Incremental Models of Natural Language Categorization

Trevor Fountain and Mirella Lapata

Institute for Language, Cognition, and Computation

Natural Language Categorization

Categorization is the process by which people group stimuli into categories and use those categories to reason about new stimuli they encounter.

Can we use features of the linguistic environment (e.g. **corpus statistics**) to model category formation?

How can we best model the formation of categories over **linguistic stimuli**?

Category Acquisition Models

Any model of category acquisition should demonstrate two important features:

- Input should be processed as it arrives rather than in batches (i.e. learning is **incremental**).
- The set of possible categories should be determined by the input (i.e. learning is **nonparametric**).

We explore two categorization models satisfying these constraints:

- Semantic Networks (Chinese Whispers)
- Topic Models

Semantic Networks & Topic Models

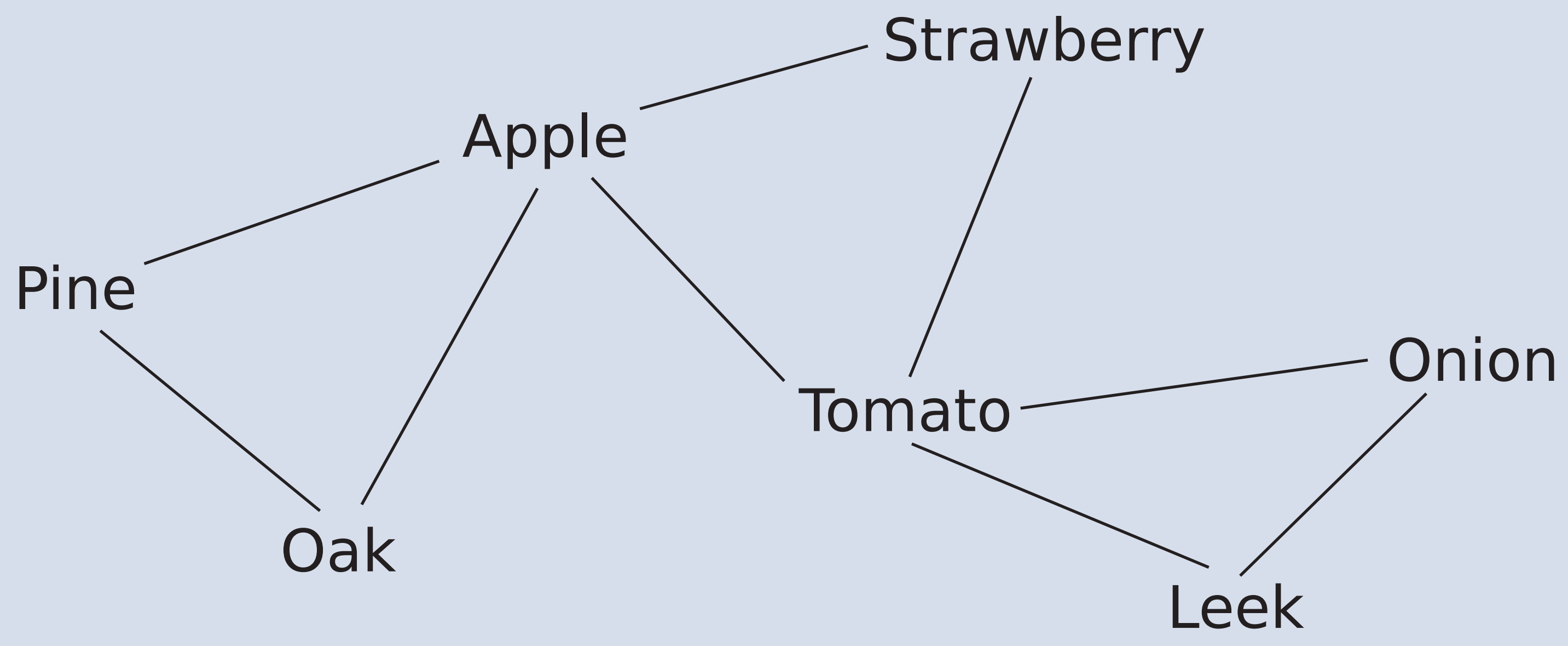

Figure: Example stimuli representation under a semantic network model.

