Incremental and Nonparametric: Modeling Category Acquisition

Trevor Fountain

School of Informatics
Institute for Language, Cognition, and Computation
The University of Edinburgh

11 May 2011

“Categorization?”

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

“Category Acquisition?”

Can we model the process by which people learn categories over a given set of stimuli?

“Natural Language?”

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

Hearst Patterns

“__A__, a kind of __X__”

“__A__, __B__, and other __X__s”

“__X__s, such as __A__ and __B__”

Co-occurrence

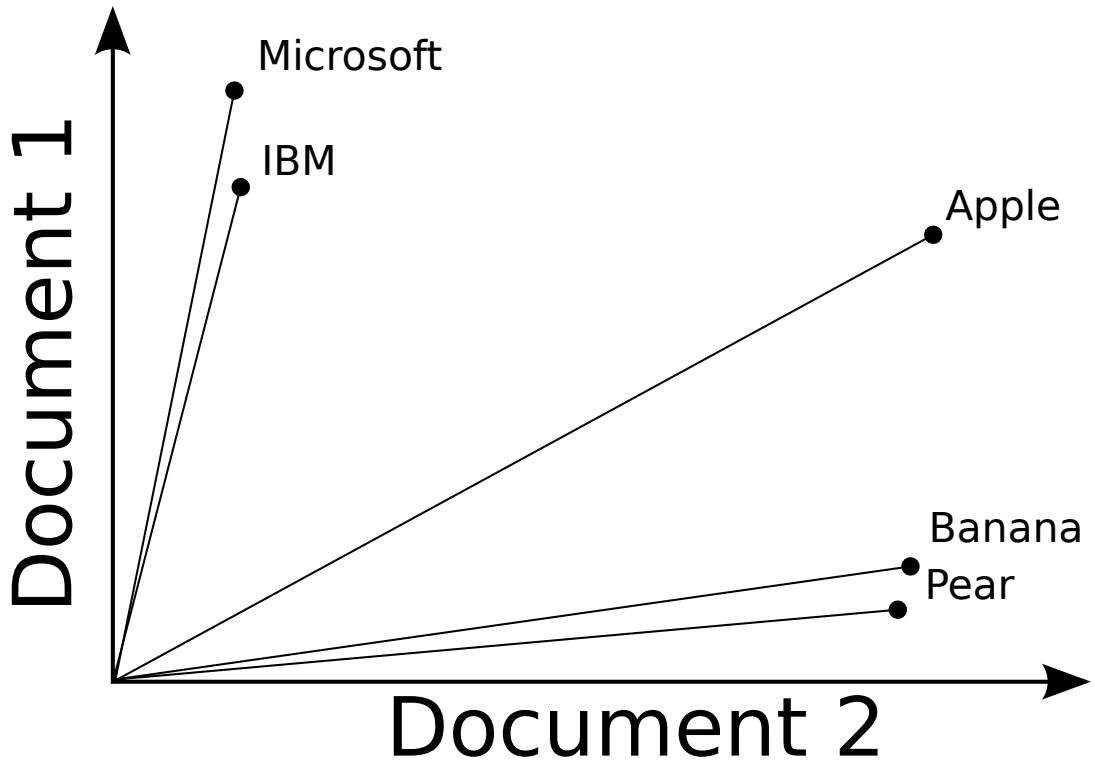
Document 1

Tech companies Google, **IBM**, **Apple** and **Microsoft** head the world's top 100 most valuable brands, according to a new global survey on...

Document 2

Online Plant Nursery featuring exotic fruit trees including **apple**, **orange** and **pear** for your garden or orchard.

Co-occurrence



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**).

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:

- Topic Models
- Semantic Networks (Chinese Whispers)

Models

Topic Models

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.

