Do CFG-Based Language Models Need Agreement Constraints?

Manny Rayner   **Beth Ann Hockey**   **Johan Boye**
Genevieve Gorrell   **John Dowding**

netdecisions   RIACS, Mail Stop 19–39   Telia Research
Wellington House, East Rd   NASA Ames Research Center   S-123 86 Farsta
Cambridge CB1 1BH, UK   Moffett Field, CA 94035-1000   Sweden

{manny.rayner, genevieve.gorrell}   {bahockey, jdowding}   johan.boye@trab.se
@netdecisions.co.uk   @riacs.edu

Abstract

Many people are now routinely building grammar-based language models for interactive spoken language applications; these language models are typically *ad hoc* semantic grammars which ignore many standard linguistic constraints, in particular grammatical agreement. We describe a series of experiments in which we took three CFG-based language models from non-trivial implemented systems, and in each case contrasted the performance of a version which included agreement constraints against a version which ignored them. Our findings suggest that inclusion of agreement constraints significantly improves performance in terms of both word error rate and semantic error rate.

1 Introduction

A key problem in building interactive spoken language systems is constructing a language model to guide speech recognition. There are two main approaches: statistical language models, and grammar-based language models. The basic idea of the statistical approach is to train the language model (most often some kind of N-gram grammar) from a domain corpus; if a sufficiently large corpus is available, experience shows that this method can yield excellent results (Cohen et al., 1995; Ward and Issar, 1995). In contrast, the grammar-based approach creates the language model directly in the form of a (most often hand-coded) grammar.

For the last decade researchers have paid more attention to the statistical alternative, and there are many theoretically attractive reasons for preferring it (Mori and Kuhn, 1991; Rosenfeld and Huang, 1992). None the less, within the last two years commercial speech technology has almost exclusively adopted the grammar-based

approach (VoiceXML Forum, 2001; W3C, 2001; Nuance Communications, 2001; SpeechWorks International, 2001; Tellme, 2001; BeVocal, 2001; HeyAnita, 2001). This focus on grammar-based methods is motivated by important practical and theoretical considerations. Most obviously, the large quantity of corpus data needed to train a statistical language model is hardly ever available; creating it, by Wizard of Oz simulation or similar methods, is extremely expensive. It is also frequently the case that the language model changes dynamically with the state of the interaction in some way that is easiest to specify in terms of rules rather than statistical regularities. For example, the dialogue may concern choice from some continually changing set of objects; the appropriate language model will have to take account of what those objects currently are, and how they can be referred to.

Another factor that must be considered is the type of dialogue strategy the system employs. Language modelling is both more critical and more difficult in mixed initiative and user initiative systems. With these dialogue strategies, the user has a much wider range of potential utterances than with system initiative. This makes the recognition problem harder and the language model larger. In what follows, we will be mainly concerned with systems that use a CFG-based language model and a mixed initiative dialogue strategy. Since these systems need to be able to respond to a fairly free range of user input, constructing the language model is usually a non-trivial task. If it is too constrained, the system will reject many of the user's utterances; if it is too loose, the system will make too many recognition errors.

One way to build grammar-based language models is just to apply the techniques developed for building other types of grammars, in

particular those used for parsing and generation. Unfortunately, it is non-trivial to transform a grammar written in a high-level linguistic formalism into a useful CFG-based language model. In fact, the only really successful piece of work we are aware of that has taken this path is the approach pioneered by the Gemini (Dowding et al., 1993; Moore, 1999) and CommandTalk (Moore et al., 1997; Stent et al., 1999) projects at SRI. Although the idea is promising, it is still not clear that it scales up well to large grammars (Rayner et al., 2000b). Rather than use the type of methods developed under Gemini, most people have instead adopted a simpler and more pragmatic approach. Grammars are developed in an *ad hoc* way directly in CFG, and without reference to linguistic principles. However dubious this may be from a theoretical point of view, experience shows that useful grammars can be quickly developed even for quite complex domains.

In the current paper, we present an empirical study which contrasts linguistically motivated and *ad hoc* approaches to language model construction. In order to be able to make clear comparisons, we focus specifically on the single topic of grammatical agreement. This is a phenomenon that is relevant at least to some extent in nearly all domains, and is easy to model in high-level grammatical frameworks. In contrast, modelling agreement directly in CFG is rather painful, and as far as we can tell most commercially deployed speech applications choose not to do so.

