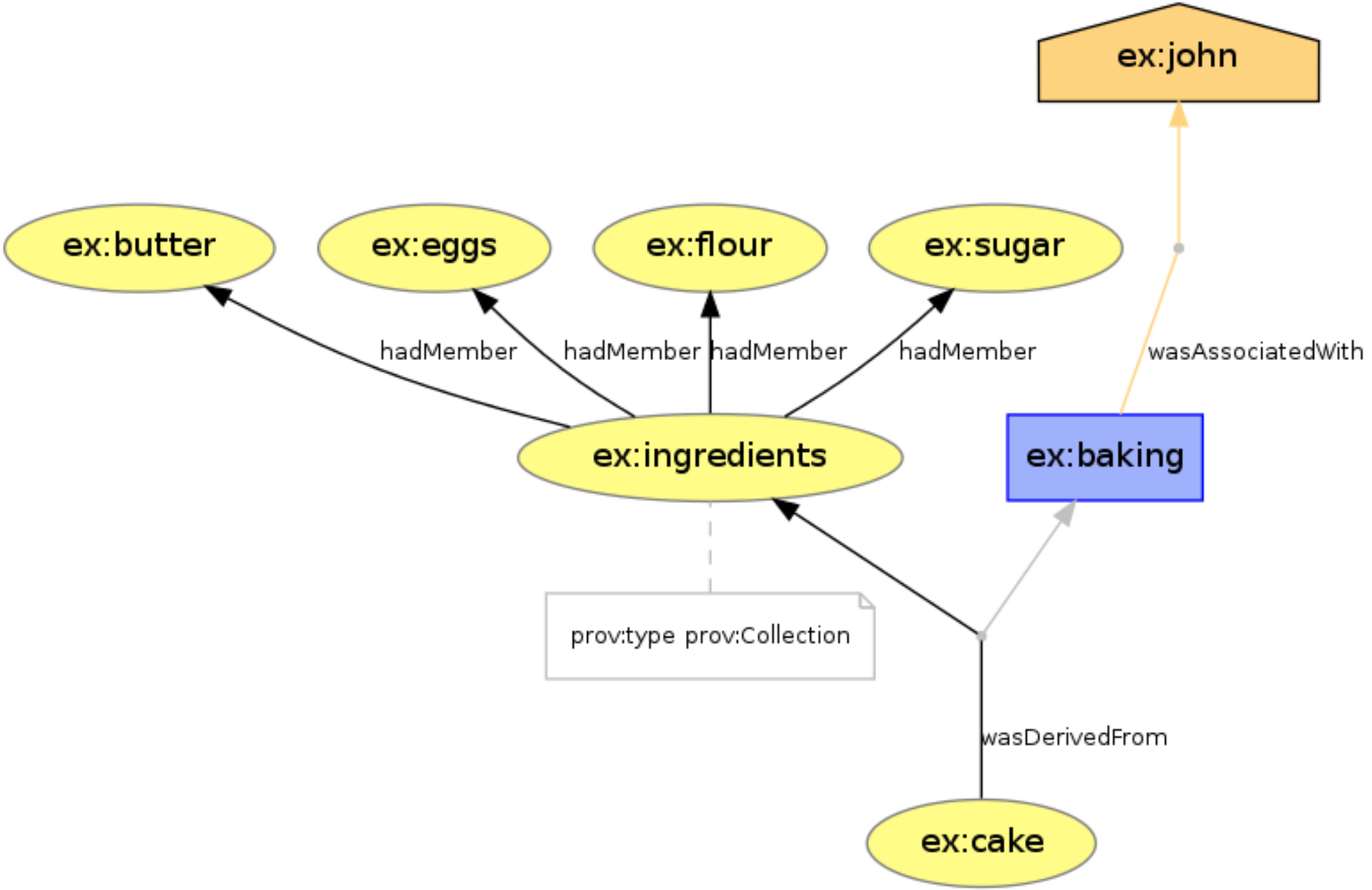
Revision 3

PROV-N

```
document
  prefix xsd_1 <http://www.w3.org/2001/XMLSchema>
  prefix ex <https://example.net/#>

  wasAssociatedWith(ex:baking, ex:john, -)
  hadMember(ex:ingredients, ex:eggs)
  hadMember(ex:ingredients, ex:flour)
  hadMember(ex:ingredients, ex:sugar)
  hadMember(ex:ingredients, ex:butter)
  agent(ex:john)
  entity(ex:butter)
  entity(ex:eggs)
  entity(ex:ingredients, [prov:type='prov:Collection'])
  entity(ex:cake)
  entity(ex:flour)
  entity(ex:sugar)
  activity(ex:baking, -, -)
  wasDerivedFrom(ex:deriv; ex:cake, ex:ingredients, ex:baking, -, -)
endDocument
```

