Squibs

Fruit Carts: A Domain and Corpus for Research in Dialogue Systems and Psycholinguistics

Gregory Aist*
Iowa State University

Ellen Campana**
Arizona State University

James Allen†
University of Rochester

Mary Swift‡
University of Rochester

Michael K. Tanenhaus§
University of Rochester

We describe a novel domain, Fruit Carts, aimed at eliciting human language production for the twin purposes of (a) dialogue system research and development and (b) psycholinguistic research. Fruit Carts contains five tasks: choosing a cart, placing it on a map, painting the cart, rotating the cart, and filling the cart with fruit. Fruit Carts has been used for research in psycholinguistics and in dialogue systems. Based on these experiences, we discuss how well the Fruit Carts domain meets four desired features: unscripted, context-constrained, controllable difficulty, and separability into semi-independent subdialogues. We describe the domain in sufficient detail to allow others to replicate it; researchers interested in using the corpora themselves are encouraged to contact the authors directly.

1. Introduction and Relation to Prior Work

Dialogue system research, like much else in computational linguistics, has greatly benefited from corpora of natural speech. With notable exceptions (e.g. the Edinburgh Maptask, Anderson et al. [1991]), these corpora consist of samples annotated with linguistic properties (e.g. POS, syntax, discourse status) setting aside the visual and

* 206 Ross Hall, Iowa State University, Ames, Iowa 50011. E-mail: `gregory.aist@alumni.cmu.edu`.
** 240c Matthews Center, Arizona State University, Tempe, Arizona 85287. E-mail: `ellen.campana@asu.edu`.
† 721 CSB, University of Rochester, Rochester, New York 14627. Email: `james@cs.rochester.edu`.
‡ 732 CSB, University of Rochester, Rochester, New York 14627. E-mail: `swift@cs.rochester.edu`.
§ 420 Meliora, University of Rochester, Rochester, New York 14627. E-mail: `mtan@bcs.rochester.edu`.

Submission received: 24 August 2009; revised submission received: 6 May 2010; accepted for publication: 20 September 2010.

pragmatic aspects of the context in which they occurred. In recent years natural language processing (NLP) researchers have been working to incorporate visual and other context into their models and systems (DeVault and Stone 2004; Gabsdil and Lemon 2004; Schuler, Wu, and Schwartz 2009). This is consistent with the growing evidence in psycholinguistics that human language production crucially depends on such aspects of context. To take this NLP research further, there is a need for more corpora that include both variation in, and annotation of, visual and pragmatic context.

There are still many open questions that span computational linguistics and psycholinguistics concerning how natural language and context are related. One core question at the intersection of these areas is how the inherent difficulty of describing an end-goal (i.e., its codability) will affect the structure and content of referring expressions and the referential strategy speakers adopt. Referential strategies are a topic of growing interest in natural language generation. In recent work, Viethen and Dale (2006) demonstrated that even when describing simple grid layouts, people adopt different referential strategies, due perhaps to proximity to landmarks (and hence codability): *the orange drawer below the two yellow drawers*, in contrast to *the yellow drawer in the third column from the left second from the top*. For systems to produce humanlike references in these situations, existing methods of reference generation will need to be modified or extended to include better models of the choice of referential strategies (Viethen and Dale 2006). Such models can also be expected to improve reference resolution: If better predictions can be made about what people will say in a given situation, automatic speech recognition language models can be tighter, NLP grammars can be smaller, and unlikely parses can be avoided, improving both speed and accuracy.

Recent psycholinguistic research suggests that codability does play a role in human reference production (e.g., Cook, Jaeger, and Tanenhaus 2009). This work has largely focused on timing, signals of production difficulty (e.g., disfluency, gesture), and the content of referring expressions (e.g., adjectives, pronouns). There has been much less consideration of how entire referential strategies might systematically vary with codability. A corpus with the correct design and structure will allow for investigation of the more well-studied aspects as well as higher-level factors such as strategy choice, and possible interactions between them.

With these considerations in mind, we designed a domain, Fruit Carts, and a set of corresponding tasks in order to elicit human language production for two purposes: 1) the testing of psycholinguistic hypotheses, specifically that object complexity modulates referential strategy, and more generally the exploration of the relationship between visual context and human–human dialogue, and 2) research and development of dialogue systems that understand language as it unfolds, taking pragmatic factors into account early in the recognition process. By designing with both fields in mind we hope to strengthen the long tradition of cross-fertilization between the disciplines (e.g., Brennan 1991), particularly for task- or game-oriented systems and domains, with a visual component.

