Word Grouping

Chaitali Dangarikar[1], Arnab Bhattacharya[1], Karthika N J[2], Chaitanya S Lakkundi[3], Ganesh Ramakrishnan[2], Annarao Kulkarni[4], Shivani V[3], Pramit Bhattacharyya[1], and Hrishikesh Terdalkar[1]

The power of words in a sentence (termed as śakti or vr̥tti) has been studied for millennia by Sanskrit scholars and grammarians. Vākyapadīyam, etc. go to astonishing depths on this. Most of the general principles for Sanskrit apply to most of the contemporary, major Indian languages.

A word or pada, for Sanskrit, is defined as a unit that can be used in a sentence. The most important semantic part consists of a root, which is either verbal, called *dhātu*, or non-verbal, called *prātipadika*. The root is joined with a suffix, called pratyaya (and sometimes a prefix). Multiple suffixes and prefixes can be attached, and the entire combination is considered a single "word" or "pada". Inflections for both verbs and non-verbs are, thus, part of the word. Case, gender, tense, person, noun, etc. are all encoded as different pratyaya that are fused with the word. In written text, words are typically separated by whitespaces (blanks, punctuation marks, etc.).

While the above is, again, largely true for most of the major Indian languages, there are notable exceptions in typographic conventions. Hindi, for example, mostly separates the vibhakti endings for case from the word root. Thus, the construct "of Rama" is written as "राम का" (rāma kā) in Hindi. The same is written without any space between the root and the suffix is languages such as Sanskrit (रामस्य / rāmasya), Bangla (রামের / rāmēra), Marathi (रामाचे / rāmāce), Kannada (ರಾಮನ / rāmana), Malayalam (രാമന്റെ / rāmanr̲e), etc. Similarly, for verbs, different tense, aspect and mood markers for verbs are written separately for many languages, for example, "किया था" (kiyā thā, had done) in Hindi, "केले होते" (kēlē hōtē, had done) in Marathi. While languages such as Kannada and Malayalam are more agglutinative and examples of such separations are rare, there are languages that are midway, such as

[1] IIT Kanpur

[2] IIT Bombay

[3] Karnataka Samskrit University

[4] C-DAC Bangalore

Bangla. For example, the fifth case denoting ablation is denoted by a suffix followed by a separately written word – "রামের থেকে" (rāmēra thēkē, from Rama).

In order to unify these, we propose the concept of "**word groups**". A *word group* is a group of words (which, for our purposes, is a whitespace separated sequence of characters) that semantically denote *a single meaning*, and is *fit to be used in a sentence* without any further modification or addition of other words (or word groups). Thus, for the above mentioned examples, "राम का" (rāma kā), "किया था" (kiyā thā) and "রামের থেকে" (rāmēra thēkē) are all single word groups. Once word groups are identified, they are treated as *single words* for further purposes.

There are several advantages of using word groups over words separated by white spaces:

1. **Unity in Semantics:** The entire word group together denotes a single meaning. The words by themselves may or may not have meanings. Even when they have a meaning, that may not be appropriate for the overall meaning. For example, consider the Bangla word group "রামের থেকে" (rāmēra thēkē). The first word, "রামের" (rāmēra), if treated separately, means "of Rama" (i.e., the sixth or the genitive case). This is clearly a wrong understanding of the usage of the word here. Hence, only when words are grouped, does this unity in semantics is established.
2. **Unity in Morphology:** Most Indian languages have similar and common word roots. The similarity in a sentence is evident only when word groups are considered as a single unit and not separately. As an example, consider the equivalents "रामात्" (rāmāt, Sanskrit), "राम से" (rāma sē, Hindi) and "রামের থেকে" (rāmēra thēkē, Bangla). Only when the entire word groups are considered, do they mean the same construct "from Rama". Individually, the word "राम" (rāma) in Hindi denotes the first or subjective case (Rama), while the word "রামের" (rāmēra) in Bangla denotes the sixth or genitive case (of Rama). Consequently, the morphological tags of the word groups remain the same. (They become unnecessarily fragmented when words are treated individually.)
3. **Unity in Translation:** The unity in morphology directly leads to unity in translation. Machine translational systems would benefit immensely if word groups are input as units (or tokens) rather than individual words.
4. **Unity in Dependency Relationships:** Dependency relationships mark the semantic connections between the units in a sentence. Relationships between word groups across most major Indian languages remain typically the same. In fact, for many simple and canonical sentences, the dependency parse tree structures are exactly the same. Unnecessarily separating a word group

into whitespace separated words destroys this unity. In addition, it requires additional and non-semantic dependency relations such as “post-positions” to connect a separately written inflection marker with the main word (e.g., “से”, sē to “राम”, rāma).

