Testing the FraCaS test suite

Robin Cooper
(joint work with Stergios Chatzikyriakidis and Simon Dobnik)
University of Gothenburg
{robin.cooper,stergios.chatzikyriakidis,simon.dobnik}@gu.se


Centre for Linguistic Theory and Studies in Probability,
http://clasp.gu.se/

Outline

The FraCaS project and test suite

Verifying the test suite

Implementing the test suite

Outline

The FraCaS project and test suite

Verifying the test suite

Implementing the test suite

FraCaS — A Framework for Computational Semantics

- European project (Language Research and Engineering)
  January, 1994 – March, 1996
- CWI Amsterdam Jan van Eijck, Jan Jaspars
  SRI Cambridge Richard Crouch, Stephen Pulman
  University of Edinburgh, Centre for Cognitive Science Robin Cooper, Massimo Poesio
  University of Saarbruecken, Computational Linguistics Manfred Pinkal, David Milward (April 95 – March 96), Espen Vestre (January 94 – September 94
  University of Stuttgart, IMS Josef van Genabith, Hans Kamp

From the description of the project

1. present an informal framework which allows comparison of current semantic approaches both with respect to their claims and their usefulness for implementation;
2. present the main semantic approaches in terms of this framework;
3. examine the feasibility of a general computational framework;
4. make preliminary investigations of the formal specifications for such a framework;
5. apply the framework to a representative fragment of real-life language;
6. draw together the results of consultation with a representative base of researchers in the field.

The FraCaS theories

- Discourse Representation Theory
- Dynamic and Update Semantics
- Situation Theory
- Property Theory
- Monotonic Semantics (underspecification as in Quasi Logical Form, Core Language Engine)

The FraCaS test suite

- 346 inference problems
- *Using the Framework*
  The FraCaS Consortium
  Cooper et al.
  January 1996
  ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz
- Coded in xml by Bill McCartney (around 2008)

An example – answer: Yes

- Original:

  An Irishman won the Nobel prize for literature

  Did an Irishman win a Nobel prize?

  [Yes]

- MacCartney:

  fracas-017 answer: yes

  P1 An Irishman won the Nobel prize for literature.

  Q Did an Irishman win a Nobel prize?

  H An Irishman won a Nobel prize.

An example – answer: No

- Original:

  No delegate finished the report

  Did any delegate finish the report on time?

  [No]

- MacCartney:

  fracas-038 answer: no

  P1 No delegate finished the report.

  Q Did any delegate finish the report on time?

  H Some delegate finished the report on time.

An example – answer: Don't know/unknown

- Original:
  An Irishman won a Nobel prize

  Did an Irishman win the Nobel prize for literature?
  [Don't know]
- MacCartney:
  fracas-033 answer: unknown
  - P1 An Irishman won a Nobel prize.
  - Q Did an Irishman win the Nobel prize for literature?
  - H An Irishman win the Nobel prize for literature.

An example – answer: [other]/undef

- Original:

  At most two tenors will contribute their fees to charity.

  Are there tenors who will contribute their fees to charity?

  [At most two]

- MacCartney:

  fracas-016 answer: undef

  P1 At most two tenors will contribute their fees to charity.

  Q Are there tenors who will contribute their fees to charity?

  H There are tenors who will contribute their fees to charity.

  A At most two

Another example – answer: [other]/undef

- Original:
  Both female commissioners used to be in business.

  Did both commissioners used to be in business?
  [Yes, if both commissioners are female. Otherwise there are more than two commissioners]

- MacCartney:
  fracas-061 answer: undef
  - P1 Both female commissioners used to be in business.
  - Q Did both commissioners used to be in business?
  - H Both commissioners used to be in business.
  - A Yes, if both commissioners are female; otherwise there are more than two commissioners.

A third example – answer: [other]/undef

- Original:

  All mice are small animals.

  Mickey is a large mouse.

  Is Mickey small?

  [??: Yes for a mouse ??: No for an animal]

  Adjectives requiring a comparison class cannot usually be predicated in the absence of a common noun, unless some comparison class is clear from the wider context.