PROV Working Group Diagrams



Sankey Diagrams

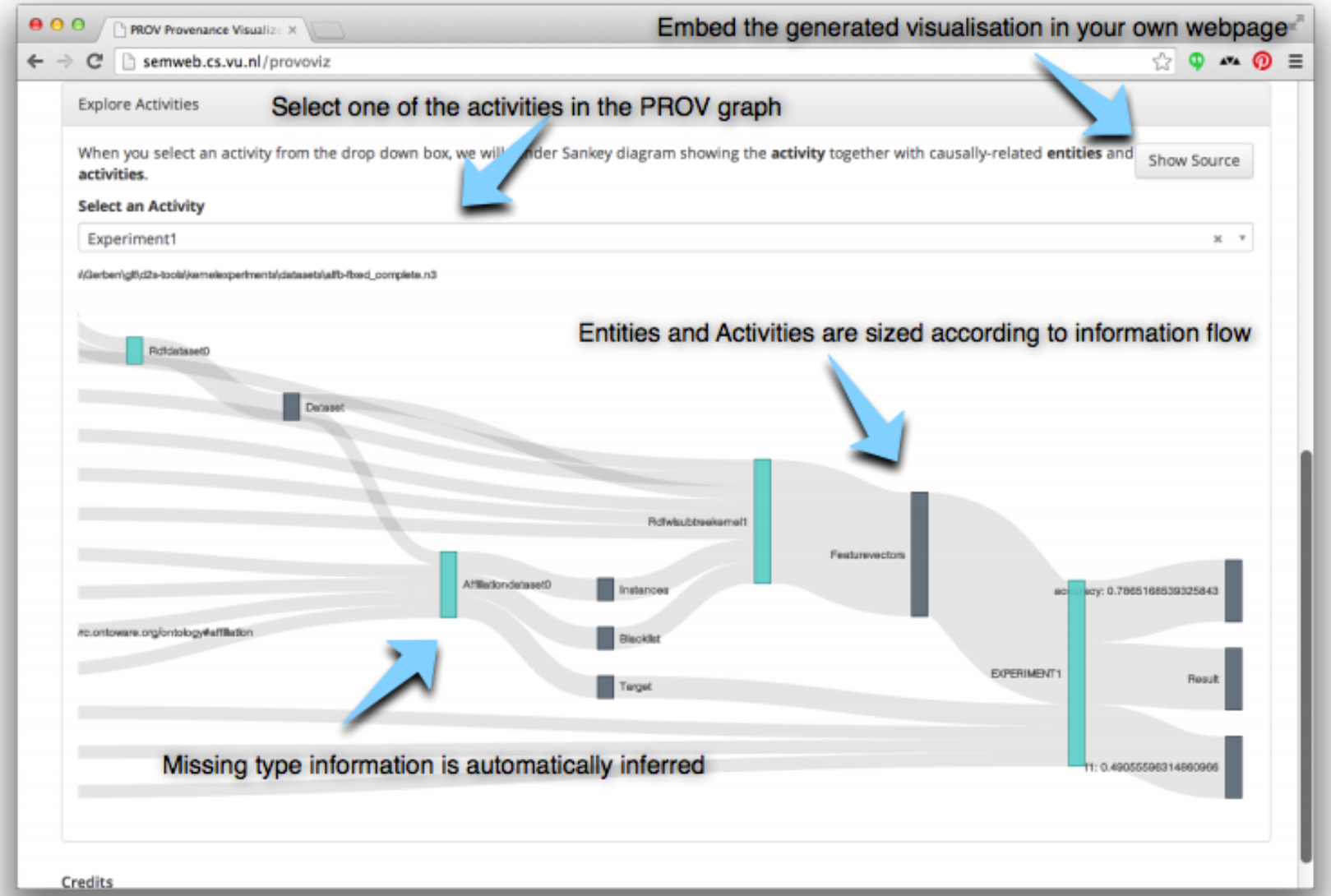
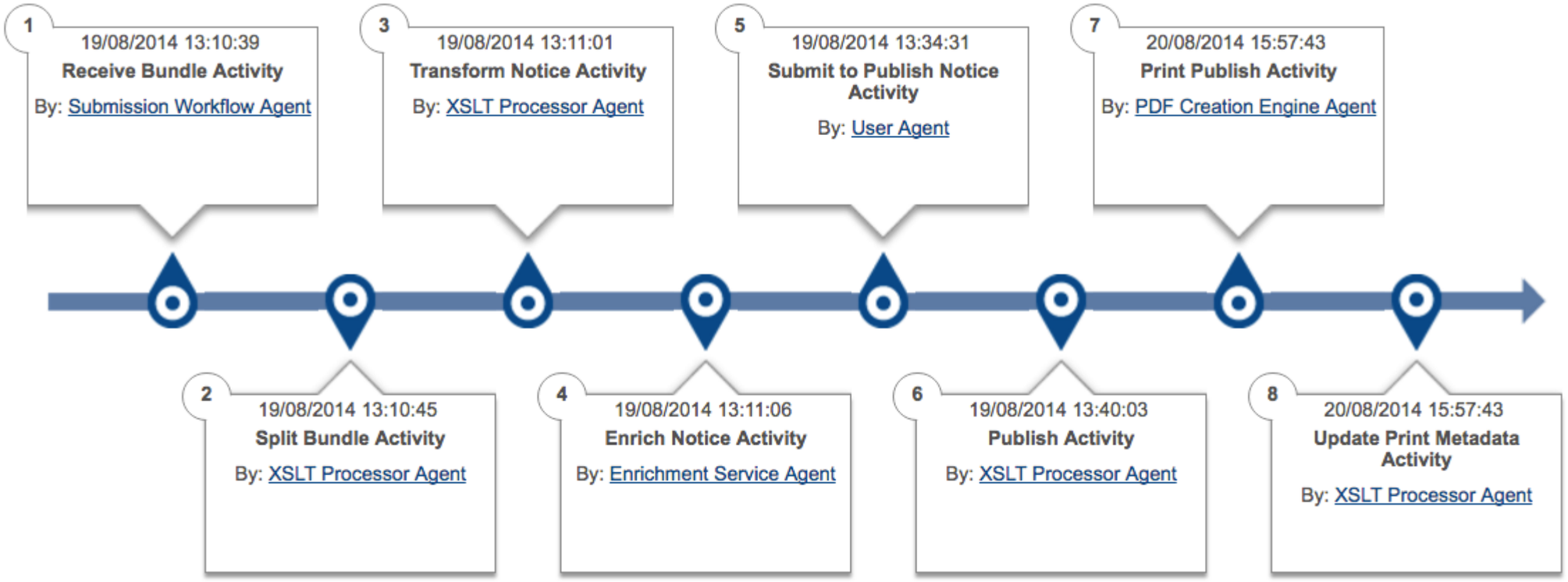


Image taken from *Hoekstra & Groth*

Custom solution from *The Gazette*



Motivation

- Increasingly important to convey provenance to humans
- Existing methods
  - Formal (PROV-N, etc.)
  - Diagrammatic (PROV Working Group, Sankey Diagrams, etc.)
  - Linguistic, including text and speech
- Advantages of linguistic forms of communication
  - Widens our repertoire, and allows for multimedia
  - Screens not suitable in some contexts (i.e. whilst driving)
  - Potentially more accessible to casual users