There are only two important examples of agreement in English: between subject and verb in clauses, and between determiner and noun in NPs. In contrast, agreement is a central phenomenon in many languages like French, Spanish, Italian, German, Swedish, Russian and Greek. For example, in French there is agreement between subject and verb and determiner and noun as in English, and also (among other things) agreement between nouns and adjectives, subjects and past participles, and subjects and some adverbials. Note that different inflected forms of a word can often sound fairly different. For example, in Swedish the plural form of an adjective normally adds an "a": thus *"röd"* ("red", singular) is a monosyllable, while *"röda"* ("red", plural) is a disyllable. In French,

the feminine form of an adjective adds an "e", which again can affect the pronunciation. For example, *"vert"* ("green", masculine) doesn't sound the "t", but *"verte"* ("green", feminine) does sound the "t".

The rest of the paper is structured as follows. In Section 2, we describe three implemented systems, two for English and one for Swedish, which all use mixed-initiative strategies and CFG-based language models. The language models for two of these systems were developed directly in CFG using *ad hoc* methods, and in particular ignoring grammatical agreement. The third system was developed in a high-level formalism and then compiled down to CFG; this grammar used a principled linguistic approach, which in particular took careful account of agreement. For each system, we constructed a second version of the language model, which embodied the converse approach to agreement. Thus for the *ad hoc* systems we constructed versions of the grammars modified to include all relevant agreement constraints, and for the theoretically motivated system we constructed a simplified version of the language model in which all the agreement constraints had been removed.

For each of the three systems, we collected and transcribed a sizeable corpus of recorded utterances. Section 3 describes experiments in which the performance of the different versions of each language model were evaluated empirically on the domain corpora. The final section discusses the significance of these results and concludes.

2 Base systems

We carried out experiments on the following three systems:

On/Off House An *ad hoc* system for English in a home automation domain.

Advanced House An *ad hoc* system for Swedish, also in a home automation domain.

Simulated Personal Satellite Assistant A linguistically motivated system for English in a robotics domain.

In the rest of the section, we describe the systems and their language models in more detail.

2.1 On/Off House

The On/Off House (OOH) system is implemented using the Nuance Toolkit platform (Nuance Communications, 2001), and offers English spoken language control, via telephone, of about 20 devices in a simulated home. Device states can only be "on" or "off". The dialogue manager is implemented in Visual C++ using the Nuance DialogueBuilder API. The mode of operation is primarily user-initiative. The system offers coverage of a fairly broad range of language, including commands ("Turn on the heater", "Turn off the light in the bathroom"), several types of questions ("Is the heater switched on?"; "What is there in the kitchen?"; "Where is the washing machine?"; "Could you tell me which lights are on?"), universal quantification ("Switch off everything in the bathroom"), conjunction ("Are the hall and kitchen lights switched on?"; "Switch off the radio, TV and computer"), ellipsis ("Turn on the cooker"... "now the microwave") and pronouns ("Switch off the stereo and the hi-fi"... "switch them on again"). The system has been tuned over four or five iterations of user testing, and performs well enough to have been successfully demonstrated in public on several occasions. All spoken utterances input to the system during development, testing and demos have been recorded and transcribed, resulting in a speech corpus of 3975 sentences.

The main focus of interest for the purposes of the present paper is the grammar. It contains 346 context-free rules; of these, 113 are "lexical" (i.e. have only terminal symbols on their right-hand sides), while the remaining 233 are "grammatical" (i.e. contain at least one non-terminal on the right-hand side).

The grammar is implemented in Nuance Toolkit Grammar Specification Language (GSL; (Nuance Communications, 1999)), and directly encodes a simple slot-value semantics, using a set of 15 slots. These include slots for the type of utterance (command/query), the type of device (light/heater/TV/...), the location (kitchen/bedroom/bathroom/...), the polarity (on/off), and the quantifier (existential/universal). Because of the quantifier slot, determiners ("a"/"the"/"all the"/...) can affect the semantic value of an utterance; this is necessary in order, for example, to distinguish
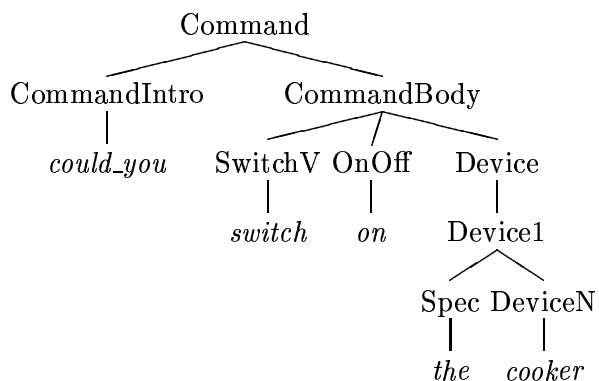

Figure 1: Parse tree for "could you switch on the cooker"

"switch on the light" from "switch on all the lights".

