Clausal Complement? Passive Subject? Grammatical Relationships for (non) Linguists

First Lastname **Firstname Lastname**

ABSTRACT

INTRODUCTION

What does X do? How is Y described? Most analysts and researchers ask themselves these questions while trying to understand a subject. So far, the field of information retrieval has tackled this problem indirectly. The analyst enters a search query, and the system returns some results. When the underlying data has structure, queries can be specific and targeted – and there is now a substantial body of work [cite] on how best to issue structured queries over *structured* data. But what if the "data" is is text – medical literature, legal records, interview transcripts? Although text is richly structured by rules of grammar and narrative, it is rarely treated as such. While there is a lot of work on extracting various kinds of structured information from text, there is little on how to make it easy for users to access and query over it. As the questions above show, however, structured queries over language can be extremely useful. In grammatical terms, they become "What are the verbs of which X is the subject?" "What are the adjectives that modify Y?"

Adapting existing structured-query interfaces to grammatical search is problematic for several reasons. The first is that the structures in language are not explicit, like columns in a table, but are implicit, and have to be extracted computationally. Only in the last decade have computational linguistic technologies become fast and accurate enough to use in the real world.

The second problem is the lack of programming experience among searchers. Current structured query languages like SPARQL have complex syntaxes that require time and effort to learn. For example, here is a SPARQL query for "What are all the country capitals in Africa?":

```
SELECT ?capital ?country
WHERE {
  ?x cityname ?capital ;
     isCapitalOf ?y .
  ?y countryname ?country ;
     isInContinent Africa .
}
```

Paste the appropriate copyright statement here. ACM now supports three different copyright statements:
• ACM copyright: ACM holds the copyright on the work. This is the historical approach.
• License: The author(s) retain copyright, but ACM receives an exclusive publication license.
• Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.
This text field is large enough to hold the appropriate release statement assuming it is single spaced.

Even if we assume that only professional analysts engaging in information-intensive work would want to issue such targeted queries, programming experience is scarce. One study on a query language for selecting phrase structures from sentences found that that only 50% of the people who wanted to use it had programming experience [].

The third problem is that there are no common-language terms for grammatical relationships even though ordinary people are perfectly capable of understanding and using them. Modern parsers use standard linguistics terminology to label their outputs, but those technical names and definitions are not always accessible to those outside the field. Take the phrases "he threw the ball" and "the ball was thrown by him". In both cases, it is clear that 'he' is is the one who 'threw' whereas, in grammatical terms, the first phrase is in the active case, and the phrase is in the passive case. The Stanford Dependency Parser [], for example, outputs two different variants of the verb-subject relation: `nsubj(he, threw)` and `nsubjpass(he, threw)`. A grammatical search system therefore has to bridge the gap between the relations that are recognizable to people and the relations that are extracted from the data

We conducted experiments to investigate how grammatical relationships between words can be made more recognizable to ordinary people. Following the principle of recognition over recall, we hypothesized that examples would help people identify grammatical relationships more accurately rather than formal terms.

We presented participants with a series of identification tasks. In each task, they were shown a list of sentences in which a particular grammatical relationship existed between highlighted words. They were asked to identify which relationship it was from a list of four options. Using a between-subjects design, we tested different strategies for presenting these options. Our goal was to see whether participants identified the relationship more accurately when the options showed example usages.

We tested two types of examples: a list of matching words and a list of matching phrases containing the relationship. These alternatives correspond to the explicitly visible and implicitly-inferred portions of a grammatical relation. The words are explicitly visible in the text, but the grammatical relationship is implicitly inferred from contextual information such as the part of speech of the verb, the relative ordering, and any accompanying words. The matching-words alternative showed only the

explicitly visible part, and the matching phrases showed the words their entire context.

Our hypothesis was the following:

> H1. Grammatical relations can be made more recognizable by showing examples of words or phrases that match.

We wanted to avoid having any specific backgrounds overrepresented because the ability to issue grammatical search queries is relevant to many fields outside linguistics and language study. We therefore chose Amazon's Mechanical Turk crowdsourcing platform as a source of study participants.

Our results confirm that showing examples significantly improves the accuracy with which grammatical relationships are recognized. Words or phrases did better than the linguistic label alone in all cases.

In each task, there was a 'query' word and a relationship. The participants were shown list of sentences containing that relationship between the query word and other words. The query word was highlighted in yellow and the matching word in pink (see Figure **??**. Their task was to identify the relationship from list of four choices. We presented the choices in three different ways – as a short label using linguistic terminology (Figure **??**), the short label accompanied by a list of words that matched (Figure**??**), and the short label accompanied by a list of phrases in which that relationship surfaced (Figure(**??**). Each task looked Figure **??** shows what three conditions of this identification task looked like for the case where the grammatical relationship was `nsubj` – verb subject `nsubj(ship, ___)` case.

For most grammatical relationships, the matching-phrases presentation resulted in the highest accuracy. Nevertheless, there were a few relations, notably adverb modifiers and noun compounds, for which the presentation matching-words presentation was significantly more recognizable than the matching-phrases presentation. In a follow-up experiment, we found that if distinctive or closed-class words enter into a grammatical relation, then showing a list of matching words makes it easier to recognize the relation. In other cases, a list of example phrases is more helpful.

The rest of this paper is structured as follows. In the next section, we summarize the previous work on issuing structured queries over linguistic information extracted from text data. Then, we motivate and describe our study design and analyze the results. Then, we describe our follow-up study design and results. To conclude we, summarize our findings and discuss its implications for grammatical search interfaces.

RELATED WORK

EXPERIMENT 1

The goal of our first experiment was to find out whether grammatical relationships between words could be made more recognizable

EXPERIMENT 2

We hypothesized that words were more helpful when the relations involved distinctive or closed-class words: adverbs are distinctive because they usually end in 'ly' – thoughtfully, helpfully, quickly, etc. Closed-class words include determiners (a, the, that, etc.), pronouns and prepositions. Our follow-up hypothesis was:

> H2. If distinctive or closed-class words enter into a grammatical relation, then showing a list of matching words makes it easier to recognize the relation. In other cases, a list of example phrases is more helpful.

We conducted a follow-up study to verify this hypothesis. This study was had the same design as the first study: a series of identification tasks in which there was a query word and a relationship. The participants were shown a list of sentences containing that relationship between the query word and other words, their task was to identify which relationship it was, from a list of 4 choices. Except this time, instead of presenting all the choices in the same way (as a list of matching words, or a list of matching phrases) we presented each choice in the way we thought would be most useful according to our hypothesis. If the hypothesis was true, participants in this 'optimal presentation' condition would outperform participants who had all the options presented in the same way.

The results from this follow-up experiment confirmed our hypothesis.