INTERSPEECH 2009 BRIGHTON

No Sooner Said Than Done? Testing Incrementality of Semantic Interpretations of Spontaneous Speech

Michaela Atterer, Timo Baumann, David Schlangen

Department of Linguistics, University of Potsdam, Germany

Abstract

Ideally, a spoken dialogue system should react without much delay to a user's utterance. Such a system would already select an object, for instance, before the user has finished her utterance about moving this particular object to a particular place. A prerequisite for such a prompt reaction is that semantic representations are built up on the fly and passed on to other modules. Few approaches to incremental semantics construction exist, and, to our knowledge, none of those has been systematically tested on a spontaneous speech corpus. In this paper, we develop measures to test empirically on transcribed spontaneous speech to what extent we can create semantic interpretation on the fly with an incremental semantic chunker that builds a frame semantics.
Index terms: incrementality, spoken dialogue systems, spontaneous speech, evaluation

1. Introduction

Incremental spoken dialogue systems build up syntactic and semantic structure on the fly while the user is still speaking, in order to compute system reactions with as little delay as possible. As [1] has shown, in highly interactive settings such systems are preferred over non-incremental systems because they can react faster and more naturally. In this paper, we examine to what extent we can build up semantic structure incrementally, using a semantic module and data from a German spontaneous speech corpus. Our data, like any spontaneous speech, contains many ungrammaticalities and hesitations, and hence the semantics component must be highly robust. Previous work on building semantics incrementally focuses more on theoretical aspects of semantic composition than on evaluating performance on spontaneous speech corpora. In this paper, we focus on developing measures for incremental semantic components and evaluating the incrementality of a particular parsing/semantics construction component on a corpus of transcribed German spontaneous speech. We want to find answers to the following questions: Using our semantic component, how much of the semantic interpretation have we built up at what percentage of the utterance? When on average do we know all, and when the first bit of relevant information? We use a robust semantic chunker that fills in slots after processing chunks of text and checking consistency with previously filled slots. The chunker was described in detail elsewhere, but not tested in terms of its incrementality.

The rest of the paper is structured as follows. In Section 2 we recapture some aspects of incremental semantic construction and review the literature on this topic. In Section 3 we propose strategies and measures for evaluating the incrementality of semantic components. Section 4 describes aspects of an incremental semantic component which form the basis for describing its evaluation using the newly proposed measures. They are then employed to this component in Section 5 and results are demonstrated. Sections 6 and 7 contain a general discussion and the conclusion.

2. Incremental semantics construction

A number of aspects have to be considered when incrementally constructing semantics for spoken language.

- When building semantics for spoken language we need to live with disfluencies and other sources for ungrammaticality
- In practical applications we need to live with incorrect ASR
- We have to decide on how strict incrementality can be.[1] ASR-hypotheses might be revised. One way to deal with this is that additions to the semantic structure be reversible. Another possibility is to reduce strictness of incrementality, i. e. to lag behind. A similar decision can be made with respect to ambiguity. Ambiguity can be resolved after a number of semantic representations have been built, or construction of the semantics can be deferred until more information is available. In other words, incrementality becomes less strict the fewer ambiguity we represent at each point in time.

Thus, in the literature roughly two types of semantics construction can be distinguished. Our system—as described later—defers adding semantic information until enough textual material is available to fill a semantic slot. An alternative is to build parallel hypotheses.

The first type is less strict in terms of incrementality (it may lag behind), but does not have to cope with as many ambiguous structures. An example for the first type is the system by [3]. They propose an incremental semantics using free variables: the sentence fraction *move a large triangle to* provokes the semantic representation `move(X,Y)` only when the word *a* has been processed and the representation `move(triangle1,Y)` after the word *to* has been processed. [4] build up Discourse Representation Structures [5] incrementally and do not explicitely deal with ambiguity either.

In contrast, [6] parallely builds up a number of semantic representations. Likewise,the system by [7], when confronted with a partial NP *the white* ⟨*?*⟩ also comes up with a number of hypotheses about which objects might be referred to at this early stage. [8] also represent ambiguous structure but try to limit it for reasons of efficiency. Evaluation is carried out in terms of number of chart edges used and parse time after typing has finished. [9] also try to limit ambiguous structure and build semantics incrementally. However, the evaluation of their system does not include evaluation of incrementality.