To give the flavour of the grammar, Figure 1 presents a slightly simplified derivation of the utterance "could you switch on the cooker". This yields the semantic representation

```
<operation command> <onoff on>
<device1 cooker> <spec1 existential>
```

The grammar fails to enforce agreement in number between subject and verb, or head noun and determiner; thus for example it will accept sentences such as "*are any light switched on?" or "*switch off all the heater".

2.2 Advanced House

Advanced House (AH) is essentially an extended Swedish version of the OOH system described in the preceding section. It is implemented using the same platforms (Nuance Toolkit and the DialogueBuilder API), and offers all the functionality provided by OOH. Apart from the change of language, there are two important enhancements. Firstly, AH supports interfaces to devices whose state can be represented as a scalar variable, such as dimmer switches and temperature sensors. Secondly, the devices in question are real as opposed to simulated; they are controlled via a Java servlet interface to a LonWorks device network (Echelon Corporation, 2001). AH is currently in an initial testing phase, and has not yet been as thoroughly debugged as OOH. All utterances input to the system during development and testing have been recorded and transcribed, yielding a corpus which at the time of writing contains 1039 sentences.

As above, our main interest is in the grammar. Coverage is analogous to that of the OOH system, with additional constructions to cover control and querying of scalar devices, e.g. *"Sänk lampan i köket till femtio procent"* (Dim the light in the kitchen to 50 percent); *"Ytterligare tio procent"* (another ten percent); *"Hur många grader är det i kylskåpet?"* (How many degrees is it in the fridge?) The grammar contains a total of 448 context-free productions, of which 192 are lexical and 256 non-lexical. It also resembles the OOH grammar in that it fails to enforce any kind of agreement constraints. In Swedish, the types of agreement relevant to this domain are between subject and adjective and between determiner and head noun; agreement is with respect to both number (singular/plural) and gender (common/neuter). Swedish also marks nouns for definiteness, and certain grammatical contexts require specifically definite or indefinite nouns; the AH grammar fails to model any of the constraints associated with definiteness. Finally, the infinitive and imperative forms of Swedish verbs are in general distinct; the grammar once again treats them interchangeably. The sentences in Figure 2 exemplify common types of utterance incorrectly accepted by the grammar.

2.3 Simulated Personal Satellite Assistant

The Personal Satellite Assistant (PSA; (NASA Ames Research Center, 2001)) is a small robot currently being developed at NASA Ames Research Center, designed for use in microgravity and targeted for deployment on the International Space Station. Because of the mobility of the robot and the problems of microgravity, English spoken language dialogue is favored as its primary interface mode. The PSA simulator (Rayner et al., 2000a) was constructed to aid in development and testing of the spoken dialogue interface. The simulated robot can be instructed to move around a diagram of the space shuttle and measure environmental factors such as temperature and carbon dioxide. There are also doors in the simulation that can be opened and closed, and fans that can be turned on and off via spoken commands. The system is implemented as a suite of about 20 agents running under the Open Agent Architecture (OAA; (Martin et al., 1998)), and runs on a high-end SUN

stäng av lampa
(switch off light-INDEF-SING)
switch off the light

släck alla lampa
(switch off all the light-INDEF-SING)
switch off all the lights

är lampan tänt
(is light-DEF-COMMON lit-NEUTER)
is the light switched on?

finns det någon element i köket
(is there any-COMMON radiator-NEUTER in kitchen-DEF)
is there any radiator in the kitchen?

sänka belysningen till hälften
(lower-INF lighting-DEF to half-DEF)
dim the lighting to a half

Figure 2: Typical ungrammatical utterances accepted by the AH grammar

workstation; it is a mature prototype, which has been publicly demonstrated on numerous occasions. The tests described below were carried out on a transcribed corpus of 6261 utterances, collected according to a well-defined protocol described in (James et al., 2000).

