Abstract

Standard models for syntactic dependency parsing take words to be the elementary units that enter into dependency relations. In this paper, we investigate whether there are any benefits from enriching these models with the more abstract notion of nucleus proposed by Tesnière. We do this by showing how the concept of nucleus can be defined in the framework of Universal Dependencies and how we can use composition functions to make a transition-based dependency parser aware of this concept. Experiments on 12 languages show that nucleus composition gives small but significant improvements in parsing accuracy. Further analysis reveals that the improvement mainly concerns a small number of dependency relations, including relations of coordination, direct objects, nominal modifiers, and main predicates.

1 Introduction

A syntactic dependency tree consists of directed arcs, representing syntactic relations like subject and object, connecting a set of nodes, representing the elementary syntactic units of a sentence. In contemporary dependency parsing, it is generally assumed that the elementary units are word forms or tokens, produced by a tokenizer or word segmenter. A consequence of this assumption is that the shape and size of dependency trees will vary systematically across languages. In particular, morphologically rich languages will typically have fewer elementary units and fewer relations than more analytical languages, which use independent function words instead of morphological inflection to encode grammatical information. This is illustrated in Figure 1, which contrasts two equivalent sentences in English and Finnish, annotated with dependency trees following the guidelines of Universal Dependencies (UD) (Nivre et al., 2016, 2020), which assume words as elementary units.

An alternative view, found in the seminal work of Tesnière (1959), is that dependency relations hold between slightly more complex units called nuclei, semantically independent units consisting of a content word together with its grammatical markers, regardless of whether the latter are realized as independent words or not. Thus, a nucleus will often correspond to a single word – as in the English verb chased, where tense is realized solely through morphological inflection – but it may also correspond to several words – as in the English verb group has chased, where tense is realized by morphological inflection in combination with an auxiliary verb. The latter type is known as a dissociated nucleus. If we assume that the elementary syntactic units of a dependency tree are nuclei rather than word forms, then the English and Finnish sentences will have the same dependency trees, visualized in Figure 2, and will differ only in the realization of their nuclei. In particular, while nominal nuclei in Finnish are consistently realized as single nouns inflected for case, the nominal nuclei in English involve standalone articles and the preposition from.

In this paper, we set out to investigate whether research on dependency parsing can benefit from making explicit use of Tesnière’s notion of nucleus, from the point of view of accuracy, interpretability and evaluation. We do this from a multilingual perspective, because it is likely that the effects of introducing nuclei will be different in different languages, and we strongly believe that a comparison between different languages is necessary in order to assess the potential usefulness of this notion. We are certainly not the first to propose that Tesnière’s notion of nucleus can be useful in parsing. One of the earliest formalizations of dependency grammar for the purpose of statistical parsing, that of Samuelsson (2000), had this notion at its core, and Sangati and Mazza (2009) presented a conversion of the Penn Treebank of English to Tesnière style...
representations, including nuclei. However, previous attempts have been hampered by the lack of available parsers and resources to test this hypothesis on a large scale. Thus, the model of Samuelsson (2000) was never implemented, and the treebank conversion of Sangati and Mazza (2009) is available only for English and in a format that no existing dependency parser can handle. We propose to overcome these obstacles in two ways. On the resource side, we will rely on UD treebanks and exploit the fact that, although the annotation is word-based, the guidelines prioritize dependency relations between content words that are the cores of syntactic nuclei, which facilitates the recognition of dissociated nuclei and gives us access to annotated resources for a wide range of languages. On the parsing side, we will follow a transition-based approach, which can relatively easily be extended to include operations that create representations of syntactic nuclei, as previously shown by de Lhoneux et al. (2019a), something that is much harder to achieve in a graph-based approach.