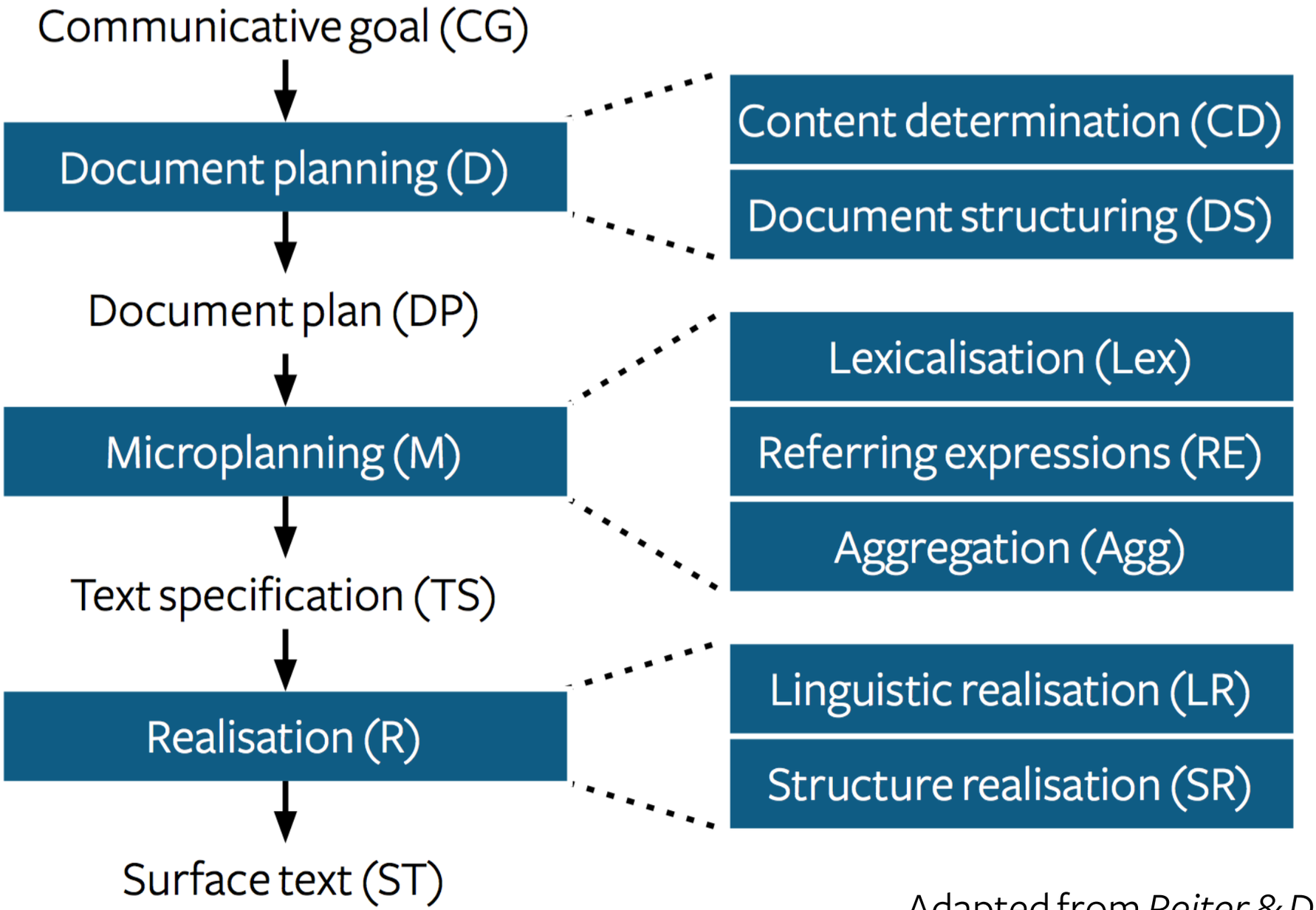
Existing linguistic solutions

- Simple string-substitution methods
  - Easy to develop
  - Tend to be either:
    - Domain specific, with richer sentences
    - Domain agnostic, with less natural-looking sentences
- We need something natural-looking and domain agnostic
  - This is not possible with simple string-substitution

Using sophisticated NLG technology

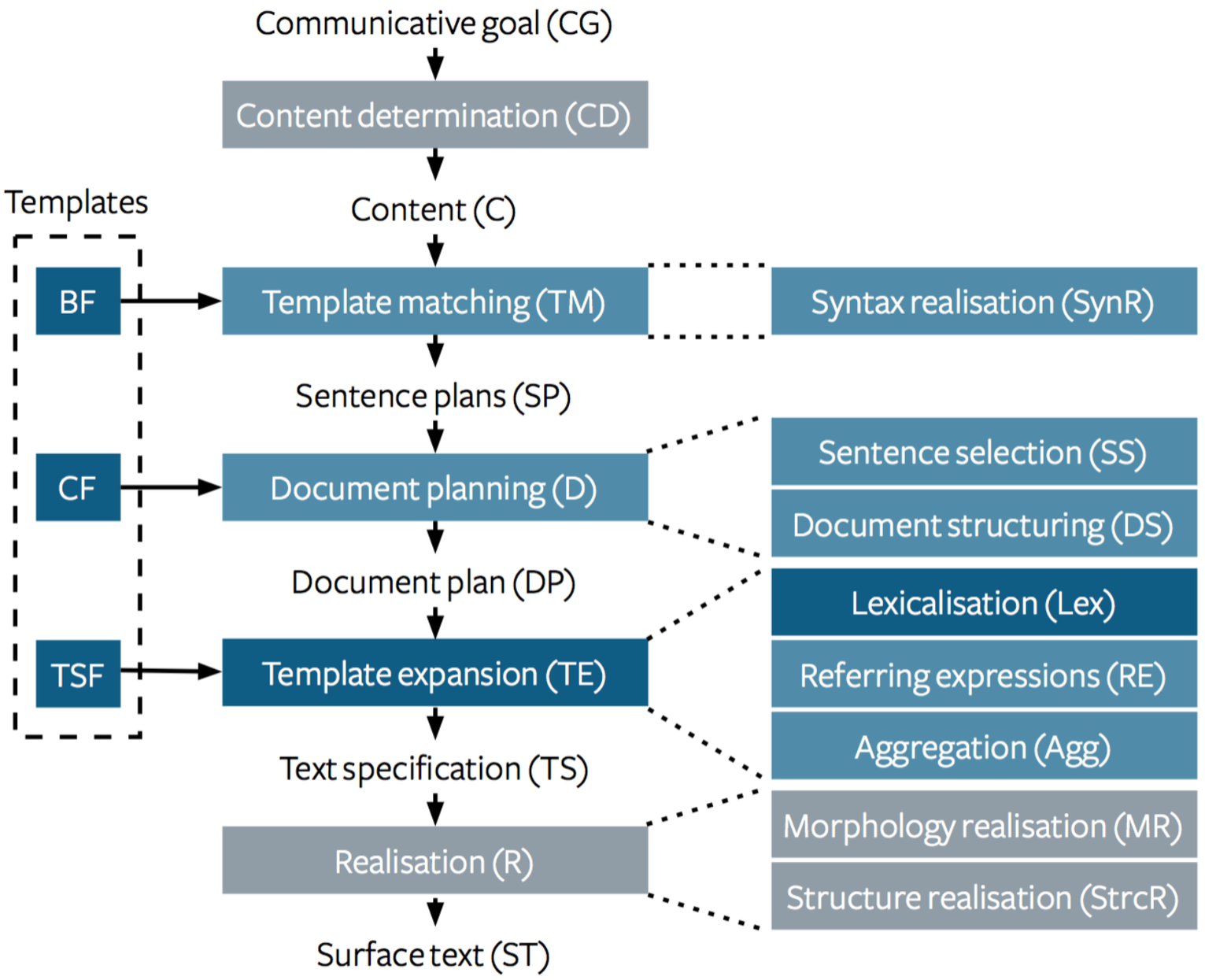
- Pros
  - Takes advantage of a large body of existing research
  - Capable of generating orthographically correct sentences
    - Verb conjugation, tenses, number agreement, etc.
  - Modular construction
  - Allows for the use of off-the-shelf components
- Cons
  - Requires a greater understanding of linguistics to develop
  - Requires a larger amount of linguistic information at runtime