Neither of the work cited above evaluates the incrementality of the semantic interpretation on a corpus. What is needed is an empirical evaluation method for systems of various kinds not depending on internal representations such as chart edges and

[1] For a definition of strictness of incrementality cf. [2] and the examples of stricter and less strict representations below.

Copyright © 2009 ISCA

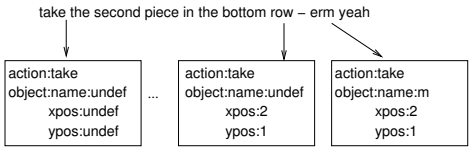

Figure 1: Example of how a frame can change over time.

applicable to varying strictness-levels of incrementality. This is what this paper aims to provide.

3. Evaluating incremental semantic components

One way to evaluate semantic representations for utterances is to compare them to a gold standard, an approved, possibly hand-built representation of the same utterance. This, however, does not capture aspects of the construction process, which in incremental semantics construction is perhaps as important as the final representation.

A more appropriate way to say something about the dynamics of the construction process is to compare semantic output of the system at several points in time. The measures we suggest refer to words in the utterances. As utterances differ in length we normalize the measures by utterance length:

- first correctly-filled representation or, in this case, first correct frame (FCR): when (at which percentage of the utterance) is the frame first completely correct?
- first finally correctly-filled representation (FFR): when (at which percentage of the utterance) is the frame completely correct and doesn't change any more until the end. Notice that FFR is different from FCR because a currently correct frame may intermittedly be changed and only later return to a correct state. Thus FCR $\leq$ FFR.
- first correctly filled slot or, more generally, first correct element (FCE): when is the first slot correctly filled (while no other slot is incorrectly filled)
- degree of correctness (DC): in our case, the percentage of slots correctly filled on average at a certain time

The measures are suitable to compare systems, even if they differ with respect to the tradeoff between ambiguity and strict incrementality. When ambiguity comes into play, FCR should hold when there is at least one correct frame. Systems with strict incrementality will hence be superiour with FCR but inferiour with FCE.[2] The measures can also be used or adapted for other semantic representations, where *elements* can be free variables and predicates instead of slots. In this case FCE would read: *when are all free variables and predicates instantiated?*

Figure 1 shows a constructed example sentence of 10 words, where the first slot is filled after the first word. Later on after the eighth word the frame is correctly filled (this would also be the frame representation given in the gold standard). Then the speaker continues to mumble something which is actually a hesitation 'erm' or 'hm', but which might be recognized as a letter name 'M' by the speech recognition. One of the puzzle pieces in the domain that we are using is also frequently referred to as 'M' or 'W'. Hence another slot in the frame semantics is filled, and the frame becomes false again according to the gold standard.

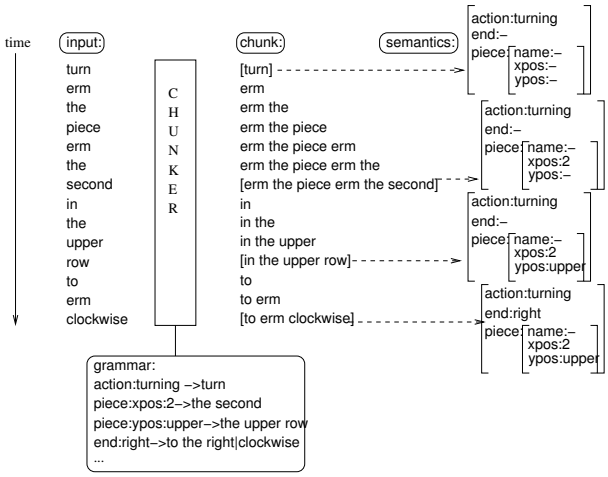

Figure 2: Basic functionality of a semantic component that incrementally breaks up an input string into semantically "valuable" parts and fills in a frame semantics.

In this example FCE will be at 1/10 of the utterance and FCR at 8/10 of the utterance. The FFR measure is not applicable for this example, because the final representation is not correct. As we keep track of the number of utterances at various fractions of utterance-length, the non-applicability will be implicitly shown by a lower over-all number (or a lower curve, cf. results section).