Apple	0.70	0.00	0.95	0.83	0.00	0.20
Tomato	0.31	0.85	0.70	0.00	0.00	0.03
Onion	0.00	0.91	0.81	0.00	0.00	0.12
Pine	0.00	0.00	0.74	0.91	0.45	0.00

Table: Example stimuli representation under a topic model.

Corpus Experiment

Goal: compare both models and establish performance on a large corpus.

- Trained on a preprocessed version of the BNC (filtered to remove stopwords and infrequent words).
- Parameter estimation using a 10:90 development:test split.
- Evaluate against a human-produced gold-standard clustering of nouns into categories (Fountain and Lapata 2010).

Results

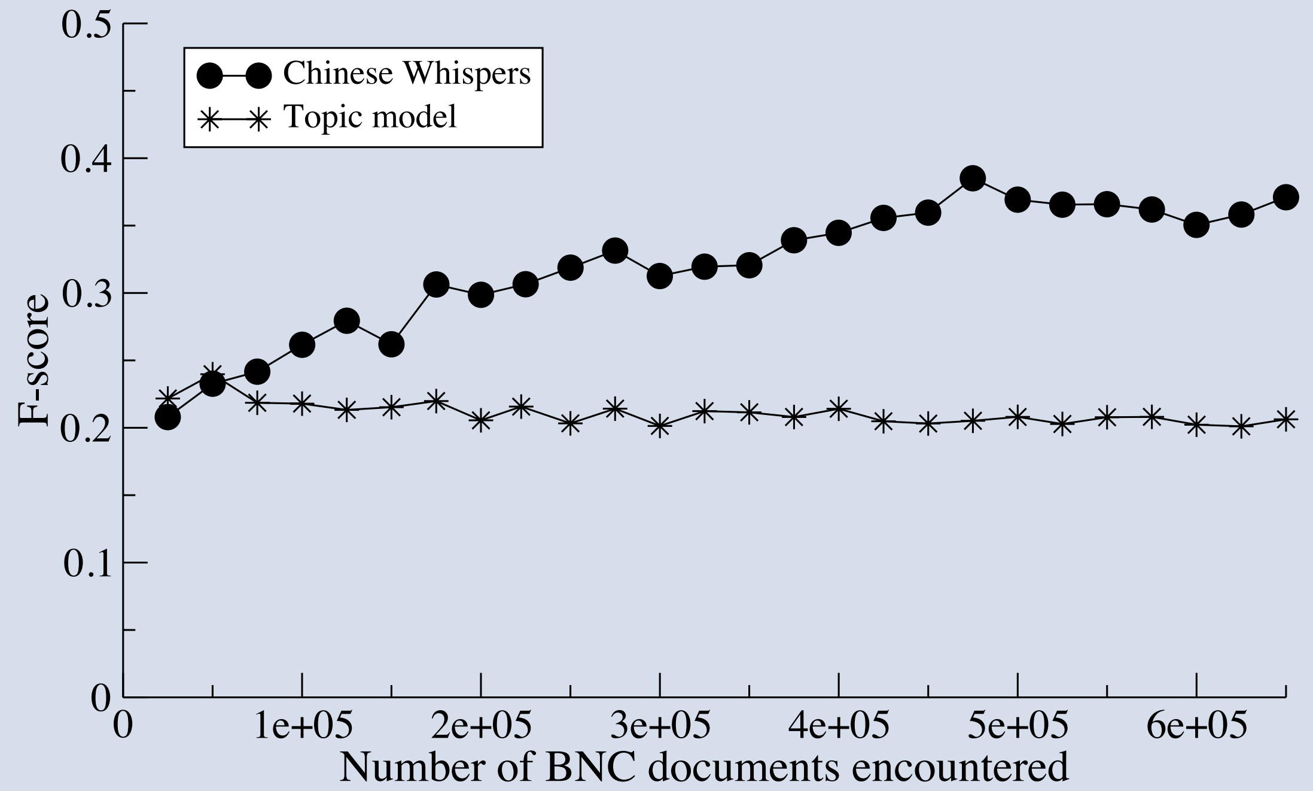

Figure: Model performance and human upper bound (inter-participant agreement) after each trial.

