Partially Saturated Referents as a Source of Complexity in Semantic Interpretation

David D. McDonald
Department of Computer Science, Brandeis University
davidmcdonald@alum.mit.edu

A significant factor in the complexity of the compressed, complex prose style used by journalists in short, targeted commercial reports (Who's News, joint ventures, earnings reports, etc.) is the fact that many of the phrases are semantically incomplete, i.e. their interpretation is dependent on information in other parts of the sentence or the in discourse context. We propose that the complexity that such partially saturated referents contribute to the overall process of semantic interpretation can be characterized by two factors we will call displacement and unpacking. This complexity source can be quantified by counting the distance, in nodes, between each phrase that has a locally incomplete interpretation and the phrase(s) that supply the terms that complete them.

In this paper we will define this phenomenon and illustrate its impact on interpretation by examining short texts excerpted from the Tipster corpus and other online sources.

1. The Problem

The goal of this paper is to precisely characterize the intuitive observation that the A sentences below are more complex than their B counterparts. (Example1a. is from article 231 of the Tipster joint venture corpus; 2a is from article 2279.) The B examples were com-posed by the author. The task is information extraction, where the goal is to determine the amount that each partner in the joint venture is contributing to the venture's total capital-ization.

1a. It will be capitalized at 130 million ringgit, which the three companies will equally shoulder.

1b The three companies will shoulder equal amounts of the venture's capitalization of 130 million ringgit.

2a. ... the joint firm, capitalized at one billion yen, will be 60 pct owned by P.T. Astra International, Inc., and 40 pct by Daihatsu.

2b. ... P.T. Astra will own 60 pct of the joint firm's capitalization of one billion yen and Daihatsu will own 40 pct.

We are trying to quantify an aspect of the semantic interpretation process—the process by which the lexical and syntactic elements of a text are mapped to a collection of typed, structured objects with respect to some model (broadly speaking, a collection of individuals and relations over them).

We presume (a) that interpretations are formed compositionally following the paths provided by the syntax; (b) that they come into existence incrementally phrase by phrase, object by object as the parser moves left to right through the text. This implies that most relations will initially be only partially saturated. And (c) that the mapping from lexico-syntactic objects to semantic objects is a matter of recognizing function-argument patterns that are indicated structurally or morphologically and ultimately driven by information provided by the lexical sources of the predicates.

Given this background, the question is what makes the A sentences more complex than the B sentences even though both convey essentially the same information.[1] The answer,

[1] Information, albeit of a different kind, is also conveyed by ordering, choice of cohesive devices, or even just following the stylistic conventions of the genre (which the B sentences do not). Quantifying the impact of

as we see it, lies in the nature of the path that that terms must take through the text's phrase structure as they are composed to form relations: the farther the distance the greater the complexity.

Compositional complexity, as we propose to call this phenomenon, is a problem that arises because speakers establish their relationship with their audience by producing texts (in the formal sense) rather than a jumbled salad of independent phrases. To this end, speakers have at their disposal a large battery of linguistic devices that give texts their cohesion by omitting information that their audience must now infer, thereby inducing the audience's attention (Halliday & Hasan 1976).

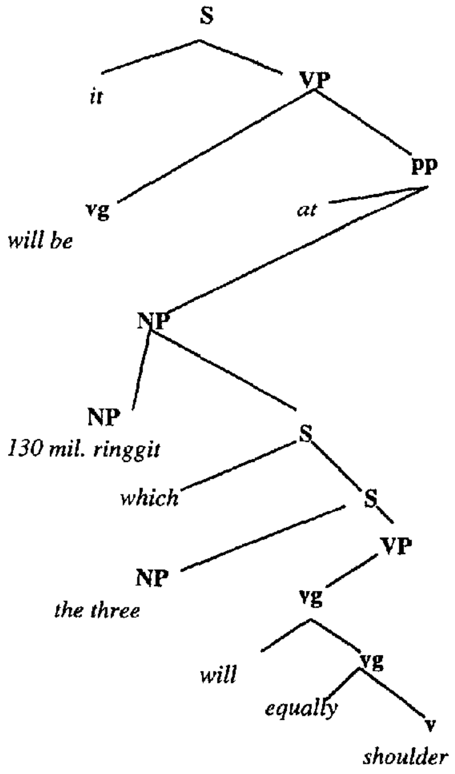
One of these devices is the use of phrases whose interpretations are locally incomplete: partially saturated. To understand such phrases, the audience (natural language understanding system) must search through the context and identify the terms that are needed to fully populate (saturate) the model-level relations these phrases denote.

