Transformation Process

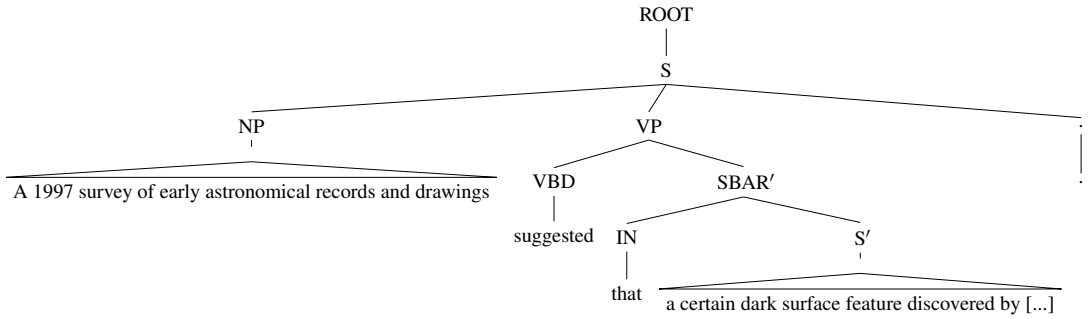
In this document, we demonstrate the transformation stage in practice and explain some of the applied rules in detail. We will use the following examples:

(1) *A 1997 survey of early astronomical records and drawings suggested that a certain dark surface feature discovered by astronomer Giovanni Cassini in 1690 may have been an impact scar.*

(2) *He helped set up a job training program, a college preparatory tutoring program, and a tenants' rights organization in Altgeld Gardens.*

1 Example 1

In the clausal disembedding layer, multi-clause sentences are split up and transformed into simplified sentences with reduced syntactic complexity. In the first sentence, our rule for identifying attributions matches and separates the clause S' contained in a subordinate clause $SBAR'$:

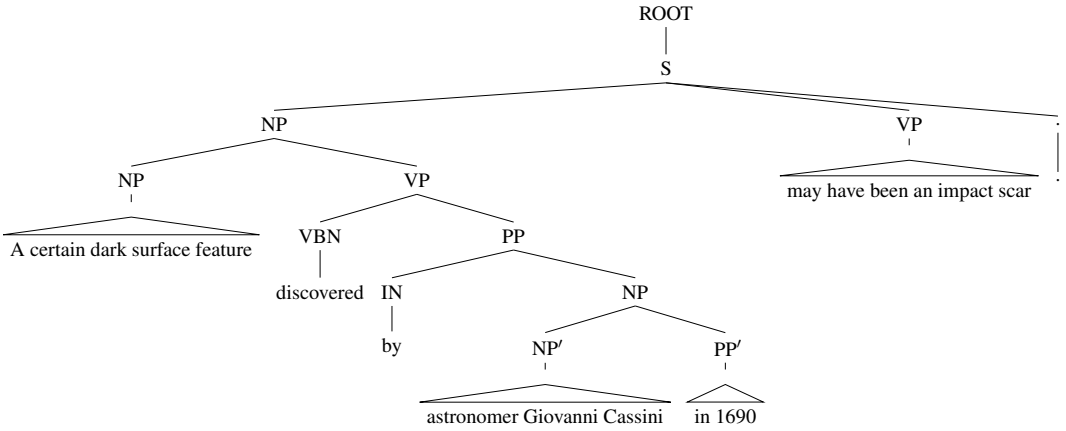


By default, a subordinate clause will be assigned as contextual, but in this case, the extracted signal phrase *suggested that* is classified as an *attribution* relation where we reverse the core and context assignment. As a result, the two following sentences are returned:

(1) Context: *A 1997 survey of early astronomical records and drawings suggested.* (ATTRIBUTION)

(2) Core: *A certain dark surface feature discovered by astronomer Giovanni Cassini in 1690 may have been an impact scar.*

Phrasal Disembedding further simplifies sentences by extracting additional context information. For the core sentence from example 1 (after clausal disembedding), we extract two contextual relations.



We obtain the first context from the prepositional phrase PP' which is converted into a verb-mediated context and classified as *temporal* due to the contained year format. The second context is extracted from the noun phrase NP': *astronomer Giovanni Cassini*, generating a context mediated by the noun *astronomer* :

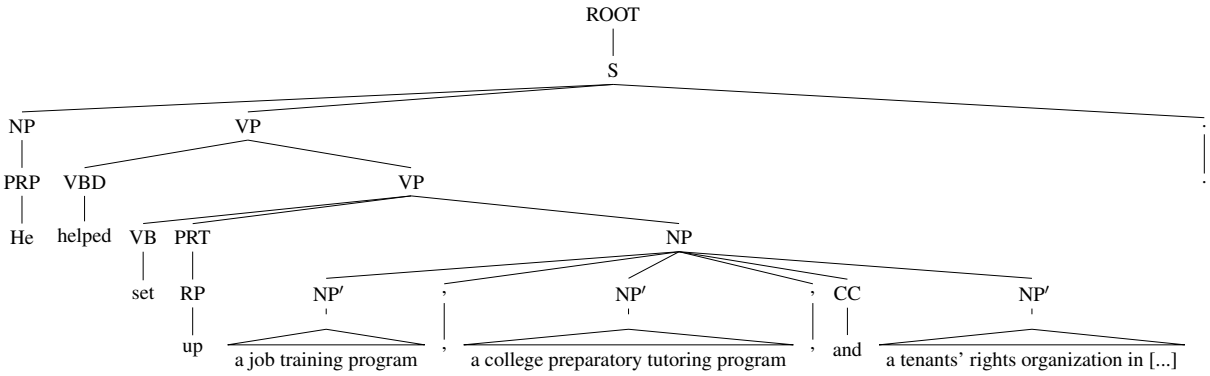
(1) Core: *A certain dark surface feature discovered by Giovanni Cassini may have been an impact scar.*

