Linguistics in Automatic Speech Recognition:

Problems and Challenges

Dalina Kallulli
University of Vienna, Austria

Talk given at
University of Ottawa, 1st March 2004

Overview

- Introduction
- Automatic Speech Recognition
  - History
  - State-of-the-Art
- More semantics in Language Modeling
  - Use WordNets
  - Use Collocation Databases
- More syntax in Language Modeling
  - Use Theoretical Linguistics
- Summary

A bit of history

- 70s: DARPA project on speech understanding
  - "Biggest speech recognition project ever" (5 years, budget $15 million)
  - Several labs (CMU, Dragon, BBN, MIT, SRI, SDC) developed competing systems and architectures
  - e.g. CMU's *Hearsay II* system integrated specialist modules via a generic common whiteboard
    - Some psychological plausibility
  - **BUT**: CMU's *Harpy* system, developed in one PhD thesis, outperformed all others based on a simple statistical precompilation of recognition options
    - "Beam search"

Historical dominance of statistical approaches

- Hidden Markov Models (HMMs) started to be used widely for speech recognition (70s)

- After the DARPA project results, statistical approaches became dominant, due to better
  - robustness
  - ease of system construction across applications
  - well-understood mathematical properties leading to efficient systems

Automatic Speech Recognition (ASR)
Bayesian Formulation

$$\arg\max_{wordsequence} P(wordsequence \mid acoustics) =$$

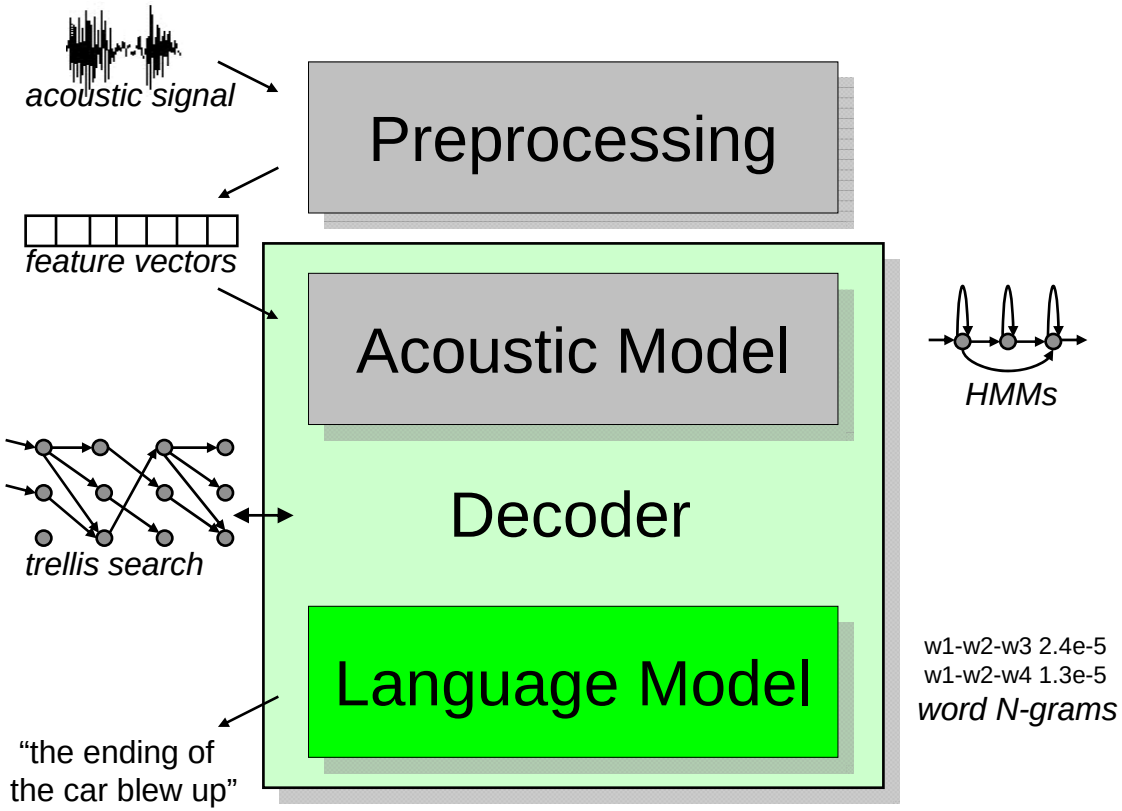
$$\arg\max_{wordsequence} \frac{P(acoustics \mid wordsequence) \times P(wordsequence)}{P(acoustics)}$$

$$\arg\max_{wordsequence} P(acoustics \mid wordsequence) \times P(wordsequence)$$