We call this aspect of the semantic interpretation process 'compositional' complexity because we assume that the bulk of the organization on the context that is searched is provided by the text's syntactic structure, and that the interpretation process overall is organized compositionally as a walk over the phrase structure the syntax defines (for us a bottom up and left to right traversal in lock-step with the parser as it establishes phrasal boundaries).

These assumptions suggest that a text will be harder to understand the greater the separation between the partially-saturated relations and their missing terms (i.e. the process of its interpretation will require more effort in terms of larger working state, using a richer type system, deploying a more complex control structure, inviting a greater chance of error, etc.). As a first approximation we will measure this complexity by counting the number of intervening syntactic nodes.

2. An Example

We will explore this notion of compositional complexity by first looking in some detail at the structure and interpretation of example 1a, *"It [the joint venture] will be capitalized at 130 million ringgit, which the three companies will equally shoulder"*, which we take to have the following syntactic structure.[2]



The first clause, *"it will be capitalized at 130 million ringgit"*, illustrates the simplest case of compositional complexity, where terms are adjacent to their targets. We assume that

this information structure, however, is beyond our present abilities.

[2] We are agnostic about what the 'true' choice of labelings and other theory-governed particulars should be; what is important is the overall shape of the tree.

the word *capitalize* in the sense used here[3] denotes a function of two arguments, where `J` is restricted to (can only be bound to) objects of type joint venture and `$` to objects of type amount of money.

λ J,$. capitalization(J, $)

In this base case the two needed terms are not separated by any intermediary syntactic nodes and we say that the text has a compositional complexity of zero.

The result of binding these two terms is the instantiation of the fully saturated relation (i) below. What is shown is an expression but it intended just as a gloss of a typed structured object. Here and the examples to follow we will abbreviate freely in the interests of space, e.g. jv indicates the object that represents the joint venture, 130-million-ringget the object rep-resenting the instance of that amount of money that is being invested in the venture, and so on. We have given expression (i) a label, `Cap-1`, to emphasize its status as an object and to provide a simple means of indicating references to it in other relations.

(i) Cap-1: capitalization(J, 130-million-ringgit)

Adopting an operational perspective, we can identify two different aspects of compositional complexity: displacement and unpacking. *Displacement* is simply the separation between a term and its binding site given their relative depths in the tree.

The need for unpacking follows from our assumption that a text is interpreted incrementally, with relations (or partial relations) forming as soon as possible in the parser's progress through the text. We also assume that the individual elements of the text become unavailable at that moment except with respect to their configuration within the relation they have become part of.

In our experience this is a valuable property. Consider the partially saturated relation below that is the denotation of the relative clause of 1a at the point when the downstairs S has been parsed ("*which the three companies will equally shoulder*"). We assume for present purposes that *shoulder* denotes a model-level category we can gloss as contributes-to-capitalization. The objects representing the three companies are glossed as just C1, C2, C3.

(ii) λ amount . contributes-to-capitalization(collection(C1, C2, C3), amount)

The agent of this relation is plain enough (those three particular companies), but what about the 'amount' that they contribute?

Syntactically, the relative clause is of course open in its direct object, which the parser will associate with the np *130 million ringgit*. But how is this syntactic open variable mirrored semantically? When thought of as a contri-bution to capitalization, the denotation of *130 million ringgit* is not simply an amount of money in Indonesian currency, which would be meaningless. The np's denotation should instead provide a link though which we can determine that the money constitutes the funding of some particular venture. This can be reflected in the restriction we place on the `amount` variable.

This is where *unpacking* comes in. We have the option to view (i) as a composite object with a first class object representing each of its variable bindings in its own right, as in (iii) which is the unreduced binding of the amount of money to the `amount` variable of the object we named `Cap-1` in (i).

```
(iii) Amt-1:
    ((λ amount . Cap-1)
      130-million-ringgit)
```

Under this view we can unpack `Cap-1` into its constituent elements and make this binding object accessible to be bound to `amount`, giving us:

```
(iv)    contributes-to-capitalization(
     collection(C1, C2, C3), Amt-1)
```

[3] This is the sense of *capitalize* where it does not have an agent; cf. "*Oracle lost $3.9 billion in market capitalization*" [Wired 8.03, pg. 272].

