

A General Computational Treatment Of The Comparative

Carol Friedman*

*Courant Institute of Mathematical Sciences
New York University
715 Broadway, Room 709
New York, NY 10003*

Abstract

We present a general treatment of the comparative that is based on more basic linguistic elements so that the underlying system can be effectively utilized: in the syntactic analysis phase, the comparative is treated the same as similar structures; in the syntactic regularization phase, the comparative is transformed into a standard form so that subsequent processing is basically unaffected by it. The scope of quantifiers under the comparative is also integrated into the system in a general way.

1 Introduction

Recently there has been interest in the development of a general computational treatment of the comparative. Last year at the Annual ACL Meeting, two papers were presented on the comparative by Ballard [1] and Rayner and Banks [14]. Previous to that a comprehensive treatment of the comparative was incorporated into the syntactic analyzer of the Linguistic String Project [15]; in addition the DIALOGIC grammar utilized by TEAM [9] also contains some coverage of the comparative.

An interest in the comparative is not surprising because it occurs regularly in lan-

guage, and yet is a very difficult structure to process by computer. Because it can occur in a variety of forms pervasively throughout the grammar, its incorporation into a NL system is a major undertaking which can easily render the system unwieldy. We will describe an approach to the computational treatment of the comparative, which provides more general coverage of the comparative than that of other NLP Systems while not obscuring the underlying system. This is accomplished by associating the comparative with simpler, more basic linguistic entities so that it could be processed by the system with only minor modifications.

The implementation of the comparative described in this paper was done for the *Proteus Question Answering System* [8]¹ (referred to hereafter as *Proteus QAS*), and should be adaptable for other systems which have similar modules. A more detailed discussion of this work is given in [7].

1.1 The Problem

The comparative is a difficult structure to process for both syntactic and semantic reasons. Syntactically the comparative is extraordinarily diverse. The following sentences illustrate a range of different types of comparative structures, some of which resemble other English structures, as noted by Sager [15]. In the examples below, sentences with the comparative that resemble other forms are followed by a

*This work was supported by the Defense Advanced Research Projects Agency under Contract N00014-85-K-0163 from the Office of Naval Research. The author's current address is: Center for Medical Informatics, Columbia-Presbyterian Medical Center, Columbia University, 161 Fort Washington Avenue, Room 1310, New York NY 10032.

¹The treatment of the comparative in the syntactic analysis component was adapted from a previous implementation done by this author for the Linguistic String Project [15].

sentence illustrating the similar form:

conjunction-like :

- 1a. *Men eat more apples than oranges.*
- 1b. *Men eat apples and oranges.*

- 2a. *More men buy than write books.*
- 2b. *Men buy and write books.*

- 3a. *We are more for than against the plan.*
- 3b. *We are for or against the plan.*

- 4a. *He read more than 3 books.*
- 4b. *He read 2 or 3 books.*

wh-relative-clause-like :

- 5a. *More guests than we invited visited us.*
- 5b. *Guests that we invited visited us.*

subordinate and adverbial :

- 6a. *More visitors came than was expected.*
- 6b. *Visitors came, which was expected.*

- 7a. *More visitors came than usual.*
- 7b. *Many visitors came as usual.*

Special Comparative Constructions :

8. *A taller man than John visited us.*
9. *John is taller than 6 ft.*
10. *A man taller than John visited us.*
11. *He ran faster than ever.*

The problems in covering the syntax of the comparative are therefore at least as complex as the problems encountered for general coordinate conjunctions, relative clauses, and certain subordinate and adverbial clauses. Incorporating conjunction-like comparatives into a grammar is particularly difficult because that structure can occur almost anywhere in the grammar. Wh-relative-clause-like comparatives are complicated because they contain an omitted noun where the omission can occur arbitrarily deep within the comparative clause.

The comparative is difficult to process for semantic reasons also because the comparative marker can occur on different linguistic categories. Adjectives, quantifiers, and adverbs can all take the comparative form, as in: *he is taller than John*, *he took more courses than John*, and *he ran faster than John*. Therefore the semantics of the comparative has to

be consistent with the semantics of different linguistic categories while retaining its own unique characteristics.