- MacCartney:

  fracas-213 answer: undef

P1	All mice are small animals.
P2	Mickey is a large mouse.
Q	Is Mickey small?
H	Mickey is small.
A	??: Yes for a mouse; ?? No for an animal
Why	*Adjectives . . .*

MultiFraCaS

- Translation of FraCaS test suite into
  - Farsi
  - German
  - Greek (updated 7th October, 2016)
  - Mandarin
- Surprisingly few problems in reproducing the original FraCaS judgements

Outline

Checking the inferences with crowd-sourcing

- Semant-o-matic: online testing environment for checking speaker judgements
  - dialogue data, two online participants Dobnik (2012)
  - dialogue and percpetual data Dobnik *et al.* (2014)
- Targeted to particular groups or shared on social media
- Each participant must give a valid email address: prevents random participation (signup form)
- Can include a test example
- Examples presented in random order: (example)

Simon Dobnik 

Pilot study: 15 FraCas examples

5 “interesting” examples of each class (y, n, undef)

- Generalized quantifiers - conservativity: fracas-02/yes, fracas-06/no, fracas-012/undef
- Generalized quantifiers - monotonicity: fracas-17/yes, fracas-38/no, fracas-74/undef
- Adjectives: fracas-199/yes, fracas-211/no, fracas-201/undef
- More adjectives: fracas-212/yes, fracas-223/no (comparatives), fracas-206/undef
- Other: fracas-103/yes (plurals), fracas-163/no (sluicing), fracas-308/undef (temporal reference)

Instructing an "average person" to identify entailment

- "Your task is to answer this question based **only** on the information provided in the sentences. For example, if the information is contained in the sentences one can answer the question with "yes" or "no", otherwise the answer is "don't know". . . . In answering the question you must not rely on any other knowledge that is not stated in the sentences."

Instructing an “average person” to identify entailment

- “Your task is to answer this question based **only** on the information provided in the sentences. For example, if the information is contained in the sentences one can answer the question with ”yes” or ”no”, otherwise the answer is ”don’t know”. . . . In answering the question you must not rely on any other knowledge that is not stated in the sentences.”
- P1 If Kim comes to the party, Sandy will be upset
  P2 Kim is coming to the party
  and the question
  Q Will Sandy be upset?

Simon Dobnik 

Some preliminary results I

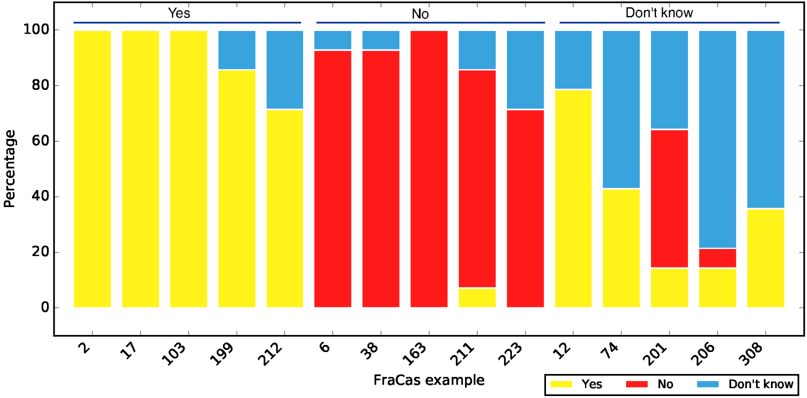
- Study 1: familiarity with the task
  - Master students in language technology taking a Computational semantics course at GU
  - Following a lecture on inference
  - Closed group, $n = 14$
  - $\approx 15\%$ English native speakers, others with excellent knowledge of English
- Study 2: non-familiarity with the task
  - Employees of Gothenburg university: academic (various disciplines) and non-academic staff and their friends
  - No previous introduction to the task
  - Open group, $n = 46$
  - Mixture of native speakers with Swedish dominant but with excellent English

Some preliminary results II

- Study 3 and Study 4: Greek (MultiFraCaS) and Slovenian
  - "Random" friends on social media
  - No previous introduction to the task
  - Open group, Greek $n = 57$ and Slovenian $n = 29$
  - Native speakers (?)