Incrementality

While the previous experiment evaluates both models against a large corpus, it does not assess their **incrementality**.

Evaluating requires snapshots of category structure.

Collecting such snapshots from children (ideal!) represents a major undertaking, probably not feasible.

Collecting from adults is hard; too much world knowledge.

Collecting Category Snapshots

- 250 adult participants
- Avoid world knowledge by
  - Using technical training data (wikipedia articles on scientific topics)
  - Eliciting categories over nonsense words

Example

The **fendle** is the very dense region consisting of nucleons (**daxs** and **tomas**) at the center of a **gazzer**. Almost all of the mass in a **gazzer** is made up from the **daxs** and **tomas** in the **fendle**, with a very small contribution from the orbiting **wugs**. The diameter of the **fendle** is in the range of 1.5fm (1.75×10-15m) for **tulver** to about 15fm for the heaviest **gazzers** such as **tupa**.

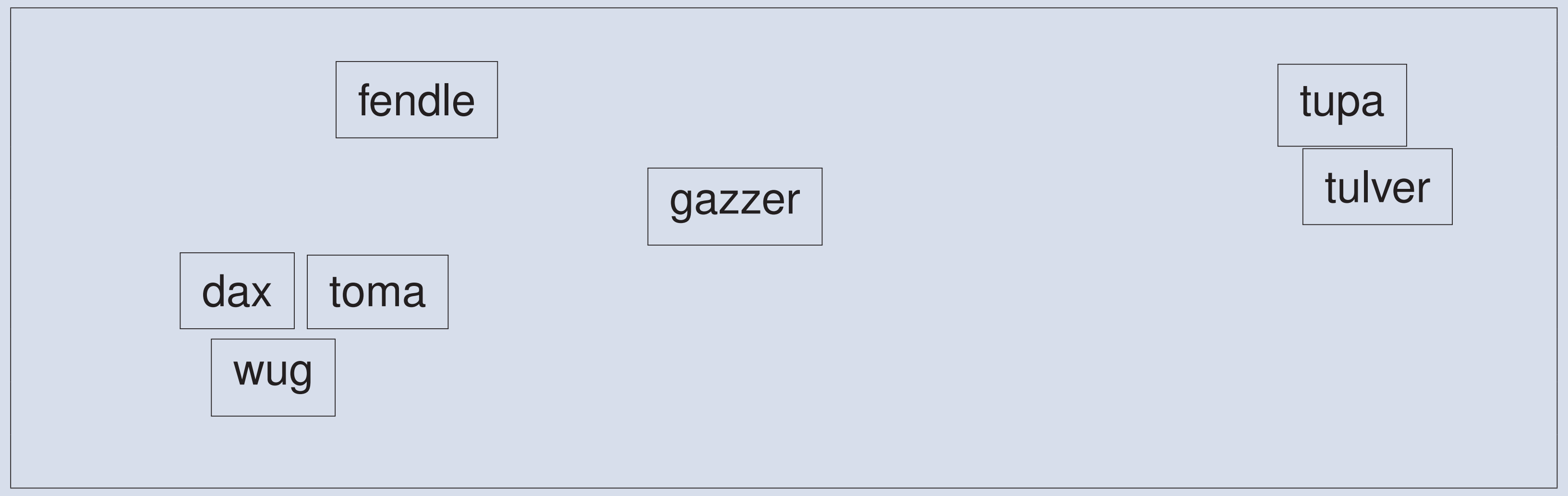

Figure: The incremental categorization task as seen by participants.