The "consensus" NLG architecture



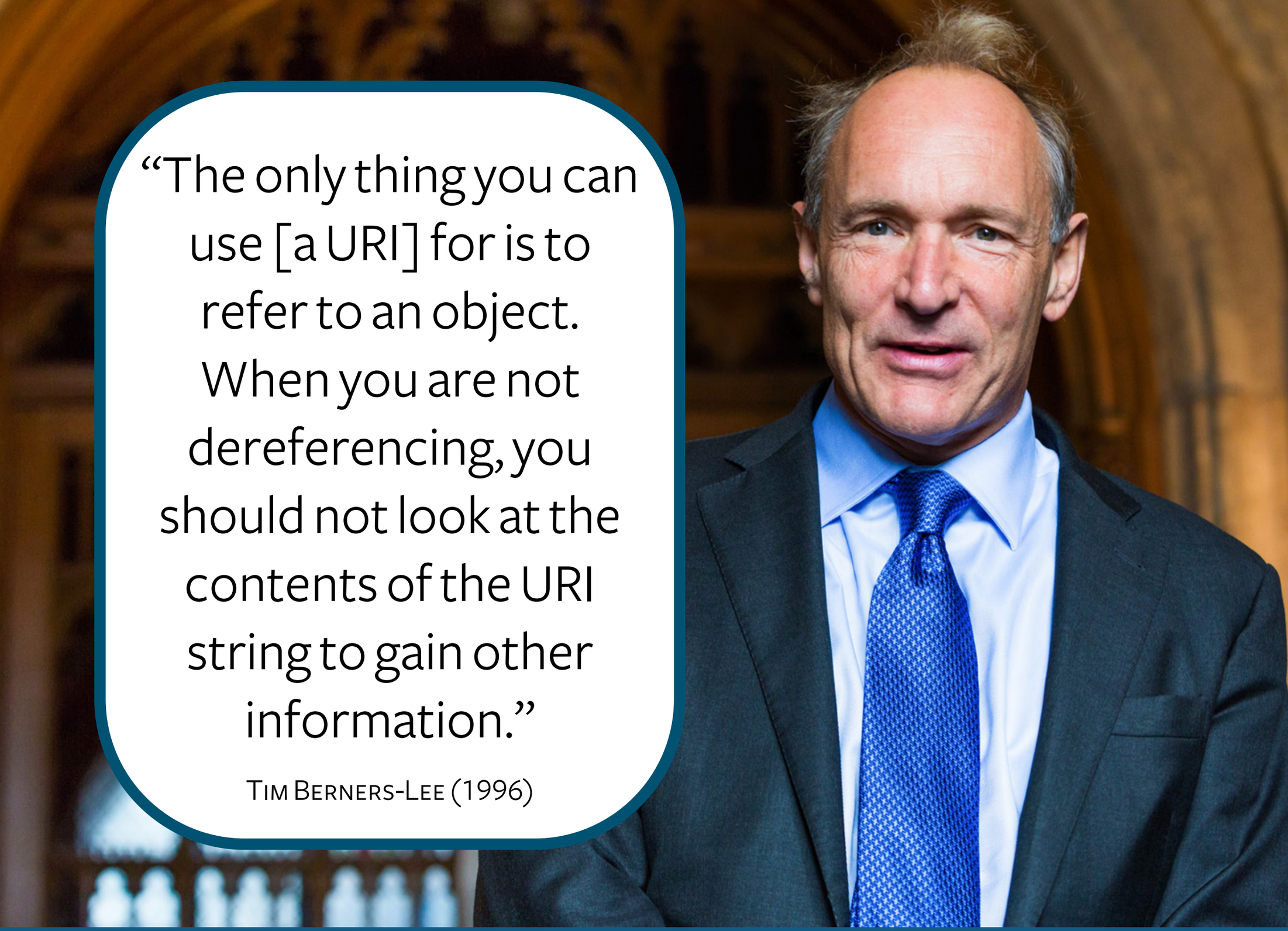
Adapted from *Reiter & Dale*

The PROVglish architecture



The PROVglish architecture

- What are templates in this architecture?
  - These are not PROV templates!
  - They look for common patterns in PROV graphs
    - A relatively small number of templates (approx 20)
  - They map those patterns onto sentence structures
  - This results in a text specification (TS)
    - An abstract representation of a sentence
  - The text specification can be interpreted by an off-the-shelf realisation engine, and turned into orthographically correct English.


"The only thing you can use [a URI] for is to refer to an object. When you are not dereferencing, you should not look at the contents of the URI string to gain other information."
Tim Berners-Lee (1996)

URIs & linguistic information

- More sophisticated NLG requires a greater degree of linguistic information
- Tokenisation
  - Regex (see paper!)
  - Works 96% of the time
- Tagging
  - Off-the-shelf maximum-entropy POS tagger from NLTK
  - Gives correct specific tag 62.7% of the time
  - Gives correct class of tag 92.3% of the time
    - This is the one we really care about

Examples — Tokenisation

- http://example.org#derek
  - derek
- http://example.org#aggregatedByRegions
  - aggregated, By, Regions
- http://example.org#civil_action_group
  - civil, action, group
- http://www.agentswitch.org/ns/APIRequestParsing_1
  - ns, API, Request, Parsing, 1
- http://www.ipaw.info/data/people/McGuinnessDeborahL
  - data, people, McGuinness, Deborah, L

Examples — Generation

— Reputation manager generated opinion 1.

  — '/rs/reputation_manager' generated '/rs/opinion/1/' by '/reputationapi/#generate_opinion_1'.