(2) Context: *in 1690* (TEMPORAL)

(3) Context: *Giovanni Cassini was an astronomer.* (NOUN_BASED)

2 Example 2

Besides other simplification rules, we also added rules for breaking up lists of entities in order to increase recall for information extraction. These rules match patterns of coordindated noun phrases within a parent noun phrase. In order to avoid inadvertently mistaking coordinated noun phrases for appositives (e.g. *He returned to Kenya for* $\boxed{\textit{a visit to his fathers birthplace}}$ *,* $\underline{\textit{a village near Kisumu in rural western Kenya}}$), we used a heuristic determined by the regular expression: $(NP)(,NP)*,?(and|or)(.+)$. Each extracted entity is embedded into the original sentence and considered as core information. The sentence from example 2:



yields the following sentences:

(1) Core: *He helped set up a job training program.*

(2) Core: *He helped set up a college preparatory tutoring program.*

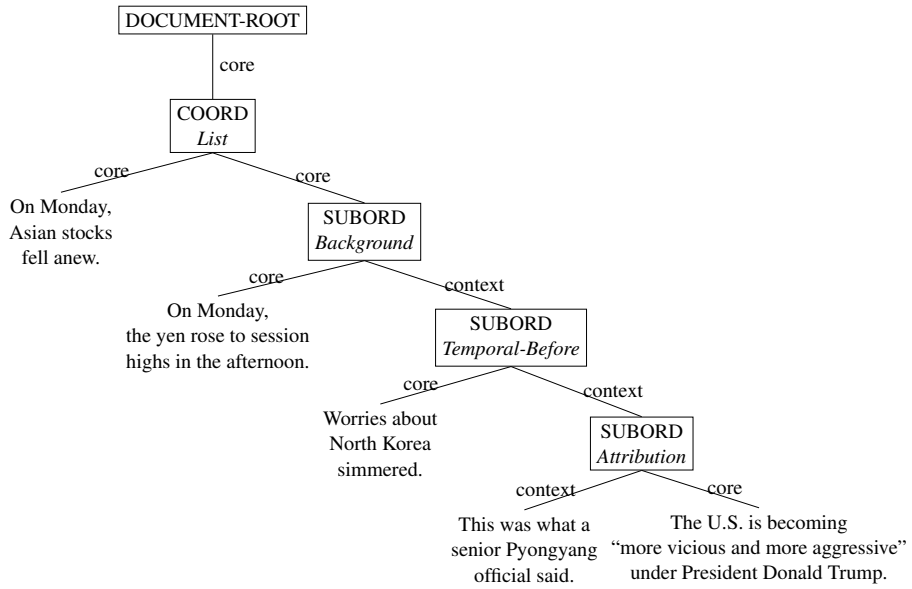
(3) Core: *He helped set up a tenants' rights organization in Altgeld Gardens.*

With the distinction between core and contextual sentences, we are able to transform the clausal simplification process into a hierarchical tree representation that is similar to the diagrams used in Rhetorical Structure Theory (RST). For this purpose, we create *coordination* nodes for rules that create two or more core sentences and *subordination* nodes for rules that create one core and one context sentence. These nodes contain the simplified sentences as child nodes, which constitute the leaves of the tree, and the classified rhetorical relation that holds between those nodes. By recursively splitting up sentence leaves in a top-down fashion, we generate a tree that spans over the whole input document.

As an example, consider the following sentence:

(4) *On Monday, Asian stocks fell anew and the yen rose to session highs in the afternoon as worries about North Korea simmered, after a senior Pyongyang official said the U.S. is becoming "more vicious and more aggressive" under President Donald Trump."*

and it's resulting tree:



With this representation, we can later enrich relational tuples that have been extracted out of the sentence leaves with rhetorical relations that occur between the corresponding sentences:

```
#1     0     Asian stocks     fell anew
     S:TEMPORAL     [This was] on Monday.
     L:LIST         #2

#2     0     the yen     rose     to session highs in the afternoon.
     S:TEMPORAL       [This was] on Monday.
     L:LIST           #1
     L:BACKGROUND     #3

#3     1     worries     simmered.
     S:SPATIAL            [This was] about North Korea.
     L:TEMPORAL_BEFORE     #4

#4     2     the U.S.     is becoming     ``more vicious and more aggressive'' under Donald Trump.
     S:ATTRIBUTION     [This was what] a senior Pyongyang official said.
     L:NOUN_BASED      #5

#5     2     Donald Trump     is     President.
```