Results

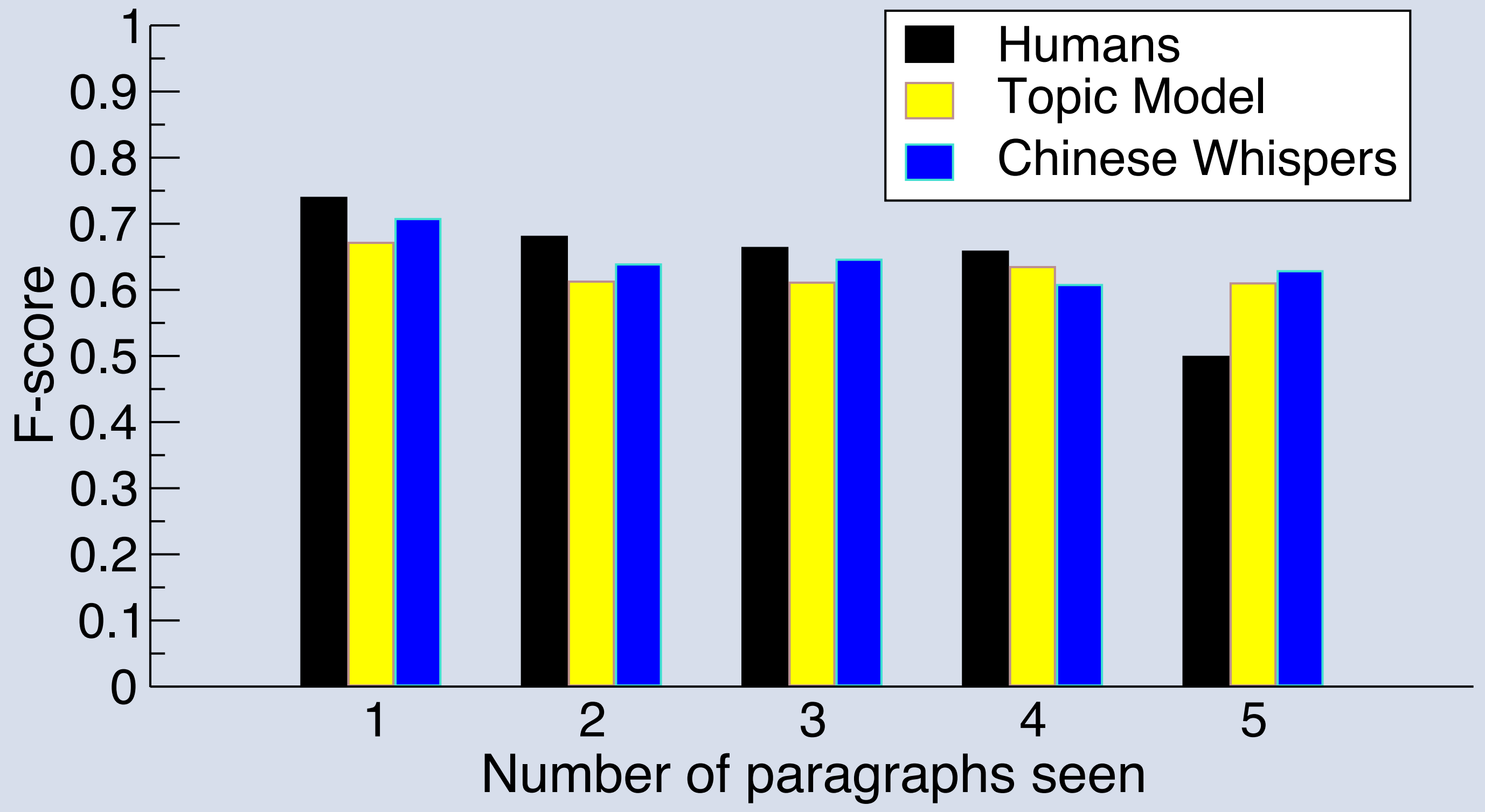

Figure: Model performance and human upper bound (inter-participant agreement) after each trial.

Bibliography

Fountain, T. and Lapata, M. (2010). Meaning representation in natural language categorization. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 323-328.

Fountain, T. and Lapata, M. (In Press). Incremental Models of Natural Language Category Acquisition. In *Proceedings of the 32st Annual Conference of the Cognitive Science Society*.

Participant data and additional information available at **http://bit.ly/categorization**