A BENCHMARK FOR RECIPE UNDERSTANDING IN AUTONOMOUS AGENTS

Creation and Analysis of a New Recipe Execution Benchmark

Robin De Haes

VRIJE
UNIVERSITEIT
BRUSSEL

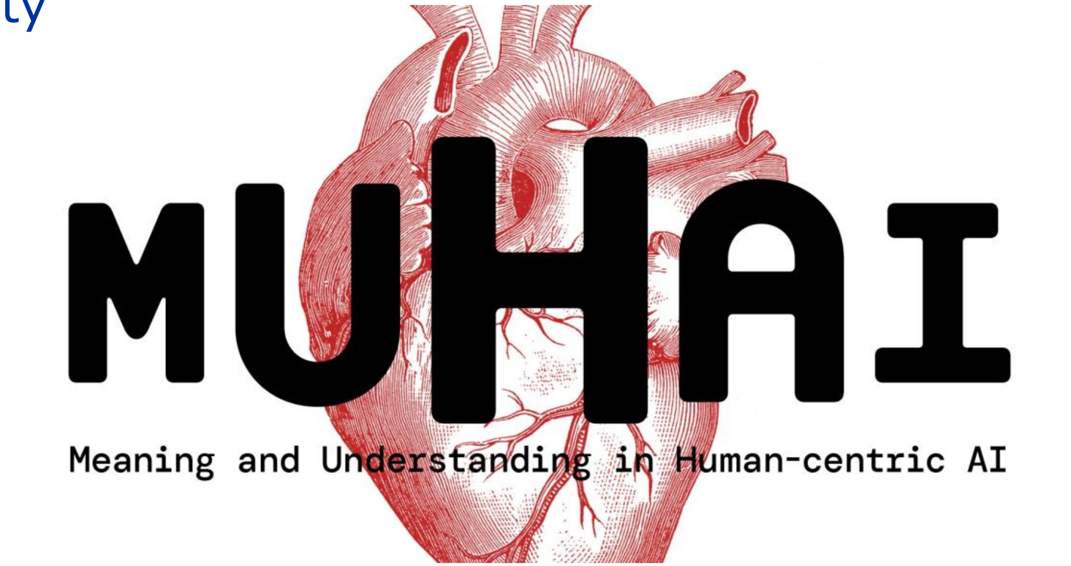
Promotor: Prof. Dr. Paul Van Eecke
Advisor: Dr. Jens Nevens

BACKGROUND

INTRODUCTION

CONTEXT & MOTIVATION

- Meaning and Understanding in Human-centric Artificial Intelligence (MUHAI)
- Deep understanding of an everyday human activity
  - Cooking
- Natural Language Understanding (NLU)
  - Instructional Language
- Benchmarking
  - Written English recipes
  - Recipe Execution



LITERATURE REVIEW
RESEARCH TRAINING

Recipe Understanding

- Challenges
- State-of-the-art
- Robotic Recipe Execution

Benchmarking in AI

- Benchmark definition
- High-quality properties
- Advantages & Disadvantages

Existing Benchmarks for Recipe Understanding

- Composition
- Analysis

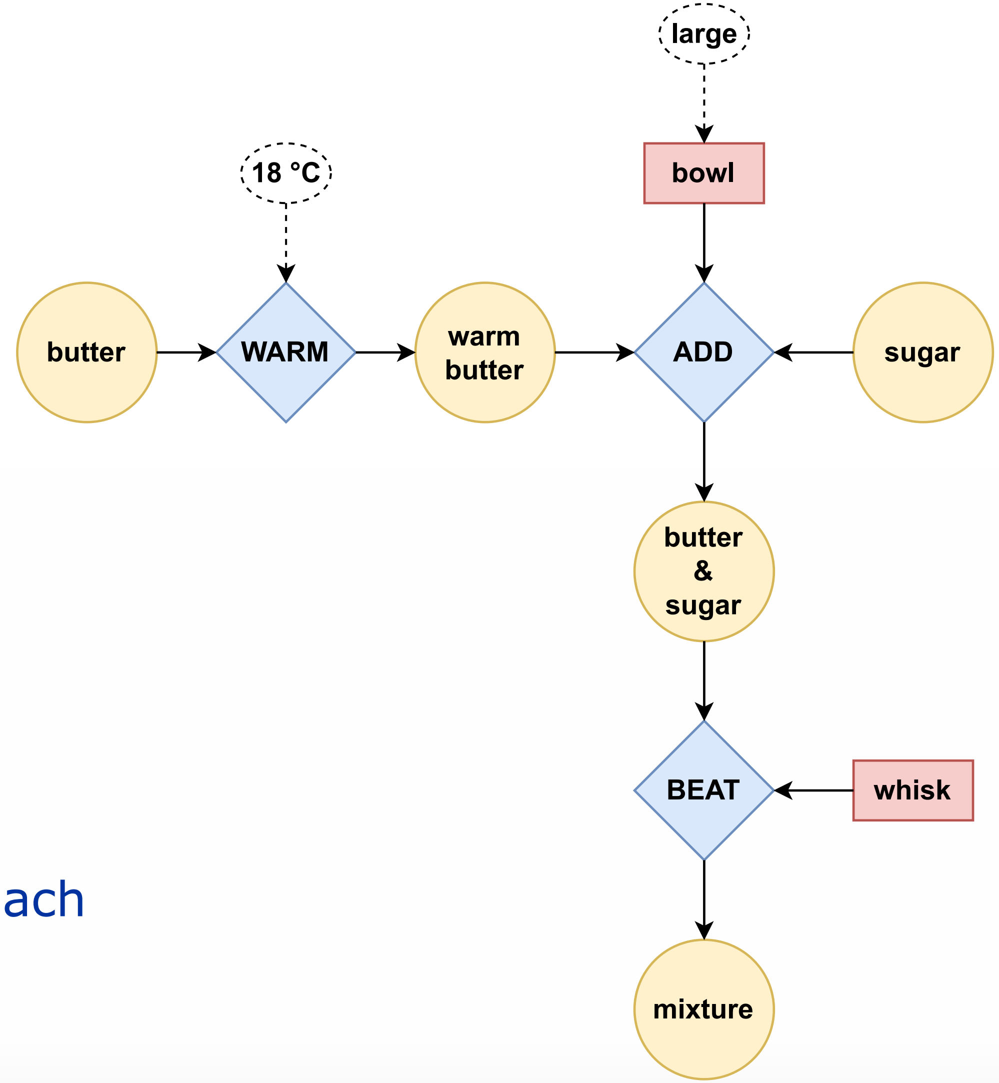
New Benchmark?