Speech recognition in the simulated PSA is once again performed using the Nuance Toolkit, together with a CFG language model. In contrast to the two previous systems, the PSA's language model is compiled from a linguistically motivated general unification grammar for English, using the SRI Gemini compiler (Moore, 1999). The unification grammar formalism provides an extremely compact description of a broad range of linguistic constructions; it currently contains 60 rules and 327 lexical entries, but after compilation expands these to over 10 000 CFG rules. Coverage includes commands ("Go to flight deck"; "Turn on the fan at storage lockers"), WH- and Y-N questions ("What is the pressure at pilot's seat?"; "Is the radiation level increasing?"), past tense and reference to past times ("What was the carbon dioxide level at crew hatch at fifteen oh five?"), numbered objects ("Start scenario three"), conjunction of

	English OOH		Swedish AH	
	Non-lex	Lex	Non-lex	Lex
No Agr.	233	113	256	192
Agr.	390	114	326	235

Table 1: Numbers of non-lexical and lexical CFG rules for the two versions of the OOH and AH systems

both NPs and clauses ("Measure pressure at crew hatch and flight deck"; "Go to crew hatch and switch on the fan"), ellipsis ("Measure pressure"... "how about temperature"), and use of pronouns ("Open the crew hatch"... "close it").

The CFG language model produced by Gemini compilation is only used for recognition. Parsing as such is performed on the recognised string, using the Gemini parser, and produces a semantic representation in a version of Quasi Logical Form (van Eijck and Moore, 1992). The unification grammar, and issues relating to its compilation to CFG form, are described in detail in (Rayner et al., 2000b).

3 Experiments

In order to carry out the experiments described here, we constructed alternate versions of each of the three systems described above. For the *ad hoc* OOH and AH systems, we manually constructed new versions of the CFG language models modified to include agreement constraints. The new versions of the grammars were substantially larger, as shown in Table 1; the worst problem, however, was that many rules from the original grammar had to be duplicated in two or more slightly differing forms, greatly complicating the task of future grammar maintenance. In contrast, it was extremely easy to build a version of the PSA unification grammar in which the agreement constraints had been removed; this only involved editing out a handful of feature specifications. We were not able to measure directly the number of context-free productions in the two versions of the PSA grammar, but the size of the generated Nuance grammar decreased from 31K lines of code to 24K, a difference in size comparable with that observed in the English OOH system.

Our basic plan was to measure performance of each grammar on its appropriate corpus using the Nuance `batchrec` tool, which returns infor-

mation about word error rate (WER), sentence error rate (SER), and semantic error rate (Sem). For the OOH and AH systems, we use the standard Nuance definition of semantic error rate as being the proportion of utterances which receive an incorrect slot-level representation. Since the PSA grammar returns strings rather than semantic representations, we measured semantic error rate for it on the QLF representations produced by the subsequent parsing stage, using a specially constructed tool. We set the Nuance recognition parameters to maximize the proportion of utterances which produced a recognition result: with normal settings, sentences outside grammar coverage are most frequently rejected. In this way, we maximize the return of information on the out-of-coverage portion of the corpus.

Since performance on in-grammar and out-of-grammar sentences is still very different, it makes sense to consider them separately. We have two grammars for each corpus, a tight grammar implementing agreement constraints and a loose grammar failing to do so. This naturally splits each corpus into the following four pieces:

Both Utterances inside coverage of both grammars.

Loose-only Utterances inside the coverage of the loose grammar, but not of the tight grammar.

Tight-only Utterances inside the coverage of the tight grammar, but not of the loose grammar. (If the grammars are correctly implemented, this set should be empty).

Neither Utterances inside coverage of neither grammar.

Table 2 shows the relevant breakdown for each of our three corpora. As expected, the "tight-only" portion of each corpus turns out to be empty.

In the context of an interactive spoken language system, the in-coverage part of the corpus is the practically interesting one. Table 3 presents performance figures for the portion of each corpus that is within coverage of both grammars. In terms of word error rate, the grammar which enforces agreement constraints scores significantly better in the two En-

OOH	Words	Sentences
Both grammars	16067	3511
Neither grammar	2161	437
Loose gram. only	205	27
Tight gram. only	0	0
Total	18433	3975

AH	Words	Sentences
Both grammars	2535	691
Neither grammar	1482	345
Loose gram. only	5	3
Tight gram. only	0	0
Total	4022	1039

PSA	Words	Sentences
Both grammars	23051	5676
Neither grammar	3937	585
Loose gram. only	10	2
Tight gram. only	0	0
Total	26998	6263

Table 2: Sizes of sub-corpora for the systems

OOH	WER	SER	Sem.
No Agr.	10.86%	25.06%	14.64%
Agr.	9.01%	19.71%	10.17%

AH	WER	SER	Sem.
No Agr.	12.47%	24.89%	12.74%
Agr.	6.71%	10.85%	10.27%

PSA	WER	SER	Sem.
No Agr.	13.73%	29.03%	27.63%
Agr.	11.47%	23.24%	22.11%

Table 3: Error rates for both versions of the systems on utterances within coverage of both grammars

