Corpus Pragmatics: Something old, something new

David Beaver
Department of Linguistics
UT Austin

TLS 2007, Austin Texas

Confession


"Swedish may export synthetic wolf urine - sprayed along roads to keep elk away - to Kuwait for use against camels." Associated Press, January 19, 1995

- Like many linguists brought up in a tradition of artificial examples (and often self-verified judgments), my work has shifted to include more corpus data and more experimental data.

Plan: three corpus vignettes

- Anaphora resolution
- Accent prediction
- Existential constructions
First vignette
Anaphora resolution

- D. Beaver, *New Directions in Discourse Optimization*, Linguistics and Philosophy, 2004

Centering

- There is a huge literature on how anaphoric expressions are resolved.
- An old approach, Centering (Grosz et al. 1983/1995) is still the most influential.
- The basic idea is simple: speakers produce texts that minimize abrupt, unexpected transitions.
- Hearers then resolve anaphora by tracking what referents have been mentioned, and how prominent they are, and assuming that the speaker will always signal transitions in what’s being talked about ahead of time.

The anaphora resolution problem

And Caleb said, “The one who attacks Kiriath-sepher and captures it, I will even give him my daughter Achsah for a wife.” And Othniel, the son of Kenaz, Caleb’s younger brother, took it; so he gave him his daughter Achsah for his wife.


How do we know what the pronouns mean?

The BFP Algorithm

- Centering attracted huge interest, but was imprecisely stated.
- The first precise statement was by Brennan, Friedman and Pollard (BFP).
- Unfortunately, the BFP model is stated in the form of an algorithm which is hard to understand at an intuitive level.

Construct the proposed anchors for Cj
(a) Create set of referring expressions (REs).
(b) Order REs by grammatical relations.
(c) Create set of possible forward orient (CF) elements. Expand REs set with every RE and for each RE, determine its agreement with all other REs for every possible referent. These expansions are a way of modeling a discretization of possibilities.
(d) Create list of possible backward orient (CB) elements. This is taken as the set from (CF) that is an optimal choice, subject to the possibility that we will not accept a CB for the current anchor.
(e) Create the proposed anchors (PC) combinations from the outer-product of the previous two steps.

Filter the proposed anchors for each anchor in our list of proposed anchors we apply the following three rules. If a passes each filter then it is still a possible anchor for the current sentence.
(a) Filter by construction. That is, if we have proposed CC(s) and CB(s) and CB(s) are all present, propose CC if we have proposed an anchor for a pronoun which is constructed with, chooses the anchor from consideration.
(b) Go through (CC,CB) (in order) those which are marked 
(c) If the proposed CB of the anchor does not equal the first concretization for that element in order. This means that the CC will be the highest ranked element of the proposed anchors, regardless of the current situation. This corresponds to constraint 1 in section 3.5.
(d) If none of the anchors marked as pronouns in the proposed CC is equal to the proposed CB then eliminate the anchor. This guarantees that if any element is to be treated as a pronoun then the CB is realized as a pronoun. If there are no proposals in the proposed CC then it is treated as in section 3.5. This rule could be incorporated as a subrule of section 3.5.

Classify and rank
(a) If any anchor in the list of proposed anchors is not the one we are currently working on.
(b) Rank each proposed anchor using the extended ranking in section 3.5. The (CC,PC) of the proposed CC and (PC) of the most highly ranked PC.

1. Centering attracted huge interest, but was imprecisely stated.
2. The first precise statement was by Brennan, Friedman and Pollard (BFP).
3. Unfortunately, the BFP model is stated in the form of an algorithm which is hard to understand at an intuitive level.

The BFP Algorithm

- Centering attracted huge interest, but was imprecisely stated.
- The first precise statement was by Brennan, Friedman and Pollard (BFP).
- Unfortunately, the BFP model is stated in the form of an algorithm which is hard to understand at an intuitive level.
Centering in Optimality Theory (COT)

• In Beaver (2004), I proposed a framework for stating Centering type theories in Optimality Theory.
• An independently motivated set of constraints exactly reproduce the predictions of the BFP Centering model under a certain constraint ranking.
• I also suggested some improvements, including an alternative ranking.