Models

Semantic Networks

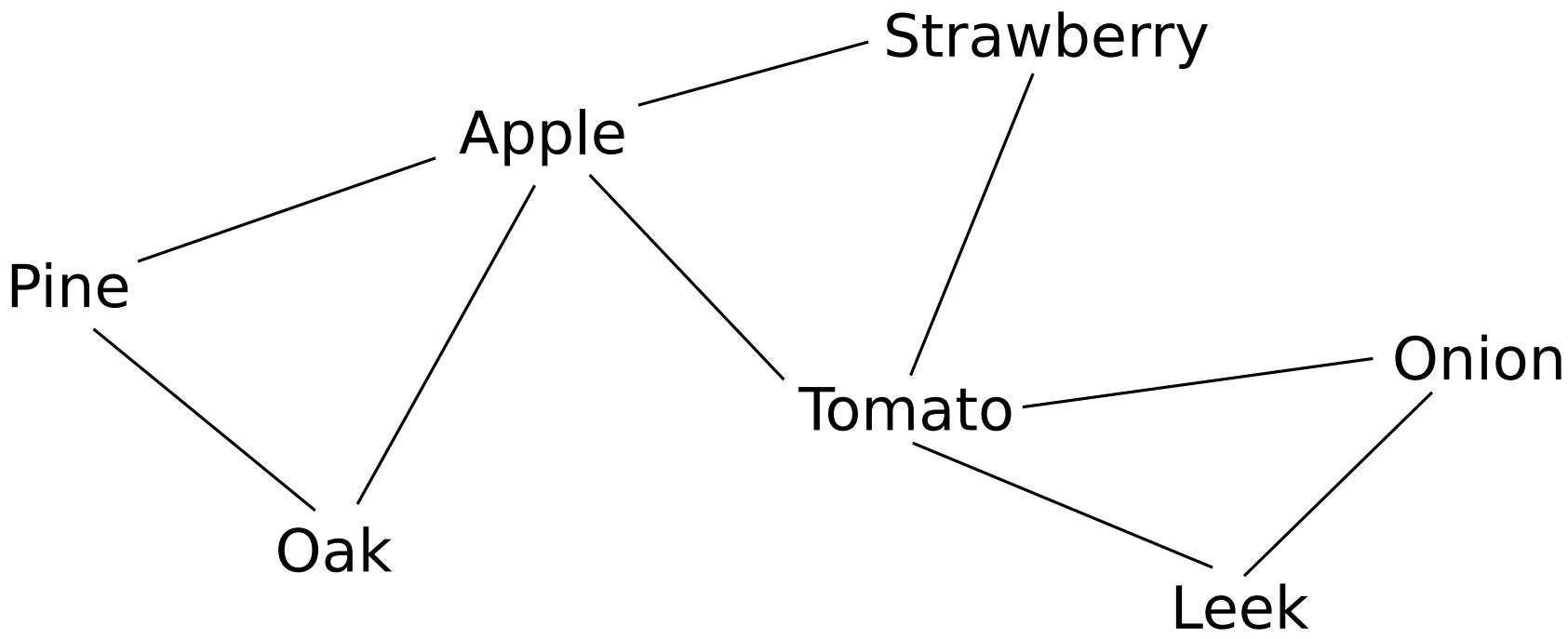


Figure: Example stimuli representation under a semantic network model.

Corpus Experiments

BNC

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).