OOH	WER	SER	Sem.
No Agr.	58.72%	95.19%	n/a
Agr.	58.54%	95.19%	n/a

AH	WER	SER	Sem.
No Agr.	68.56%	98.84%	n/a
Agr.	66.80%	98.55%	n/a

PSA	WER	SER	Sem.
No Agr.	52.78%	95.90%	n/a
Agr.	53.49%	95.04%	n/a

Table 4: Error rates for both versions of the systems on utterances within coverage of neither grammar

	WER	SER	Sem.
No Agr.	14.15%	74.07%	22.22%
Agr.	20.98%	100.00%	40.74%

Table 5: Error rates for both versions of the OOH system on utterances within coverage of loose grammar only

glish language systems (relative improvements of 17% for OOH and 16.5% for PSA), a difference which increases to a striking 46% for the Swedish language system. In view of the fact that agreement is considerably more important in Swedish than in English, this disparity is not surprising.

A similar pattern is displayed in the sentence error rate figures, with relative improvements of about 20% for the two English systems, and 56% for the Swedish one. For completeness, we also present in Table 4 results for utterances outside coverage of either grammar; these display little interesting variation. Finally, Table 5 gives the results for the 27 OOH utterances inside coverage of the loose but not the tight grammar. Unsurprisingly, the loose grammar performs much better on this set.

The most critical measure of performance for systems of this kind is the semantic error rate, which is only meaningful for the in-coverage portion of the corpus. At first sight, the figures in Table 3 also appear to show a marked improvement in favour of the tight grammar, with proportional reductions of 31% for OOH, 20% for AH, and 20% for PSA. We were somewhat surprised to see OOH scoring so much better than the other two systems, and carried out a detailed item-by-item comparison of the OOH data. This revealed that a substantial proportion of the OOH improvement was essentially spurious; for technical reasons relating to the way in which NP conjunction was modelled, many sentences where the tight grammar won were examples like

switch on the light and the TV

where the loose grammar incorrectly recognised

switch on the light and TV

Although this technically counts as a reduction of the semantic error rate, it is obviously of little practical importance. After eliminating all examples of the above type, we were left with a residue of 47 utterances where one

grammar was right and the other wrong; of these, the tight grammar was correct in 37 cases and the loose one in the remaining 10. A more realistic estimate of the absolute reduction in semantic error rate for the OOH system as a result of correctly modelling agreement would thus be $(37 - 10)/3511$, or 0.7%, giving a relative reduction of about 5%. Although undramatic, this margin is significant at the $P = 0.0005$ level according to the McNemar sign test (McNemar, 1947). The following examples show typical instances of the tight grammar (T) outscoring the loose one (L):

T: turn them off
L: is them off

T: put the computer on
L: are the computer on

T: what is switched on in the kitchen
L: what fridge are in the kitchen

We carried out a similar item-by-item comparison of the tight and loose grammars for the Swedish AH system. This time, there were 20 utterances for which the result in one grammar was clearly correct and the other clearly incorrect, dividing 19-1 in favour of the tight grammar. The revised estimate of reduction in semantic error rate is thus $(19-1)/691 = 2.6\%$ absolute, or a more substantial 20% relative. This result is also significant at the $P = 0.0005$ level according to the McNemar test.

4 Conclusions and further directions

When we began work on the experiments described here, we felt that there were two competing positions concerning the question of whether or not it was important to include agreement constraints in CFG-based language models. From a research-oriented theoretical standpoint, we believed that it would be impossible to ignore a linguistic phenomenon as central as grammatical agreement without incurring some significant penalty. Practical implementation experience however pointed in the opposite direction: most commercial systems fail to take account of agreement, and achieve adequate performance using mostly semantic and domain constraints. In addition,

it is certainly the case that that people sometimes break agreement constraints in spontaneous spoken dialogue. It seemed possible to us that this would happen sufficiently often that a language model which failed to enforce agreement constraints would outperform one that did.

With the results in front of us, we think that both sides can lay some claim to being right. There is indeed a very significant improvement in performance when agreement constraints are added to a grammar. In particular, we found no evidence to suggest that it would be advantageous to refrain from modelling agreement in order to get better recognition of ungrammatical utterances. On the other hand, the improvement due to inclusion of agreement constraints is manifested most strongly at the surface level, as measured by WER and SER. Although the difference is still strongly significant at the level of semantic representation, the absolute increase in semantic accuracy is fairly small. A practical system builder can reasonably wonder whether these gains are enough to motivate the considerable extra implementation burden created by adding agreement to a hand-coded CFG language model.