Example

1. Jane likes Mary.
2. She often visits her for tea.
3. The woman is a compulsive tea drinker.

The COT constraints (ranked so 1 is strongest)

1. AGREE Anaphoric expressions must agree with antecedents for number and gender.
2. DISJOINT Co-arguments of a verb are disjoint.
3. PRO-TOP The topic is pronominalized.
4. FAM-DEF The referent of each definite NP is familiar and no new information is provided by the definite.
5. COHERE The topic of the current sentence is the topic of the previous one.
6. ALIGN The topic is in subject position.

She often visits her for tea in COT

<table>
<thead>
<tr>
<th></th>
<th>AGREE</th>
<th>DISJOINT</th>
<th>PRO-TOP</th>
<th>FAM-DEF</th>
<th>COHERE</th>
<th>ALIGN</th>
</tr>
</thead>
<tbody>
<tr>
<td>k = i, l = j</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>k = i</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>k = i, l ≠ {i, j}</td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k ≠ {i, j}, l = i</td>
<td>*</td>
<td></td>
<td>* *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k ≠ {i, j}, l = j</td>
<td>*</td>
<td></td>
<td></td>
<td>* *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k ≠ {i, j}, k ≠ l</td>
<td>* *</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>k = l ≠ {i, j}</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td>**</td>
<td></td>
</tr>
</tbody>
</table>
Empirical basis of Centering/COT

- Centering was motivated by a small number of example discourses, primarily artificial.
- There has been much post hoc empirical work justifying various aspects of Centering (experimental, corpus, cross-linguistic comparison).
- But the form in which Centering was originally stated meant that it was hard to see what the space of such theories was, and to separate individual empirical claims.
- In this sense, standard Centering does not consist of a set of easily testable hypotheses.
- The constraints in COT are much more directly testable, and hence falsifiable. (Unfortunately.)

Advantages of COT

- Independent motivation: Every constraint is based on rules from prior linguistic literature.
- Modular: effects of individual constraints are easily seen, and constraints can be independently modified, removed, or added.
- Reversible: can model production as well as comprehension. This allows applications to text generation (not discussed here).

Testing COT

- They tested the implementation on the 1M word WSJ portion of the Penn Treebank, annotated for anaphoric coreference, and found 80% correct resolution for COT, close to the ceiling for current models.
More generally…

• Objective measures of linguistic theories cannot be based on hand-picked, or, worse, hand-crafted examples. The only objective measures come from performance on text and speech generated by writers and speakers unconstrained by the linguist.

• But a theory that maximizes performance over a set of artificial examples is unlikely to provide the best model on naturally occurring texts.

A place for artificial examples?

• On the other hand, there may still be a place for artificial examples. In this case, many parts of a theory motivated by artificial data survived a fairly rigorous test, and *might* form part of later theories.

• And modular statements of theories, however motivated, provide an avenue for development in the face of individual counterexamples or bulk testing on naturally occurring data.

• So maybe the moral is not that we should stop using artificial examples to develop theories, but that we should stop only after we’ve also tested those theories on natural texts.
Second vignette
Accent and Information Status


Plan for this section

- How does *information status*, such as *old* vs. *new* (previously mentioned vs. not), affect what is accented in speech?
- We start with accented pronouns, and yet more problems with Beaver (2004).
- Then we turn to a more general consideration of how accents are distributed in speech.