2 The Underlying System

Proteus QAS answers natural language queries relevant to a domain of student records. It is highly modular and contains fairly standard components which perform:

1. A syntactic analysis of the sentence using an augmented context-free grammar consisting of a context-free component which defines the grammatical structures, a restriction component which contains well-formedness constraints between constituents, and a lexicon which classifies words according to syntactic and semantic categories.
2. A syntactic regularization of the analysis using Montague-style compositional translation rules to obtain a uniform operator-operand structure.
3. A domain analysis of the regularized structure to obtain an interpretation in the domain.
4. An analysis of the scope of the quantifiers.
5. A translation to logical form.
6. Retrieval and answer generation.

The syntactic analyzer also covers general coordinate conjunction by containing a conjunction metarule mechanism which automatically adds a production containing conjunction to certain context-free definitions.

3 The Syntactic Analysis of the Comparative

In Section 1.1 it was shown that the comparative resembles other complex syntactic structures. This observation suggests that the comparative could be treated as general coordinate conjunctions, wh-relative clauses, and certain subordinate and adverbial clauses

by the syntactic analysis component of the system. If the system can already handle these structures, the extension for the comparative is straightforward. This approach has the advantage of utilizing the system's existing machinery to process comparative structures which are very complex and diverse; in this way a minimal amount of effort results in extensive coverage. For example, to cover conjunction-like comparative structures, the production containing possible conjunctions was modified to include *than*; to include relative-clause-like comparatives, the production containing words which can head relative clauses was also modified to include *than*. Analogous minor grammar changes were made for the other types of similar structures shown above. Using this approach, a comprehensive comparative extension was obtained by a trivial modification of only a small number of grammar productions.

Thus, a conjunction-like comparative structure such as Sentence 1a. in Section 1.1 would be analyzed as consisting of an object which contains a conjoined noun phrase *more apples CONJ () oranges* where the value of *CONJ* is *than*, and where a quantifier phrase similar to *more* has been omitted which occurs with *oranges*. A relative-clause type of comparative structure such as Sentence 5a. would be analyzed as a relative clause *than we invited ()* adjoined to *more guests*. Those constructions that are unique to the comparative, as shown in Sentences 8 through 11, have to be uniquely defined. For example, the comparative clause in Sentence 8 is defined as a clause where the predicate is omitted, whereas the comparative clause in Sentence 9 is defined as a measure phrase.

Although the comparative syntactically resembles other structures, this type of similarity does not carry over to the underlying structure or to the semantics of the comparative, as will be discussed shortly.

There are also some syntactic differences between the comparative and the structures it resembles. For example, the comparative has zeroing patterns that are somewhat different from those associated with conjunctions:

- + *John slept more than Mary [slept].*
- *John slept and Mary [slept].*

The comparative constructions also have scope marker constraints that are not applicable to non-comparative structures. These differences are handled by special add-on constraints that specifically deal with the comparative, and do not interfere with the other restrictions.

The treatment of the comparative marker is complicated because it can occur in a large number of different locations in the head clause², as illustrated by a few examples below:

- He wanted to travel to more countries than he was able to.*
- He is taller than Mary.*
- He ate 3 more apples than Mary did.*
- He ate more in the fall than in the winter.*

Because the comparative marker can occur in such a variety of locations and also be deeply embedded in the head clause, it cannot be conveniently handled in the BNF component of the grammar. Instead, the constraint component deals with this problem by means of special constraints that assign and pass up the comparative marker; other constraints test that the comparative clause is in the scope of the marker.

4 Underlying Structure

Basically, linguists such as Chomsky [3,4], Bresnan [2], Harris [10], and Pinkham [13] agree on fundamental aspects concerning the underlying structure of the comparative. They regard its underlying structure as consisting of two complete clauses where information in the comparative clause which is identical to information in the head clause is required to be zeroed.

Harris' work is particularly suitable for computational purposes because he claims that one underlying structure is the source of

²This phrase was used by Bresnan [2] to refer to the clause of the comparative that contains the comparative marker.

all comparative forms. We modified his interpretation somewhat to obtain a more convenient form for computation. In our version, the underlying structure contains a main clause where the comparison is the primary relation; each quantity in the relation contains an embedded clause specifying the quantity being compared. An example of this form is shown below for the sentence *John ate more apples than Mary*, which resembles a conjunction-like comparative structure where the verb phrase has been omitted:

$$\begin{aligned} N_1 & [\text{John ate } N_1 \text{ apples}] > \\ N_2 & [\text{Mary ate } N_2 \text{ apples}] \end{aligned}$$

This form is also appropriate for all the different comparative forms shown in Section 1.1. For example, the underlying form for a relative-clause-like comparative, such as Sentence 5a. is:

$$\begin{aligned} N_1 & [N_1 \text{ guests visited us}] > \\ N_2 & [\text{we invited } N_2 \text{ guests}] \end{aligned}$$