RECIPE UNDERSTANDING

LITERATURE REVIEW

- Recipes are popular in AI
  - High availability
  - Practical applications

  BUT: referential & reasoning challenges
- Graph-based representations are common
  - Implicit or explicit
  - Dependencies
  - Multiple modalities
- Recipe execution requires an integrated approach
  - Robustness by improving subsystems



Illustrative graph, created for this presentation

Benchmark components

1. Abstract & concrete task(s)
2. Dataset(s)
3. Evaluation method(s)
4. *Community adoption?*

leading to

- Relevancy
- Reproducibility
- Verifiability
- Fairness
- Usability

(von Kistowski et al., 2015)

BENCHMARKING

LITERATURE REVIEW

Pros

- Communal resources
- Empirical comparisons
- Historically proven catalyst

Cons

- Introduction of biases
- Benchmark overfitting & saturation
- Competition over exploration

RECIPE CORPORA

LITERATURE REVIEW

- CURD (Tasse & Smith, 2008)

  = 260 recipes with domain-specific FOL annotations, but
  - limited design documentation
  - no evaluation tools

- RISeC (Jiang et al., 2020)

  = 260 recipes with PropBank annotations, but
  - limited design documentation
  - no evaluation tools

- r-FG Corpus (Yamakata et al., 2020)

  = 300 recipes with directed acyclic graphs, but
  - low accessibility
  - no evaluation tools

⇨ Low community adoption

BENCHMARK DESIGN

MUHAI RECIPE EXECUTION BENCHMARK

TASKS

Abstract task

= Understanding a recipe deeply enough to execute it

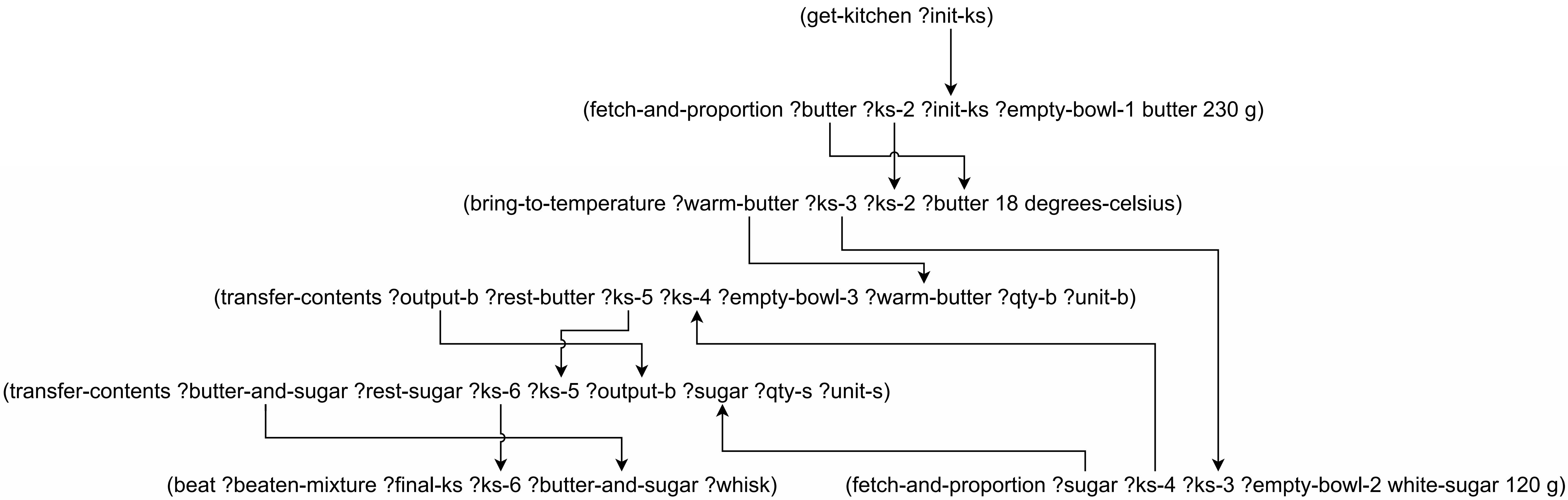
Concrete task

= Parsing a written English recipe text into an executable semantic network

- MUHAI Cooking Language

MUHAI COOKING LANGUAGE

SEMANTIC NETWORK



(get-kitchen ?init-ks)

(fetch-and-proportion ?butter ?ks-2 ?init-ks ?empty-bowl-1 butter 230 g)

(bring-to-temperature ?warm-butter ?ks-3 ?ks-2 ?butter 18 degrees-celsius)

(transfer-contents ?output-b ?rest-butter ?ks-5 ?ks-4 ?empty-bowl-3 ?warm-butter ?qty-b ?unit-b)

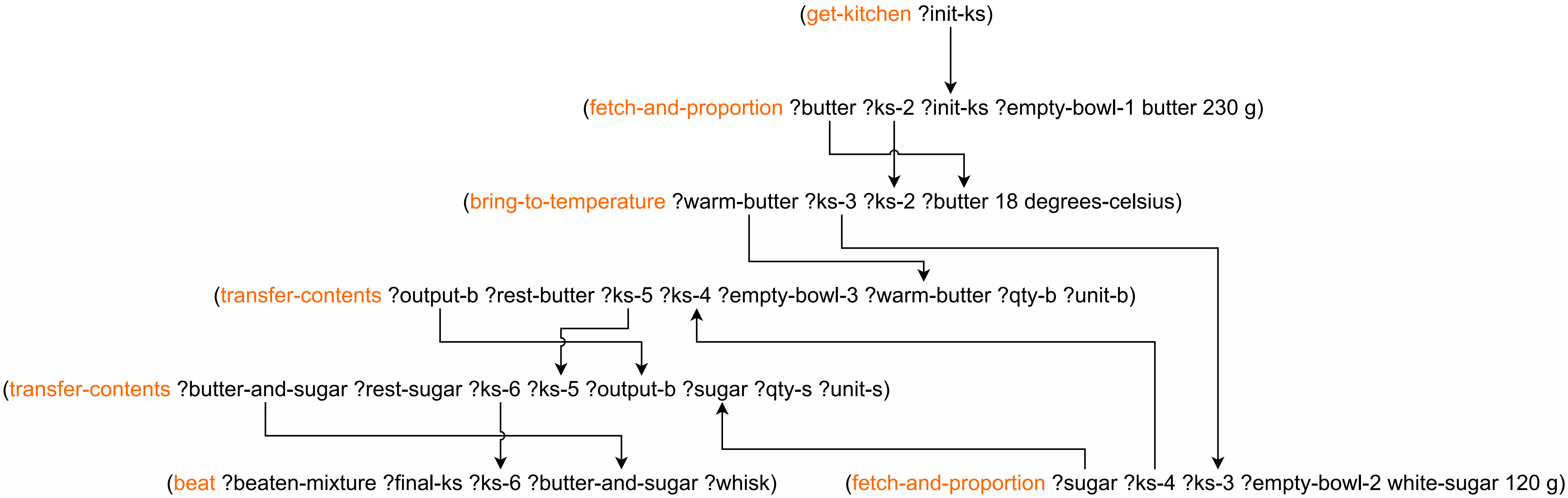
(transfer-contents ?butter-and-sugar ?rest-sugar ?ks-6 ?ks-5 ?output-b ?sugar ?qty-s ?unit-s)