2 Related Work

Dependency-based guidelines for syntactic annotation generally discard the nucleus as the basic syntactic unit in favor of the (orthographic) word form, possibly with a few exceptions for fixed multiword expressions. A notable exception is the three-layered annotation scheme of the Prague Dependency Treebank (Hajič et al., 2000), where nucleus-like concepts are captured at the tectgrammatical level according to the Functional Generative Description (Sgall et al., 1986). Bārzdinš et al. (2007) propose a syntactic analysis model for Latvian based on the x-word concept analogous to the nucleus concept. In this grammar, an x-word acts as a non-terminal symbol in a phrase structure grammar and can appear as a head or dependent in a dependency tree. Nesporte et al. (2010) compare this model to the original dependency formalism of Tesnière (1959). Finally, as already mentioned, Sangati and Mazza (2009) develop an algorithm to convert English phrase structure trees to Tesnière style representations.

When it comes to syntactic parsing, Järvinen and Tapanainen (1998) were pioneers in adapting Tesnière’s dependency grammar for computational processing. They argue that the nucleus concept is crucial to establish cross-linguistically valid criteria for headedness and that it is not only a syntactic primitive but also the smallest semantic unit in a lexicographical description. As an alternative to the rule-based approach of Järvinen and Tapanainen (1998), Samuelsson (2000) defined a generative statistical model for nucleus-based dependency parsing, which however was never implemented.

The nucleus concept has affinities with the chunk concept found in many approaches to parsing, starting with Abney (1991), who proposed to first find chunks and then dependencies between chunks, an idea that was generalized into cascaded parsing by Buchholz et al. (1999) among others. It is also clearly related to the vibhakti level in the Paninian computation grammar framework (Bharati and Sngal, 1993; Bharati et al., 2009). In a similar vein, Kudo and Matsumoto (2002) use cascaded chunking for dependency parsing of Japanese, Tongchim
et al. (2008) show that base-NP chunking can significantly improve the accuracy of dependency parsing for Thai, and Durgar El-Kahlout et al. (2014) show that chunking improves dependency parsing of Turkish. Das et al. (2016) study the importance of chunking in the transfer parsing model between Hindi and Bengali, and Lacroix (2018) show that NP chunks are informative for universal part-of-speech tagging and dependency parsing.

In a more recent study, de Lhoneux et al. (2019b) investigate whether the hidden representations of a neural transition-based dependency parser encodes information about syntactic nuclei, with special reference to verb groups. They find some evidence that this is the case, especially if the parser is equipped with a mechanism for recursive subtree composition of the type first proposed by Štønetorp (2013) and later developed by Dyer et al. (2015) and de Lhoneux et al. (2019a). The idea is to use a composition operator that recursively combines information from subtrees connected by a dependency relation into a representation of the new larger subtree. In this paper, we will exploit variations of this technique to create parser-internal representations of syntactic nuclei, as discussed in Section 4. However, first we need to discuss how to identify nuclei in UD treebanks.

3 Syntactic Nuclei in UD

UD1 (Nivre et al., 2016, 2020) is an ongoing project aiming to provide cross-linguistically consistent morphosyntactic annotation of many languages around the world. The latest release (v2.7) contains 183 treebanks, representing 104 languages and 20 language families. The syntactic annotation in UD is based on dependencies and the elementary syntactic units are assumed to be words, but the style of the annotation makes it relatively straightforward to identify substructures corresponding to (dissociated) nuclei. More precisely, UD prioritizes direct dependency relations between content words, as opposed to relations being mediated by function words, which has two consequences. First, incoming dependencies always go to the lexical core of a nucleus.2 Second, function words are normally leaves of the dependency tree, attached to the lexical core with special dependency relations, which we refer to as functional relations.3

Figure 3 illustrates these properties of UD representations by showing the dependency tree for the English sentence This killing of a respected cleric will be causing us trouble for years to come with functional relations drawn below the sentence and other relations above. Given this type of representation, we can define a nucleus as a subtree where all internal dependencies are functional relations, as indicated by the ovals in Figure 3. The nuclei can be divided into single-word nuclei, whitened, and dissociated nuclei, grayed. The latter can be contiguous or discontiguous, as shown by the nucleus of a cleric, which consists of the two parts colored with a darker shade.

