POS comes with parsing: a refined word categorisation method

Balázs Indig[1,2] and Noémi Vadász[1,3]

[1]MTA–PPKE Hungarian Language Technology Research Group, Budapest, Hungary
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
[3]Pázmány Péter Catholic University, Faculty of Humanities and Social Sciences
{indig.balazs,vadasz.noemi}@itk.ppke.hu
http://nlpg.itk.ppke.hu/

AnaGramma [1,3] is a performance-based, psycholinguistically motivated parsing system with a strictly left-to-right, word-by-word approach and a *supply-and-demand framework*, as it tries to handle utterances by following the patterns of human language processing as much as possible. Each word *supplies* its lexical representation and morpho-syntactic information. According to this these features *demands* are issued (e.g. nouns want to be verbal arguments or adjuncts and verbs have an obligatory need for their arguments). At the end of the utterance all the demands should be fulfilled either from the sentence or with default mechanisms.

We present multiple Hungarian examples on how the words statistically correlate with these supply-and-demand processes. We show that this classification is quite different from the classical view of part of speech (POS) classes. It is rather similar to Word-Expert-Parsing [4] where every word has its own features regardless of its POS tag. We created main and subcategories according to the syntactic and semantic features of the words, based on their statistical behaviour. These categories can start different processes during the parsing: the main category such as the *verb* inherits its processes to the subcategories *finite* and *infinite* beside their own processes. This framework enables us to mix features of classical categories within one word as the processes are issued by the features leaving the classical categories in place.

For example, participle is a classical category incorporating the properties of adjective and verb. It has its own argument structure as participles are derived from a verb. Also it can have two different roles[1] in the sentence: (1) *NP modifier*, when it is a supply for the noun's demand within the limits of the NP containing them, (2) *nominal predicate*, when it demands its supplies from the sentence. During parsing both the verbal and adjectival features start their processes.

In our framework the postposition can be a special case marker, even though the former is a word and the latter is a morpheme. From the view of parsing, after the noun with a zero case-morpheme – which "suggests nominative case" – postposition behaves as a case marker (other than the a priori nominative case) and closes the noun phrase (NP) making it ready as an argument of the main verb.

[1] In addition, participles can have a lexicalised noun form.

Furthermore, the overlap between the class of adverbs and preverbs also corresponds to the aforementioned theory. Adverbs can have argument frame modifying role depending on the actual verb. The decision whether the actual element is classified as an adverb or a preverb is a verb-driven process, therefore it can not be determined a priori by the token itself. During parsing the verb must search for its frame modifier (regardless of the modifier's "pre-issued" POS tag) before it selects the right argument frame (see the details in [2]).

We illustrate how the dissected features that build the classical POS categories start supplies and demands – as the basic principle of AnaGramma – still, our method conforms to the existing theories.

References

1. Indig, B., Prószéky, G.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. Alkalmazott nyelvtudomány 15(1-2), 29–44 (2015)
2. Indig, B., Vadász, N.: Windows in human parsing – how far can a preverb go? In: Marko, T., Bekavac, B. (eds.) Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016). Dubrovnik, Croatia (sept 2016), (Accepted, In press)
3. Prószéky, G., Indig, B., Vadász, N.: Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In: Bence, K. (ed.) "Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére, pp. 223–232. MTA Nyelvtudományi Intézet, Budapest (2016)
4. Small, S.L.: Word expert parsing. In: Proceedings of the 17th Annual Meeting on Association for Computational Linguistics. pp. 9–13. ACL '79, Association for Computational Linguistics, Stroudsburg, PA, USA (1979), http://dx.doi.org/10.3115/982163.982167