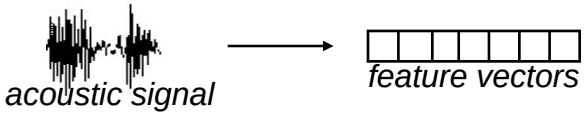
$P(acoustics \mid wordsequence)$ ... Acoustic Model

$P(wordsequence)$ ... Language Model

Speech Recognition Schema: State-of-the-Art



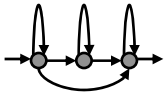
Acoustic Preprocessing



(this presentation focuses on language modeling, not acoustic modeling)

- Start from analog acoustic signal
- Discretize, quantize
- Derive a “frame” every 10-30ms:
  - By calculating a weighted mean in a time window longer than the frame, derive a vector of features that describe the speech signal
- Model characteristics of human hearing
- Common feature extraction methods:
  - MFCC (mel-frequency cepstral coefficients)
  - PLP (perceptually linear prediction coefficients)

Acoustic Modeling



HMMs

- Common methods:
  - Hidden Markov Models (HMMs) – dominant
  - Artificial Neural Networks (ANNs)

- Hidden Markov Models
  - model a unit (e.g. a phone) as a doubly embedded stochastic process:
    - a network of different states with stochastic transitions
    - each state has a probability distribution of values of the acoustic feature vectors
    - each model is parametrized automatically by training with many samples of the unit
    - at system runtime, compare the match of each sound with the Hidden Markov Model of each unit

What is a Language Model

w1-w2-w3 2.4e-5
w1-w2-w4 1.3e-5
word N-grams

- A language model is a probability distribution over word sequences
- P("the engine of the car blew up") $\approx$ 0.001
- P("the ending of the car blew up") $\approx$ 0

Principle of N-gram Language Models

w1-w2-w3 2.4e-5
w1-w2-w4 1.3e-5
word N-grams

- Hard to compute

  P("the engine of the car blew up")

- Decompose probability

  P("the engine of the car blew up") =

  P("the") × P("engine|the") × P("of|the engine") ×
  P("the|the engine of") × P("car|the engine of the") ×
  P("blew|the engine of the car") ×

  P("up|the engine of the car blew")

Trigram Language Models

w1-w2-w3 2.4e-5
w1-w2-w4 1.3e-5
word N-grams

- Assumption: each word depends only on the previous two words (altogether three words)

  P("blew|... the engine of the car") ≈ P("blew|the car"

  P("up|... the engine of the car blew") ≈ P("up|car blew")

- How do we find probabilities?
- Get text, start counting!

  P("blew | the car") ≈ C("the car blew") / C("car blew")

State-of-the-Art: N-gram Extensions

w1-w2-w3 2.4e-5
w1-w2-w4 1.3e-5
word N-grams

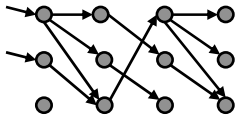
- Backoff
- Longer N-grams
- Class-based N-grams
- Skipping models

Cluster Models

w1-w2-w3 2.4e-5
w1-w2-w4 1.3e-5
word N-grams

- Put similar words into classes
  - E.g. *banana, apple* → *fruit*
- Probability of a word calculated by combining class probability and conditional probability of the word in that class
- P(banana | ripe) = P(FRUIT | ripe) × P(banana | ripe FRUIT)