This definition of nucleus in turn depends on what we define to be functional relations. For this study, we assume that the following 7 UD relations4 belong to this class:

- Determiner (det): the relation between a determiner, mostly an article or demonstrative, and a noun. Especially for articles, there is considerable cross-linguistic variation. For example, definiteness is expressed by an independent function word in English (the), by a morphological inflection in Swedish (flicka-), and not at all in Finnish.
- Case marker (case): the relation between a noun and a case marker when it is a separate syntactic word and not an affix. UD takes a radical approach to adpositions and treats them all as case markers. Thus, in Figure 1, we see that the English adposition from corresponds to the Finnish elative case inflection.
- Classifier (clf): the relation between a classifier, a counting unit used for conceptual classification of nouns, and a noun. This relation is seen in languages which have a classification system such as Chinese. For example, English three students corresponds to Chinese 三个学生, literally “three [human-classifier] student”.
- Auxiliary (aux): the relation between an auxiliary verb or nonverbal TAME marker and a verbal predicate. An example is the English verb group will be causing in Figure 3, which

---

1https://universaldependencies.org
2Except in some cases of ellipsis, like she did, where the auxiliary verb did is “promoted” to form a nucleus on its own.
3A more detailed description of the relations is available in the UD documentation at https://universaldependencies.org.
This killing of a respected cleric will be causing us trouble for years to come.

Given the definition of nucleus in terms of functional UD relations, it would be straightforward to convert the UD representations to dependency trees where the elementary syntactic units are nuclei rather than words. However, the usefulness of such a resource would currently be limited, given that it would require parsers that can deal with nucleus recognition, either in a preprocessing step or integrated with the construction of dependency trees, and such parsers are not (yet) available. Moreover, evaluation results would not be comparable to previous research. Therefore, we will make use of the nucleus concept in UD in three more indirect ways:

- **Evaluation**: Even if a parser outputs a word-based dependency tree in UD format, we can evaluate its accuracy on nucleus-based parsing by simply not scoring the functional relations. This is equivalent to the Content Labeled Attachment Score (CLAS) previously proposed by Nivre and Fang (2017), and we will use this score as a complement to the standard Labeled Attachment Score (LAS) in our experiments.

- **Nucleus Composition**: Given our definition of nucleus-internal relations, we can make parsers aware of the nucleus concept by differentiating the way they predict and represent dissociated nuclei and dependency structures, respectively. More precisely, we will make use of composition operations to create internal representations of (dissociated) nuclei, as discussed in detail in Section 4 below.

- **Oracle Parsing**: To establish an upper bound on what a nucleus-aware parser can achieve, we will create a version of the UD representation which is still a word-based dependency

---

5In addition to separating the cc relation from the rest, such a division might include distinguishing nominal nucleus relations (det, case and clf) from predicate nucleus relations.

6Our use of CLAS differs only in that we include punctuation in the evaluation, whereas Nivre and Fang (2017) excluded it.
tree, but where nuclei are explicitly represented by letting the word form for each nucleus core be a concatenation of all the word forms that are part of the nucleus. We call this oracle parsing to emphasize that the parser has oracle information about the nuclei of a sentence, although it still has to predict all the syntactic relations.

4 Syntactic Nuclei in Transition-Based Dependency Parsing

A transition-based dependency parser derives a dependency tree from the sequence of words forming a sentence (Yamada and Matsumoto, 2003; Nivre, 2003, 2004). The parser constructs the tree incrementally by applying transitions, or parsing actions, to configurations consisting of a stack $S$ of partially processed words, a buffer $B$ of remaining input words, and a set of dependency arcs $A$ representing the partially constructed dependency tree. The process of parsing starts from an initial configuration and ends when the parser reaches a terminal configuration. The transitions between configurations are predicted by a history-based model that combines information from $S$, $B$ and $A$.

For the experiments in this paper, we use a version of the arc-hybrid transition system initially proposed by Kuhlmann et al. (2011), where the initial configuration has all words $w_1, \ldots, w_n$ plus an artificial root node $r$ in $B$, while $S$ and $A$ are empty. There are four transitions: Shift, Left-Arc, Right-Arc and Swap. Shift pushes the first word $b_0$ in $B$ onto $S$ (and is not permissible if $b_0 = r$). Left-Arc attaches the top word $s_0$ in $S$ to $b_0$ and removes $s_0$ from $S$, while Right-Arc attaches $s_0$ to the next word $s_1$ in $S$ and removes $s_0$ from $S$. Swap, finally, moves $s_1$ back to $B$ in order to allow the construction of non-projective dependencies.