4. Incremental semantic component

Our semantic processing module has been presented in more detail elsewhere [10], but its incrementality has not been investigated and evaluated. In this section we present some of its aspects relevant for evaluating incrementality. We call it a semantic *chunker* because it is based on the idea of semantic units or chunks inspired by the notion of so-called sense-units [11], which correspond to phonological phrases. The original notion of phonological phrases is that they are roughly the lexical heads of a phrase with their preceding function words up to the next head. A phrase consisting of only one word can be united with the preceding one [12]. The chunker collects word material until there is enough semantic information in it to change the state of the semantic frame. Then the current chunk is closed, stored in memory, and the next incoming text is processed until it forms a content-full unit again. Content-full units are defined in a grammar via regular expressions. Thus, the chunker roughly collects non-content-full material up to the content-full material (e. g. *horizontally* or *cross*). We say roughly, because non-content-material can be included in the grammar rules for greater robustness. In the domain, we are working on, puzzle pieces that have letter names can be turned, flipped, moved, etc. The grammar writer might for instance decide to refer to a piece called *M* with $(the|a)$ *M* because that way a confusion with a hesitation *erm* becomes less likely with speech recognition output. Example 1 shows an example of a grammar rule.

(1) ACTION: grasping, END: empty→`nimm|nehme.?`

[2] Recall that FCE does not allow any wrong entries and thus punishes ambiguous representations.

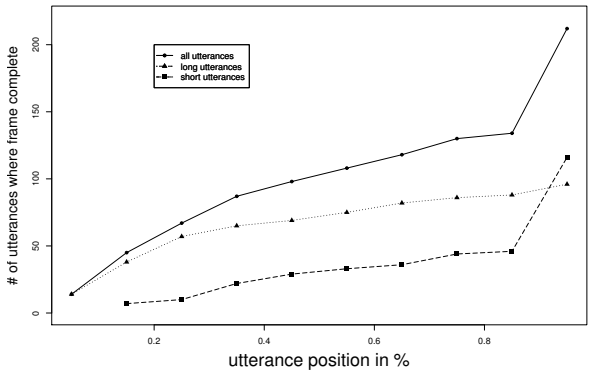

Figure 3: At each relative point in time the graphs show the number of utterances where the frame is correct or was correct at an earlier stage (FCR).

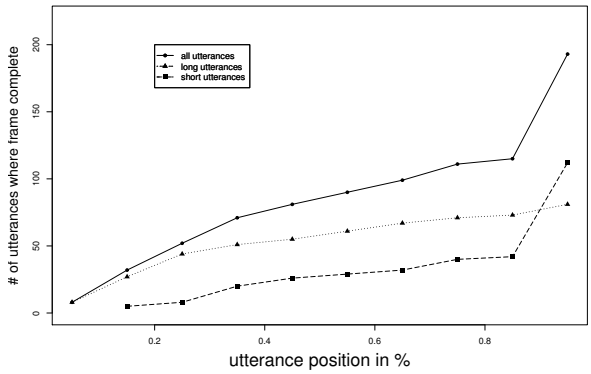

Figure 4: At each relative point in time the graphs show the number of utterances where the frame is finally correct or was finally correct at an earlier stage (FFR).

On the righthand side is a regular expression for the German word *nimm* (take) in a number of variants. To the left are the slots that are filled. If the word occurs the action slot is filled with a grasping action and the end slot is filled with *empty*, because grasping has no end position as opposed to flipping (*horizontal*, *vertical*), turning (*left*, *right*), moving (*into the leg*, *head*, etc. of a figure).

Moreover, the semantic chunker can, to some extent, account for selectional restrictions. The filling of slots with certain entries can be prevented by other "neighbouring" slots if they have already been filled with material that does not fit with the new material. In our domain we have turning and flipping actions. If the action is identified as flipping the end of the action can be *horizontal*, but if the action has been identified as a turning or grasping action, *horizontal* can not be entered in the frame any more, because pieces cannot be *turned* horizontally.

Figure 2 visualizes the basic functionality of the chunker. For more details refer to the aforementioned paper.

5. Evaluating an incremental semantic component

5.1. Gold standard

We used the transcribed speech data from [13]. For our experiments, we created a semantic gold standard by having a human subject annotate each utterance from our transcribed corpus with a 5-slot frame. The human gold standard emulates an 'ideal chunker', or perhaps rather a human chunker. Slots for which no material was present in the utterance remained undefined. In the following we show 2 examples with rough English translations:

- *spiegel es dann einmal vertikal* – flip it then once vertically[3]

$$\begin{bmatrix} \text{ACTION:} & \text{flipping, END: vertical} \\ \text{OBJECT:} & \begin{bmatrix}\text{NAME: pro, XPOS: undef, YPOS: undef}\end{bmatrix} \end{bmatrix}$$