We identified four important features to build into the domain. First, the language produced should be completely unscripted: Participants should be able to perform the task with a general description of what to do (e.g., *Give instructions on how to make the map on the screen look like the map in your hand*) and zero prior examples of what to say. For psycholinguistics, this makes the language natural speech rather than speech that is restricted by the instructions or by prior examples. For dialogue systems, this makes the language "untrained" rather than the result of careful training, meaning that systems will be processing language that is representative of what speakers are likely to produce when they use the system, especially without extensive training. Second, the language should be fairly well constrained by context. For psycholinguistics, this makes the language more straightforward to analyze and also more directly tied to the visual context and thus amenable to "visual world" studies that use eye movements to examine real-time production (Griffin and Bock 2000) and comprehension (Tanenhaus et al. 1995). For dialogue systems, this makes the language more amenable to automatic processing and also facilitates the integration of different types of knowledge into the recognition process. Third, it should be possible to vary the difficulty of the tasks. For psycholinguistics, this makes hypotheses about the effect of task difficulty on language production amenable to study. For dialogue systems, this allows the resulting corpora to have a combination of relatively easy tasks ("low-hanging fruit") and more difficult NLP challenges. Fourth, the domain should support the collection of dialogues that are separable into partially or semi-independent subdialogues, with limited need for reference to previous subdialogues. For psycholinguistics, this makes each subdialogue a separate trial, allowing for analyses where trials are treated as random effects in mixed-effect regression models or repeated measures in ANOVAs. For dialogue systems, this limits the likelihood that errors in processing one subdialogue will spill over and affect processing of subsequent subdialogues. For both research areas, this separability constraint enables within-subject experiments with each subdialogue as a trial.

In purpose and approach, Fruit Carts is most similar to the Map Task (Anderson et al. 1991); both are simultaneously a set of experiments on language and a corpus used for developing language processing systems. Map Task dialogues "are unscripted [but] the corpus as a whole comprises a large, carefully controlled elicitation exercise" (Anderson et al. 1991, page 352) that has been used in many computational endeavors as well. Fruit Carts was guided by our twin goals of furthering the development of spoken language systems, and providing a psycholinguistic test bed in which to test specific hypotheses about human language production. Fruit Carts differs from Map Task in terms of dynamic object properties and in terms of the information available to the speaker and hearer. In the Map Task, objects have fixed properties that differ between giver and follower, yet remain constant while the path is constructed. In Fruit Carts, objects have properties that can be changed: position, angle, and color. This allows for a wide variety of linguistic behavior which in turn supports detailed exploration of continuous understanding by humans and machines. In the Map Task, the participants' screens differ, whereas in Fruit Carts the speaker and hearer share the same visual context, which simplifies the analysis and interpretation of results (Figure 1).

2. Fruit Carts Domain and Tasks

The **Fruit Carts domain** has three screen areas: a map, an object bin, and a controls panel. Each area was designed in part to elicit the types of expressions that require continuous understanding to approximate human behavior such as progressive restriction of a reference set throughout the utterance.

The map contains named regions divided by solid lines, with three flags as landmarks. The region names did not appear on the screen, to preclude use of spelling in referring expressions (*the C in Central Park*). Names were chosen to be phonetically distinct. To support progressive restriction of potential regions, regions whose initial portions overlap are adjacent (*Morn* identifies Morningside and Morningside Heights) and some regions have flags and others not (*put the square on the flag in...* identifies the regions with flags.) No compass is displayed, in an attempt to limit the directions elicited to *up*, *down*, *left*, and *right* and not *north*, *south*, and so on.