Accented pronouns

- Pronouns are sometimes accented, and this is known to correlate with changes in reference in some cases:

  2. John, called Bill a Republican and then
     a) He insulted him
     b) HE insulted HIM
     (Lakoff 1971)

Pronouns and accent

- It has often been argued that accent on a pronoun is used by speakers to signal the information status of the referent.
- Specifically, accent is claimed to signal that a pronoun should not be resolved to the most salient entity.
- In Beaver (2004), I show that the effect of accent on pronominal resolution preferences could be seen as a case of *partial blocking*, the process whereby a marked form is given a marked interpretation.
Accented pronouns in Beaver (2004)

- Under further assumptions about the interpretation process (technically: bidirectional optimization), COT predicts that accent on pronouns will shift interpretation to an otherwise dispreferred resolution.
- Only one very general, independently motivated constraint is needed (cf. Schwarzschild 1999):
  AvoidF: Avoid focus, i.e. don’t accent something unless you need to.
- There’s only one major problem with this analysis…

Results

<table>
<thead>
<tr>
<th></th>
<th>-SHIFT</th>
<th></th>
<th>+SHIFT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean F0</td>
<td>max F0</td>
<td>mean F0</td>
<td>max F0</td>
</tr>
<tr>
<td>-PRO</td>
<td>180 ± 54</td>
<td>221 ± 65</td>
<td>179 ± 53</td>
<td>211 ± 62</td>
</tr>
<tr>
<td>+PRO</td>
<td>173 ± 77</td>
<td>183 ± 78</td>
<td>199 ± 69</td>
<td>212 ± 85</td>
</tr>
</tbody>
</table>

Effect of topic shift on frequency (hz) for pronouns and non-pronouns

- Subjects reading texts in which there was a switch in pronoun reference did not reliably accent pronouns.
- Though f0 tends to change in the direction Beaver (2004) et al predict, this is not significant, and, topic shifted pronouns still have lower max f0 than nouns in definite descriptions, and the small f0 effects probably have other causes (paragraph effects).

The problem with the analysis

- It’s probably wrong!
- Production experiments conducted as I was writing Beaver (2004), but published earlier (Wolters and Beaver 2001) already indicated that the data was not clearly as Kameyama and others had described.
- 20 subjects read texts with pronouns and definite descriptions that either involved topic shifts or not, and we measured fundamental frequency (f0) and other parameters.

Conclusions from Wolters & Beaver (2001)

``Any formal semantic theory of accented pronouns needs to deal with the fact that in many cases, this accent may be optional… secondly, most of the accented pronouns in our corpus data could be interpreted as cues to some sort of contrast.” (W&B 2001)
- We also concluded that we needed a design based on spontaneous rather than read speech.
Conclusions from empirical work on accented pronouns

- There is a long history of linguists providing models of how accent on pronouns affects reference.
- This includes the formally precise demonstration in Beaver (2004) of how such an effect might be derived from independent principles.
- The main problem with all these models is that there is no good reason to believe that accent on pronouns does affect reference.

Kothari’s corpus study

- Extending work she conducted in the Stanford Synthesis project, which I co-Pi-ed, Kothari (2007) studied accenting of pronouns in a corpus of spontaneous speech.
- She used 19 dialogues from Switchboard - telephone conversations hand-annotated for accent, contrastiveness and coreference.
- There were 834 3rd person pronouns in the sample, and Kothari considered how accenting related to the antecedent (e.g. position, form), and the pronoun (e.g. position, overall rate of pronominal accenting of the speaker, contrastiveness).

Kothari’s results

- Factors connected to properties of the antecedent (which correlate with switch of reference) were not significant factors in any of her statistical models.
- The only reliably significant factors were whether the pronoun was used contrastively, whether the pronoun was in subject position, and speaker effects.

Models of old/new and accent

- Accent on pronouns is not related to information status of the antecedent in the standardly understood way.
- But is information status in general related to accent?
- The standard hypothesis: discourse new constituents must contain an accent, but discourse old constituents need not.
The Synthesis project

- The Synthesis project studied accent distribution in the same section of Switchboard.
- The goal was to build a model of prominence for use in e.g. TTS systems, and at the same time test some of the ideas in the linguistic literature.
- 14,555 words were annotated for information status, contrast and concrete/non-concrete distinctions, all features that linguistic literature suggests are predictive of prominence.