Corpus Experiments

BNC

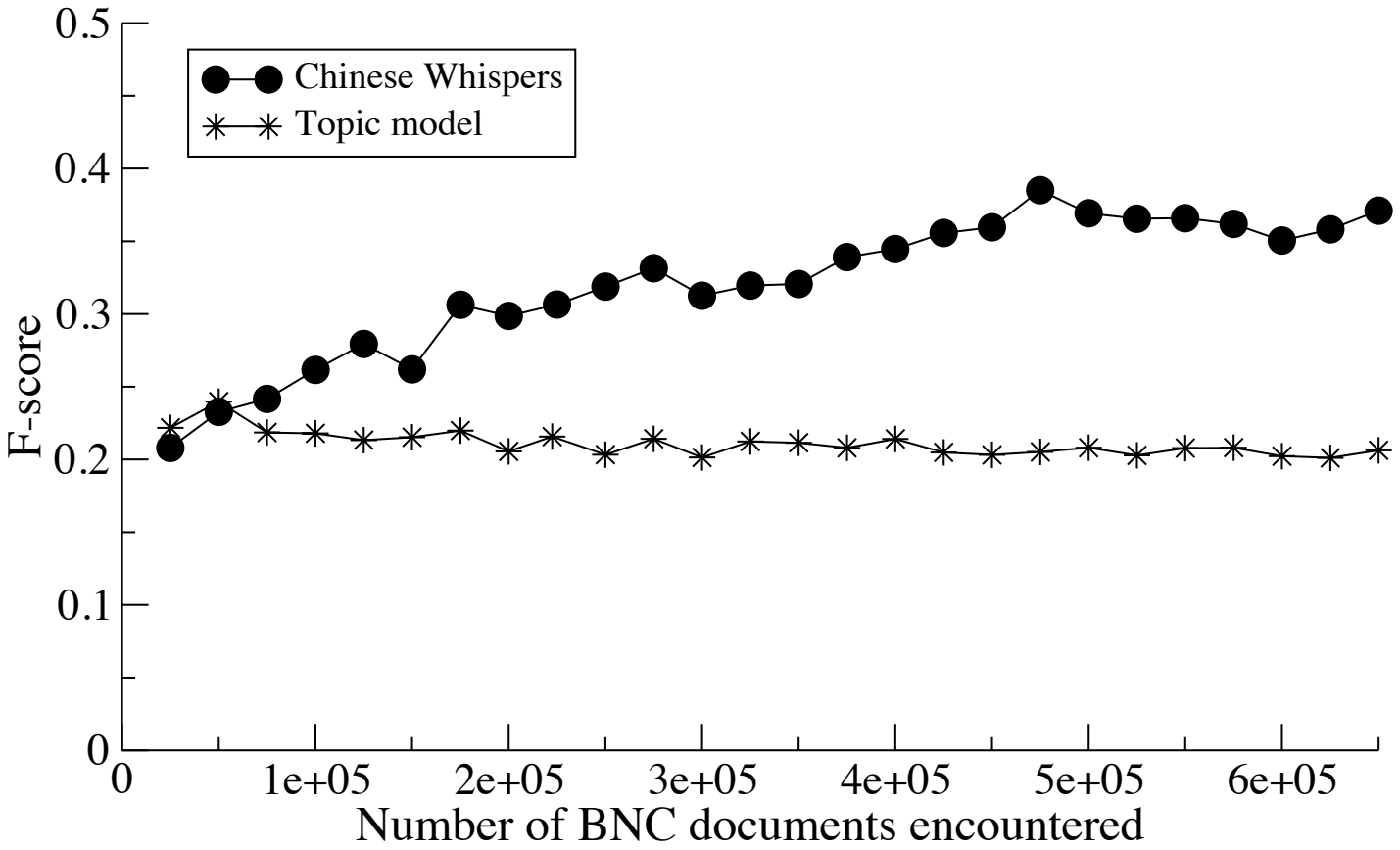


Figure: Model performance after encountering *N* training documents.

Corpus Experiments

BNC

Conclusions

- Semantic network model does a better job of acquiring categories over a large corpus
- Still a difficult task

Corpus Experiments

CHILDES

- BNC is a corpus of *adult* speech – probably not representative of child-directed speech.
- Ideally, we'd see the same effect in applying both models to a corpus of child-directed speech.

Corpus Experiments

CHILDES

Same setup as BNC experiment:

- Preprocess CHILDES to remove stopwords, transcription errors, and child speech.
- Evaluate against same gold standard categories.
- Group corpus according to child age.