— Derek illustrated chart 1.

  — '/derek' generated '/chart1' by '/illustrate'.

— 2 agent posted ride requests 1.

  — '/rideshare/#/users/agent2' generated '/rideshare/#/rideRequests/1' by '/rideshare/#post_ride_request_100339'.

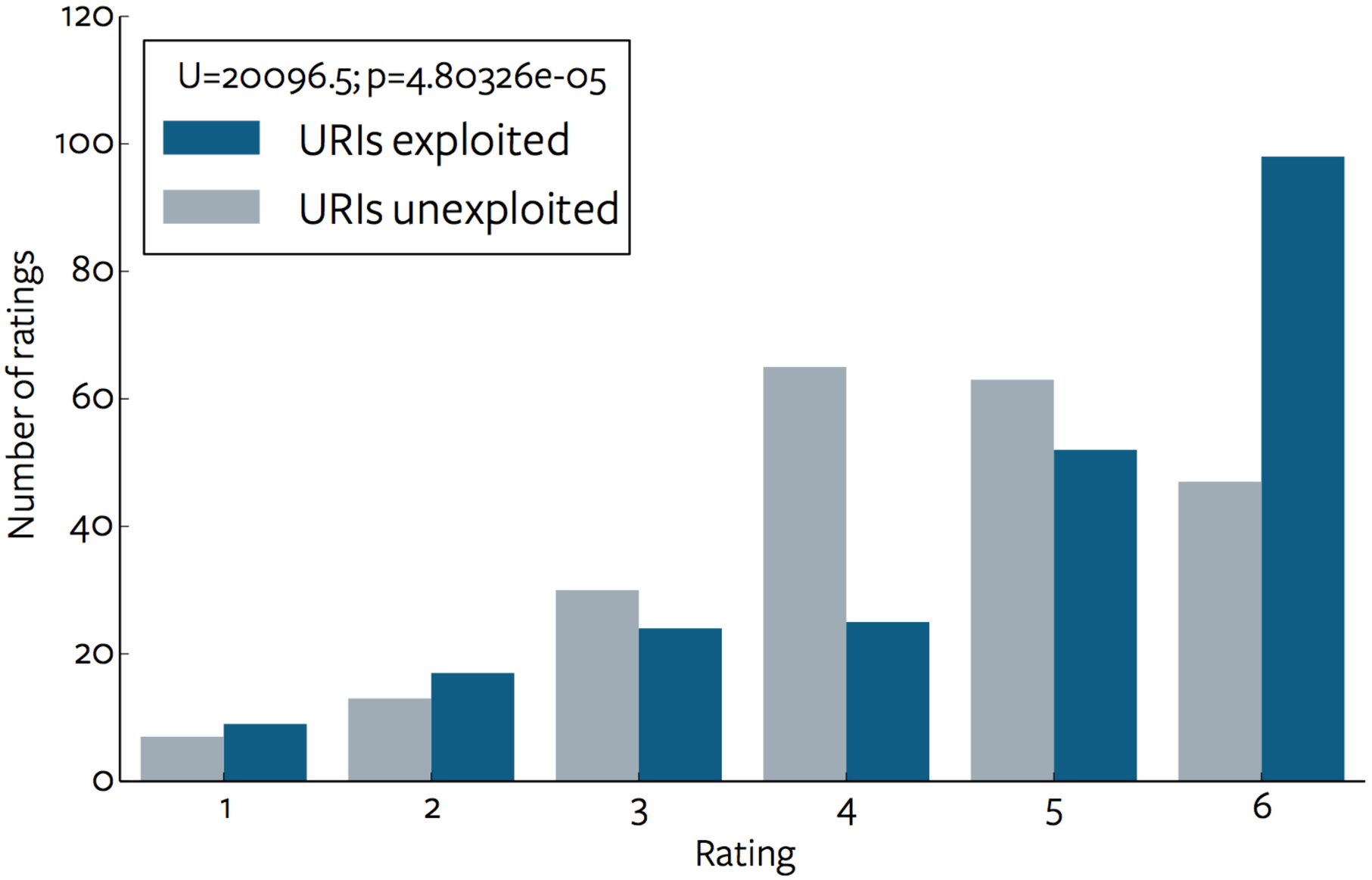
Evaluation

- Participants shown two sentences side-by-side
  - One generated exploiting linguistic information in URIs
  - One generated not exploiting this information
  - Which sentence appears on the left randomised each time
- Direct sentence comparison across 3 dimensions:
  - Grammatical correctness — *no difference expected*
  - Fluency — *improvement expected*
  - Comprehensibility — *improvement expected*
- 15 participants each reviewing 15 sentence-pairs
  - N=225