Our implementation of this transition-based parsing model is based on the influential architecture of Kiperwasser and Goldberg (2016), which takes as input a sequence of vectors $x_1, \ldots, x_n$ representing the input words $w_1, \ldots, w_n$ and feeds these vectors through a BiLSTM that outputs contextualized word vectors $v_1, \ldots, v_n$, which are stored in the buffer $B$. Parsing is then performed by iteratively applying the transition predicted by an MLP taking as input a small number of contextualized word vectors from the stack $S$ and the buffer $B$. More precisely, in the experiments reported in this paper, the predictions are based on the two top items $s_0$ and $s_1$ in $S$ and the first item $b_0$ in $B$. In a historical perspective, this may seem like an overly simplistic prediction model, but recent work has shown that more complex feature vectors are largely superfluous thanks to the BiLSTM encoder (Shi et al., 2017; Falenska and Kuhn, 2019).

The transition-based parser as described so far does not provide any mechanism for modeling the nucleus concept. It is a purely word-based model, where any more complex syntactic structure is represented internally by the contextualized vector of its head word. Specifically, when two substructures $h$ and $d$ are combined in a Left-Arc or Right-Arc transition, only the vector $v_h$ representing the syntactic head is retained in $S$ or $B$, while the vector $v_d$ representing the syntactic dependent is removed from $S$. In order to make the parser sensitive to (dissociated) nuclei in its internal representations, we follow de Lhoneux et al. (2019a) and augment the Right-Arc and Left-Arc actions with a composition operation. The idea is that, whenever the substructures $h$ and $d$ are combined with label $l$, we replace the current representation of $h$ with the output of a function $f(h, d, l)$. We can then control the information flow for nuclei and other constructions through the definition of $f(h, d, l)$.

**Hard Composition:** The simplest version, which we call hard composition, is to explicitly condition the composition on the dependency label $l$. In this setup, $f(h, d, l)$ combines the head and dependent vectors only if $l$ is a functional relation and simply returns the head vector otherwise:

$$f(h, d, l) = \begin{cases} \vec{h} \circ \vec{d} & \text{if } l \in F \\ \vec{h} & \text{otherwise} \end{cases} \tag{1}$$

We use $\vec{x}$ to denote the vector representation of $x$. In the baseline parser, $\vec{h}$ is always identical to the contextualized representation $v_h$ of the head word $w_h$, but after introducing composition operations we need a more abstract notation.
where \( l \) we do not restrict the application to functional re-
dition and how transition-based parsers can be made
tactic nuclei can be identified in the UD annota-
5 Experiments

In the previous sections, we have shown how syn-
tactic nuclei can be identified in the UD annota-
how to a generalized version of soft composition, where
sensitive to these structures in their internal repre-
We now proceed to a set of experiments investigat-
ing the impact of nucleus composition on a diverse

5.1 Experimental Settings

We use UUParser (de Lhoneux et al., 2017a; Smith et al., 2018), an evolution of the transition-based dependency parser of Kiperwasser and Goldberg (2016), which was the highest ranked transition-based dependency parser in the CoNLL shared task on universal dependency parsing in 2018 (Zeman et al., 2018). As discussed in Section 4, this is a greedy transition-based parser based on the extended arc-hybrid system of de Lhoneux et al. (2017b). It uses an MLP with one hidden layer to predict transitions between parser configurations, based on vectors representing two items on the stack \( S \) and one item in the buffer \( B \). In the baseline model, these items are contextualized word representations produced by a BiLSTM with two hidden layers. The input to the BiLSTM for each word is the concatenation of a randomly initialized word embedding and a character-based representation produced by running a BiLSTM over the character sequence of the word. We use a dimensionality of 100 for the word embedding as well as for the output of the character BiLSTM.

For parsers with composition, we considered various composition operators \( \circ \) and functions \( g \). For the former, we tested vector addition, vector concatenation,11 and perceptron. For the latter we tried a multi-layer perceptron with different activation functions. Based on the results of the preliminary experiments, we selected vector addition for the composition operator \( \circ \) and the perceptron with sigmoid activation for the soft composition function \( g \). The inputs to the perceptron consist of two token vectors of size 512 and a relation vector of size 10. The token vectors are the outputs of the BiLSTM layer of the parser and the relation vector is trained by a distinct embedding layer.