Corpus Annotations

- Manual features:
  - Info. Status (new/old/mediated)
  - Contrast
  - Animacy
- Automatic features:
  - Part-of-speech
  - Unigram / bigram probability
  - TF.IDF (relative frequency in the dialogue)
  - Stopword (special, high frequency word)
  - Lead word value (relative frequency in start of conversation)
  - Verb specificity
  - Accent ratio (how often word is accented in a large corpus)

Results: good news?

- Here are the results for nouns:

<table>
<thead>
<tr>
<th></th>
<th>old</th>
<th>mediated</th>
<th>new</th>
</tr>
</thead>
<tbody>
<tr>
<td>accented</td>
<td>156 (61%)</td>
<td>752 (63%)</td>
<td>307 (73%)</td>
</tr>
<tr>
<td>unaccented</td>
<td>99 (39%)</td>
<td>437 (37%)</td>
<td>113 (27%)</td>
</tr>
</tbody>
</table>

More results

- However, when building decision-tree classifiers using leave-one-out selection, the best accent classifiers did not include information status.
- The best classifiers (>77% accuracy) made greatest use of Accent Ratio, and limited use of some linguistic features (animacy, contrast).
- But leaving out even these linguistic features reduced performance by <0.5%.
- Overall performance is very near a ceiling of interspeaker variation identified in a separate study: for six subjects reading the same text, a speaker on average agreed on accents for only 82% of words, <5% above the accuracy of our best classifier.
Why Old/New doesn’t help predict accent

- It seems unlikely that information status will play a major role in applications like TTS.
- But why is Accent Ratio (AR) so effective, and why doesn’t information status help?
- Effects of information status are masked by word choice:
  - a speaker who wants to code something old, chooses words with low AR, e.g. pronouns.
  - And a speaker who wants to code something new chooses words with high AR, e.g. low frequency nouns.

Discussion of accent results

- Again, corpus work hasn’t yielded what linguists would have expected: information status correlates with accent, but not usefully.
- Machines to generate accents need not use information status. So it seems unlikely that humans use accent primarily to signal information status.
- Speculation:
  - Accent is best seen as a way to boost phonological discriminability and hence enhance morpho-lexical distinctions. It is a tendency to make unpredictable material more readily perceived. (Further work: test this!)
  - Contrastive accents might be a special case where this tendency is exploited as a signal that a distinction is being made.

Third vignette
Existential Constructions

- D. Beaver, I. Francez and D. Levinson, 2005, *Bad subject: (non-)canonicality and NP distribution in existentials*, SALT 15
- And further unpublished/ongoing work with:
  - E. Destrue (U.T. Austin)
  - G. Bouma (U. Groningen)
  - I. García Álvarez, T.F. Jaeger (then, Stanford U.)
  - A. Bratkievich, L. Mikkelsen (U.C. Berkeley)

Introduction to the existential project

- I’m midway through a project to understand existential constructions cross-linguistically as examples of non-canonical argument realization.
- I’ll present the basic ideas and preliminary results.
Existential (e) vs Canonical (c) constructions:

- **copular**
- **coda**
- **expletive**
- **(existential) pivot**

**c)** No mountain is high enough!

**e)** There ain't no mountain high enough!

### Non-canonicality

- Many existential constructions involve non-canonical sentence forms:
  - Expletive/null grammatical subject; non-standard position of primary argument (Germanic, Romance)
  - Reduced verbal paradigms or inconsistent agreement (Hebrew, Spanish, English)
  - Marked case (Russian, Hebrew, Finnish)

### Existential Constructions

1. Beidh go leor bia ann. (Irish, McCloskey 2006)
   - There'll be plenty of food!

2. Er zijn alle mijnopsporing programma's. (Dutch, web)
   - There are all the mine-clearing programs.

3. Disafortunadamente hay los estudiantes que no tienen motivación. (Spanish, web)
   - Unfortunately, there are (the) students that lack motivation.