The underlying form for a sentence such as *a man taller than John visited us* is slightly different because the comparative structure itself is embedded in a noun phrase. The main clause is *a man visited us*, and the comparative structure is a clause adjoining *a man*, whose underlying structure is:

$$\begin{aligned} N_1 & [\text{the man is } N_1 \text{ tall}] > \\ N_2 & [\text{John is } N_2 \text{ tall}] \end{aligned}$$

The notion that there is one underlying form for all comparatives has important implications for a computational treatment:

- Regularization procedures can be written to transform all comparative structures into one standard form consisting of a comparative operator and two complete clauses which specify the quantities being compared.
- In the standard form, each clause of the comparative operator is a simpler structure which can be processed using basically the usual procedures of the system. This means that further processing does not have to be modified for the comparative.

This process can be illustrated by a simple example. When the sentence *more guests than we invited visited us* is regularized, a structure consisting of an operator connecting two complete clauses is obtained:

$$\begin{aligned} (> (\text{visited} (\text{er guests}) (\text{us})) \\ (\text{invited} (\text{we}) (\text{than guests}))) \end{aligned}$$

The symbols *er* and *than*, shown above, roughly correspond to quantities being compared, and in subsequent processing they are each interpreted as denoting a certain type of quantity. Notice that each clause of the comparative is also in operator-operand form where generally the verb of a sentence is considered the operator and the subject and object (and sometimes sentence adjunct phrases) are considered the operands³. Each of the two clauses can be processed in the usual manner provided that *er* and *than* are treated appropriately. This will be described further in Section 5 which contains a discussion of semantics and the comparative.

The regularization process was modified to be a two phase process. The first phase uses ordinary compositional translation rules to perform the standard regularization so that the surface analysis is transformed into a uniform operator-operand form. The compositional regularization procedure is effective for fairly basic sentence structures but not for complex ones such as the comparative. The compositional rules associated with comparative structures only include labels categorizing the type of comparative structure. The second phase, written specifically for the comparative, completes the regularization process by filling in the missing elements, permuting the structures to obtain the correct operator-operand form, and supplying the appropriate quantifiers *er* and *than* to the items being comparativized. An example of this process is shown for the relative-clause type of comparative in *more guests than we invited visited us*, where the comparative clause *than we invited* is analyzed syntactically as being a right adjunct modifier of *guests*.

³However, if the predicate is an adjectival phrase, the adjective is considered the operator and the verb be the tense carrier. Thus, ignoring tense information, the regularized form of *John is tall* is: (*tall (John)*).

Phase 1: (visited (more guests
 (reln-than
 (invited (we) ()))))
 (us))

Phase 2: (> (visited (er guests) (us))
 (invited (we) (than guests)))

Another example is shown below for a conjunction-like comparative, such as *John ate more apples than oranges*:

Phase 1: (ate (John)
 (conj-than (more oranges)
 (() oranges)))

Phase 2: (> (ate (John) (er apples)
 (ate (John) (than oranges)))

There are a few key points that should be made concerning the regularization procedures. The Montague-style translation rules could not readily be used to regularize the comparative constructions as they were defined in the context-free component. To use the rules, the grammar would have to be modified substantially because the translation of the comparative is different and more complex than that of the structures it resembles. In particular, it would then not be possible to use the general conjunction mechanism to obtain coverage of that type of comparative structure. In the case of the usual relative clause, the regularized form is also substantially different from the regularized form of the relative-clause type of comparative shown above. For a typical relative clause, such as *that we invited () in guests that we invited visited us*, the regularized form occurs as a clause embedded in the main clause as follows:

(visited (guests (invited (we) ())
 (us)))

The second important point is that because of regularization further processing of sentences containing a comparative is significantly simplified and only minor changes are required specifically for the comparative. In *Proteus QAS*, as well as other NLP Systems, several other processing components are needed after syntactic regularization until the final result is obtained. Therefore a significant result of our approach is that subsequent

components do not have to be modified for the comparative. As long as the underlying system can handle adjectives, degree expressions, quantifiers, and adverbs, the remainder of the processing of sentences with the comparative is basically no different than the processing of ordinary sentences because at that point the comparative is represented as being composed of fundamental linguistic entities.

5 Semantics of the Comparative

Semantically the comparative denotes the comparison of two quantities relative to a certain scale. This interpretation is consistent with work in formal semantics ([12,11], [6,5]), although our formalism is not the same.

