Detecting Linguistic Function in HPSG Grammars

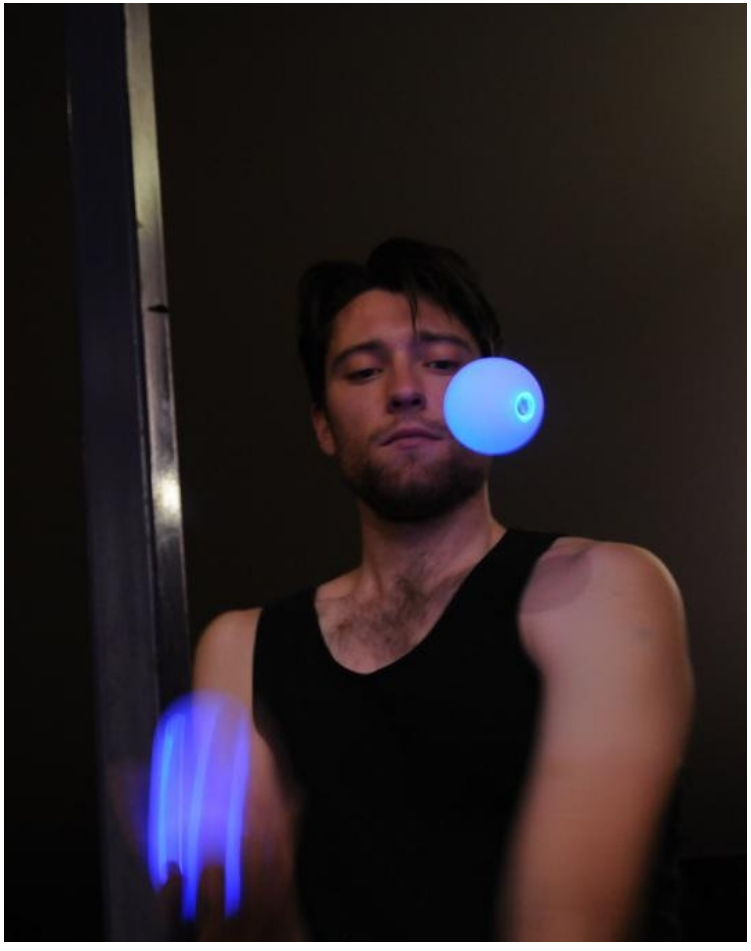
Ned Letcher

Supervisors:
Tim Baldwin
Emily Bender

THE UNIVERSITY OF
MELBOURNE

Hello (DELPH-IN) world!

(a) Coffee



(b) Juggling

Figure: Things Ned likes

Motivation

Precision grammars are valuable linguistic resources

Two specific uses:

1. Leveraging existing implementations for inspiration
2. Grammar engineering for linguistic documentation

But linguistic function is relatively opaque:

- Grammars don't wear their linguistic phenomena on their sleeve
- Phenomena usually implemented using multiple types
- Types often contain constraints on multiple phenomena
- TDL is hard to read!

Discoverability

Precision grammars contain tried and tested solutions

But...

