**METHOD**

The desired efficient and compact representation of hierarchical translation equivalence can be achieved by Hierarchical Alignment Trees (HATs). HATs [Sima’an and Maillette de Buy Wenniger, 2013] are recursive synchronous tree pairs with nodes corresponding to phrase pairs induced by word alignments, structured to form a minimally branching factorization of these phrase pairs. They extend Normalized Decomposition Trees (NDTs) [Zhang et al., 2008] by providing explicit labels for the type of reordering occurring at the nodes, as well as maintaining the internal word alignments for atomic (non-decomposable) phrase pairs.

Atomic phrase pairs are phrase pairs that do not subsume smaller phrase pairs. Starting from such atomic phrase pairs larger phrase pairs are recursively built by combining a minimal number of smaller subsumed phrase pairs into larger units. Figure 4 above gives an illustration of this process, where in Step 5 the simple atomic phrases of our sentence and target sentences induced by the word alignments are composed into a bigger monotone unit.

Disruptions are discontinuous when no phase-based factorization into parts is possible (see the top node in Figure 1 / Figure 6 for an example). In such cases first the spans corresponding to proper subsumed phrase pairs are added as normal child nodes below the new phrase pair and finally the discontinuous parts are added directly as terminal productions/children below it as well. Building HATs in this recursive way, every node is labeled with a set-permutation label. Given the local alignment at a node (phrase pair), this label specifies the relative mapping occurring directly below this node. The relative mapping is specified as an ordered list of sets of relative target mapping positions, one such set of positions for each relative position in the source phrase.

In the case of non-atomic m-n mappings there are recurring target position in the mapping set of different source positions and/or multiple target positions occurring in the mapping set(s) of some source positions. The set-permutation labels can be clustered into coarser categories of mapping complexity. We distinguish the following five cases, ordered by increasing complexity:

1. **Atomic**: If the alignment does not allow the existence of smaller (child) phrase pairs: a subset of alignment positions that is not connected to the other positions while also forming a contiguous sequence on the source and target does not exists.
2. **Monotonic**: If the alignment can be split into two monotonically ordered parts.
3. **Inverted**: If the alignment can be split into two inverted parts.
4. **PET (Permutation Tree)**: If the alignment can be factored as a permutation of more than 2 parts.
5. **HAT (Hierarchical Alignment Tree)**: If the alignment cannot be factored as a permutation of parts, but the phrase does contain at least one smaller phrase pair.

Typically there are multiple HATs for a word alignment, corresponding to different possible minimally branching factorizations of monotone phrases. These alternative HATs can be efficiently computed and stored as a chart using a CYK-parser like chart parsing algorithm that parses the alignment and builds a hypergraph of HATs in the process.

A categorization of the complexity of the HAT as a whole is determined based on the complexity categories of the alignment mappings at its nodes. Binary Inversion-Transduction Trees (BITTs) is the least complex class consisting of only binary HATs that can be built for binary-permutations [Huang et al., 2009], any HAT that contains only Monotonic and/or Inverted nodes belongs to this class. If a HAT contains at least one PET node but no HAT nodes it belongs to the category called PETs corresponding to general permutations [Zhang et al., 2008]. Finally the occurrence of at least one HAT node implies the set HATs which captures all possible many-to-many mappings.

**RESULTS**

The database selection window of the HATs visualization shows broad complexity categories for search. The ratio of different subsets of HATs in the corpus: BITTs, PETs and HATs is shown in Table 1. One way to characterize the complexity of hierarchical translation equivalence is to look what types of composition operators and associated types of Hierarchical Alignment Trees (HATs) aligned sentence pairs induce. The ratio of different subsets of HATs captures a coarse notion of the level of alignment complexity of real data. The table below shows this ratio for three aligned language pairs from Europarl, with approximately 1 million sentence pairs per language pair. Word alignment is done using GIZA++. This gives a taste of results obtained in our other studies. Much more extensive results are reported in [Sima’an and Maillette de Buy Wenniger, 2013].

**REFERENCES**


**SOURCE CODE**

The source code and compiled executables are available at: [https://bitbucket.org/teamwildtreechase/hatparsing](https://bitbucket.org/teamwildtreechase/hatparsing).