(beat ?beaten-mixture ?final-ks ?ks-6 ?butter-and-sugar ?whisk)

(fetch-and-proportion ?sugar ?ks-4 ?ks-3 ?empty-bowl-2 white-sugar 120 g)

MUHAI COOKING LANGUAGE

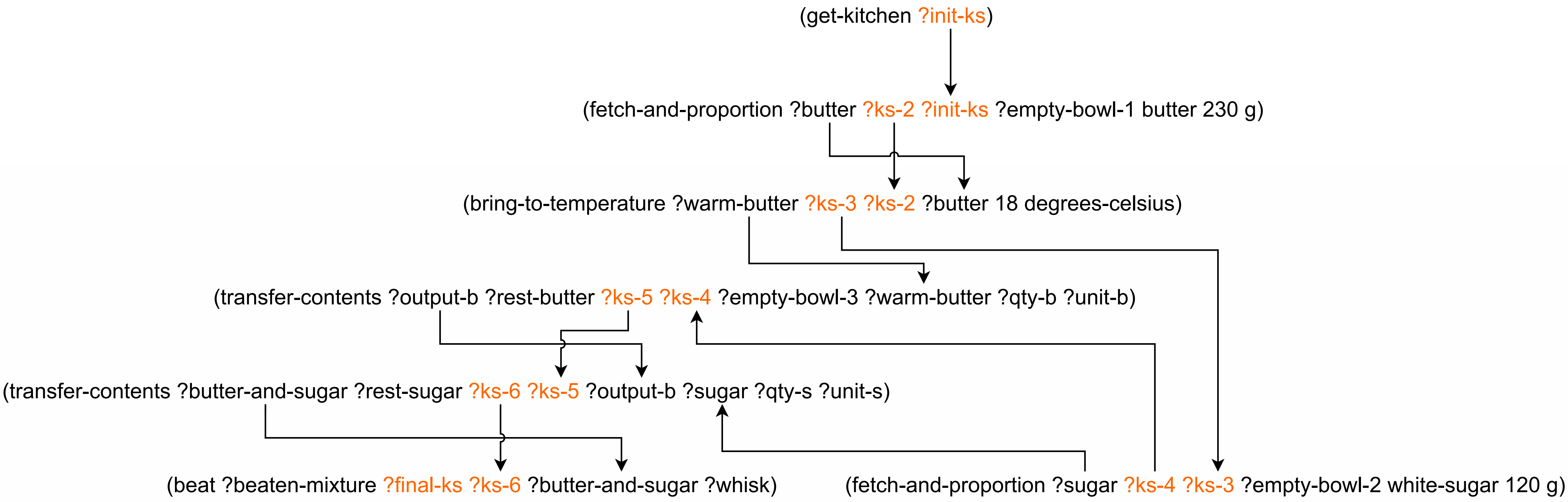
PRIMITIVES



```
(get-kitchen ?init-ks)
(fetch-and-proportion ?butter ?ks-2 ?init-ks ?empty-bowl-1 butter 230 g)
(bring-to-temperature ?warm-butter ?ks-3 ?ks-2 ?butter 18 degrees-celsius)
(transfer-contents ?output-b ?rest-butter ?ks-5 ?ks-4 ?empty-bowl-3 ?warm-butter ?qty-b ?unit-b)
(transfer-contents ?butter-and-sugar ?rest-sugar ?ks-6 ?ks-5 ?output-b ?sugar ?qty-s ?unit-s)
(beat ?beaten-mixture ?final-ks ?ks-6 ?butter-and-sugar ?whisk)
(fetch-and-proportion ?sugar ?ks-4 ?ks-3 ?empty-bowl-2 white-sugar 120 g)
```

MUHAI COOKING LANGUAGE

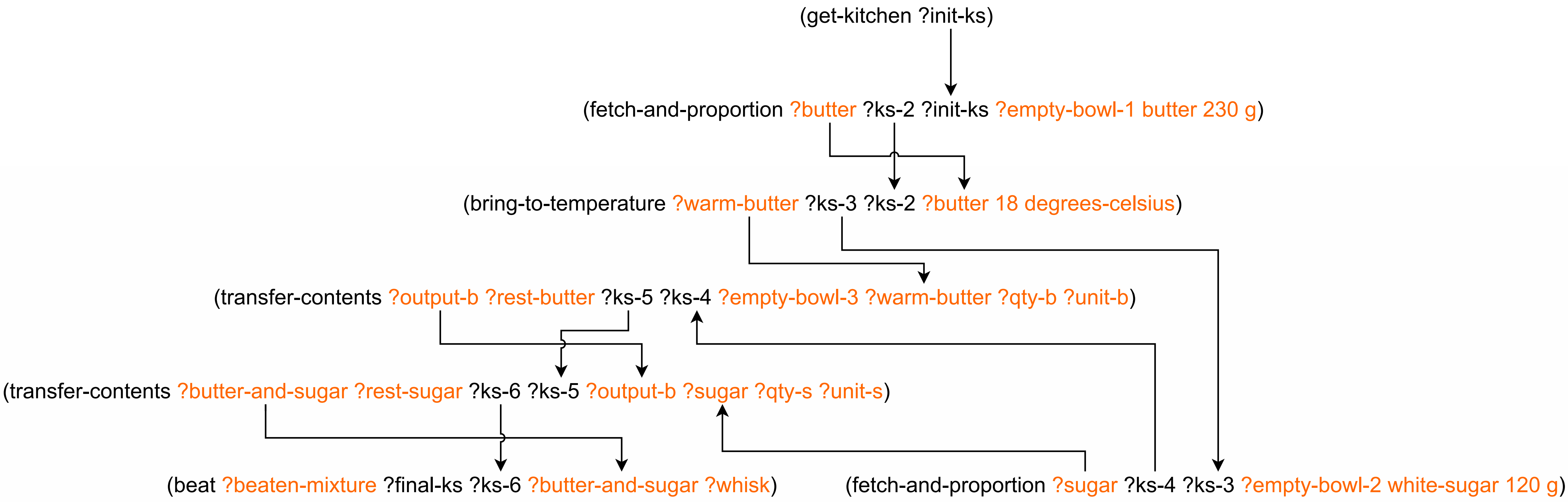
KITCHEN STATES



```
(get-kitchen ?init-ks)

(fetch-and-proportion ?butter ?ks-2 ?init-ks ?empty-bowl-1 butter 230 g)

(bring-to-temperature ?warm-butter ?ks-3 ?ks-2 ?butter 18 degrees-celsius)

(transfer-contents ?output-b ?rest-butter ?ks-5 ?ks-4 ?empty-bowl-3 ?warm-butter ?qty-b ?unit-b)

(transfer-contents ?butter-and-sugar ?rest-sugar ?ks-6 ?ks-5 ?output-b ?sugar ?qty-s ?unit-s)

(beat ?beaten-mixture ?final-ks ?ks-6 ?butter-and-sugar ?whisk)

(fetch-and-proportion ?sugar ?ks-4 ?ks-3 ?empty-bowl-2 white-sugar 120 g)
```