Decoder

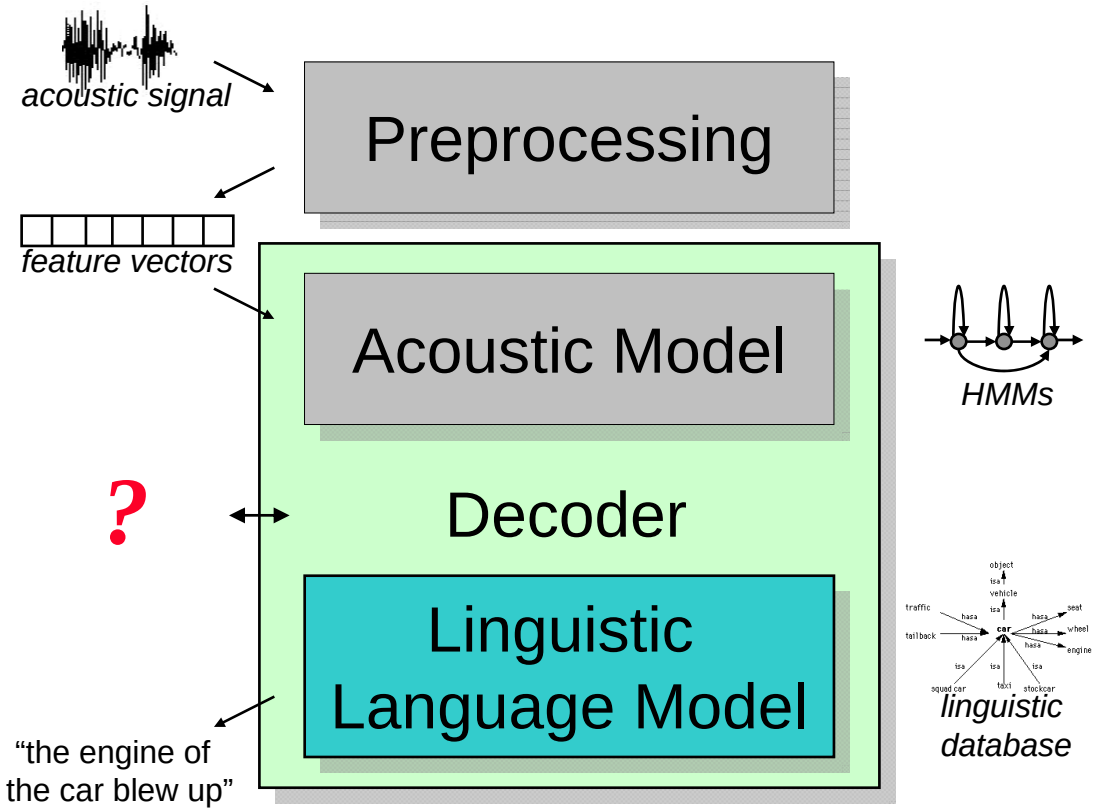


- To work in real time, a speech recogniser needs an efficient search algorithm
  - Evaluates the acoustic model and the language model in an integrated algorithm
  - Puts constraints on the structure and information formats of acoustic and language models