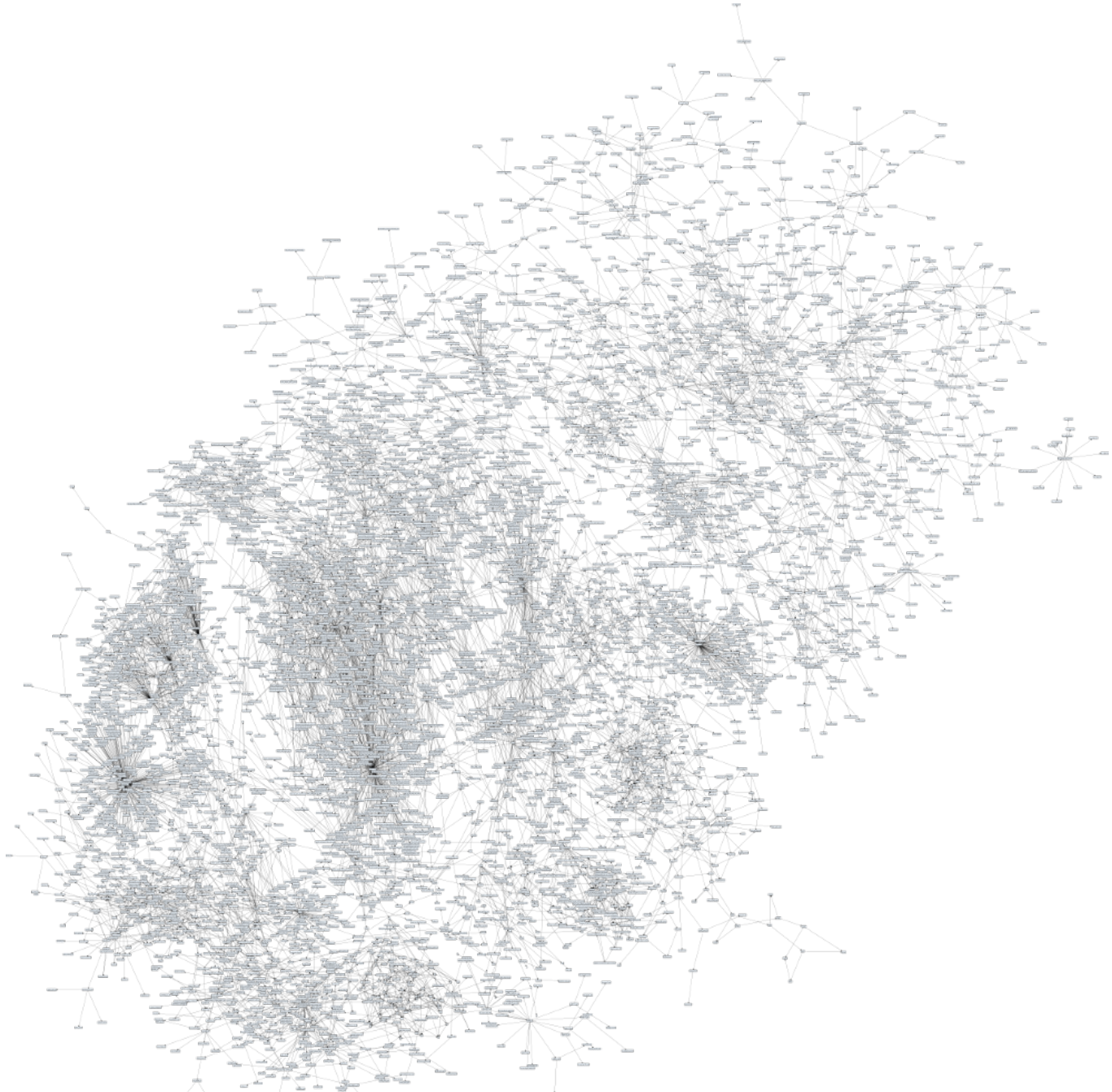
Discoverability of phenomena is poor:

- Don't know which phenomena grammars cover
- Requires familiarity with each grammar

The ideal tool:

- A search interface to locate relevant grammar fragments
- Make the poor grammar engineer's life easier

ERG Types



Language Documentation

Established precision grammars are valuable resources:

- Distillation of much work from descriptive linguistics
- Yielded analyses useful as examples of linguistic phenomena

But...

No way to associate analyses with phenomena.

What would be nice:

- Means of labelling grammar components with phenomena
- Embed precision grammar treebanks within descriptive grammars
- Enrich descriptive grammars with on-demand examples

Possible Approach

Manual annotation of grammar fragments

- But hard to apply this to existing grammars
- Requires convincing grammar engineers

Desirable to find an automated approach.

Proposal

Investigate techniques for automatically detecting linguistic phenomena

In particular:

1. Labelling grammar fragments
2. Measuring constructional similarity across grammars

With an eye towards being used in aforementioned applications

Step 1: Creating Phenomena Corpora

Two cross-linguistic corpora of items:

- annotated with linguistic phenomena
- GOLD used for annotation

Data sources:

1. ODIN
   - But bias towards morpho-syntactic phenomena
   - IGT is not a formal standard — heterogeneous dataset
2. DELPH-IN treebanks

Chosen languages:

- English, Spanish, German, Portuguese, Wambaya, Japanese

Questions

1. How is linguistic phenomena defined?
   - "I'll know it when I see it" is problematic

2. What type of phenomena will we use?
   - Focus on constructions in phenomena catalogue
   - Implementationally "interesting" phenomena?

Grammar Labelling

Associating grammar components with linguistic phenomena:

- Parse items from phenomena corpus with relevant grammars
- Use parser output as input into machine learning algorithms
- Possible outcome: clusters of types associated with a phenomenon
- Even better: clusters of type constraints

Challenges:

- Supertypes not found in AVMs
- Architectural and linguistic differences between grammars
- Which features work for different types of grammars?

Needed for evaluation:

- Gold standard data
- Requires manual creation
- Could come from phenomena catalogue

Constructional Similarity

- Similarity of analysis vs similarity of phenomenon
- Using similarity of analysis presents distinct use-case

Constructional similarity:

- Compare the underlying analysis
- Unsupervised task
- Identification of phenomena amenable to similar analysis

How to proceed?

Hands-on Grammar Engineering

Extending a grammar to handle new constructions

- Gain familiarity with grammar engineering
- Trial techniques developed in the course of the project
- Use the Grammar Matrix as a starting point (obviously)
- Chosen language: French

Summary

We propose to investigate...

Techniques for detecting linguistic phenomena in precision grammars

1. Labelling of grammar fragments
2. Constructional similarity

Motivation:

Increase utility of precision grammars

- Discoverability of implementations of phenomena
- Grammar engineering for language documentation