MUHAI COOKING LANGUAGE

FOOD, TOOLS & SPECIFIERS



(get-kitchen ?init-ks)

(fetch-and-proportion ?butter ?ks-2 ?init-ks ?empty-bowl-1 butter 230 g)

(bring-to-temperature ?warm-butter ?ks-3 ?ks-2 ?butter 18 degrees-celsius)

(transfer-contents ?output-b ?rest-butter ?ks-5 ?ks-4 ?empty-bowl-3 ?warm-butter ?qty-b ?unit-b)

(transfer-contents ?butter-and-sugar ?rest-sugar ?ks-6 ?ks-5 ?output-b ?sugar ?qty-s ?unit-s)

(beat ?beaten-mixture ?final-ks ?ks-6 ?butter-and-sugar ?whisk)

(fetch-and-proportion ?sugar ?ks-4 ?ks-3 ?empty-bowl-2 white-sugar 120 g)

MUHAI RECIPE EXECUTION BENCHMARK

DATASET

Data

- 30 recipes, from 5 different sources
- Salads & baked goods
- Input: XML
- Annotation: File with flat network

6 fully explained example solutions

```xml
<recipe>
    <id>easy-banana-bread</id>
    <title>Easy Banana Bread</title>
    <ingredients>
        <ingredient>
                60 grams butter, room temperature
        </ingredient>
        ...
    </ingredients>
    <instructions>
        <instruction>
            Cream together butter, eggs and sugar.
        </instruction>
        ...
    </instructions>
</recipe>
```

easy-banana-bread.xml

MUHAI RECIPE EXECUTION BENCHMARK

DATASET

Data

- 30 recipes, from 5 different sources
- Salads & baked goods
- Input: XML
- Annotation: File with flat network

6 fully explained example solutions

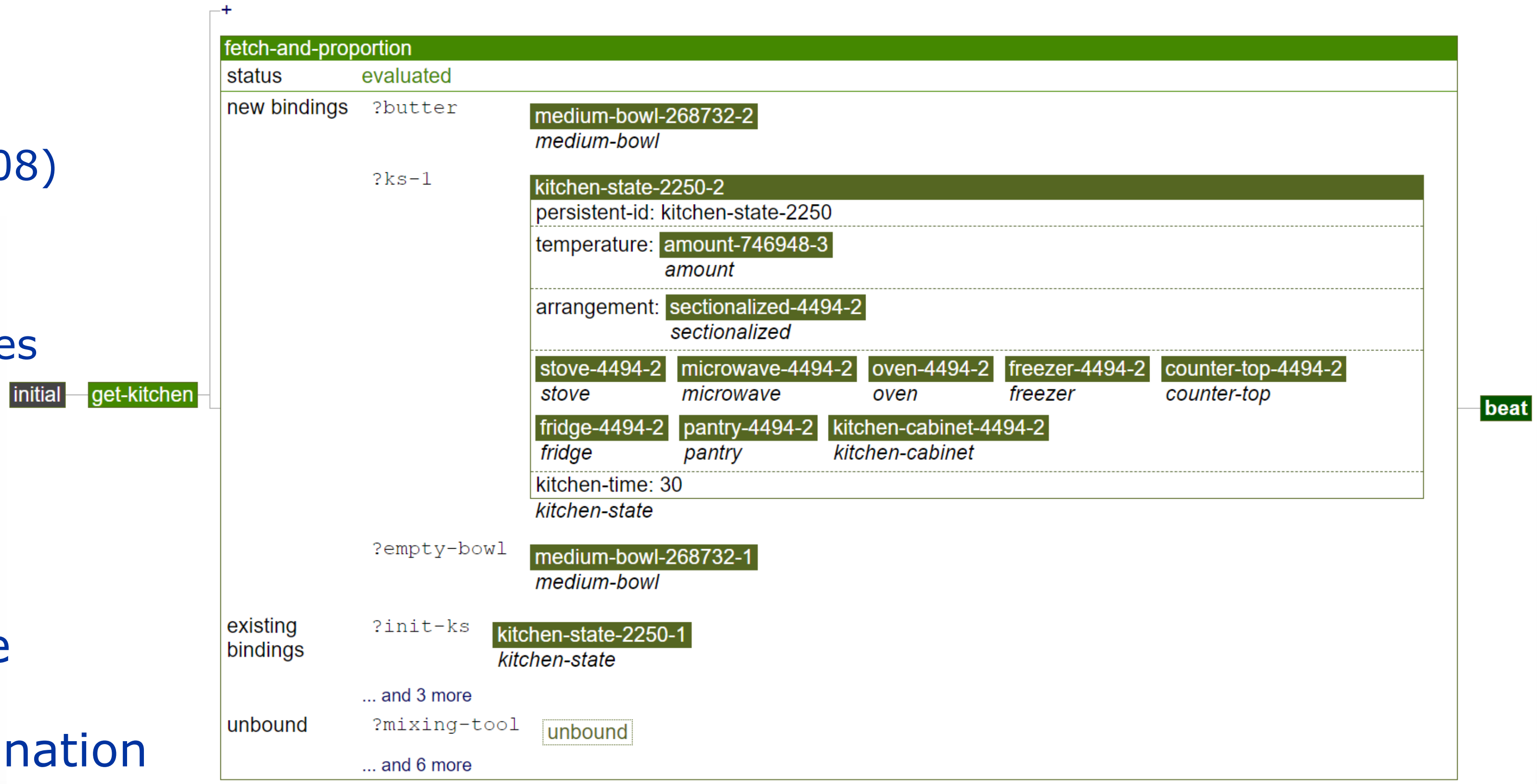
```
#easy-banana-bread
(get-kitchen ?init-ks)
(fetch-and-proportion ?butter ?ks-2
              ?init-ks ?empty-bowl-1
              butter 60 g)
(bring-to-temperature ?warm-butter ?ks-3
              ?ks-2 ?butter
              ?room-temp-qty ?room-temp-unit)
…
(beat ?creamed-mixture ?ks-5
              ?ks-4 ?output-e ?whisk)
…
(bake ?banana-bread ?final-ks
              ?ks-10 ?pan-with-batter ?oven
              60 minute 165 degrees-celsius)
```

easy-banana-bread.solution

MUHAI RECIPE EXECUTION BENCHMARK

SIMULATION-BASED EVALUATION

- Simulator
  - Babel: IRL (Loetzsch et al., 2008)
  - Portable
  - Qualitative simulation
  - Implementation for all primitives
  - Hierarchical ontology
  - Affordances
  - Temporality
- Visualization via web interface
- Multiperspective metric combination