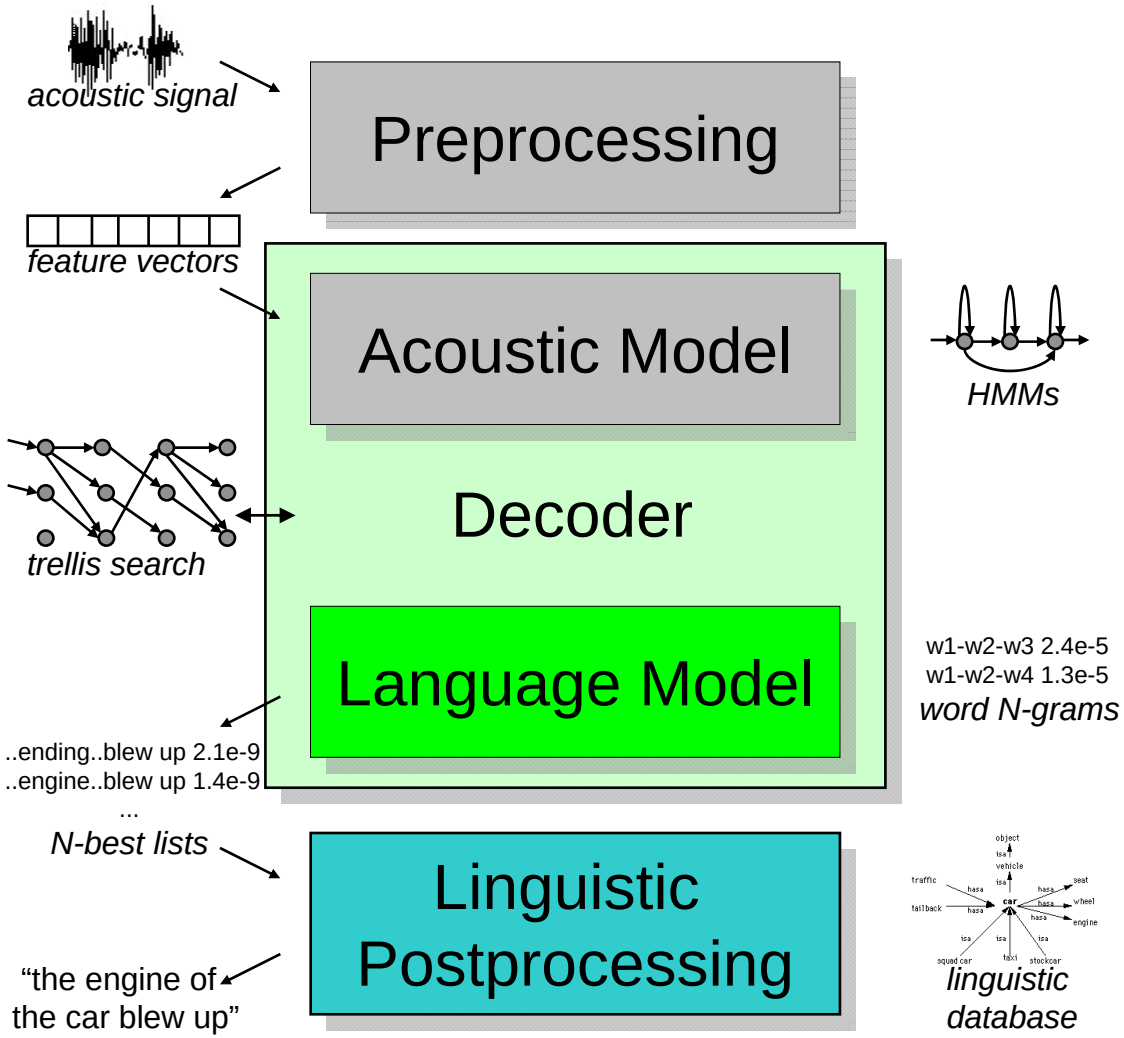
- Trellis search (or Viterbi search)
- Stack search (or Tree search)

- All approaches use heuristics to avoid having to compute all possible paths
  - No guarantee that the best result will be found