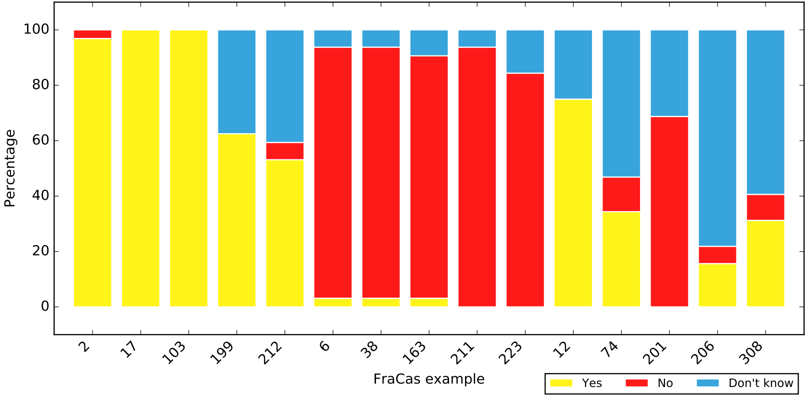
Simon Dobnik 

Study #1: familiarity with task



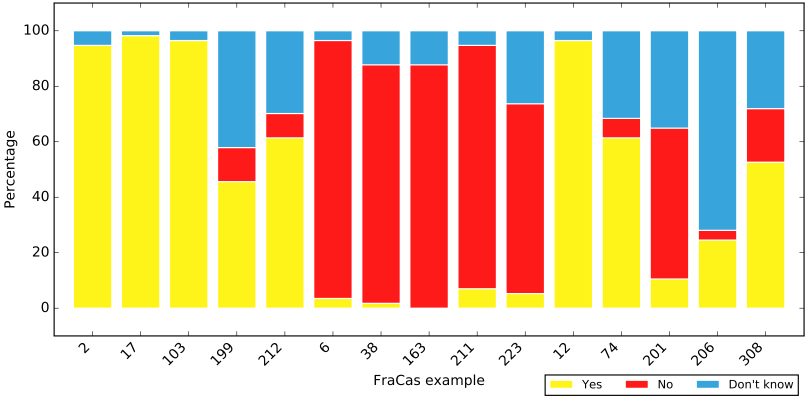
Simon Dobnik

Study #2: unfamiliarity with task

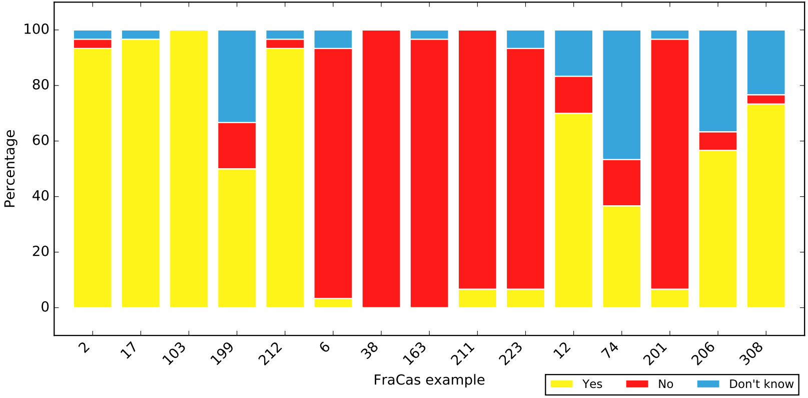


Simon Dobnik

Study #3a: Greek, unfamiliarity with task



Study #3b: Slovenian, unfamiliarity with task



Simon Dobnik

Examples tested

Yes	2, 17, 103, 199, 212
No	6, 38, 163, 211, 223
Don't know	12, 74, 201, 206, 308

Towards probabilistic inference

fracas-002 answer: yes

P1 Every Italian man wants to be a great tenor.