- *also zweite Reihe drittes Teil* – so second row third piece

$$\begin{bmatrix} \text{ACTION:} & \text{undef, END: undef} \\ \text{OBJECT:} & \begin{bmatrix}\text{NAME: undef, XPOS: 3, YPOS: -2}\end{bmatrix} \end{bmatrix}$$

500 utterances were annotated in this way, of which 100 were used for grammar development, and 400 for testing. Following our procedure, some of the frames remained completely empty; e. g. the utterance *That is difficult* is represented by an empty frame, as it contributes nothing to the slot values. There were 68 such utterances which were disregarded in the evaluation (leaving 332 utterances) because the chunker's performance on these is much better than on contentful sentences. After all, we are mostly interested in evaluating the chunker on sentences for which meaningful semantics exist.

For some of the numbers given below, we also divide the utterances by their length: Short utterances contain 10 or less words, long utterances 11 or more. There are 171 short utterances and 161 long utterances in the test corpus.

5.2. Results

Figure 3 shows FCR for the whole test corpus, and for short and long utterances respectively. Figure 3 shows that 63 % of the utterances that are represented by a correct frame according to gold standard at some point, are represented like this before the end of the utterance. It also shows that it is especially long utterances whose semantic content can be obtained at a relatively early stage (72 % FCR after half the utterance). With utterances up to 10 words, the majority can only be fully represented towards the utterance's end. One of the reasons for this is that short utterances are more concise (cf. Example 2), while longer utterances tend to contain more self-corrections and additional dispensable material (cf. Example 3).

(2) ***spiegel*** ***es*** *dann* *einmal* ***vertikal***
flip **it** then once **vertically**

(3) *und* *wird* *in* *den* *Fuß* *des* *lin-* *in* *den* ***linken*** ***Fuß***
and is in the foot of the lef- in the **left** **foot**
des *Kamels* *so* ***eingebaut*** *dass* *es* *genau* *passt*
of the camel so **put** that it exactly fits

To quantify the amount of dispensable material, we conducted an analysis of our test data examining the last chunk in an utterance in cases where none of its information was used, i. e. in such cases where it was considered dispensable material. For short utterances (up to 10 words) the mean length of this chunk was 6.4 words (note that in principle the whole utterance can be the last chunk, if there is no information or the chunker does not find information), (sd = 2.6). For long utterances it was much greater: 19.0 (sd = 8.8). This strengthens our above statement that short utterances are more concise and longer utterances contain additional dispensable material.

[3] Translations for German examples are given word-by-word.

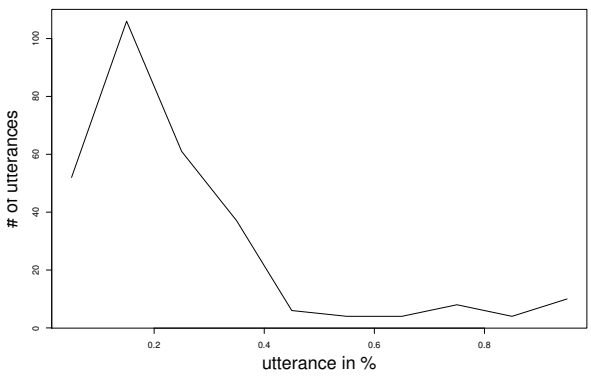

Figure 5: The fraction of the utterance in which the first correct entry occurred on average (FCE).

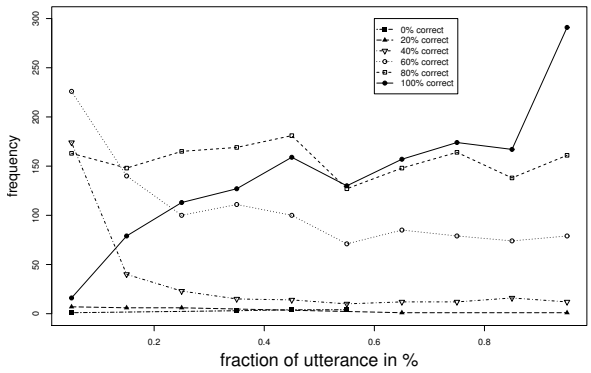

Figure 6: Degree of correctness (DC).

Figure 4 shows FFR, i. e. those utterances, where the frame changes again after it has been recognized correctly for the first time are subtracted from the numbers in Figure 3. As hoped, we obtain roughly the same picture for this measure as before; the curves are only slightly lower. Not unexpectedly, it is the long utterances that contribute to the slight decrease, because it is "additional" speech material after a semantically sufficient statement that can change an already correct frame.