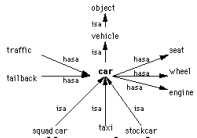
Integrated Linguistic Processor



Speech Recognition Schema: Linguistic Postprocessing



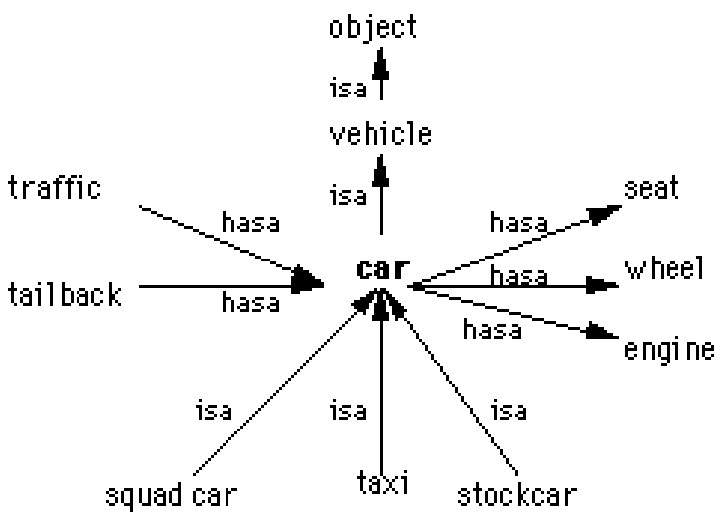
Proposals for More Semantics

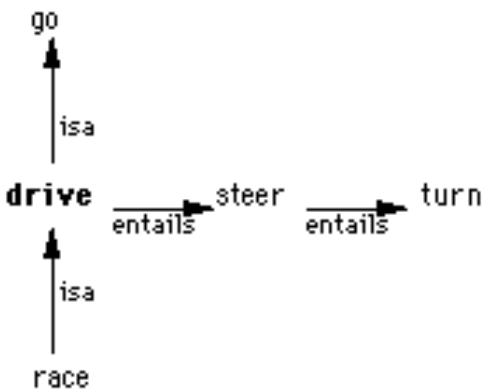


linguistic database

- Word nets
  - Lexical databases that store information about notional words of a language. In particular, information about semantic relations such as synonymy, hyponymy / hyperonymy, aggregation, entailment

    (e.g. Princeton WordNet, Miller et al. 1990, http://www.cogsci.princeton.edu/~wn/index.shtml)





Relations in WordNet

- WordNet designed to model lexical memory rather than lexical knowledge (i.e. it excludes much of a speaker's knowledge about both semantic and syntactic properties of lexical items).
- Synonymy
  - E.g. {*board, plank*}, {*board, committee*}; {*shut, close*} etc.
- Hypernymy / hyponymy
  - E.g. {*plant, tree, maple*}
- Meronymy / holonymy (HASA)
  - E.g. {*car, wheel*}, {*door, handle*}
- Different types of (lexical) entailment
  - E.g. *"He is snoring"* → *"He is sleeping"*
  - Troponymy (or coextensiveness)
    - E.g. *"He is limping"* → *"He is walking"*
  - Backward presupposition
    - E.g. *suceed – try, untie – tie*
  - Causal entailment
    - E.g. *give – have, raise - rise*

N-best List Example

1: SILENCE es ist SILENCE wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
2: SILENCE das ist SILENCE wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
3: SILENCE hätte SILENCE wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
4: SILENCE wer das wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
5: SILENCE es ist SILENCE vor den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
6: SILENCE rettet vor den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
7: SILENCE es das wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
8: SILENCE er das wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
9: SILENCE es ist SILENCE von den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
10: SILENCE es SILENCE wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
11: SILENCE es ist SILENCE wohl den meisten den neun und achtzig Menschen an Bord das Leben SILENCE
12: SILENCE hätten SILENCE wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
13: SILENCE das SILENCE wohl den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
14: SILENCE rettet zu_den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
15: SILENCE das ist SILENCE vor den meisten der neun und achtzig Menschen an Bord das Leben SILENCE
..
29: SILENCE rettet wohl den meisten der neun und achtzig Menschen an Bord das Leben
'SILENCE saves likely (for) most (of) the nine-and-eighty persons on board the life'
...
99: SILENCE das ist SILENCE wohl den meisten der neun und achtzig Menschen Antwort das Leben SILENCE
100: SILENCE das ist SILENCE wohl den meisten in neun und achtzig Menschen an Bord das Leben SILENCE

Using word nets

- Modeling of non-local collocations
  - When inspecting “n-best” lists, it is striking that the position of a complete phrase hypothesis in the ranked list is quite unrelated to its semantic coherence (i.e. to whether the resulting sentence includes a set of word choices at the uncertain positions that are likely to coocur in a common sentence)
- Example

  “The engine of the car blew up”

  should be favored over:

  “The ending of the car blew up”

Collocation Databases

- E.g. Princeton WordNet at one point contained 51.500 single words and 44.100 multi-word collocations.
  - E.g. *attorney general, stand in line, line of products, custom duty*
- Distinguish local collocations from
- Non-local collocations

  *(1) The **pilot** suffering the heart attack **landed** the plane safely.*

  *(2) The **butcher** from Albuquerque **slaghtered** chickens.*

  (3) Dass so viele andere Schiffe in der Nähe sind,

  that so many other ships in the near are

  ***rettet** wohl den meisten der 89 Menschen an Bord*

  saves well the most of 89 people on board

  *das **Leben**.*

  the life

Non-local Collocations

- Non-local collocations are not captured by trigrams!
  - Therefore they have a strong constraining influence on ASR
- The (partly unsolved) technical problem is how to collect, estimate, and process non-local collocations
- Dilemma:
  - Either there is too little data to estimate collocation strength
  - Or the resulting models are too complex to process with todays computer systems

More syntax

Why?