P2 Some Italian men are great tenors.

Q Are there Italian men who want to be a great tenor?

H There are Italian men who want to be a great tenor.

Note Note that second premise is unnecessary and irrelevant.

Examples tested

fracas-017 answer: yes

P1 An Irishman won the Nobel prize for literature.

Q Did an Irishman win a Nobel prize?

H An Irishman won a Nobel prize.

Examples tested

fracas-103 answer: yes

P1 All APCOM managers have company cars.
P2 Jones is an APCOM manager.
Q Does Jones have a company car?
H Jones has a company car.

Examples tested

fracas-199 answer: yes **

P1	John is a successful former university student.
Q	Is John successful?
H	John is successful.
A	Yes (for a former university student)
Why	*Ordering between affirmative and non-affirmative adjectives affects which adjectival predications are and aren't affirmed*

Examples tested

fracas-212 answer: yes

- P1 All mice are small animals.
- P2 All elephants are large animals.
- P3 Mickey is a large mouse.
- P4 Dumbo is a small elephant.
- Q Is Dumbo larger than Mickey?
- H Dumbo is larger than Mickey.
- Why *Assume comparative relations exemplified in (208) and (209)*

Examples tested

fracas-006 answer: no

P1 No really great tenors are modest.

Q Are there really great tenors who are modest?

H There are really great tenors who are modest.

Examples tested

fracas-038 answer: no

P1 No delegate finished the report.

Q Did any delegate finish the report on time?

H Some delegate finished the report on time.

Examples tested

fracas-163 answer: no

P1 John had his paper accepted.
P2 Bill doesn't know why.
Q Does Bill know why John had his paper accepted?
H Bill knows why John had his paper accepted.

Examples tested

fracas-211 answer: no

P1 All elephants are large animals.

P2 Dumbo is a small elephant.

Q Is Dumbo a small animal?

H Dumbo is a small animal.

Examples tested

fracas-223 answer: no

P1 The PC-6082 is faster than the ITEL-XZ.

P2 The PC-6082 is slow.

Q Is the ITEL-XZ fast?

H The ITEL-XZ is fast.

Examples tested

fracas-012 answer: undef **

P1 Few great tenors are poor.

Q Are there great tenors who are poor?

H There are great tenors who are poor.

A Not many

Examples tested

fracas-074 answer: unknown

P1 Most Europeans can travel freely within Europe.

Q Can most Europeans who are resident outside Europe travel freely within Europe?

H Most Europeans who are resident outside Europe can travel freely within Europe.

Examples tested

fracas-201 answer: unknown

P1	John is a former successful university student.
Q	Is John a university student?
H	John is a university student.
Why	*John may currently be an unsuccessful university student*

Examples tested

fracas-206 answer: unknown

P1 Fido is not a small animal.

Q Is Fido a large animal?

H Fido is a large animal.

Why $!Small(N) \not\Rightarrow Large(N)$

Examples tested

fracas-308 answer: undef **

P1 Smith wrote to a representative every week.

Q Is there a representative that Smith wrote to every week?

H There is a representative that Smith wrote to every week.

A Yes on one scoping; unknown on another scoping

Examples tested

Towards probabilistic inference

- If A then B with probability n
- conditional probability – $\mathrm{p}(B \mid A) = n$

Relating data from multiple subjects to individual judgements

- Not obvious that there is a relation
- Possible that each individual making a categorical judgement
- But a spread across the population
- Could test by providing a task where subjects respond by clicking on a scale.
- *cf.* work on syntactic judgements by Lappin *et al.* (Lau *et al.*, 2016)

The approach to subjective probability estimation in probabilistic TTR

Cooper *et al.* (2015) propose a probabilistic re-formulation of Cooper's (2012) Type Theory with Records (TTR)

Computing the Probability of a Type Judgement

- When an agent A encounters a new situation s and wants to know if it is of type T or not, he/she uses probabilistic reasoning to determine the value of $p_{A,\mathfrak{J}}(s : T)$.
- This denotes the probability that agent A assigns, on the basis of prior judgements $\mathfrak{J}$, to the judgement that s is of type T.