initial → get-kitchen → fetch-and-proportion → beat

fetch-and-proportion

status evaluated

new bindings

?butter medium-bowl-268732-2 *medium-bowl*

?ks-1 kitchen-state-2250-2

persistent-id: kitchen-state-2250

temperature: amount-746948-3 *amount*

arrangement: sectionalized-4494-2 *sectionalized*

stove-4494-2 *stove* microwave-4494-2 *microwave* oven-4494-2 *oven* freezer-4494-2 *freezer* counter-top-4494-2 *counter-top*

fridge-4494-2 *fridge* pantry-4494-2 *pantry* kitchen-cabinet-4494-2 *kitchen-cabinet*

kitchen-time: 30

kitchen-state

?empty-bowl medium-bowl-268732-1 *medium-bowl*

existing bindings

?init-ks kitchen-state-2250-1 *kitchen-state*

... and 3 more

unbound

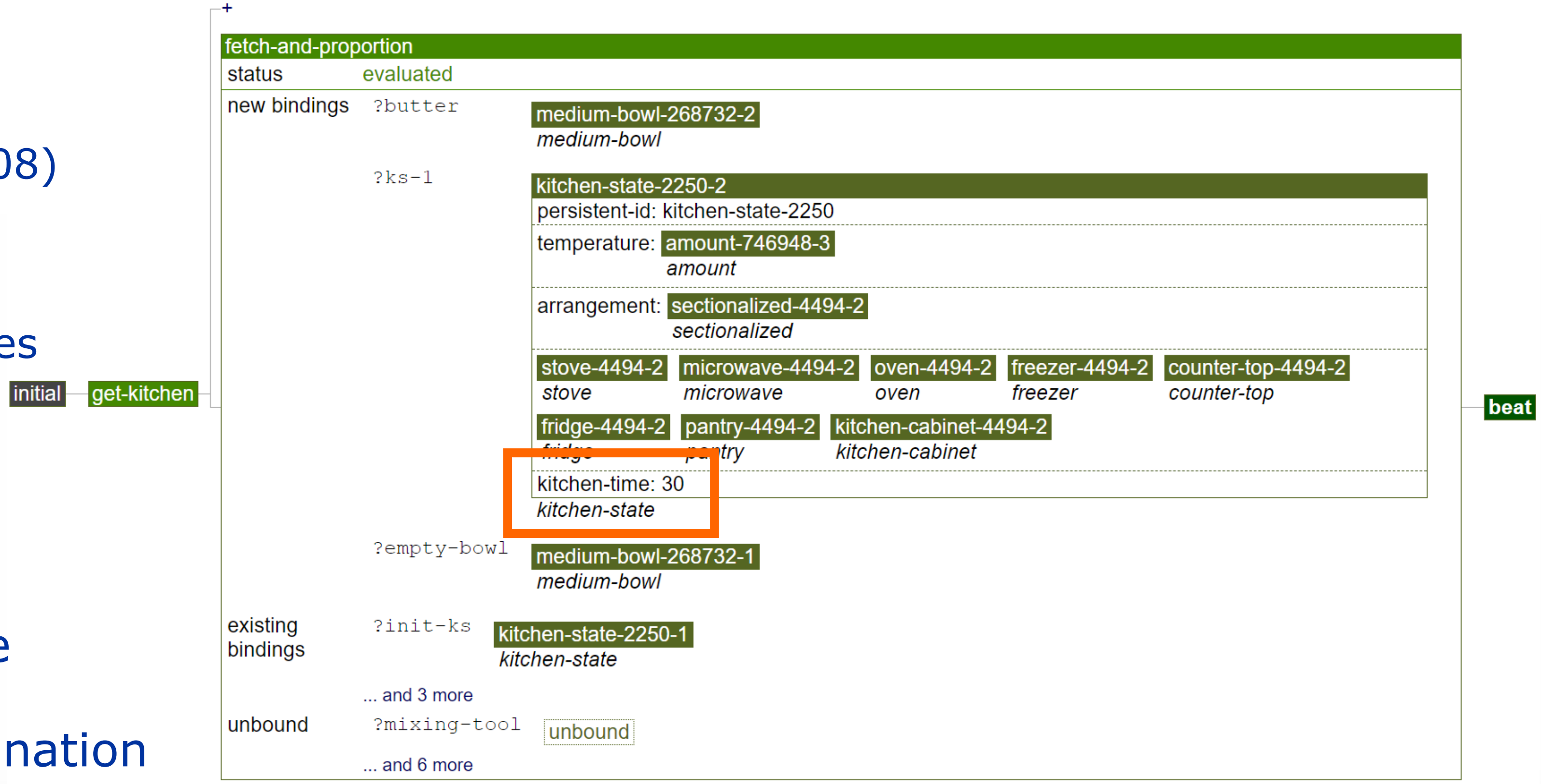
?mixing-tool unbound

... and 6 more

MUHAI RECIPE EXECUTION BENCHMARK

SIMULATION-BASED EVALUATION

- Simulator
  - Babel: IRL (Loetzsch et al., 2008)
  - Portable
  - Qualitative simulation
  - Implementation for all primitives
  - Hierarchical ontology
  - Affordances
  - Temporality
- Visualization via web interface
- Multiperspective metric combination



fetch-and-proportion

status evaluated

new bindings
?butter medium-bowl-268732-2 *medium-bowl*
?ks-1 kitchen-state-2250-2
persistent-id: kitchen-state-2250
temperature: amount-746948-3 *amount*
arrangement: sectionalized-4494-2 *sectionalized*
stove-4494-2 *stove* microwave-4494-2 *microwave* oven-4494-2 *oven* freezer-4494-2 *freezer* counter-top-4494-2 *counter-top*
fridge-4494-2 *fridge* pantry-4494-2 *pantry* kitchen-cabinet-4494-2 *kitchen-cabinet*
kitchen-time: 30
kitchen-state
?empty-bowl medium-bowl-268732-1 *medium-bowl*