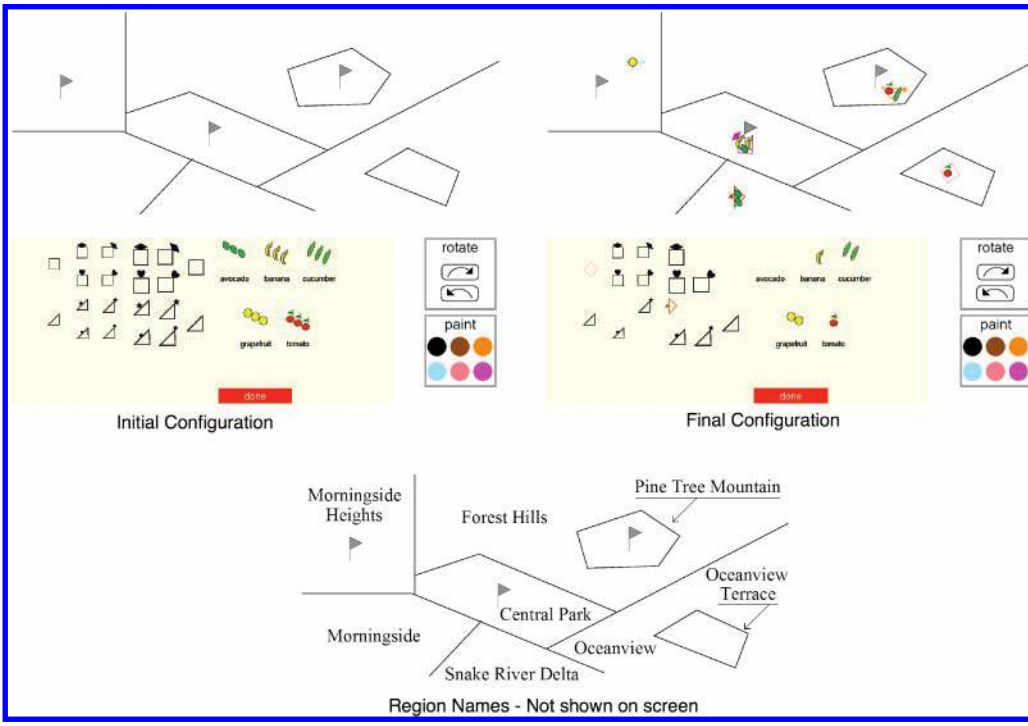

Figure 1
Example initial and final configurations for Fruit Carts domain and corpus. The region names were available to both director and actor (on paper) but were not shown on screen. The final configuration shown is the actual screen after the five dialogues from the participant whose third, fourth, and fifth dialogues are shown in Appendix A.

The object bin contains fruits and carts, by analogy with food vendor carts (e.g., hot dog stands). The fruits are avocados, bananas, cucumbers, grapefruits, and tomatoes, all botanically fruits. We chose fruits because they were nameable, especially with a label, and visually different from the carts. The carts are either squares or triangles, in two sizes, with an optional tag that for squares is either a diamond or a heart and for triangles is either a star or a circle. Adjectives (e.g., *large*, *small*) are commonly used in natural language descriptions and there is a growing body of psycholinguistic research, mostly with scripted utterances, that has used adjectives to investigate real-time language processing (Sedivy et al. 1999; Brown-Schmidt, Campana, and Tanenhaus 2005). Here, to support progressive restriction of potential carts, each component is easy to name but the entire shape requires a complex description rather than a prenominal modifier—or at least strongly prefers one, as no examples to the contrary were observed in the Fruit Carts corpus described later in this article. That is, whereas a square with stripes could be either *the square with stripes* or *the striped square*, a square with a diamond on the corner is *the square with a diamond on the corner* but not **the corner-diamonded square*.

The controls panel contains left and right rotation arrows and six paint colors (black, brown, orange, blue, pink, and purple) chosen to be distinct from the colors of the fruit.

Five tasks are included in Fruit Carts, all performed by using a mouse. To CHOOSE a cart, the user clicks on it. To PLACE it on the map, the user drags it there. To PAINT the cart, the user clicks on the desired color. Painting is a uniformly easy control task. To ROTATE the cart, the user presses and holds down the left or right rotation button.

The goal of the rotation tool was to allow arbitrary rotations and to elicit utterances that were in response to visual feedback, such as *rotate it a little to the right*, *more*, *stop*. Finally, to FILL the cart, the user drags fruit to it.

3. Fruit Carts Corpus

For the dual goals of gathering a corpus of utterances for dialogue system research, and testing the hypothesis that object complexity modulates referential strategy in human language production, we designed a set of goal maps that systematically manipulated:

POSITION. Each cart was in a high-codability "easy" position, such as centered on a flag or in a region; or a low-codability "hard" position, such as off-center.