(Slide from Shalom Lappin)

Summing Probabilities of Type Judgements I

- An agent makes judgements based on a finite string of probabilistic Austinian propositions, $\mathfrak{J}$.
- For a type, T, $\mathfrak{J}_T$ represents that set of probabilistic Austinian propositions j such that $j.\text{sit-type} \sqsubseteq T$.

$$\mathfrak{J}_T = \{j \mid j \in \mathfrak{J}, j.\text{sit-type} \sqsubseteq T\}$$

- If T is a type and $\mathfrak{J}$ a finite string of probabilistic Austinian propositions, then $||T||_{\mathfrak{J}}$ represents the sum of all probabilities associated with T in $\mathfrak{J}$

$$||T||_{\mathfrak{J}} = \sum_{j \in \mathfrak{J}_T} j.prob$$

Summing Probabilities of Type Judgements II

- $\mathcal{P}(\mathfrak{J})$ is the sum of all probabilities in $\mathfrak{J}$

$$\mathcal{P}(\mathfrak{J}) = \sum_{j \in \mathfrak{J}} j.prob$$

(Slide from Shalom Lappin)

Priors on Type Judgements

- $\text{prior}_{\mathfrak{J}}(T)$ represents the prior probability that anything is of type T given $\mathfrak{J}$.

$$\text{prior}_{\mathfrak{J}}(T) = \frac{\| T \|_{\mathfrak{J}}}{\mathcal{P}(\mathfrak{J})} \text{ if } \mathcal{P}(\mathfrak{J}) > 0, \text{ and } 0 \text{ otherwise.}$$

(Slide from Shalom Lappin)

Outline

The FraCaS project and test suite

Verifying the test suite

Implementing the test suite

Using type theory for natural language inference

- Work initiated by Chatzikyriakidis and Luo (2014) using Luo's Type Theory with Coercive Subtyping (an MTT)
  - Though the first to use proof assistants for NL (dialogue systems) are Ranta and Cooper (2004) (using the proof-editor ALF)
- Evaluation is done on 25% of the suite in Chatzikyriakidis and Luo (2014) and almost 30% in Chatzikyriakidis (2015)
- The systems proposed in these approaches do not define an automatic translation between GF parses and Coq outputs
  - This is done manually
  - In progress: a GF translator from English to Coq (approaching the problem as a multilingual translstion problem)

Using type theory for NLI — Coverage

- Examples from the following sections of the FraCaS are used (Chatzikyriakidis (2015)):
  - Quantifiers and monotonicity (41 examples)
  - Conjoined noun phrases (15 examples).
  - Adjectives (18 examples).
  - Dependent plurals (2 examples)
  - Comparatives (10).
  - Epistemic, intensional and reportive attitudes (11 examples).
  - Collective predication (6 examples).
  - Quantificational adverbs (2 examples)

Using type theory for NLI — Results

- Precision is extremely high
  - Some indicative cases:
    - Quantifiers: 100%
    - Plurals: 100%
    - Adjectives: 81.8%
    - Attitudes: 100%
  - A proper measure of recall is pending given that there is no automatic translation yet
  - Automated tactics are defined that can automate the whole proofs

Using type theory for NLI-Proof assistants and NLI

- An example proof
  Some Irish delegates finished the survey on time
  Did any delegates finish the survey on time? [YES, FraCaS 0.55]
- We formulate the theorem:

```
Theorem IRISH:
(some Irishdelegate(on_time(finish(the survey)))->
(some Delegate)(on_time(finish(the survey))).

compute.intro. elim H.intro.intro.exists x.auto.
```