Since the comparative marker can occur with adjectives, quantifiers, and adverbs, we would like to integrate its semantic treatment with the semantics of those fundamental linguistic categories and also remain true to the semantics and syntax of the comparative. This can be done by noting that once the comparative is regularized, the comparative marker becomes a higher order operator connecting two clauses and what remains of the marker within each clause functions as a quantitative phrase. For example, the regularized form for *is John taller than Mary* is:

(> (tall (DEG er) (John))
 (tall (DEG than) (Mary)))

In this form *er* and *than* are each interpreted as a type of degree phrase that occurs with adjectives. In a question answering application such as that of *Proteus QAS*, each clause of the above form is equivalent to the regularized form of *how tall is John*, where *how* is also interpreted as a degree phrase modifying *tall*:

(tall (DEG how) (John))

The interpretation of a sentence containing the comparative is therefore reduced to the interpretation of two similar simpler clauses, each containing an adjective operator and an

operand which is a degree phrase. Issues concerning the correct scale and criteria of comparison for adjectives are non-trivial, but are generally not different from those issues concerning adjectives not being comparativized. For example, determining the scale and criteria that should be used to interpret *is John more reliable than Jim* raises similar issues to those for *how reliable is Jim*.

The semantic treatment of adverbs generally parallels that of adjectives; the interpretation of quantifiers in the comparative form is also equivalent to the interpretation of certain interrogatives. For example, the regularized form of *did John take more courses than Mary* consists roughly of the two clauses *John took er courses* and *Mary took than courses*, which is treated analogously to *how many* in *how many courses did John take*.

6 Quantifier Analysis

An interesting problem involving the comparative concerns the scope of quantifiers when there is a higher order sentential operator such as the comparative. The problem is not discussed much in the literature, but was discussed by Rayner and Banks [14] when they described their treatment of quantifiers for *everyone spent more money in London than in New York*. The basic issue is whether the quantifier *every* in *everyone* should be given wider scope than the comparative itself, in which case it is applicable to both clauses of the comparative. Our approach addresses this problem in a general way by adding a preliminary phase to the standard quantifier analysis. Our approach has several key features:

- The replication of a quantified noun phrase does not lead to impossible scoping combinations, as frequently happens when these phrases are replicated for the purpose of obtaining a complete clause.
- Our approach is applicable to all general higher order operators connecting two clauses.
- The scope of quantifiers is determined in a late stage of processing so that com-

mittment is not done prematurely.

- A procedure using pragmatics and domain knowledge can easily be incorporated into the system as a separate component to aid in scope determination.

In *Proteus QAS*, the scope of quantifiers is determined subsequent to the regularization and domain analysis components in a manner similar to other NLP Systems, as described by Woods [16]. The basic quantifier analysis procedure initially handled simple clauses, and therefore had to be modified to accommodate scope determination when a sentence contains a higher order operator such as a comparative or a coordinate conjunction. A preliminary quantifier analysis phase was added to find and label quantifiers which have a wider scope than the comparative. In addition, minor modifications were made to the component which translates the regularized form to logical form, in order to handle the translation of wider scope quantifiers.

Generally, in the case of the comparative, the criteria used for determining whether or not a quantifier should have a wider scope involves the location of the quantifier relative to the comparative marker in the surface form. Usually, a preference is given to the wider scope interpretation if the quantifier precedes the marker. Using this approach, the sentence *everyone spent more money in London than in New York* is first interpreted syntactically as consisting of two complete clauses, which are roughly *everyone spent er money in London* and *everyone spent than money in New York*. The semantics of each clause is interpreted the same as that of a simpler sentence *how much money did everyone spend in London*. The preliminary quantifier analysis phase prefers the reading where the scope of *everyone* is wider than the comparative operator because *everyone* precedes *more*. The sentence is translated to logical form so that the quantified expression $\forall X : \text{person}(X)$ occurs outside the comparative operator, and therefore has scope over both clauses of the comparative. The interpretation is roughly:

$$\forall X: \text{person}(X) \rightarrow (\text{spent}(X) \text{ (er money)} \\ \quad \quad \quad \text{(in London))} \\ \quad \quad \quad (\text{spent}(X) \text{ (than money)} \\ \quad \quad \quad \text{(in New York)))}$$

A different scope interpretation is obtained for *more students read than wrote a book*, where the two clauses are *er students read a book* and *than students wrote a book*. The narrow scope interpretation of *a* in *a book* is obtained because *a* follows *more*. In this case, the quantified expressions for each clause of the comparative are completely independent of the other.