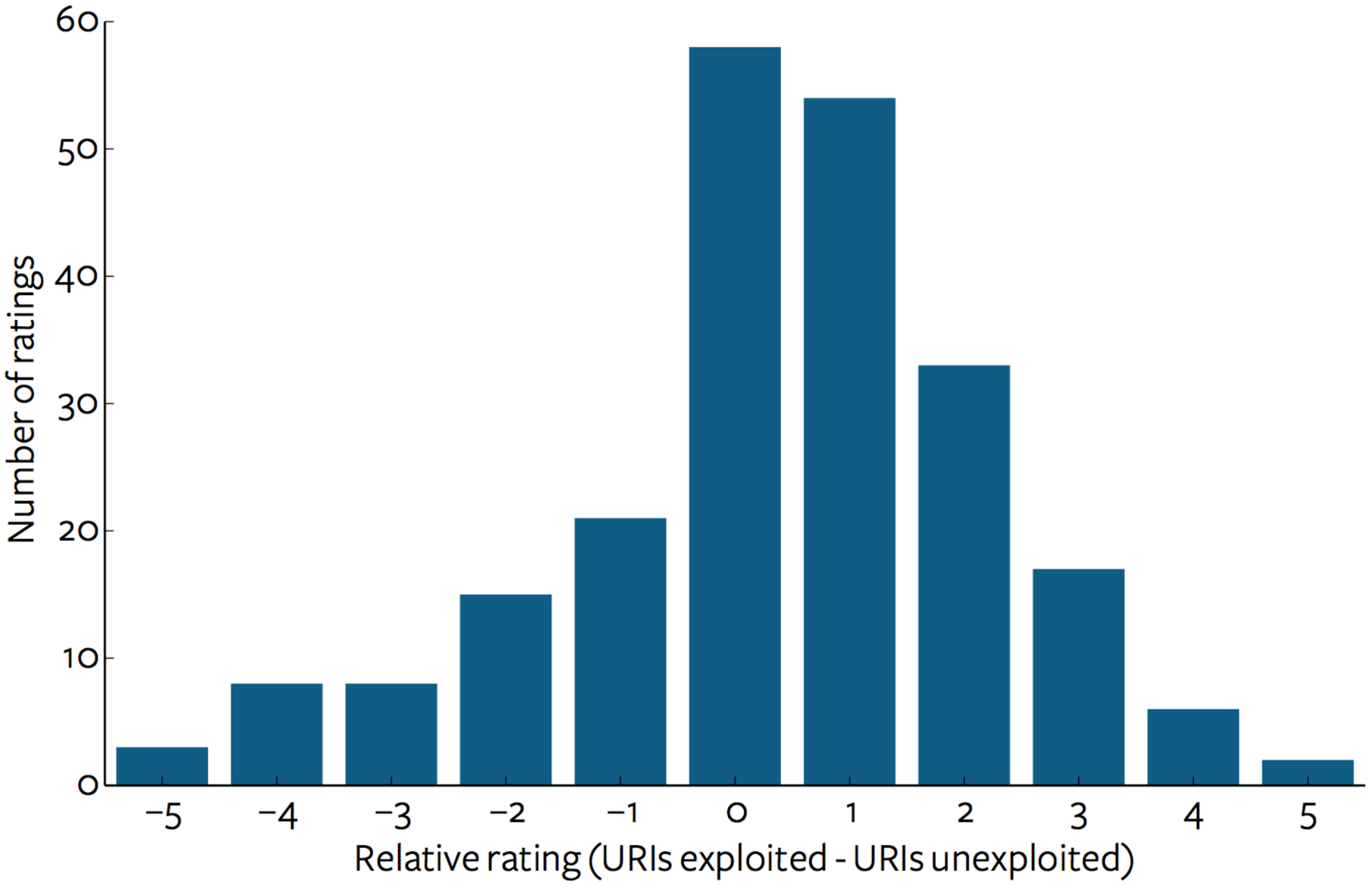
Evaluation

- Participants asked which explanation they thought was better
  - Sentences with URIs exploited favoured 56.5% of the time
  - Sentences with URIs unexploited favoured 29.3% of the time
  - Neither sentence favoured 14.2% of the time