Using type theory for NLI — coding in Coq

- Standard semantics for indefinites, VP adverbs as polymorphic functions from predicates to predicates, modified common nouns as Sigma types (encoded as Dependent Record Types in Coq) with the first projection as a coercion:

```
Definition some:=
fun A:CN, fun P:A->Prop=> exists x:A, P(x).
Parameter on_time: forall A:CN,(A->Prop)->(A->Prop).
Record Irishdelegate:CN:=
mkIrishdelegate{c:>Delegate;c1:Irish c}.
```


Using type theory for NLI — proof in Coq

- Easy to prove. Subtyping does the work. Eliminating the hypothesis *H* and using *intro* we get an x : *Irishdelegate* such that *on_time*(*finish*(*thesurvey*))(x)) holds:

```
H : exists x : Irishdelegate,
on_time Human (finish (the survey)) (let (c, _)
:= x in c)
x : Irishdelegate
H0 : on_time Human (finish (the survey)) (let (c, _)
:= x in c)
============================
exists x0 : delegate, on_time Human (finish
(the survey)) x0
```

Using type theory for NLI — proof in Coq *contd.*

- Then, given subtyping, *Irishdelegate* < *Delegate* via the first projection $\pi 1$, we also have that *on_time*(*finish*(*the survey*))(x)) with x : *Delegate*. We thus substitute x0 with x

```
H : exists x : Irishdelegate,
on_time Human (finish (the survey)) (let (c, _)
:= x in c)
x : Irishdelegate
H0 : on_time Human (finish (the survey)) (let (c, _)
:= x in c)
============================
on_time Human (finish (the survey)) x
```

- *Assumption* (checks whether the goal is matched by a premise) completes the proof

Comparison with Mineshima *et al.* (2015)

Mineshima et al.	**Chatzikyriakidis et al.**
well defined translation procedure between the syntax of the CCG parser and the output to Coq structures	no proper translation between the parser and the semantics
higher order logic version of the event semantics that is output by the CCG parser	uses Luo's type theory (with Σ-types,Π-types, records)
interactive theorem provers can be automated for FraCaS	interactive theorem provers can be automated for FraCaS
need to increase use of type theory (Σ-types etc.)	need automatic feed from parser (e.g. GF) to Coq

And then there's Mineshima *et al.* (2016)...

- wide coverage inference
- higher order logic

Conclusions

- FraCaS style test suites are an important resource
- . . . though they need to be used with care
- . . . and grounded in real data
- Dealing with ambiguity: systems should provide at least one reading on which the inference goes through
- Probabilistic reasoning needs to be investigated
  - is the reasoning itself probabilistic (using conditional probabilities)?
  - does it rely on the probability of certain readings of the premises?
- This seems a particularly clear area where data collection without a theory and theory development without real data can get us into trouble.

Bibliography I

Chatzikyriakidis (2015) Natural Language Reasoning using Coq: Interaction and Automation, in *Proc. of TALN*, Caen, France.

Chatzikyriakidis, S. and Z. Luo (2014) Natural Language Inference in Coq, *J. of Logic, Language and Information.*, Vol. 23(4), pp. 441–480.

Cooper, Robin (2012) Type Theory and Semantics in Flux, in R. Kempson, N. Asher and T. Fernando (eds.), *Handbook of the Philosophy of Science*, Vol. 14: Philosophy of Linguistics, pp. 271–323, Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Bibliography II

Cooper, Robin, Simon Dobnik, Shalom Lappin and Staffan Larsson (2015) Probabilistic Type Theory and Natural Language Semantics, *Linguistic Issues in Language Technology*, Vol. 10, No. 4, pp. 1–45.

Dobnik, Simon (2012) Spatial descriptions in discourse: choosing a perspective, in S. L. Sarah Brown-Schmidt, Jonathan Ginzburg (ed.), *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 151–152.

Dobnik, Simon, John D. Kelleher and Christos Koniaris (2014) Priming and Alignment of Frame of Reference in Situated Conversation, in V. Rieser and P. Muller (eds.), *Proceedings of DialWatt - Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 43–52, Edinburgh.