All parsers are trained for 50 epochs and all reported results are averaged over 10 runs with different random seeds. Altogether we explore five different parsers:

\[ f(h, d, l) = h \circ g(h, d, l) \]

where \( l \) can be any dependency label. In this approach, the if-clause in Equation 2 and 3 is eliminated and the parser itself learns in what conditions the composition should be performed. In particular, if the composition operator is addition, and \( g \) is a perceptron with a sigmoid activation on the output layer (as in Equation 2), then \( g \) operates as a gate that controls the contribution of the dependency elements \( h, d, \) and \( l \) to the composition. If the composition should not be performed, it returns a vector close to zero.

5 Experiments

In the previous sections, we have shown how syntactic nuclei can be identified in the UD annota-
and how transition-based parsers can be made

Soft Composition: The soft composition is similar to the hard composition, but instead of applying the composition operator to the head and dependent vectors, the operator is applied to the head vector and a vector representation of the entire dependency arc \((h, d, l)\). The vector representation of the dependency arc is trained by a differentiable function \( g \) that encodes the dependency label \( l \) into a vector \( \vec{l} \) and maps the triple \((\vec{h}, \vec{d}, \vec{l})\) to a vector space, i.e., \( g : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n \) where \( n \) and \( m \) are the dimensionalities of the word and label spaces, respectively. An example of \( g \) is a perceptron with a sigmoid activation that maps the vector representations of \( h, d \) and \( l \) to a vector space:

\[ g(h, d, l) = \sigma(W(\vec{h} \circ \vec{d} \circ \vec{l}) + b) \tag{2} \]

where \( \circ \) is the vector concatenation operator. The soft nucleus composition is then:

\[ f(h, d, l) = \begin{cases} h \circ g(h, d, l), & \text{if } l \in F \\ \vec{h} & \text{otherwise} \end{cases} \tag{3} \]

The parameters of the function \( g \) are trained with the other parameters of the parser.

Generalized Composition: To test our hypothesis that composition is beneficial for dissociated nuclei, we contrast both hard and soft composition to a generalized version of soft composition, where we do not restrict the application to functional relations. In this case, the composition function is:

\[ f(h, d, l) = g(h, d, l) \]

where \( l \) can be any dependency label. In this approach, the if-clause in Equation 2 and 3 is eliminated and the parser itself learns in what conditions the composition should be performed. In particular, if the composition operator is addition, and \( g \) is a perceptron with a sigmoid activation on the output layer (as in Equation 2), then \( g \) operates as a gate that controls the contribution of the dependency elements \( h, d, \) and \( l \) to the composition. If the composition should not be performed, it returns a vector close to zero.

\[ 11 \text{The concatenation operator requires special care to keep vector dimensionality constant. We double the dimensionality of the contextual vectors and fill the extra dimensions with zeros. We then replace the zero part of the second operand with the first operand’s non-zero part at composition time.} \]

\[ \mathbb{R}^n, \text{ e.g., vector addition } \vec{h} + \vec{d}, \text{ where } n \text{ is the dimensionality of the vector space.} \]
Table 1: Parsing accuracy for 5 parsing models evaluated on 12 UD treebanks. Language family includes genus according to WALS for large families (AA = Afro-Asian, IE = Indo-European). LAS = Labeled Attachment Score. CLAS = Content Labeled Attachment Score.