existing bindings
?init-ks kitchen-state-2250-1 *kitchen-state*
... and 3 more

unbound
?mixing-tool unbound
... and 6 more

METRICS

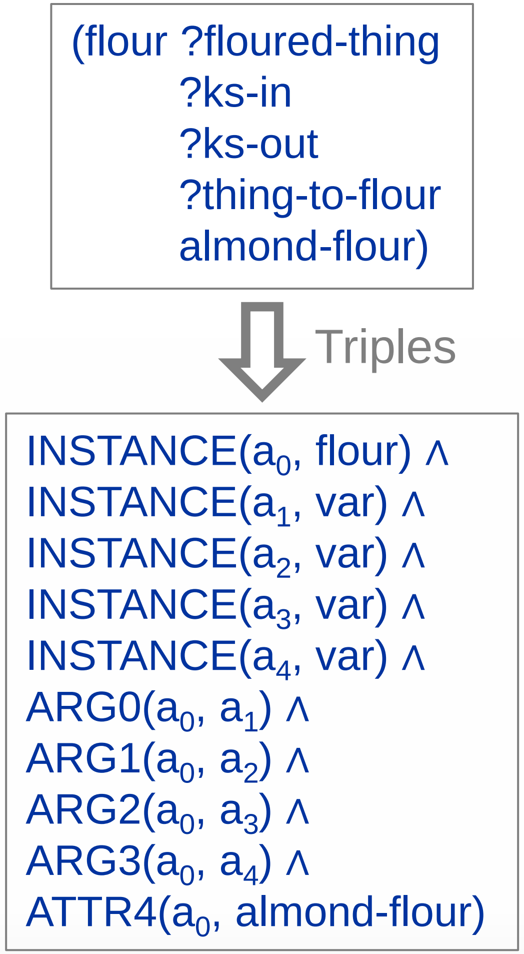
NON-SIMULATION- & SIMULATION-BASED

Non-simulation-based

- Smatch (Cai & Knight, 2013)
  - Semantic overlap of two semantic networks
  - Maximum F-score of matching triples
  - Good adoption rate, but resource-intensive and structural

Simulation-based

- Goal-condition Success
- Dish Approximation Score
- Recipe Execution Time



```
(flour ?floured-thing
       ?ks-in
       ?ks-out
       ?thing-to-flour
       almond-flour)
```

Triples

INSTANCE(a_0, flour) ∧
INSTANCE(a_1, var) ∧
INSTANCE(a_2, var) ∧
INSTANCE(a_3, var) ∧
INSTANCE(a_4, var) ∧
ARG0(a_0, a_1) ∧
ARG1(a_0, a_2) ∧
ARG2(a_0, a_3) ∧
ARG3(a_0, a_4) ∧
ATTR4(a_0, almond-flour)

SIMULATION-BASED METRICS

GOAL-CONDITION SUCCESS

= Ratio of reached goal-conditions to minimally required goal-conditions

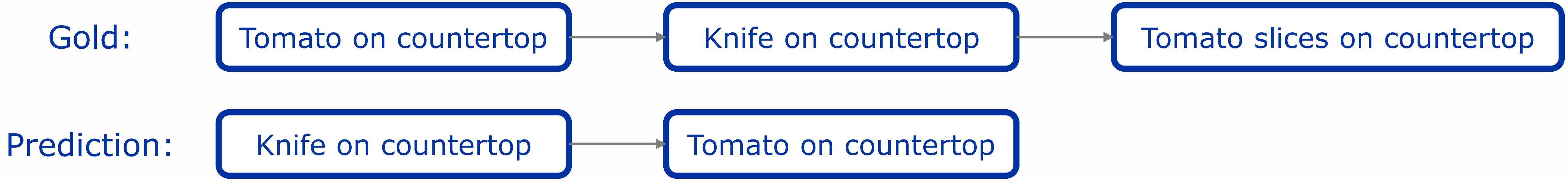
- Between 0 and 1
- Estimates the number of correctly identified steps
- Very strict
- Sequence-independent

SIMULATION-BASED METRICS

GOAL-CONDITION SUCCESS

= Ratio of reached goal-conditions to minimally required goal-conditions

“Cut the tomato with a knife.”



Gold: Tomato on countertop → Knife on countertop → Tomato slices on countertop

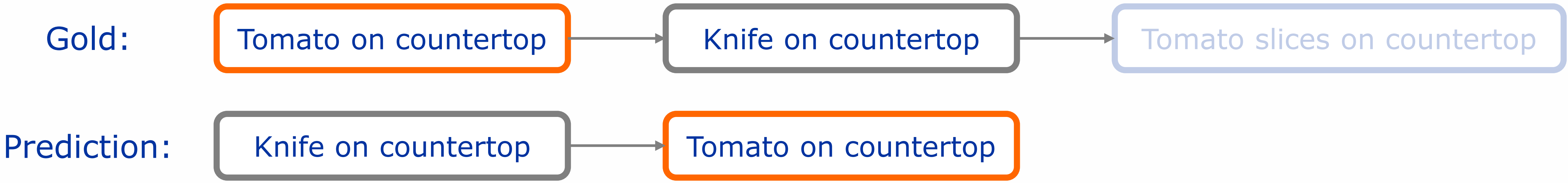
Prediction: Knife on countertop → Tomato on countertop

SIMULATION-BASED METRICS

GOAL-CONDITION SUCCESS

= Ratio of reached goal-conditions to minimally required goal-conditions

“Cut the tomato with a knife.”



Gold: Tomato on countertop → Knife on countertop → Tomato slices on countertop

Prediction: Knife on countertop → Tomato on countertop

⇒ Goal-condition success = 2/3

SIMULATION-BASED METRICS

DISH APPROXIMATION SCORE

= Similarity estimate between two dishes

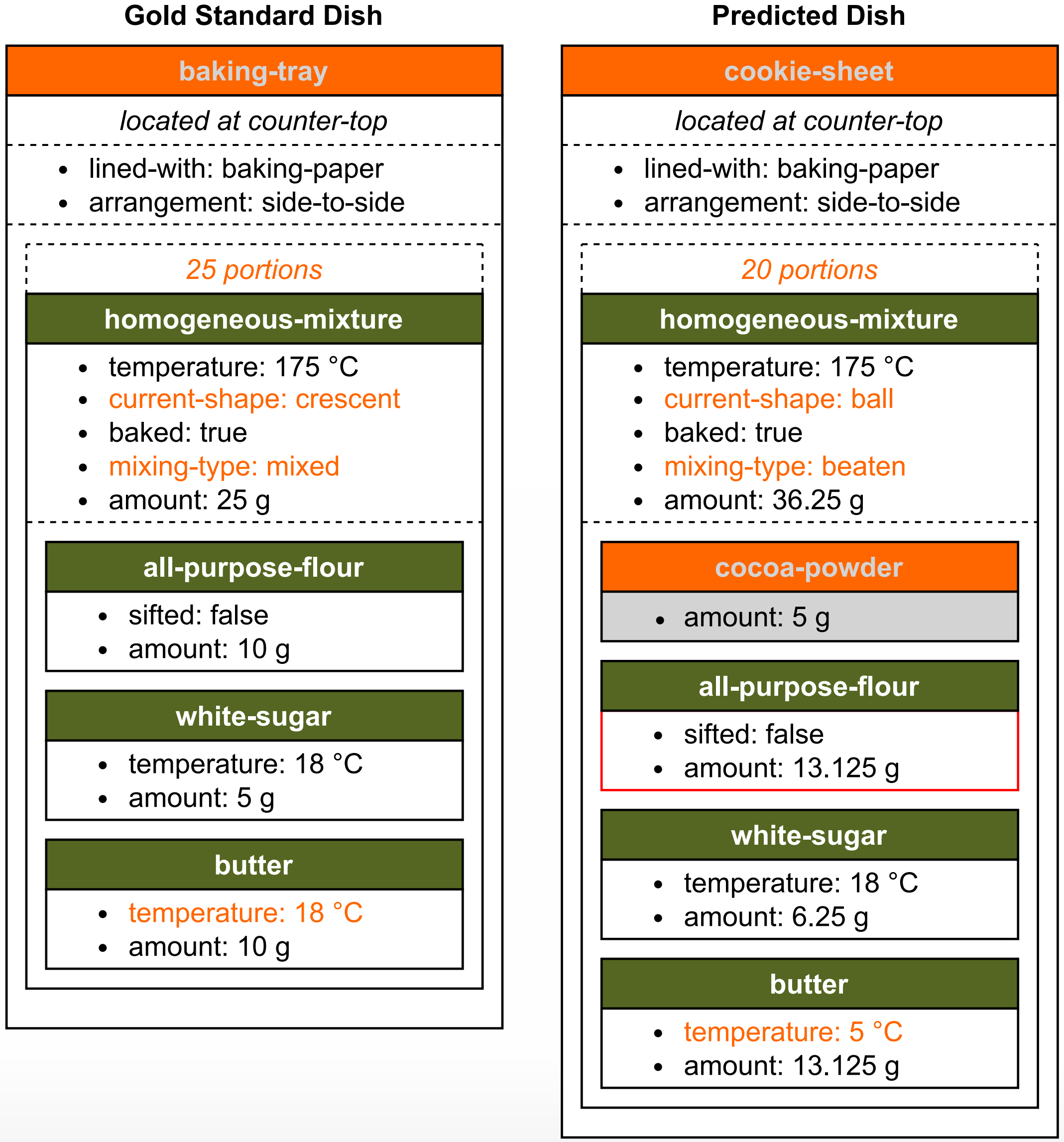
- Between 0 and 1
- Estimates how close to the expected dish we are
- “Taste test”: imperfect and subjective, but useful