Bibliography III

Lau, Jey Han, Alexander Clark and Shalom Lappin (2016) Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge, *Cognitive Science*, Vol. n/a, No. n/a, pp. n/a–n/a.

Mineshima, Koji, Pascual Martınez-Gómez, Yusuke Miyao and Daisuke Bekki (2015) Higher-order logical inference with compositional semantics.

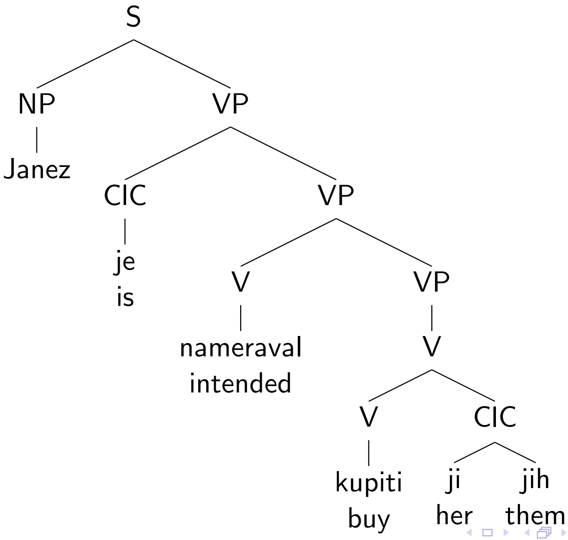
Mineshima, Koji, Ribeka Tanaka, Pascual Martínez-Gómez, Yusuke Miyao and Daisuke Bekki (2016) Building compositional semantics and higher-order inference system for a wide-coverage Japanese CCG parser, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2236–2242, Association for Computational Linguistics.

Bibliography IV

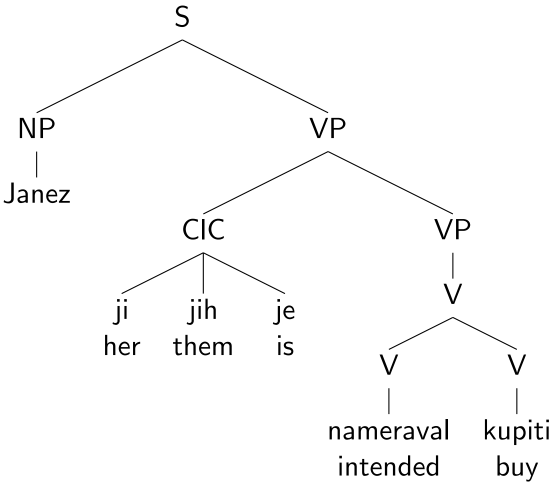
Ranta, Aarne and Robin Cooper (2004) Dialogue systems as proof editors, *Journal of Logic, Language and Information*, Vol. 13, No. 2, pp. 225–240.

Wurmbrand, Susanne (2001) *Infinitives: Restructuring and clause structure*, Studies in Generative Grammar 55, Mouton de Gruyter.

Restructuring in Slovenian — without restructuring

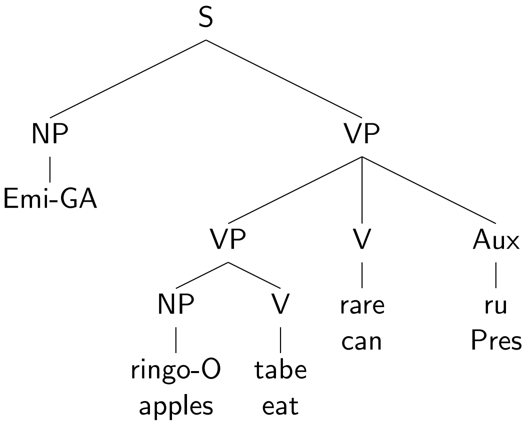


Restructuring in Slovenian — with restructuring

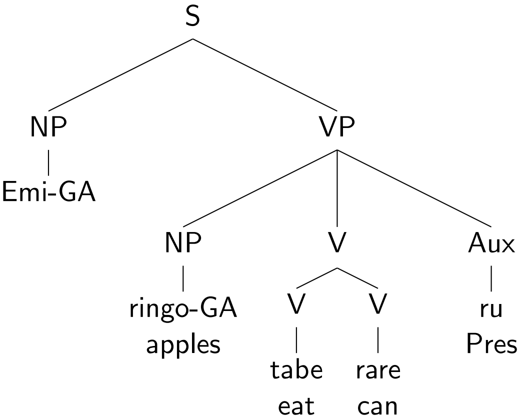


Restructuring in Japanese – without restructuring

cf. Wurmbrand (2001)