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
<th>Family</th>
<th>Size</th>
<th>LAS</th>
<th>Base</th>
<th>Hard</th>
<th>Soft</th>
<th>Gen</th>
<th>Ora</th>
<th>CLAS</th>
<th>Base</th>
<th>Hard</th>
<th>Soft</th>
<th>Gen</th>
<th>Ora</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>PADT</td>
<td>AA-Semitic</td>
<td>242K</td>
<td>78.0</td>
<td>78.0</td>
<td>78.4</td>
<td>78.5</td>
<td>80.5</td>
<td>74.4</td>
<td>75.0</td>
<td>75.1</td>
<td>75.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>BDT</td>
<td>Basque</td>
<td>70K</td>
<td>73.5</td>
<td>73.6</td>
<td>74.1</td>
<td>75.0</td>
<td>78.9</td>
<td>70.9</td>
<td>71.1</td>
<td>71.5</td>
<td>72.6</td>
<td>75.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>GSD</td>
<td>Sino-Tibetan</td>
<td>49K</td>
<td>72.1</td>
<td>72.1</td>
<td>72.8</td>
<td>72.8</td>
<td>80.4</td>
<td>68.8</td>
<td>68.6</td>
<td>69.5</td>
<td>69.5</td>
<td>74.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>EWT</td>
<td>IE-Germanic</td>
<td>188K</td>
<td>82.5</td>
<td>82.4</td>
<td>82.7</td>
<td>82.6</td>
<td>85.6</td>
<td>78.8</td>
<td>78.7</td>
<td>79.0</td>
<td>78.9</td>
<td>81.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>TDT</td>
<td>Uralic</td>
<td>168K</td>
<td>78.5</td>
<td>78.7</td>
<td>79.5</td>
<td>80.1</td>
<td>84.2</td>
<td>77.5</td>
<td>77.7</td>
<td>78.6</td>
<td>79.1</td>
<td>81.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>HTB</td>
<td>AA-Semitic</td>
<td>108K</td>
<td>81.5</td>
<td>81.5</td>
<td>81.9</td>
<td>82.3</td>
<td>83.3</td>
<td>74.5</td>
<td>74.5</td>
<td>75.1</td>
<td>75.6</td>
<td>75.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>HDTB</td>
<td>IE-Indic</td>
<td>294K</td>
<td>87.9</td>
<td>88.2</td>
<td>88.4</td>
<td>89.4</td>
<td>89.8</td>
<td>83.9</td>
<td>84.1</td>
<td>84.5</td>
<td>85.8</td>
<td>85.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>ISDT</td>
<td>IE-Romance</td>
<td>252K</td>
<td>87.4</td>
<td>87.6</td>
<td>87.8</td>
<td>87.8</td>
<td>89.0</td>
<td>81.5</td>
<td>81.6</td>
<td>82.1</td>
<td>82.1</td>
<td>82.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>GSD</td>
<td>Japanese</td>
<td>61K</td>
<td>93.4</td>
<td>93.4</td>
<td>93.6</td>
<td>93.5</td>
<td>94.1</td>
<td>89.7</td>
<td>89.8</td>
<td>90.1</td>
<td>89.9</td>
<td>90.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>GSD</td>
<td>Korean</td>
<td>21K</td>
<td>75.1</td>
<td>75.0</td>
<td>75.4</td>
<td>75.8</td>
<td>76.6</td>
<td>74.9</td>
<td>74.8</td>
<td>75.2</td>
<td>75.7</td>
<td>75.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>Talbanken</td>
<td>IE-Germanic</td>
<td>61K</td>
<td>76.9</td>
<td>77.2</td>
<td>77.5</td>
<td>78.0</td>
<td>82.9</td>
<td>73.2</td>
<td>73.4</td>
<td>73.9</td>
<td>74.5</td>
<td>78.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>IMST</td>
<td>Turkic</td>
<td>40K</td>
<td>55.6</td>
<td>55.7</td>
<td>56.3</td>
<td>56.6</td>
<td>58.6</td>
<td>53.4</td>
<td>53.6</td>
<td>54.3</td>
<td>54.6</td>
<td>54.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td>78.5</td>
<td>78.6</td>
<td>79.0</td>
<td>79.4</td>
<td>82.0</td>
<td>75.1</td>
<td>75.2</td>
<td>75.7</td>
<td>76.1</td>
<td>77.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Base(line):** No composition.
- **Hard:** Baseline + hard composition.
- **Soft:** Baseline + soft composition.
- **Gener(eralized):** Baseline + gen. composition.
- **Ora(cle):** Baseline trained and tested on explicit annotation of nuclei (see Section 3).