HEADING. Each cart was at an "easy" angle, an integer multiple of 45 degrees from its original orientation; or a "hard" angle, a non-multiple of 45 degrees.

CONTENTS. Each cart contained an "easy" set of objects, fruit of the same type, such as three tomatoes; or a "hard" set of objects, such as two bananas and a grapefruit.

COLOR. Each cart was painted a uniformly "easy" color to provide a control condition.

One person (the director) gave directions to the other (the actor) on how to carry out the task. The director wore a headset microphone that collected speech data; the actor in this set-up wore a head-mounted eye-tracker that collected eye movements. The director (a subject) sat just behind the actor (a confederate); both viewed the same screen. Twelve subjects participated, each of whom specified twenty objects to place on the map; thus, a total of 240 dialogues were collected. The recordings were transcribed word-for-word by a professional transcription service that also provided sentence boundaries. The corpus has been labeled for referential strategy at the utterance level (Aist et al. 2005) and subsequently with referring expressions, spatial relations, and actions in order to support word-by-word incremental interpretation (Gallo et al. 2007); see Appendix A.

4. Analysis with Respect to Desired Features

How well does the Fruit Carts domain meet the desired features described earlier?

1. Unscripted. Subjects were generally able to complete the task with only the instructions to make the screen match their paper map, and no prior examples of what to say, although one subject systematically did not issue instructions to paint the shapes.

2. Constrained. Generally speaking, subjects used the language we expected, such as *square*, *triangle*, and so forth, or high-frequency synonyms such as *box* for a square cart (from the first dialogue of the participant in Appendix A, omitted for space) or *dot* for a circle tag (Appendix A, [D3]). There were examples of participants using unexpected expressions, such as calling an avocado a *lime*, despite the on-screen label. Yet overall the language was well constrained by the context.

3. Support for varying of task difficulty. As the Fruit Carts corpus showed, location, heading, and contents of carts can be systematically varied; later corpora, outside the scope of this article, have varied the number of carts placed together in order to construct simple or compound objects, in order to test the hypothesis that higher-level task and goal knowledge (e.g. a tower is being built from several blocks) modulates language production, and to support further dialogue system research.

4. Support for collection of semi-independent subdialogues. Here the Fruit Carts domain excels. Due to the presence of multiple separate objects and regions, different subdialogues can make use of different objects, regions, properties, and so forth. By contrast, a domain revolving around construction of a single complex target, such as a landscaping plan, would have licensed substantial amounts of reference to previously placed objects including objects not in place at the time the dialogue began—making subdialogues dependent on each other in terms of accuracy, correctness, and so forth. As Appendix A illustrates, these Fruit Carts data contain relatively few such references. This is analogous to the difference between a math exercise set that contains several independent exercises, and a set where each exercise builds on previous answers.

5. Use in Research

For **dialogue systems research**, the Fruit Carts domain has already been useful in developing dialogue systems that understand language continuously while taking pragmatics into account. For example, using Fruit Carts, incorporating pragmatic feedback about the visual world early in the parsing process was shown to substantially improve parsing efficiency as well as allowing parsing decisions to accurately reflect the visual world (Aist et al. 2006). Also using Fruit Carts, a dialogue system using continuous understanding was shown to be faster than, and preferred to, a counterpart that used a traditional pipeline architecture but was otherwise identical (Aist et al. 2007).

For **psycholinguistic research**, Fruit Carts has also been used for studying the relationship between bi-clausal structure and theme complexity (Gallo et al. 2008) and testing hypotheses regarding the relationship of information in a message, resource limitations, and sentence production (Gallo, Jaeger, and Smyth 2008).

6. Discussion and Conclusions

Fruit Carts also has a number of other advantages as well as some limitations.

First, Fruit Carts provides ample temporary or **local ambiguity** in its utterances, a central challenge for continuous understanding systems and a classic target of research in psycholinguistics (for a review see Altmann [1998]). In a typical sequence such as *okay take a ... small triangle with a dot on the corner* (Appendix A, [D3]), most of the content words and some of the function words serve to resolve local ambiguity:

okay take... – uniquely identifies an action

...a ... small... – restricts (partially disambiguates) referential domain to half of the shapes

...triangle... – further restricts the referential domain to the triangles

...with... – further restricts the referential domain to carts with tags

...a dot... – further restricts the referential domain to carts with circles

...on the corner – uniquely identifes one of the twenty carts

Likewise, *flag* in *right ... um ... side of the uh ... flag in pine tree mountain* [D5] restricts regions to flagged regions.