Corpus Experiments

CHILDES

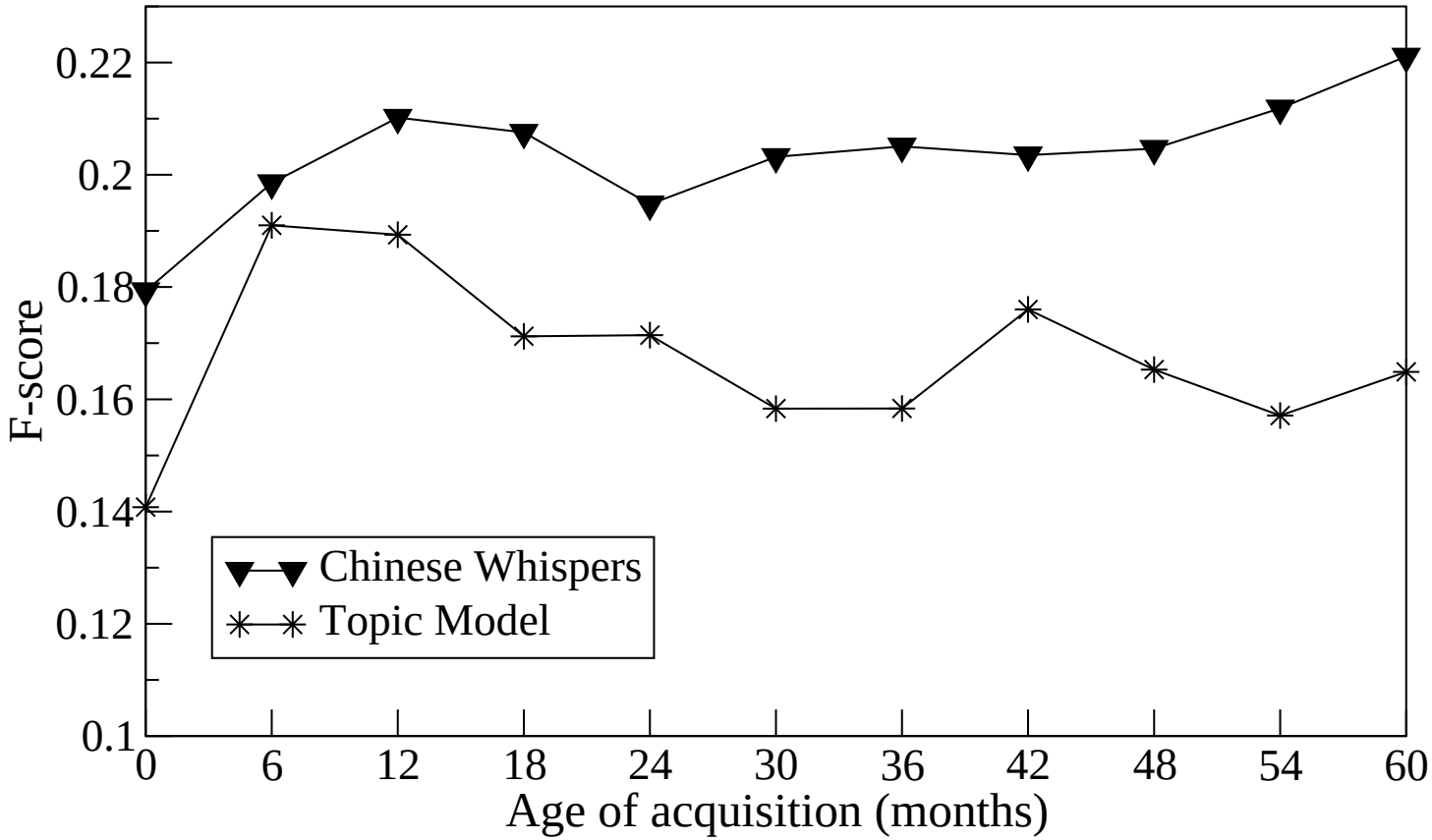


Figure: Model performance on child-directed speed

Corpus Experiments

CHILDES

Conclusions

- Small corpus size and sparsity of target words makes it difficult to draw strong conclusions
- ...But semantic network model still somewhat superior
- Randomizing order of age bins (not shown) destroys the improvement-over-time effect

Corpus Experiments

Incrementality

- While the previous experiment evaluates both models against a large corpus, it does not assess how well they capture **incrementality**.
- Evaluating incrementality 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.

Corpus Experiments

Incrementality

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$\times$10-15m) for **tulver** to about 15fm for the heaviest **gazzers** such as **tupa**.

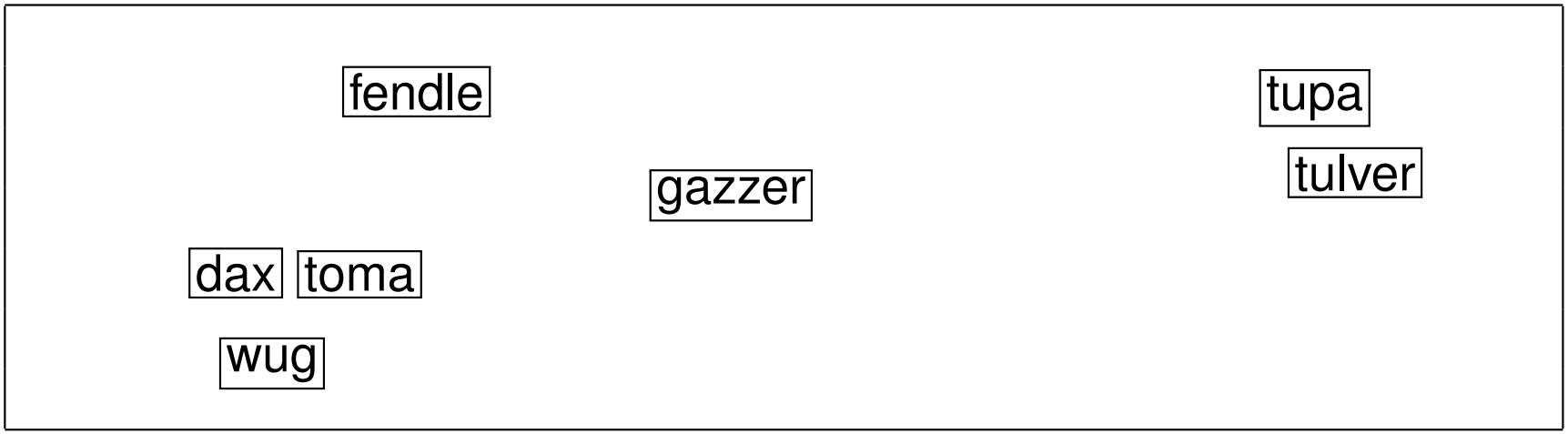


Figure: The incremental categorization task as seen by participants.