Our experiments are carried out on a typologically diverse set of languages with different degrees of morphosyntactic complexity, as shown in Table 1. The corpus size is the total number of words in each treebank. We use UD v2.3 with standard data splits (Nivre et al., 2018). All evaluation results are on the development sets.\(^\text{12}\)

### 5.2 Results

Table 1 reports the parsing accuracy achieved with our 5 parsers on the 12 different languages, using the standard LAS metric as well as the nucleus-aware CLAS metric. First of all, we see that hard composition is not very effective and mostly gives results in line with the baseline parser, with significant improvements only for Finnish and Hindi (two-tailed t-test, \(p = .05\)). By contrast, soft composition improves accuracy for all languages and the improvements are statistically significant for both LAS and CLAS except for LAS in English \((p = 0.06)\). The average improvement is 0.5 and 0.6 percentage points for LAS and CLAS, respectively, which indicates that most of the improvement occurs on nucleus-external relations thanks to a more effective internal representation of dissociated nuclei. Generalized composition, finally, where we allow composition also for non-functional relations, is again significantly better than the baseline for all languages with respect to CLAS and for all languages except English with respect to LAS \((p = 0.18)\). For six languages (Basque, Finnish, Hebrew, Hindi, Korean and Swedish), generalized composition also gives significant improvements over soft composition (for both LAS and CLAS), but for the remaining languages there is little or no improvement.\(^\text{13}\) A tentative conclusion is therefore that, while composition is almost universally beneficial when applied to nucleus-internal relations, its usefulness for other types of syntactic constructions is more sensitive to language-specific properties and needs further study.

Before we try to analyze the results in more detail, it is worth noting that most of the improvements due to composition are considerably below the improvements of the oracle parser.\(^\text{14}\) However, it is important to keep in mind that, whereas the behavior of a composition parser is only affected

\(^{12}\)Since we want to perform an informative error analysis, we avoid using the dedicated test sets.

\(^{13}\)For Japanese, there is even a significant decrease in CLAS.

\(^{14}\)The exceptions are found in the CLAS results for generalized composition, which are better than the oracle parser for Hindi, equal for Korean, and fairly close for Japanese and Turkish.
after a nucleus has been constructed, the oracle parser improves also with respect to the prediction of the nuclei themselves. This explains why the oracle parser generally improves more with respect to LAS than CLAS, by about one percentage point on average but by as much as 2.5 points for Chinese and 1.6 points for Turkish.

Figure 4 visualizes the impact of hard, soft and generalized nucleus composition for different languages, with a breakdown into (a) all relations, which corresponds to the difference in LAS compared to the baseline, (b) nucleus-external relations, which corresponds to the difference in CLAS, and (c) nucleus-internal relations. Overall, these graphs are consistent with the hypothesis that using composition to create parser-internal representations of (dissociated) nuclei primarily affects the prediction of nucleus-external relations, as the (a) and (b) graphs are very similar and the (c) graphs mostly show very small differences. There are, however, a few notable exceptions. For Finnish, all three composition methods clearly improve the prediction of nucleus-internal relations as well as nucleus-external relations, by over 1 F-score point for generalized composition. Similar patterns, with smaller magnitude, can be seen for Chinese and Hindi, possibly also for Basque and Swedish. Conversely, for Turkish, especially hard and generalized composition has a detrimental effect on the prediction of nucleus-internal relations. Further research will be needed to explain what causes these deviant patterns.

To better understand the impact of nucleus composition, we now look at the effect on individual dependency relations. We focus on the soft composition model, which most directly implements the
Table 2: Improvement (or degradation) in labeled F-score, weighted by relative frequency, for the 10 best UD relations in the 5 languages with greatest LAS improvements over the baseline (soft composition).