3. Measurements

Now that we have illustrated the character of the complexity involved, what kind of numbers should be put to this so that we can compare different text quantitatively? With no literature to guide us here we should start with a simple calculus. We will add one 'point' for each node that intervenes between the partial relation and each term that it is missing, and one for each variable binding that must be unpacked from an already formed relation.

Under this analysis, the displacement of the 'amount' term contributes two points for the two nodes that intervene between the location of the verb and the relative pronoun.[4] We add another point for unpacking given that the amount of money per se does not fit the restrictions we imposed on the AMT of a contributes-to-capitalization and we need to unpack the denotation of the upper clause to get at the binding we need. This gives us a total of three points of compositional complexity for saturating the relation created by *shoulder*.

What other kinds of costs have we ignored so far? One definite cost is establishing what category (function, predicate) *shoulder* actually denotes since unless that is known the type constraints on its variable bindings will be untenably vague. (Consider that in this domain it will be quite common to see the phrase *to shoulder debt.*)

Another, possibly debatable, cost is whether to distribute the denotation of the "*the three companies*" across the capitalization to create three individual relations. Just like one could elect to ignore the fact that a multi-term relation can be seen as a set of individual variable bindings until one of those bindings is needed to do work in another part of the text's interpretation, the distribution of this conjunction could remain a latent option until it was needed to make explicit some other semantic relation.

We do need to distribute the companies conjunction in example 1a because of the other relation-generating lexical head that we have yet to consider: *equally*. (Recall that the text of 1a is "*It will be capitalized at 130 million ringgit, which the three companies will equally shoulder*".) In isolation (before being specialized to the situation of joint venture capitalization, another cost), *equal* denotes a completely unsaturated relation:

```
λ  collection( partition(measurable-
   stuff)) . equal ( elements-of (
   collection ( partition
   (measurable-stuff))))
```

Admittedly this choice of semantics may already be biased to the joint ventures problem, but it's thrust is to say that there must be some stuff that has been partitioned into some indeterminate number of portions; in aggregate these portions form a collection; and that all of these portions are in some respect equal.

Here *equal* is predicated of whatever the *shoulder* clause denotes so the process of forming its interpretation must meet and follow the process of forming that clause's interpretation as it percolates up the headline of the relative clause and into the main clause.

Equal is open in something of type collection where that collection is a partition of something. The first collection to be seen moving up the headline at a remove of two nodes (the main verb and the vp) is the conjunction of companies. Because (a) equal is predicating the equality of some aspect of each of the elements of the collection and (b) the companies per se do not have textually obvious things that might be partitioned, we can make sense of this only by distributing not just the companies but the companies qua their participation in the contribution-to-capitalization relation.

[4] We assume the parser carries the denotation of the relativized np down to the spec posi-tion; doing that certainly permits an easier analysis of the relative clause since it allows it to take on the surface pattern of, e.g., topicalization.

This gives us the three latent contribution-to-capitalization relations (at only the cost of the distribution construction, which is probably cheap). As part of that distribution construction we must also partition the amount of the contribution (object (iii)) into three parts. This entails unpacking those relations to expose their amount bindings. The equals relation then boils to down to a predication[5] over those three binding objects, viz.

```
(v) Contrib-1: contributes-to-capital-
    ization (C1, Cap-1, Amt-2)
(vi) Contrib-2:  contributes-to-capi-
    talization (C2, Cap-1, Amt-3)
(vii) Contrib-3:   contributes-to-cap-
    italization (C3, Cap-1, Amt-4)
(viii) Amt-2: λ amount . contributes-
    to-capitalization   (C1,   Cap-1,
    amount)
(ix) Amt-3: λ amount  . contributes-
    to-capitalization   (C2,   Cap-1,
    amount)
(x) Amt-4: λ amount . contributes-to-
    capitalization (C3, Cap-1, amount)
(xi) equal (Amt-2, Amt-3, Amt-4)
```

In terms of our computational complexity metric, the interpretation of the *equally* modifier has contributed two points for the displacement between it and the conjunction of companies and then (modulo the distribution cost) one point for unpacking the relation the companies are participating it to isolate the amount binding(s).

This gives example 1a a total compositional complexity of 6: its three relation sources, capitalized, shoulder, and equally contributing zero, three, and three counts respectively; four of the counts reflecting the distance that displaced elements from their binding sites, and two reflecting the effort to dip into, or 'unpack', already created relations in order to select or reify one of the elements within them.

Contrast 1a, with its complexity of six, with 1b, which has a compositional complexity of zero (though the rather severe departure of this artificially constructed sentence from the normal stylistic patterning must have a cost to human readers).