Corpus Experiments

Incrementality

- 13 source documents compiled from technical Wikipedia articles (e.g. medicine, physics, biology)
- 3-5 paragraphs per document, 4-6 sentences per paragraph
- Average 9 target (nonsense) words per document
- 250 (adult) participants

Corpus Experiments

Incrementality

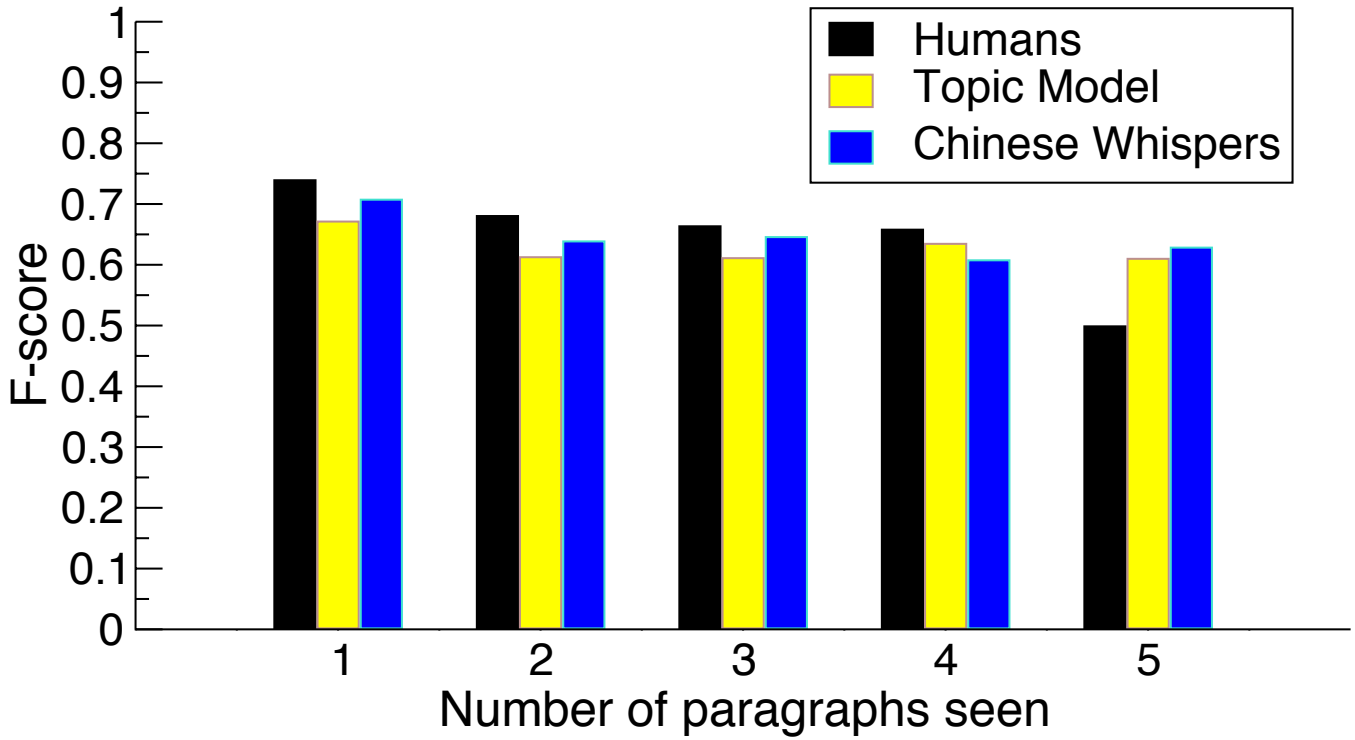


Figure: Model performance and human upper bound (inter-participant agreement) after each trial.

Corpus Experiments

Incrementality

Conclusions

- Both models do a pretty good job of modeling human performance.
- Differences not statistically significant.

Questions