7 Concluding Remarks

We have presented a method for incorporating general comparatives into a system without unduly complicating the system. This is done in the syntactic analysis component by treating the comparatives the same as similar structures so that features of the syntactic analyzer that already exist may be utilized. The various comparative structures are then regularized so that they are in a standard form consisting of a comparative operator and two complete clauses that contain a quantity *er* or *than* which is interpreted by the semantic component as a quantity such as *how*, *how many*, or *how much*, as appropriate. A preliminary quantifier analysis component was added to determine whether a sentence containing a higher order operator has any quantifiers which have a wider scope than the operator, and to label those that do. The remainder of the processing is done as usual except for minor modifications.

The treatment of the comparative that we have presented is more extensive and general than that of other NLP Systems to date, and also is simple to implement. Only a small number of productions of the BNF component were changed to cover the comparative structures described in this paper. In addition, three restrictions were modified for the comparative, and a set of separate add-on restrictions were included to handle comparative zeroing patterns and scope marker requirements. Special regularization procedures were

written to regularize the different comparative forms so that the standard Montague-style compositional translation rules could be used prior to the comparative regularization phase.

Although we can process many forms of the comparative, there is still substantial work that remains which involves comparative sentences where the comparative clause itself has been omitted, as in *New York banks are starting to offer higher interest rates*. In some cases the comparison is between two different time periods; in other cases the comparison involves different types of like objects, such as *the interest rates of New York banks* compared to *the interest rates of Florida banks*. The context can often be an aid in helping to recover the missing information, but the recovery problem is still quite a challenge. Sentences with this type of anaphora are very interesting because they occur surprisingly regularly in language, and yet the recovery possibilities are more limited and more controlled than those occurring in discourse in general. Possibly these type of sentences can provide us with clues as to what elements are significant for the recovery of the missing information.

Acknowledgements

I would like to thank Ralph Grishman, Naomi Sager, and Tomek Strzalkowski for their help and comments.

References

- [1] B. Ballard. A general computational treatment of comparatives for natural language question answering. In *Proc. of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, 1988.
- [2] Joan W. Bresnan. Syntax of the comparative clause construction in English. *Linguistic Inquiry*, IV(3):275–343, 1973.
- [3] Noam Chomsky. *Aspects of the Theory of Syntax*. M.I.T. Press, Cambridge, Mass., 1965.

- [4] Noam Chomsky. On wh-movement. In P. Culicover, T. Wasow, and A. Akmaijan, editors, *Formal Syntax*, pages 71–132, Academic Press, New York, 1977.
- [5] M.J. Cresswell. *Logics and Language*. Methuen, London, 1973.
- [6] M.J. Cresswell. The semantics of degree. In B.H. Partee, editor, *Montague Grammar*, pages 261–292, Academic Press, New York, 1975.
- [7] C. Friedman. *A Computational Treatment of the Comparative*. PhD thesis, New York University, 1989. Reprinted as PROTEUS Project Memorandum 21, New York University, Courant Institute of Mathematical Science, Proteus Project, New York, 1989.
- [8] R. Grishman. *PROTEUS Parser Reference Manual*. PROTEUS Project Memorandum 4, New York University, Courant Institute of Mathematical Science, Proteus Project, New York, July 1986.
- [9] B. Grosz, D. Appelt, P. Martin, and F. Pereira. Team: an experiment in the design of transportable natural-language interfaces. *Artificial Intelligence*, 32(2):173–243, 1987.
- [10] Zellig Harris. *A Grammar of English On Mathematical Principles*. John Wiley and Sons, New York, N.Y., 1982.
- [11] Ewan Klein. The interpretation of adjectival comparatives. *Journal of Linguistics*, (18):113–136, 1982.
- [12] Ewan Klein. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, (4):1–45, 1980.
- [13] J. Pinkham. *The Formation of Comparative Clauses in French and English*. Garland Publishing, New York, 1985.
- [14] M. Rayner and A. Banks. Parsing and interpreting comparatives. In *Proc. of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 49–60, 1988.
- [15] Naomi Sager. *Natural Language Information Processing: A Computer Grammar of English and Its Applications*. Addison-Wesley, Reading, Mass., 1981.
- [16] W.A. Woods. Semantics and quantification in natural language question answering systems. *Advances in Computers*, 17:1–87, 1978.