- Long-distance dependencies
  - E.g. agreement, anaphora, ...
- Acoustic confusability of function words
  - Syntax contributes structural constraints top-down
- Word-class n-grams
  - A special case of cluster models using part-of-speech information
- Subcategorization frames (Chomsky 1965)
  - important constraints between verbs and their arguments
  - **BUT**: descriptions in categorical linguistics are often counterexemplified ...

Subcategorization Frames

Pollard & Sag (1994, 105—108)

regard: ___ NP[acc] *as* {NP, AdjP}
consider: ___ NP[acc] {AdjP, NP, VP [inf]}
think: {___ CP[that], ___ NP[acc] NP}

(1) a. We consider Kim to be an acceptable candidate.
b. We consider Kim an acceptable candidate.
c. We consider Kim quite acceptable.
d. We consider Kim among the most acceptable candidates.
e. *We consider Kim as an acceptable candidate.
f. *We consider Kim as quite acceptable.
g. *We consider Kim as among the most acceptable candidates.
h. ?*We consider Kim as being among the most acceptable candidates.

Subcategorization Frames (Counter)exemplified – Manning (2003)

(2) a. The boys consider her as family and she participates in everything we do.
b. Greenspan said, “I don‘t consider it as something that gives me great concern.”
c. “We consider that as part of the job,” Keep said.
d. Although the raiders missed the playoffs for the second time in the past three seasons, he said he considers them as having championship potential.
e. Culturally, the Croats consider themselves as belonging to the “civilized” West, ...

(3) a. *We regard Kim to be an accepable candidate.
b. We regard Kim as an acceptable candidate.

(4) a. As 70 to 80 percent of the cost of blood tests, like perscriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.
b. Conservatives argue that the Bible regards homosexuality to be a sin.

More examples

(5)

a. *We regard Kim to be an acceptable candidate.

b. We regard Kim as an acceptable candidate.

(6) Counterexamples

a. As 70 to 80 percent of the cost of blood tests, like perscriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.

b. Conservatives argue that the Bible regards homosexuality to be a sin.

Proposal for Using Theoretical Linguistics

- Take a theoretical syntactic/semantic analysis
  - E.g. Kallulli (1999) on non-active morphology
- What types of linguistic information used in the analysis will be available in a speech recognition task?
  - Part-of-speech, an annotated lexicon, a morphology component...
- Can this be used for reranking hypotheses in the N-best list?
  - E.g. If there are two noun phrases, a non-active verb requires that one is dative.
- Test whether this analysis describes a frequent enough phenomenon to aid speech recognition

Digression: Analysis of non-active/reflexive morphology in Kallulli (1999)

(1) *Beni theu dritaren.*
Ben.nom broke window.acc
"Ben broke the window"

(2) *Benit iu thye dritarja.*
Ben.dat nact.broke window.nom
a. "Ben accidentally broke the window"
b. "The window broke to Ben"
c. "Ben's window broke"

(3) *Ana hëngri një akullore.*
Ana.nom ate an ice-cream
"Ana ate an ice-cream"

(4) *Anës i hahej një akullore.*
Ana.dat eat.nact an ice-cream
a. "Ana felt like eating an ice-cream"
b. "One could eat Ana's icecream"

Summary

- ASR today makes only limited use of current linguistic knowledge
- I propose to exploit lexical knowledge to check semantic consistency of recognition hypotheses
- I propose to study whether results from theoretical syntax can be used and deliver improvements in ASR
- In early stages of such work, decoder efficiency can be neglected for the sake of long term research progress

References

- Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.
- Fellbaum, C. (1998), *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
- Kallulli, D. (1999) Non-active morphology in Albanian and event (de)composition. In I. Kenesei (ed.) *Crossing Boundaries*. Amsterdam: Benjamins.
- Manning, C. (2003) Probabilistic syntax. In R. Bod et al. (eds.) *Probabilistic Linguistics*. Cambridge, Mass.: MIT Press.
- Pollard, C. and I. Sag (1994) *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.