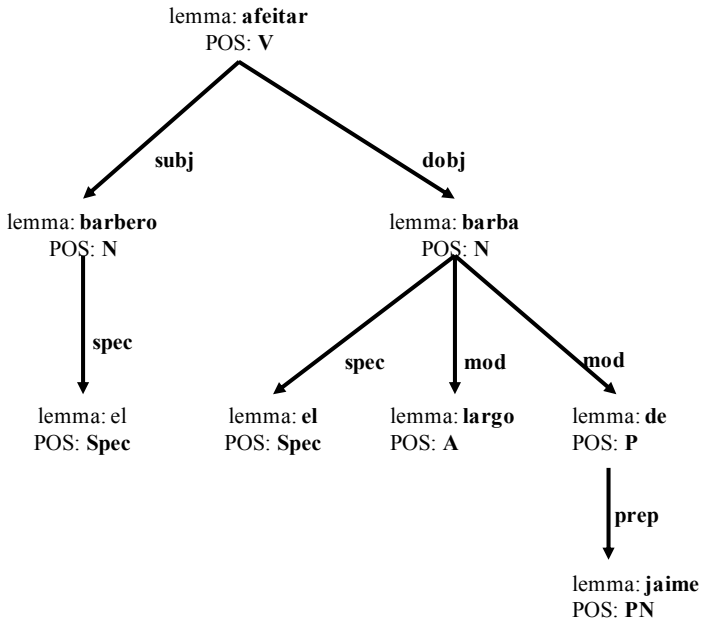

Figure 3.2: Dependency parsed sentence: *El barbero afeita la larga barba de Jaime* ('The barber shaves off James's long beard')

replaced by the *pn_n* POS tag[19].

For each context obtained from the dependency structure, three different dependency contexts are generated: one that makes all the elements of the context explicit, that is, the *dep_dir*, *dep_lab* and *context_lemma* (for example, [<:dobj: afeitar_v]); another in which the *dep_lab* is generalized by the variable 'oth' (for example, [<:oth:afeitar_v])[20] and, finally, one context that generalizes the *context_lemma* by substituting it for the variable '*' (for example, [<:dobj:*_v])[21]. The three lemmas represented in example (2) do not share any context, therefore they could not be semantically related in our model. Instead, applying the generalization of contexts, we obtained a relationship between lemma_1 and lemma_2 in example (3), and between lemma_1 and lemma_3 in example (4). In example (3), the *dep_lab* is generalized, whereas in example (4) the *context_lemma* is generalized.

[19]Since the POS tagger does not distinguish between subclasses of proper names (person, organization, place, etc.), the grouping of all with the *pn_n* tag gives better results. We used proper nouns in the *context_lemma* configuration, but not as words in the lemma *i*. Similarly, stopwords are not included in lemma *i*.

[20]The tag 'oth' (*other*) means that the dependency label is not specified.

[21]The symbol '*_v' means that a verb occurs in this position, but we do not specify which one it is.

2. lemma$_1$ [$<$: $subj$: $robar\_v$[22]]
   lemma$_2$ [$<$: $dobj$: $robar\_v$]
   lemma$_3$ [$<$: $subj$: $hurtar\_v$[23]]

3. lemma$_1$ [$<$: oth : $robar\_v$]
   lemma$_2$ [$<$: oth : $robar\_v$]
   lemma$_3$ [$<$: oth : $hurtar\_v$]

4. lemma$_1$ [$<$: $subj$: $*\_v$]
   lemma$_2$ [$<$: $dobj$: $*\_v$]
   lemma$_3$ [$<$: $subj$: $*\_v$]

In this way, the generalization of contexts allows us to take into account contexts that are similar (they share two, but not all of the elements, of their context), but not identical. Therefore, we can distinguish between those lemmas that share the same or similar context, and those that have a completly different context. By adding these contexts that are similar but not identical we add new knowledge, that is, knowledge not directly present in the corpus. This new knowledge is used to generate the Unattested-Patterns.

3.3.4 Clustering

Once we described the X matrix, we proceeded to the third step detailed in Figure 3.1 that is devoted to the clustering of this matrix. The motivation of the clustering process is to find, for each lemma in the matrix, all semantically related words (lemmas). This will allow us to create new Unattested-Patterns after the linking and filtering cluster processes. To perform this clustering step, we used the CLUTO toolkit [Karypis, 2002][24], which is used to cluster a collection of objects (in our case, lemmas) into a predetermined number of clusters labeled k. We applied a methodology based on Caliński and Harabasz [1974] and using cosine similarity and CLUTO's $\mathcal{H}_2$ metric to estimate the optimal amount of clusters.

We experimented with a number of different clustering configurations. The variables we took into account were: a) the number of most frequent lemmas,

[22] 'to_rob'

[23] 'to_steal'

[24] We use VCLUSTER program provided in the toolkit, which computes the clustering using one of five different approaches. Four of these approaches are partitional, whereas the fifth approach is agglomerative.

with the 10,000 to 15,000 most frequent lemmas giving the best results; b) the inclusion of proper nouns or their substitution for their POS; and c) considering the lemmas with and without their POS.

We evaluated the results of these configurations manually and opted for 15,000 lemmas with proper nouns grouped according to their POS tag (pn_n) and with the POS tag assigned to the lemmas. This configuration gave an optimal k of 1,500 clusters applying the Caliński and Harabasz [1974] method and the $\mathcal{H}_2$ metric.

The inclusion of POS improves the internal consistency of the clusters. Since the POS tagger does not distinguish between subclasses of proper names (person, organization, place, etc.), grouping them according to the *pn_n* tag also gives better results. Regarding the number of lemmas, all results obtained using between 10,000 and 15,000 lemmas gave satisfactory results. The choice of the number of lemmas determines the number and the content of the clusters. In all cases, the quality of clusters obtained was acceptable. We consider a cluster as acceptable when all or almost all words contained in it share one of the following relations: synonymy, hypernymy, or hyponymy. This would allow for the use of one or more configurations for the obtention of the final lexico-syntactic patterns (see Section 3.3.6).

Using CLUTO with the selected configuration, we obtained a set of clusters $C = \{c_i : 1 \leq i \leq k\}$ from matrix X. Formally, the content of each cluster $c_i \in C$ is defined in 3.2, where le is a set of related lemmas and ctx is a set of contexts. Each lemma_pos only belongs to one cluster (i.e., it can only be defined in one le), whereas a context_lemma can be in several contexts (ctx) of different clusters.

$$c_i =< le, ctx > \tag{3.2}$$

Formally, a context (called $context\_cluster$) in ctx is described as follows:

$$context\_cluster =< [dep\_dir : dep\_lab : context\_lemma], score > \tag{3.3}$$

where $dep\_dir, dep\_lab, context\_lemma$ corresponds to the definition of a context as shown in Section 3.3.3.2. The $score$ is the sum of the different scores given by CLUTO[25].

For example, Table 3.1[26] describes the lemmas, le, and the most scored contexts, ctx, in cluster number 421_n (one of the clusters obtained in the corpus analyzed).

[25]The sum of the twenty-five most descriptive and discriminative scores given automatically by CLUTO.

[26] The translation to English of Tables 1 and 2, as well as additional examples and clusters are available at `http://clic.ub.edu/corpus/`