We propose word groups as follows:

1. **Inflectional Unity:**
   - **Noun and Inflectional Morphemes:** Group nouns with their inflectional morphemes/suffixes (e.g., को, ने, से) to achieve inflectional unity.
   - **Example:** "राम ने किताब पढ़ी थी" (Ram read the book) 5 word sentence would be tokenized as "राम##ने किताब पढ़ी##थी" (having only 3 words).
   - **Grouping of Postpositions with Nouns in Hindi:** In Hindi, certain अव्यय (avyaya) words such as "के लिए," "के साथ," "के पास," etc., are treated as vibhakti pratyayas (case suffixes) in specific grammatical contexts. However, since "लिए," "साथ," "पास," etc., are themselves अव्यय and do not function as case markers, they should not be grouped with the preceding noun with its postpostion. For example "राम के साथ" (Rām ke sātha) → {राम##के} साथ and not as {राम##के##साथ}

     This rule clarifies that while certain अव्यय phrases may appear to function like case markers in Hindi, they should not be grouped with the nouns they accompany. This distinction is crucial for maintaining grammatical accuracy and clarity in sentence construction.
2. **Semantic Unity:**
   - **Verbs and Auxiliary Verbs:** Group verbs and auxiliary verbs to achieve semantic unity.
   - **Example:** "जा रहा हूँ" (am going) 3 words would be tokenized as 1 word "जा##रहा##हूँ".
3. **Named Entities:**
   - **Compound Treatment:** Treat named entities as a compound unit.
   - **Example:** "भारतीय प्रौद्योगिकी संस्थान" these 3 words would be tokenized as 1 wor"भारतीय##प्रौद्योगिकी##संस्थान".
4. **Word-Splitting:** In languages like Marathi, words like बरोबर, बद्दल, पुढे, etc. are often attached to the preceding noun, with the अकारान्त noun taking an आ suffix, आकारान्त noun taking ए suffix before these words. For example: राम बरोबर -> रामाबरोबर, शाळा बद्दल -> शाळेबद्दल To map these Marathi constructs to their corresponding Sanskrit उपपद (upapada) equivalents, it would be beneficial to split these words into separate tokens. This would allow for better alignment and

mapping between Marathi and Sanskrit. For example: राम बरोबर -> रामाबरोबर, शाळा बद्दल -> शाळेबद्दल. This splitting would result in the following tokens: रामा##बरोबर, and शाळे##बद्दल.

By splitting these constructs, we can better align them with their corresponding Sanskrit उपपद equivalents. This alignment can be useful for tasks like machine translation, cross-lingual information retrieval, and comparative linguistics between Marathi and Sanskrit. It's important to note that this splitting should be done carefully, considering the context and the specific usage of these words in Marathi. In some cases, the attached words may have different meanings or functions depending on the context.

Interlingua Mapping: Establishing a mapping mechanism between the dependency structures of different Indian languages by using Sanskrit grammar as an interlingua ensures that syntactic and semantic information is represented uniformly. This approach facilitates seamless translation and interpretation across languages.

Acknowledgments

Project Details

The work is supported in part as part of a project by the National Language Translation Mission (NLTM): BHASINI scheme of the Ministry of Electronics and Information Technology (MeitY), Govt. of India, the ŚIKṢĀ: Study Centre for Indian Knowledge System for Holistic Advancement at the Indian Institute of Technology Kanpur. We also thank various colleagues and experts; discussions with them helped refine and correct many of the concepts.

We would like to thank our Project Investigator Prof. Pawan Goyal from IIT Kharagpur. We also thank A.V.S.D.S Mahesh (IIT Kanpur) and Pralay Manna (IIT Kanpur) for their valuable contributions to this project.