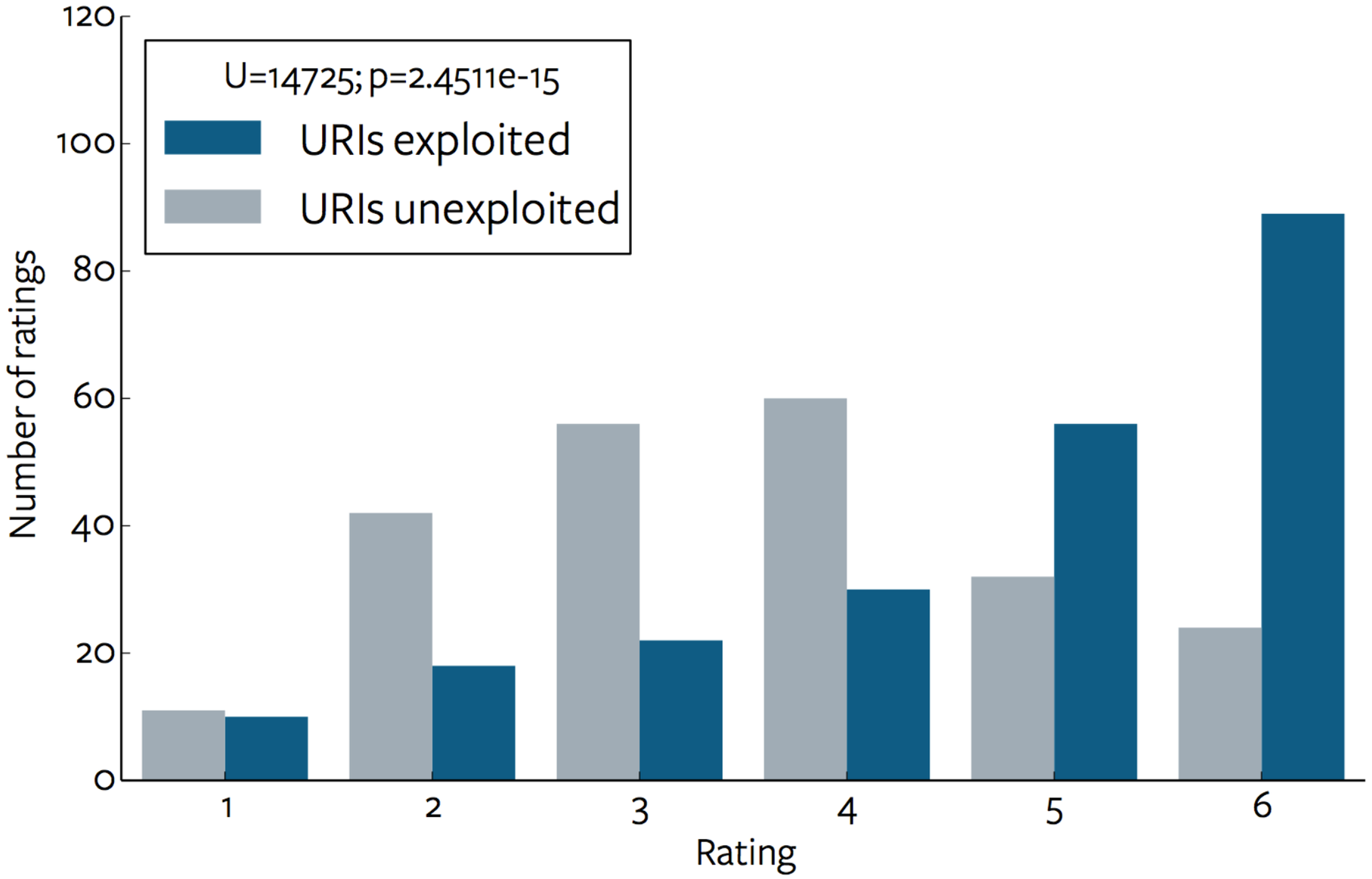
Grammatical correctness



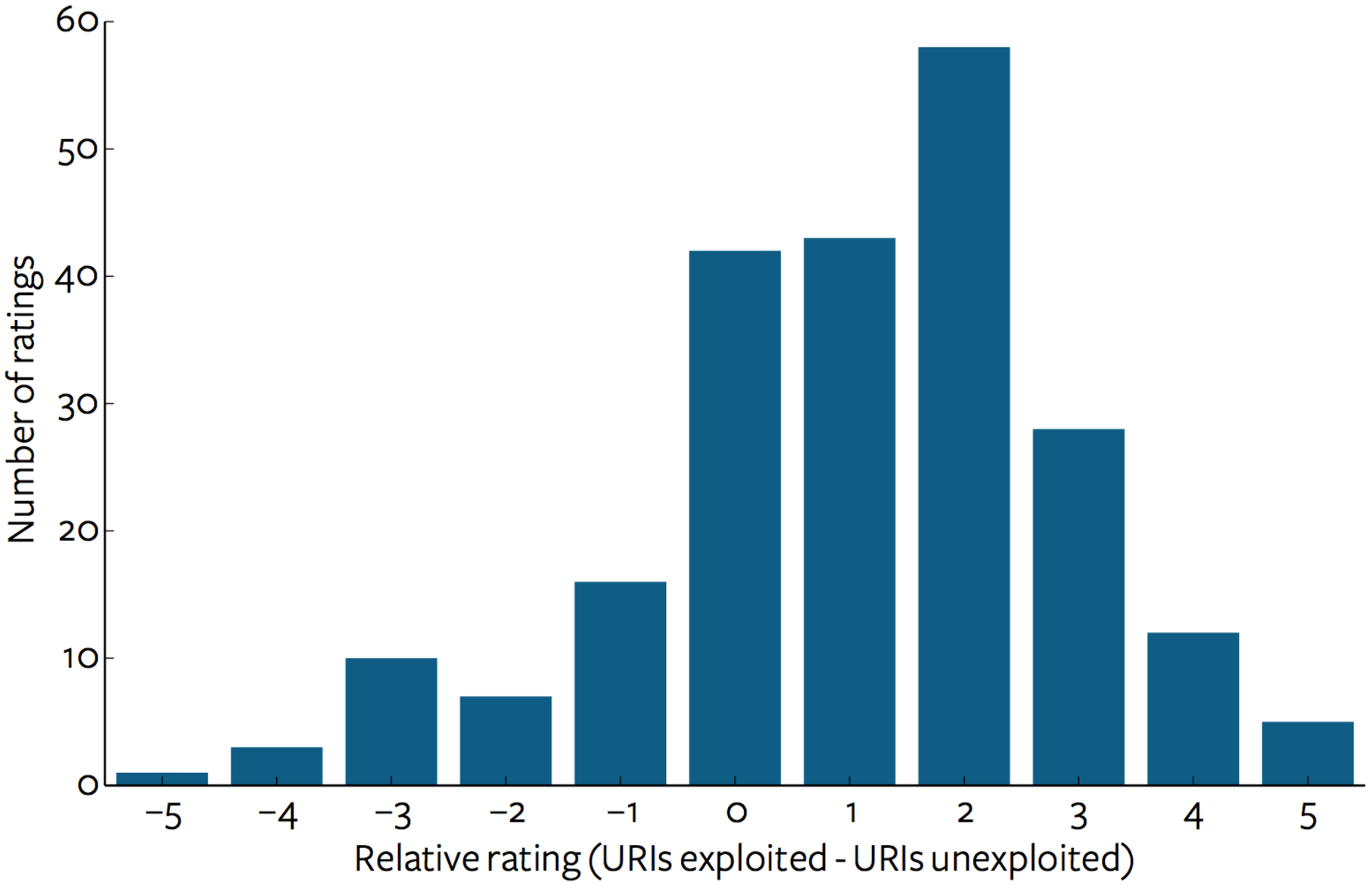
Grammatical correctness



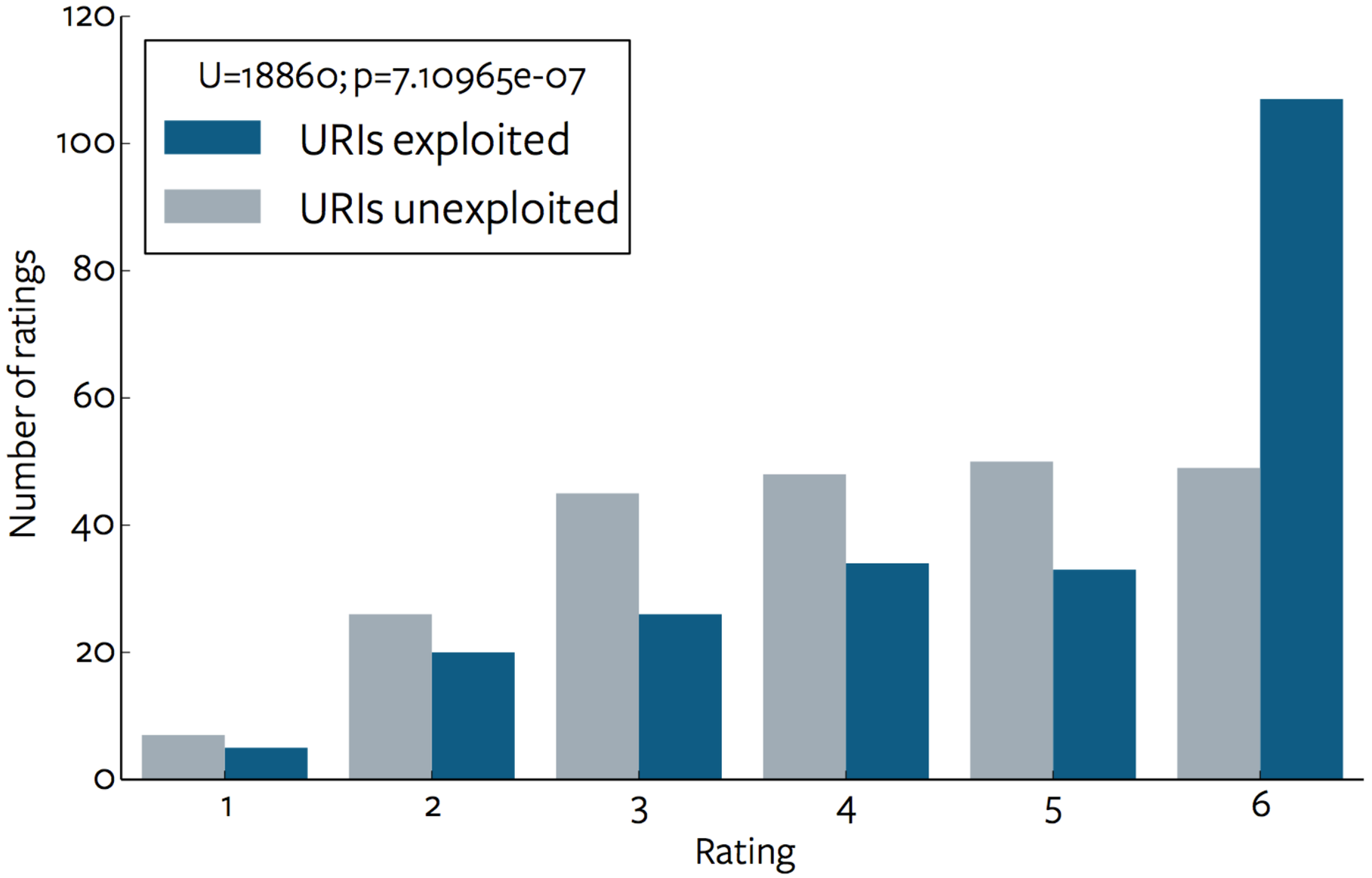
Fluency


U=14725; p=2.4511e-15
URIs exploited
URIs unexploited
Number of ratings
120
100
80
60
40
20
0
1
2
3
4
5
6
Rating

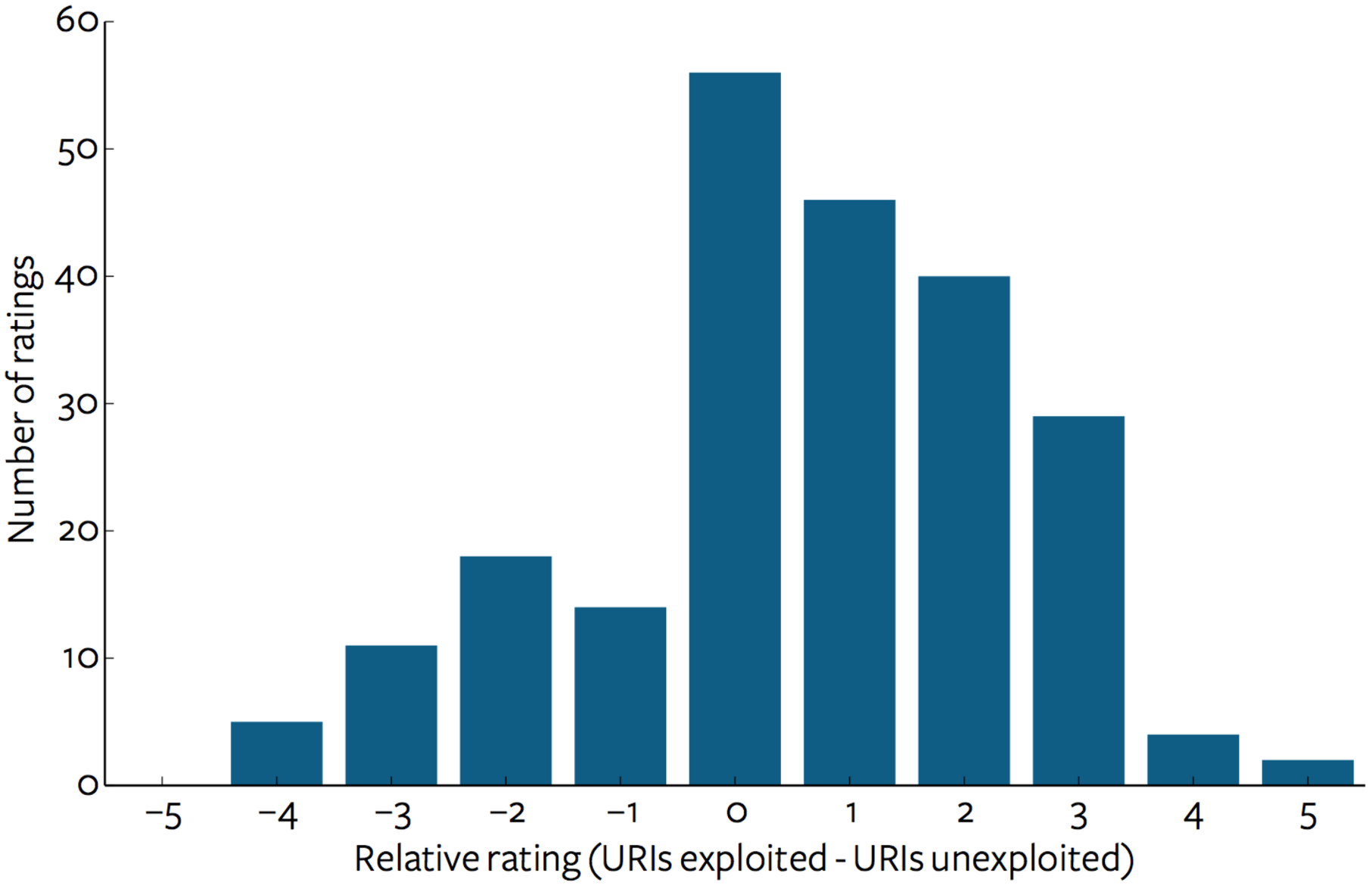
Fluency


60
50
40
30
20
10
0
Number of ratings
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative rating (URIs exploited - URIs unexploited)

Comprehensibility



Comprehensibility



Conclusions & future work

- URIs valuable source of lexical and other linguistic information
- PROVglish template-based generation
  - Domain-generic approach to text generation
  - Significant improvement in terms of:
    - Grammatical correctness — *unexpected!*
    - Fluency
    - Comprehensibility
- Good stepping-stone for future research
  - Extend to generating full paragraphs and documents
  - Based on PROV summaries?

References

— Hoekstra, R., and Groth, P., PROV-O-Viz — Understanding the Role of Activities in Provenance, in Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science, Ludscher, B., and Plale, B., Eds, Springer, (2015).

— Reiter, E., and Dale, R., Building Natural Language Generation Systems. Cam- bridge University Press, (2000).

— For others, see paper.

UNIVERSITY OF Southampton

Thank you for listening

Any questions?

Acknowledgements and Licence

— Picture of Tim Berners-Lee by Paul Clarke, used under CC-BY-SA 4.0 licence.

Research was sponsored by US Army Research laboratory and the UK Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the U.S. Government, the UK Ministry of Defence, or the UK Government. The US and UK Governments are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.