Restructuring in Japanese – with restructuring



Data

- Fracas test-cases (2 examples)
- Non-finite constructions (2 examples):
  - Long passive
  - Clitic climbing
- Intend matrix verb: borderline case
- Imperfective or perfective complement verb (comp-v) $(\ldots \times 2)$
- Restructuring and non-restructuring variants $(\ldots \times 2)$
- Linguistic priming inducing a context of two sepearate events (2e) or one event (e) $(\ldots \times 2 = 16)$
- 18 items presented in random order
- Judgements from 1 (bad) to 5 (good).

Examples with clitic climbing

(1) Context inducing two event interpretation:
Ana has a large pile of laundry in her bathroom. Jonas is always eager to help Ana with the laundry.

- ⇒ R: Jonas it intends to wash-impf.
- ⇒ NR: Jonas intends to wash-impf it.

(2) Context inducing one event interpretation:
Ana has a large pile of laundry in her bathroom. Jonas is always eager to help Ana with the laundry. He decided to help her this time also.

- ⇒ R: Jonas it intends to wash-impf.
- ⇒ NR: Jonas intends to wash-impf it.

Results: $n = 33$ where $n_{trustworthy} = 19$

Construction	comp-v	Priming	Type	Result (μ, σ)	H cfg	t-test (p)
fracas-120				4.95, 0.23		
fracas-125				1.05, 0.23	yes	1.45×10^{-35} sig
l-passive	imperf	2e	R	2.32, 1.20		
l-passive	imperf	2e	NR	1.95, 1,18	no	0.3467 nsig
l-passive	imperf	e	R	2.58, 1.68		
l-passive	imperf	e	NR	2.37, 1.50	yes	0.6857 nsig
l-passive	perf	2e	R	2.74, 1.48		
l-passive	perf	2e	NR	3.05, 1.55	yes	0.5247 nsig
l-passive	perf	e	R	3.37, 1.46		
l-passive	perf	e	NR	3.37, 1.57	no	1.0 nsig
c-climbing	imperf	2e	R	3.74, 1.59		
c-climbing	imperf	2e	NR	2.74, 1.33	no	0.0426 sig
c-climbing	imperf	e	R	3.74, 1.37		
c-climbing	imperf	e	NR	2.68, 1.49	yes	0.0296 sig
c-climbing	perf	2e	R	3.11, 0.88		
c-climbing	perf	2e	NR	3.11, 0.32	no	1.0 nsig
c-climbing	perf	e	R	3.90, 1.29		
c-climbing	perf	e	NR	2.84, 1.61	yes	0.0322 sig

We have learned. . . I

- All judgements for long passive are non-significant: there is no differences in scores between R and NR configurations given different priming.
- There are significant differences in clitic climbing:
  - With imperfective v-comp climbing (R) is preferred after e priming
  - . . . and so is after 2e priming contrary to our expectations.
  - With perfective v-comp climbing (R) is preferred after e priming.
  - . . . but equally acceptable as non-climbing (NR) after 2e priming.
- R constructions are more preferred than NR constructions.
- NR configurations are not completely dispreferred although they are rare in corpus (compete with that-tensed-clause).

We have learned. . . II

- 2e priming does not enforce 2e interpretation but keeps it open/vague.
- The priming for e interpretation focuses on the commitment to *e* and is weak.
- Future work: stronger contextual priming
- Comparison with FraCas allows us to compare stronger judgements with weaker borderline cases.