Second, Fruit Carts also elicits substantial **variation in referential strategy.** Some utterances could be grounded independent of context, up to pronominal reference. For example, the hypothetical utterance *Move a large plain square to the flag in Central Park* has a fully specified action, object, and goal, as do *rotate it about 45 degrees* (Appendix A, [D4]), and *and um make that orange* [D5]. We labeled this category "all-at-once." For other utterances, grounding relied on the surrounding context—dialogue and/or task. For example, *um a little to the left* [D4] contains a direction (*left*) but might rely on the last action to identify the intended action as rotation or movement, and on the selected shape on the screen to identify the object. We labeled this category "continuous." Some utterances exhibited "both" all-at-once and continuous properties, or properties of "neither" category. The continuous utterances contained 21% fewer words (mean, 8.72 vs. 6.85) than the all-at-once and contained shorter words, too (mean, 3.95 letters vs. 3.74). About one-third of the utterances were labeled as "continuous"; speakers produced more continuous utterances as task experience increased (Aist et al. 2005).

Finally, Fruit Carts is relatively abstract: The carts are basic shapes such as squares and triangles, and the fruit are chosen for language research purposes. On the one hand, this is desirable because it reduces the possibility of confounding effects from prior knowledge. On the other hand, it would be interesting for future work to extend Fruit Carts-style domains to more realistic object construction and placement tasks.

Appendix A: Example Dialogues

Referential strategy. These dialogues [D3]–[D5] are the third, fourth, and fifth dialogues from one subject, screen one. For conciseness, "..." concatenates some adjacent utterances. All-at-once sections are marked in **bold** and continuous sections in *italics*.

[D3] okay **take a** ... **small triangle with a dot on the corner**

and ... um ... put it ... **it should be in um** ... **kinda the uh** ... **center right side of morningside heights**

uh morningside heights ... **oh** ... *um a little further in* ... uh ... towards the um *oh wait a little back sorry* ... *uh that's good*

and then rotate it to the right until the l- hypotenuse is str- fa- *yeah like that* <laughter>

and then make that blue

and put a uh grapefruit in it so that

it ... **it's touching the left side but sticking out of the top**

oh it should be inside the triangle and touching

um a little ... *over* ... *or down and over a little bit* ... *uh yeah that's good*

um <breath> ... now ... uh

[D4] **take a square** ... **and put it in um** ... **oceanview terrace**

and pretty much in the center

um i don't know which one it i- i guess the s- try the smaller one

um and then uh

rotate it about 45 degrees

um ... oh ... *like one more turn* ... yeah

um and make that ... **pink**

and then put a uh tomato ... **in the** ... *um a little to the left* ...*okay*

good ... um ... it ... i’m not sure if it should be a bigger one that triangle or not

um you can try the bigger triangle ... i mean not the bigger triangle the bigger square ... i think maybe it should be the ... yeah i think it should be the bigger square

<mumble> ... put the *yeah right there*

[D5] and then um ... and put ... um <breath> ... <mumble> ... then put uh get a uh ... <mumble>

take the uh large triangle with the star

and um ... **put that** ... **um to the** ... **right** ... **um** ... **side of the uh** ... **flag in pine tree mountain**

er the right side

and ... <laughter> *um down a little*

um ... **then rotate it so that** ... **the** ... **the hypotenuse is** ... **almost** ... **horizontal but** ... **tilted a little sli-** *like one more rotat- yeah*

and um make that orange

um maybe a little closer to the flag

and down ... *yeah that should be good*

kay um and then put a uh tomato in the right ... er in the left **corner and then a cucumber in the right corner of it**

um ... the tomato should be a l- er um ... not ... quite ... in the corner th- *yeah that’s good and the cucumber should be a little down*

a little more yeah um oh wait that’s a little too much ... uh that sh- *um that’s good*

okay ... that’s it <laughter> ... <laughter>

oh you wanna see this ... <laughter>

i think that’s good ... okay <laughter>

Incremental disambiguation. This example, adapted from Gallo et al. (2007), shows annotation to support disambiguation, here, in *the small box in Morningside*. These are word-level annotations in the smallest possible semantic units, marked at the point of disambiguation with no lookahead, and following the speaker’s intentions (Gallo et al. 2007).