1b The three companies will shoulder equal amounts of the venture's capitalization of 130 million ringgit.

1b garners this minimal cost by placing each contributing term right next the partial relation that provides its binding site, notably pushing the capitalization clause of 1a down to the rightmost and lowest position in the sentence's phrase structure.

Example two presents a challenge to a standard compositional model of interpretation that assumes that the denotation of the syntactic head provides the basis for interpreting the head's syntactic arguments.

2a. ... the joint firm, capitalized at one billion yen, will be 60 pct owned by P.T. Astra International, Inc., and 40 pct by Daihatsu.

The syntactic head of the conjunct "*40 pct by Daihatsu*" has to be the percentage, yet there is no way to fashion a plausible rule of interpretation that binds a company to a percentage. Instead, both terms must be passed up through the conjunction node to the ownership clause (1 count) and then unpack the interpretation of that clause to extract the capitalization value and the joint venture (2 counts, one for each term). Given that the capitalization of the joint venture was given in an appositive off the subject, the ownership clause itself required two extra counts for its construction, one to unpack the capitalization and a second for the displacement of the first parent company (P.T. Astra) away from the verb in its agentive by-phrase (though that count is debatable since the grammar might explicitly subcategorize for it).

Complexity of this kind is ubiquitous in business reporting. Consider this excerpt from

[5] The amounts of money that the companies are contributing is given abstractly rather than calculated out since that appears to be the preferred level at which it should be represented for reasoning in this domain

the beginning of a quarterly earnings report (PRNewsWire 1/21/00 5:21 p.m.):

3. *Gensym Corp. < descriptive appositives> today reported that revenues for its fourth quarter ended December 31, 1999 were $9.1 million, . . . The net loss for the fourth quarter of 1999 was . . .*

The sentence that reports the loss does not say what company lost the money—to do so would be unnecessarily redundant and reduce the text's cohesion. Yet the increased tightness of the text leaves us with an partially saturated relation as the immediate referent of that sentence, open in its company variable, which must be actively filled in from context. Moreover this example is somewhat unusual in that it provides a syntax-supported explicit indicator of whose fourth quarter reporting period it is in the first of the two sentences; usually it would be stated "*for the fourth quarter . . .*" and the reporting-period object would also have been left with an unbound variable.

4. Modeling

Up to this point we have deliberately not discussed the question of how one would actually derive these compositional complexity counts automatically. We have instead provided a prose description of the process for a very few examples and many questions of just what constitutes a displacement or how one might know that a relation reached in the traversal should be unpacked remain unanswered.

The glib answer is that you fire up your natural language understanding system, add some reporting facilities to it, and apply it to the texts in question. Today at least that procedure is unlikely to work since texts of the sort we have been discussing are largely beyond the state of the art for information extraction engines without some deliberate, do-main-specific engineering.

A more germane answer would look to some resource of hand-annotated texts and then provide suitable definitions for displacement and unpacking that, given some debugging, could then be applied automatically even if there was not system that could as yet replace the knowledge of the human annotator.

But this answer too is not available to us simply because such resources do not yet exist. Besides the obvious fact that efforts at providing semantic annotations of corpora are only just now getting underway, an additional problem is that the study of the semantic phenomenon that is the focus of this paper, unsaturated, model-level relations, is uncommon in the field and for good reason.

An examination of the full text of the articles in, e.g., the Tipster Joint Ventures corpus will show that full phrases (maximal projections) that are unsaturated at the moment they are delimited by the parser and then given a semantic interpretation are unusual. A casual examination of the text in this section did not turn any up. In the full text from which example 1a was taken (which appears at the end of this paper) turns up only two more instances (reductions around the word *sales*). It is also worth noting that the original Tipster effort elected to drop attempts to extract capitalization information, as, indeed, these are among the more linguistically complex constructions in the corpus.

Partially saturated relations abound in financial texts such as quarterly earnings reports or stock market reports. Our own interest in this phenomena stems from our recent focus on such texts as well as the utility of the perspective shifts this kind of semantic object provides for work in the tactics of natural language generation (i.e. microplanning).

Without further, collective study of this class of semantic constructions any annotation effort would have a considerable startup cost as it arrived at candidate representations for its annotators to use as well as a subjective cost in convincing the rest of the community that they had made reasonable, practical choices that other project could adapt to their own purposes.