4. Havia a Terra e havia os planetas. (Portuguese, web)
   - 'There is the Earth and there are the planets.'

5. Na pisjmo ovte na poluchila. on letter-acc answer-gen no come-pst
   - 'There came no answer to the letter.'

6. Muzar she-be-ha'aretz en et kol ha-matkhon.
   - 'Its strange that they don't have the whole recipe.'

7. Der er kommet alle mulige undskyldninger.
   - 'Every possible apology came.'
Distributional Restrictions

- These involve a categorical denial of definite and strong quantificational pivots.
- Such claims are often caveated: definite or strong pivots involve a special *presentational use/reading of the construction.*

Canonical subj. properties

- Silverstein (1976), Keenan (1976) and Aissen (1999) suggest universal tendencies for grammatical subjects, expressible as scales:
  1. Agent > [order over other 0-roles...] > no 0-role
  2. Sentence (aboutness) topic > non-sentence-topic
  3. Discourse topic > discourse old > hearer old > hearer new > dependent element > no referent introduced
  4. Animate > inanimate > abstract, type/kind denoting
  5. Short > long

Basic Idea

- Standard theories of existentials treat them as isolated constructions in a grammatical vacuum.
- We start with an alternative paradigmatic / competition approach related to Lambrecht (2000) and Mikkelsen (2002).
- Perhaps pivots are not realized as syntactically canonical subjects because they lack subject properties semantically and pragmatically.
- The distributional facts result from a competition between forms: *when the sole argument is a bad subject, it appears as a pivot.*

The empirical study

- We conducted quantitative corpus and web studies of English, Danish, Dutch, Spanish, Brazilian Portuguese, Russian, and, currently, French.
- We compare frequencies of canonical and existential constructions for a given NP type.
- E.g. is “No X is in the Y” more or less common than “There is no X in the Y”? How about for “a X” or “every X”?
Summary of Results

- We find strongly correlated results across languages, though none of the patterns are predicted by any existing theory of existentials.
- Crosslinguistically, the same scale of NP types recurs:
  - Local pro > non-local pro > proportional NPs > definite descriptions, demonstratives > prototypical indefinites > neg NPs
- Theories postulating a categorical definiteness effect are falsified. However, we do find a strong quantitative effect: a factor of 20+ difference in existential ratios between typical indefinites and typical definites or strong quantifiers.

Summary of Results (cont.)

- We find clear anti-definiteness effects, especially for negative determiners. This is predicted by the competition model, but not by any standard syntagmatic account.
- There are huge genre effects, a factor of >10 between spoken and written English.
- We also highly significant weight effects across NP types: longer NPs are much more likely to appear in pivot position than in canonical subject position, and vice versa.
Discussion of Existential Data

- Theoretical linguists don’t mind making very complex generalizations, so long as they are categorical.
- Gradient phenomena, in which many factors have a cumulative or probabilistic effect are alien to that tradition.
- It could be that distributional effects in existential constructions are inherently non-categorical.

General conclusion

- Remember that syntactician? Well, here’s one important generalization that’s come out of corpus research:
- Much of the time, generalizations based on artificial examples and our own intuitions are wrong!

Back to information status?

- There is another possibility: there could be some factor that is not directly identifiable using current searches, but that categorically predicts distribution in existentials.
- Ward & Birner (1995) in fact suggest a possible factor: information status. They claim that existential pivots must be hearer new.
- Although they (uniquely in prior work on existentials) did use corpus examples, I’m skeptical about their specific proposal.
  - First, it is easy to find naturally occurring counterexamples.
  - Second, my recent experiences suggest that even if an old-new distinction does turn out to be significant, it is unlikely to be the whole story.

- The surprising thing is that people (yes, me too, at times) who have not tested their theories on large amounts of naturally occurring data, can have the temerity to think that they have found “important generalizations.”
- But new sources of data cannot be ignored.
A1. Existential ratios for selected Spanish NPs

A2. Existential ratios for 7 NP types by language