SIMULATION-BASED METRICS

DISH APPROXIMATION SCORE

- Weighted sum
  - 2% presentation
  - 98% food composition

- Focus on
  - Correctness of base ingredients
  - Correctness of intermediary food products



Gold Standard Dish

baking-tray

located at counter-top

- lined-with: baking-paper
- arrangement: side-to-side

25 portions

homogeneous-mixture

- temperature: 175 °C
- current-shape: crescent
- baked: true
- mixing-type: mixed
- amount: 25 g

all-purpose-flour

- sifted: false
- amount: 10 g

white-sugar

- temperature: 18 °C
- amount: 5 g

butter

- temperature: 18 °C
- amount: 10 g

Predicted Dish

cookie-sheet

located at counter-top

- lined-with: baking-paper
- arrangement: side-to-side

20 portions

homogeneous-mixture

- temperature: 175 °C
- current-shape: ball
- baked: true
- mixing-type: beaten
- amount: 36.25 g

cocoa-powder

- amount: 5 g

all-purpose-flour

- sifted: false
- amount: 13.125 g

white-sugar

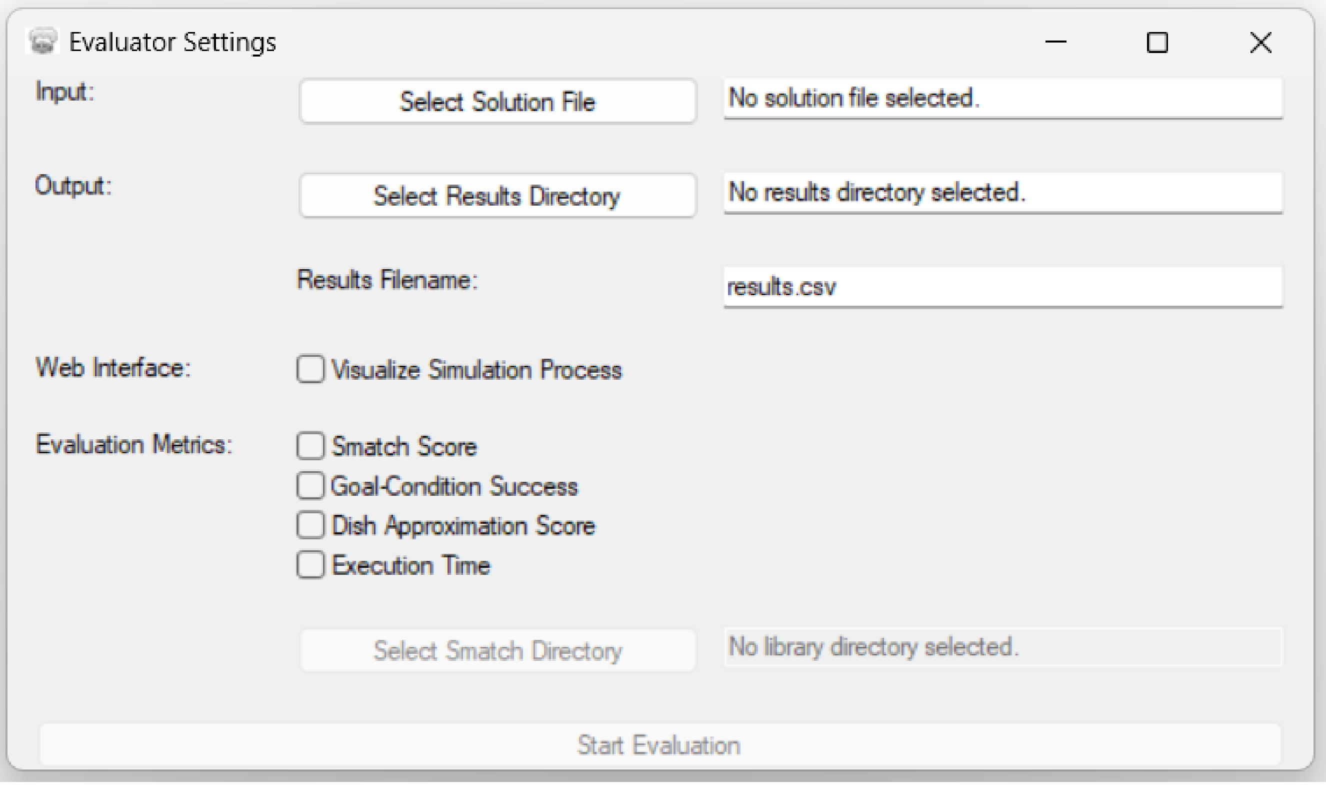
- temperature: 18 °C
- amount: 6.25 g

butter

- temperature: 5 °C
- amount: 13.125 g

EVALUATION TOOLS
IN PRACTICE

- Python Library Smatch
- Simulator Executable
  - Command-line Interface
  - Graphical User Interface
- Babel toolkit component
  - Open source
  - Extensible



No solution file selected.