<table>
<thead>
<tr>
<th>Finnish</th>
<th>Chinese</th>
<th>Swedish</th>
<th>Turkish</th>
<th>Basque</th>
</tr>
</thead>
<tbody>
<tr>
<td>conj</td>
<td>0.09</td>
<td>obl</td>
<td>conj</td>
<td>0.06</td>
</tr>
<tr>
<td>obl</td>
<td>0.06</td>
<td>case</td>
<td>root</td>
<td>0.08</td>
</tr>
<tr>
<td>nmod</td>
<td>0.06</td>
<td>amod</td>
<td>obj</td>
<td>0.07</td>
</tr>
<tr>
<td>obj</td>
<td>0.04</td>
<td>ccomp</td>
<td>ccomp</td>
<td>0.02</td>
</tr>
<tr>
<td>root</td>
<td>0.03</td>
<td>obl</td>
<td>obj</td>
<td>0.04</td>
</tr>
<tr>
<td>nsubj</td>
<td>0.02</td>
<td>obj</td>
<td>root</td>
<td>0.04</td>
</tr>
<tr>
<td>advcl</td>
<td>0.02</td>
<td>acl</td>
<td>conj</td>
<td>0.03</td>
</tr>
<tr>
<td>amod</td>
<td>0.02</td>
<td>advmod</td>
<td>nmod</td>
<td>0.03</td>
</tr>
<tr>
<td>ccomp</td>
<td>0.02</td>
<td>det</td>
<td>acl</td>
<td>0.02</td>
</tr>
<tr>
<td>acl</td>
<td>0.01</td>
<td>root</td>
<td>advcl</td>
<td>0.01</td>
</tr>
</tbody>
</table>

idea of nucleus composition (as opposed to composition in general) and which also provides the most stable improvements across languages. Figure 5 shows the improvement (or degradation) for individual UD relations, weighted by relative frequency and averaged over all languages. The most important improvements are observed for conj, obj, nmod, and root. The conj relation holds between conjuncts in a coordination structure; obj is the direct object relation; the nmod relation covers all nominal modifiers inside noun phrases, including prepositional phrase modifiers; and the root relation is assigned to the main predicate of a sentence. In addition, we see substantial improvements (of at least 1 F-score point) for four major clause relations – nsubj (nominal subjects), advcl (adverbal clauses), ccomp (complement clauses), and obl (oblique modifiers) – as well as smaller improvements for a long tail of additional relations. Of these, only aux, clf, and det are nucleus-internal relations and these three relations show very marginal improvements, so the results further support the hypothesis that richer internal representations of (disassociated) nuclei primarily improve the prediction of nucleus-external dependency relations, especially major clause relations.

It is important to remember that the results in Figure 5 are averaged over all languages and may hide interesting differences between languages. A full investigation of this variation is beyond the scope of this paper, but Table 2 presents a further zoom-in by presenting statistics on the top 10 relations in the 5 languages where LAS improves the most compared to the baseline. To a large extent, we find the same relations as in the aggregated statistics, although with slightly different orders, but there are also interesting language-specific patterns. For Chinese one of the relations tied for second place is the nucleus-internal case relation. For Swedish one of the relations tied for second place is amod (adjectival modifier), which does not show positive improvements on average (although it is in the top ten also for Finnish and Turkish). These differences definitely deserve further investigation.

6 Conclusion

We have explored how the concept of syntactic nucleus can be used to enrich the representations of a transition-based dependency parser, relying on UD treebanks for supervision and evaluation in experiments on a wide range of languages. We conclude that the use of composition operations for building internal representations of syntactic nuclei, in particular the technique that we have called soft composition, can lead to small but significant improvements in parsing accuracy for nucleus-external relations, notably for relations of coordination, direct objects, nominal modifiers and main predicates. For some languages, further improvements can be obtained by applying composition also to the nucleus-external relations themselves (generalized composition). In future work we want to study the behavior of different types of nuclei in more detail, in particular how the different internal relations of nominal and verbal nuclei contribute to overall parsing accuracy. We also want to analyze the variation between different languages in more detail and see if it can be explained in terms of typological properties.


Acknowledgments

We thank Daniel Dakota, Artur Kulmizev, Sara Stymne and Gongbo Tang for useful discussions and the EACL reviewers for constructive criticism. The research presented in this paper was supported by grant 2016-01817 of the Swedish Research Council. We acknowledge the computational resources provided by CSC in Helsinki and Sigma2 in Oslo through NeIC-NLPL (www.nlpl.eu).

References


tics (ÚFAL), Faculty of Mathematics and Physics, Charles University.