the

anchor(A1)

definite(A1)

small

size(A1, small)

box

objectType(A1, square)

in

anchor(A2)

spatialRelation(A2, inside)

location(A1, A2)

Morningside

anchor(A3)

name(A3)

objectReferent(A3, MorningsideRegion3)

ground(A2, A3)

Message structure. The following example, adapted from Gallo et al. (2008), shows annotation for the purpose of exploring the link between message structure and complexity of the theme.

original: *take a square with a ... square with a heart on the corner*

clean: *take a square with a heart on the corner*

action: SELECT

verb: *take*

theme: *a square with a heart on the corner*

theme disfluency: Yes

theme pause: Yes

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant 0328810. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This publication was partially supported by grant HD 27206 from the NIH. The contents of this report are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

Aist, G. S., J. Allen, E. Campana, L. Galescu, C. Gómez Gallo, S. Stoness, M. Swift, and M. K. Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Proceedings of the 9th International Conference on Spoken Language Processing*, pages 1922–1925, Pittsburgh, PA.

Aist, G. S., J. Allen, E. Campana, C. Gómez Gallo, S. Stoness, M. Swift, and M. K. Tanenhaus. 2007. Incremental dialogue system faster than and preferred to its nonincremental counterpart. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pages 761–766, Nashville, TN.

Aist, G. S., E. Campana, J. Allen, M. Rotondo, M. Swift, and M. K. Tanenhaus. 2005. Variations along the contextual continuum in task-oriented speech. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, pages 79–84, Stresa.

Altmann, G. T. M. 1998. Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2(4):146–152.

Anderson, A., M. Bader, E. Bard, E. Boyle, G. M. Doherty, G. M. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

Brennan, S. E. 1991. Conversation with and through computers. *User Modeling and User-Adapted Interaction*, 1:67–86.

Brown-Schmidt, S., E. Campana, and M. K. Tanenhaus. 2005. Real-time reference resolution in a referential communication task. In J. C. Trueswell and M. K. Tanenhaus, editors, *Processing World-situated Language: Bridging the Language-as-action and Language-as-product Traditions*. MIT Press, Cambridge, MA, pages 153–171.

Cook, S. W., T. F. Jaeger, and M. K. Tanenhaus. 2009. Producing less preferred structures: More gestures, less fluency. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, pages 62–67, Amsterdam.

DeVault, D. and M. Stone. 2004. Interpreting vague utterances in context. In *Proceedings of COLING*, pages 1247–1253, Geneva.

Gabsdil, M. and O. Lemon. 2004. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*, pages 79–84, Barcelona.

Gallo, C. Gómez, G. Aist, J. Allen, W. de Beaumont, S. Coria, W. Gegg-Harrison, J. Paulo Pardal, and M. Swift. 2007. Annotating continuous understanding in a multimodal dialogue corpus. In *Proceedings of the 2007 Workshop on the Semantics and Pragmatics of Dialogue*, pages 75–82, Rovereto.

Gallo, C. Gómez, T. F. Jaeger, J. Allen, and M. Swift. 2008. Production in a multimodal corpus: How speakers communicate complex actions. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2917–2920, Marrakech.

Gallo, C. Gómez, T. F. Jaeger, and R. Smyth. 2008. Incremental syntactic planning across clauses. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 845–850, Washington, DC.

Griffin, Z. M. and K. Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11:274–279.

Schuler, W., S. Wu, and Lane Schwartz. 2009. A framework for fast incremental interpretation during speech decoding. *Computational Linguistics*, 35(3):313–343.

Sedivy, J. E., M. K. Tanenhaus, C. G. Chambers, and G. N. Carlson. 1999. Achieving incremental interpretation through contextual representation: Evidence from the processing of adjectives. *Cognition*, 71:109–147.

Tanenhaus, M. K., M. J. Spivey-Knowlton, K. M. Eberhard, and J. E. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Viethen, J. and R. Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 63–70, Sydney.