No results directory selected.

results.csv

□ Visualize Simulation Process

□ Smatch Score
□ Goal-Condition Success
□ Dish Approximation Score
□ Execution Time

No library directory selected.

DISCUSSION & CONCLUSION

DISCUSSION

BENCHMARK PROPERTIES

	Pros	Cons
Relevance	• Domain-specific representation language • Separation of training and test data collection • Multiperspective, transferable evaluation	• Limited amount of test data • Results depend on simulation detail • Only Western, English recipes
Reproducibility	• Deterministic and consistent results	/
Verifiability	• Automatic evaluation tools	/
Fairness	• No promotion of specific approaches • Portable tools	• Test data is publicly available: Risk of cheating
Usability	• Portable and open source • Well-documented • Publicly available online	• Challenging task requiring substantial resource investments

CONCLUSION

SUMMARY

- Recipe benchmarks
  - Progress in recipe understanding is still needed
  - Steer research and foster field progression
  - Limitations and low adoption rates

- MUHAI Recipe Execution Benchmark
  - Reuse well-functioning concepts
  - Overcome limitations
  - Focus on evaluation that transfers well to the real world
  - BUT trade-offs are inevitable

CONCLUSION

FUTURE WORK

- Simulator improvements
  - University of Bremen
- Dataset extensions
  - Dish categories
  - Cultures
  - Languages
- Community adoption

REFERENCES

BIBLIOGRAPHY

[1] Cai, S., & Knight, K. (2013). Smatch: an Evaluation Metric for Semantic Feature Structures. In *Proceedings of the 51st Annual Meeting of the ACL* (pp. 748–752). ACL.

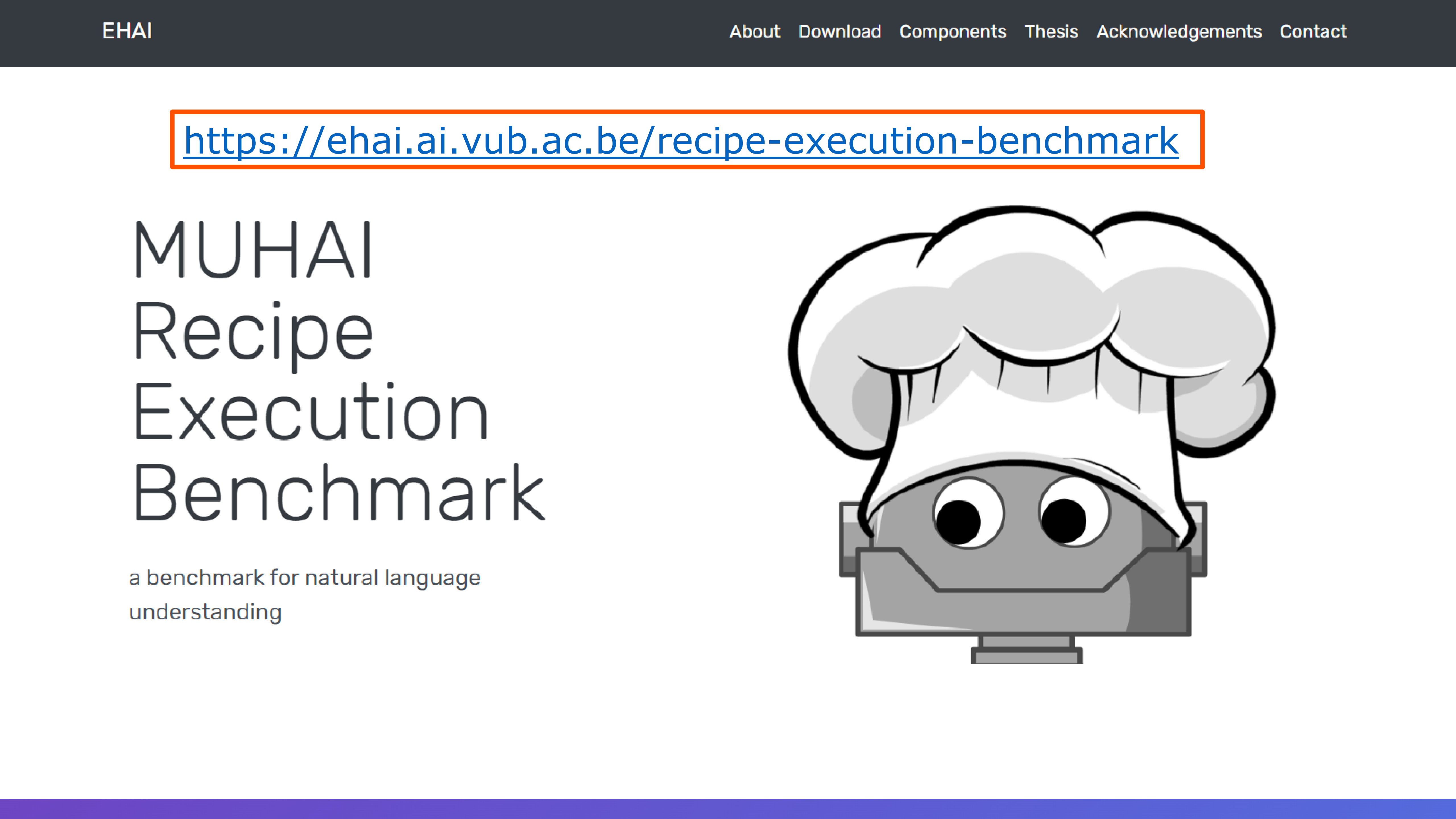
[2] von Kistowski, J., Arnold, J. A., Huppler, K., Lange, K.-D., Henning, J. L., & Cao, P. (2015). How to Build a Benchmark. In *Proceedings of the 6th ICPE* (pp. 333–336). ACM.

[3] Loetzsch, M., Wellens, P., De Beule, J., Bleys, J., & van Trijp, R. (2008). *The Babel2 Manual* (tech. rep. No. AI-Memo 01-08). AI-Lab VUB. Brussels, Belgium.

[4] Tasse, D., & Smith, N. A. (2008). *SOUR CREAM: Toward Semantic Processing of Recipes* (tech. rep. No. CMU-LTI-08-005). Carnegie Mellon University. Pittsburgh, PA, USA.

[5] Jiang, Y., Zaporojets, K., Deleu, J., Demeester, T., & Develder, C. (2020). Recipe instruction semantics corpus (RISeC). In *Proceedings of AACL-IJCNLP 2020* (pp. 821–826). ACL.

[6] Yamakata, Y., Mori, S., & Carroll, J. A. (2020). English Recipe Flow Graph Corpus. In *Proceedings of the 12th LREC* (pp. 5187–5194). ELRA.



EHAI

About Download Components Thesis Acknowledgements Contact

https://ehai.ai.vub.ac.be/recipe-execution-benchmark

MUHAI Recipe Execution Benchmark

a benchmark for natural language understanding

QUESTIONS?