Figure 5 shows FCE. Of the 332 frames, 292 are shown in the graph. The rest did not contain a first correct entry. A first entry was only considered correct if there were no wrong entries at the same time. The figure shows when the first bit of useful information is added to the frame: If this happens it usually happens during the first 40 % of the utterance.

Figure 6 shows the degree of correctness after a certain average fraction of the utterance has been seen. We can see that at around 40 % of the utterance, low degrees of correctness (0 %, 20 %) have been reduced, while high degrees of correctness (100 %) have risen.

6. General discussion

We evaluate the incrementality of the semantics produced by a semantic chunker on transcribed spontaneous speech data. Our analysis gives an idea of what semantic content can be expected at what time. It is a point in favour of incremental semantic interpretation, because it shows that on average we can obtain considerable knowledge about what the speaker says during the first 40 % of her utterance. Of course, we can only claim this for the corpus we are using, and for our frame semantic interpretation. Most of the measures, however, can be used or adapted for other semantic representations (e. g. *when are all free variables and predicates instantiated?*). What is needed is standardized corpora and annotations which allow researchers to compare semantic components. Only then is an empirical comparison between different semantic components and approaches possible. With our measures and corpus annotation we have not found a solution to this problem, but hopefully done a small step towards fruitful discussions and endeavours into that direction.

7. Conclusion

We used transcribed spontaneous speech data to evaluate the incrementality of the semantics produced by a robust semantic chunker. Our analysis shows that on average we have considerable knowledge of what a speaker says at as little as 40 % of her utterance. We contribute a step towards an empirical evaluation of incrementality in semantic components.

Acknowledgements
This work was funded by a DFG grant in the Emmy Noether programme.

8. References

[1] G. Aist, J. Allen, E. Campana, C. G. Gallo, S. Stoness, M. Swift, and M. K. Tanenhaus, "Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods," in *Decalog 2007*, Trento, Italy, 2007.

[2] J. Nivre, "Incrementality in deterministic dependency parsing," in *Proceedings of the ACL Workshop Incremental Parsing: Bringing Engineering and Cognition Together*, F. Keller, S. Clark, M. Crocker, and M. Steedman, Eds. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 50–57.

[3] G. Aist, S. Stoness, and J. Allen, "Steps towards incremental semantics for spoken dialog systems," in *The Third Midwest Computational Linguistics Colloquium MCLC-2006*, Urbana, USA, 2006.

[4] K. Bücher, G. Görz, and B. Ludwig, "Corega tabs: Incremental semantic composition," in *Procs of KI-2002 Workshop on Applications of Description Logics (CEUR Proceedings vol. 63)*, Aachen, Germany, 2002.

[5] H. Kamp and U. Reyle, *From Discourse to Logic*. Dordrecht: Kluwer, 1993.

[6] W. Schuler, "Interleaved semantic interpretaions in environment-based parsing," in *Proc of COLING-2002*, Taipei, Taiwan, 2002.

[7] T. Brick and M. Scheutz, "Incremental natural language processing for HRI," in *HRI '07: Procs of the ACM/IEEE international conference on Human-robot interaction*, New York, USA, 2007, pp. 263–270.

[8] C. P. Rosé, A. Roque, and D. Bhembe, "An efficient incremental architecture for robust interpretation," in *Proceedings of HLT 2002, Second International Conference on Human Language Technology Research*, San Francisco, USA, 2002.

[9] S. C. Stoness, J. Tetreault, and J. Allen, "Incremental parsing with reference interaction," in *In ACL Workshop on Incremental Parsing*, 2004, pp. 18–25.

[10] M. Atterer and D. Schlangen, "Rubisc – a robust unification-based incremental semantic chunker," in *Proceedings of 2nd International Workshop on Semantic Representation of Spoken Language (SRSL) 2009*, 2009.

[11] E. Selkirk, *Phonology and Syntax. The relation between sound and structure*. Cambridge, USA: MIT Press, 1984.

[12] M. Nespor and I. Vogel, *Prosodic Phonology*. Dordrecht: Foris Publications, 1986.

[13] A. Siebert and D. Schlangen, "A simple method for resolution of definite reference in a shared visual context," in *Proc of SIGdial*, Columbus, Ohio, 2008.