Barring a well-financed project to supply a suitably annotated corpus, we think that the proper way to proceed towards the goal of a suitable formalization is along the lines of the original, glib answer to this problem, namely to build a parser and interpretation system that operates at a sufficient level of generalization that it would require only a minimal effort to provide the lexicon and conceptual model needed to examine texts in a given domain. We have been personally engaged in such a project over the last few years, albeit at a very slow pace given the constraints we are working under, and have made a fair amount of progress, some of which is described in McDonald (in press).

5. Final Observations

That texts with partially saturated relations are more complex to process is, we think, undeniable. It also seems to us a simple matter of examination to conclude that the cost is proportional to the factors we have identified: the distance by which relation elements have been displaced from each other and the cost of unpacking already completed relations to find needed terms that those relations have already in some sense consumed. On the other hand, that this cost is measured in integer values based on simple phrase node counts is entirely debatable. As other aspects of the semantic interpretation process are quantified this component of the total measure will at least need to be combined with some proportionality constant to make all the numbers comparable.

More interesting is the fact that some node transitions will certainly be different from others in their practical implementation and this should probably be factored into the cost calculation. Consider this sentence from article 1271 of the Tipster joint venture corpus.

4. *Inoda Cement Co., . . . , said Tuesday its U.S. subsidiary has formed an equally owned cement joint venture . . . with Lone Star Industries Inc. . .*

The process that completes the 'equal ownership' relation will have to reach up through three nodes to get to the first of the two owner companies. But it will certainly be different (more elaborate) to pass this partial relation through a node that is itself creating a relation (the vp headed by *form*) as compared with passing it through report verbs like *said* or raising verbs like *expects to* that add relatively little information.

What the composition cost comes to in practice is, of course, a matter of the architecture of the parser and semantic interpretation engine that is being deployed. For some it may be a matter of adding additional mapping patterns that recognize the specific local configurations that denote partially saturated relations ('the <ordinal> quarter') and having heuristics for searching the discourse context for their missing elements.

Systems with rich descriptive resources for lexicalized grammars such as TAGs could define specific auxiliary trees for relational heads that can appear in non-standard locations (e.g. *equally*) and tie them into map-ping rules that might try to do the work over the derivation trees that these parsers produce. The conjunction problem presented by example two would be amenable to a syntactic treatment in a categorial grammar, though the range of semantic types that can be combined in this arbitrary way might make that quite difficult in general.

Finally, we must say that for us this whole idea of viewing the local interpretation of the interior phrases of a sentence as partially saturated relations and viewing their completion as a matter of passing these partial interpretations through the tree is the result of many years of research and development on a system where such relations are first class objects with the same ontological status as conventional individuals. In our system (see McDonald in press) the goal is to keep the syntactic processing simple and to move the onus of the interpretation effort onto to the semantic level by having more than one referent move up the headline as the phrase structure is created. The partially saturated relations are given an active role in seeking the arguments that they need. This introduces a bias into our observations in this paper and could, possibly, be creating a mountain where systems with quite different architectures might only see a molehill.

References:

Halliday, Michael A. K., and Ruqaiya Hasan (1976) **Cohesion in English**, Longman, London.

McDonald, David D. (in press) "Issues in the Representation of Real Texts: The Design of Krisp", in Iwanska and Shapiro (eds.) Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language, AAAI Press, pgs 71-104.

Appendix: The complete text of example 1a

<doc>
<docno> 0231 </docno>
<DD> August 9, 1990, Thursday </DD>
<SO> Copyright © 1990 Jiji Press Ltd.; </SO>
<TXT>

Mazda Motor Corp. and Sanyo Electric Co. of Japan and Ford Motor Co. of the United States have agreed to set up a joint venture by the end of this year to produce car audio equipment in Malaysia, they said Thursday. The new company, whose name is not decided yet, will produce radios, stereos, compact disc players and tuners used for cars. *It will be capitalized at 130 million ringgit, which the three companies will equally shoulder*. The three plan to construct a 21,000-square-meter plant in the Prai Industrial Estate of Penang. The joint venture with a startup workforce of 500 will kick off production by June 1992, with sales expected to reach some 10 billion Yen. By the mid-1990s, it will increase the number of employees to 2,000 and sales to 30 billion Yen. Output at the Malaysian company will be supplied to the three companies. Mazda will use the products for its cars to produced in and after the second half of next year, while Ford will mount them on its cars for sales in the Far East. Sanyo plans to sell Malaysian - made products in Japan and other countries.

</TXT>
</doc>