If the CFG model is compiled from a high-level description, however, there is little additional work to be done, and it is obviously desirable to include agreement constraints in the grammar. This can be interpreted as a strong argument in favour of using high-level linguistic descriptions as opposed to hand-coded CFG models; partly for these reasons, we have in fact recently rewritten the AH grammar in a unification-grammar formalism, compiling it down to CFG using a Gemini-like tool. This work, and further issues arising from it, will be reported soon elsewhere.

Acknowledgements

References

BeVocal, 2001. *BeVocal*. http://www.bevocal.com/index.html. As of 7 February 2001.

M. Cohen, Z. Rivlin, and H. Bratt. 1995. Speech recognition in the ATIS domain using multiple knowledge sources. In *Proceedings of the Spoken Language Systems Technology Workshop*, pages 257–260.

J. Dowding, M. Gawron, D. Appelt, L. Cherny, R. Moore, and D. Moran. 1993. Gemini: A natural language system for spoken language understanding. In *Proceedings of the Thirty-First Annual Meeting of the Association for Computational Linguistics*.

Echelon Corporation, 2001. *LON-WORKS CORE TECHNOLOGY*. http://www.echelon.com/Products/Core/. As of 7 February 2001.

HeyAnita, 2001. *HeyAnita*. http://heyanita.com/html/main/index.html. As of 7 February 2001.

F. James, M. Rayner, and B.A. Hockey. 2000. Accuracy, coverage, and speed: What do they mean to users? In *CHI 2000 Workshop on Natural-Language Interaction*.

D. Martin, A. Cheyer, and D. Moran. 1998. Building distributed software systems with the open agent architecture. In *Proceedings of the Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, Blackpool, Lancashire, UK.

Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

R. Moore, J. Dowding, H. Bratt, J. Gawron, Y. Gorfu, and A. Cheyer. 1997. CommandTalk: A spoken-language interface for battlefield simulations. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 1–7.

R. Moore. 1999. Using natural language knowledge sources in speech recognition. In Keith Ponting, editor, *Speech Pattern Processing*. Springer-Verlag.

R. De Mori and R. Kuhn. 1991. Some results on stochastic language modelling. In *Speech and Natural Language Workshop*, pages 225–230.

NASA Ames Research Center, 2001. *Personal Satellite Assistant (PSA) Project*. http://ic.arc.nasa.gov/ic/psa/. As of 7 February 2001.

Nuance Communications, 1999. *Nuance Speech Recognition System Developer's Manual version 6.2*. 1380 Willow Road, Menlo Park, CA 94025.

Nuance Communications, 2001. *Nuance Home*. http://www.nuance.com. As of 7 February 2001.

M. Rayner, B.A. Hockey, and F. James. 2000a. A compact architecture for dialogue management based on scripts and meta-outputs. In *Proceedings of Applied Natural Language Processing (ANLP)*.

M. Rayner, B.A. Hockey, F. James, E.O. Bratt, S. Goldwater, and J.M. Gawron. 2000b. Compiling language models from a linguistically motivated unification grammar. In *Proceedings of COLING 2000*.

R. Rosenfeld and X. Huang. 1992. Improvements in stochastic language modeling. In *Speech and Natural Language Workshop*, pages 107–111.

SpeechWorks International, 2001. *SpeechWorks International*. http://www.speechworks.com. As of 7 February 2001.

A. Stent, J. Dowding, J. Gawron, E. Bratt, and R. Moore. 1999. The CommandTalk spoken dialogue system. In *Proceedings of the Thirty-Seventh Annual Meeting of the Association for Computational Linguistics*, pages 183–190.

Tellme, 2001. *Tellme*. http://www.tellme.com/. As of 7 February 2001.

J. van Eijck and R. Moore. 1992. Semantic rules for English. In H. Alshawi, editor, *The Core Language Engine*. MIT Press.

VoiceXML Forum, 2001. *VoiceXML Forum*. http://www.voicexml.org/index.html. As of 7 February 2001.

W3C, 2001. *Speech Recognition Grammar Specification*. http://www.w3.org/TR/grammar-spec. As of 7 February 2001.

W. Ward and S. Issar. 1995. The CMU ATIS system. In *Spoken Language System Technology Workshop*, pages 249–251.