Mapping the meaning of Hungarian impulsative constructions

Ágnes Kalivoda
Institute for Lexicology, ELTE Research Centre for Linguistics
kalivoda.agnes@nytud.hu

Robert Malouf
Department of Linguistics and Asian/Middle Eastern Languages, San Diego State University
rmalouf@sdsu.edu

Farrell Ackerman
Department of Linguistics, University of California, San Diego
fackerman@ucsd.edu

1 Introduction

This paper presents a corpus-based investigation of the Hungarian impulsative construction, which uses the deverbal nominalizing suffix *-hAtnék*[1] to derive a noun denoting an immediate, often involuntary urge to perform the action denoted by the verb (Kalivoda et al., 2026).

In its original use, the impulsative construction (Cathcart, 2011) denoted basic physiological impulses, as in (1). These verbs formed a clear semantic cluster grounded in embodied experience, and historically they represent the earliest and most entrenched instances of the construction.

(1) a. *küzd-ött-em az* ***i-hatnék-kal***
fight-PST-1SG the drink-URGE-INS
'I was fighting against an urge to drink'

b. *sír-****hatnék****-om támad-t*
cry-URGE-1SG.POSS arise-3SG.PST
'I got an urge to cry'

However, the construction is now fully productive: speakers freely extend it to increasingly complex and figurative uses (2). The construction has moved beyond its original lexical niche, and it currently functions as a general schematic resource for expressing spontaneous desire or impulse toward any kind of activity.

(2) a. ***gomb-nyomogat-hatnék-ja*** *támad*
button-press-URGE-3SG.POSS arise.3SG
'an urge to press buttons seizes him/her'

b. *rá-jö-tt a* ***tuningol-hatnék***
PV-come-PST.3SG the tune-URGE
he/she got that urge to do some tuning

This trajectory raises a distributional question: is the semantic expansion of the construction reflected in, or constrained by, the statistical associations between the construction and its lexical fillers? To investigate this, we conducted two collostructional analyses on the largest available corpus of Hungarian.

[1]The suffix has two allomorphs *-hatnék* and *-hetnék* distributed following the rules of vowel harmony. Upper case vowel *A* reflects vowel harmonic allophony.

2 Text source

The data used in this study comes from the Hungarian portion of the HPLT Monolingual Datasets 3.0 (Aulamo et al., 2023), a large-scale multilingual corpus covering 198 languages and derived from web crawls spanning 2012 to 2024. The full dataset amounts to approximately 50 terabytes of compressed data, making it one of the largest publicly available multilingual resources to date.

The Hungarian subset totals 731 GB of uncompressed raw text, distributed across 9 files in JSONL format. Each document is accompanied by rich metadata, including web register labels and document quality scores.

3 Data collection

Given the size of the corpus, parsing it in full to retrieve occurrences of a relatively rare construction was not feasible. Instead, we adopted the following workflow:

1. We implemented a fast, parallelized keyword-in-context (KWIC) extractor that distributes work across CPU cores. Each worker parses JSONL documents and applies user-defined regular expressions to the text field, extracting each match together with a 200-character context window. The patterns are designed to accommodate the wide range of suffixes that can attach to *-hAtnék* nouns in agglutinative Hungarian.

2. We parsed the resulting hits and their contexts using HuSpaCy (Orosz et al., 2023) to obtain dependency structures.

3. We queried these dependency parses with the Treesearch (Malouf, 2025) tool to identify impulsatives and their head predicates.

4. We obtained detailed morphological analyses of the impulsatives using emMorph (Novák et al., 2016), separating the root verb from the derivational *-hAtnék* suffix.

5. For the planned collostructional analyses, we also needed corpus-wide frequency counts for each root verb and head predicate. We generated all possible inflected forms of these items using Hunmorph-foma[2] and retrieved their frequencies via string search over the full corpus.

The resulting dataset consists of 13,526 records, each annotated with linguistic features and metadata. For the present study, the most relevant variables are the root verbs inside the impulsative construction and the predicates co-occurring with *-hAtnék* nouns, along with their frequencies both within the construction and across the corpus as a whole.

4 Data analysis

To examine the semantic organization and usage tendencies of the *-hAtnék* construction, we conducted two collostructional analyses (Stefanowitsch and Gries, 2003; Schmid and Küchenhoff, 2013; Perek, 2015). Constructions are pairings of form and meaning. For simple lexical constructions, this is straightforward. For highly schematic constructions (e.g., the English double object construction) it is more difficult to isolate the contribution of the schema. We can approach this problem from two directions: any lexical items that fill a slot in a construction schema must have a meaning that is compatible with the meaning of the schema; and, a schema gains its meaning from the lexical items that tend to fill its slots (Hilpert, 2006).

Collostructional analysis uses measures of association to identify lexical items that fill slots in a construction with greater than chance frequency. Specifically, to compute the association between the verb *iszik* 'drink' and the impulsative construction, we would set up a 2×2 table:

	-hAtnék	¬*-hAtnék*	Total
iszik	a	b	$a+b$
¬*iszik*	c	d	$c+d$
Total	$a+c$	$b+d$	n

Here a is the frequency of *iszik* in this construction, $a+b$ is the total frequency of all forms of *iszik*, $a+c$ is the total frequency of all instances of the impulsative, and n is the total size of the corpus. The collostructional strength of this association is the degree to which these counts deviate from what we would expect if verb choices and construction choice were statistically independent.

5 Results

The first analysis investigated which verbs are most strongly associated with the [V-*hAtnék*]$_{\text{N}}$ configuration, that is, which verbal bases are most strongly attracted by the -hAtnék derivation.[3].

These verbs pattern into the following semantic fields, each of these illustrated with a few examples:

- bodily functions and expulsion: *okád, hány* ('vomit'), *pisil, vizel* ('urinate'), *köp* ('spit')
- emotional/expressive release: *nevet* ('laugh'), *sír* ('cry'), *röhög, kacag* ('laugh/giggle'), *üvölt* ('howl')
- ingestion: *eszik, zabál* ('eat/gorge'), *iszik* ('drink'), *cigiz* ('smoke')
- motion and escape: *megy* ('go'), *menekül* ('flee'), *szökik* ('escape'), *fut* ('run'), *csavarog* ('roam')
- speech and self-display: *mesél* ('tell'), *beszél* ('speak'), *káromkodik* ('curse'), *dicsekszik* ('boast'), *kérkedik* ('brag')

The construction thus typically selects for verbs denoting barely-controllable or uncontrollable bodily, emotional, or social impulses.

The second analysis examined main verbs that co-occur in the [V-*hAtnék*]$_{\text{N}}$ + V pattern. The most strongly attracted verbs form a remarkably coherent semantic class. The construction overwhelmingly recruits verbs that frame the urge as an external force impinging on the experiencer: *támad* ('attack, arise'), *rájön* ('come over someone'),

[2] https://github.com/r0ller/hunmorph-foma

[3] Only collexemes attracted at the highest level of significance ($p < .0001$, log-likelihood $\geq$ ca. 20) are discussed here.

rátör (‘break upon’), *elfog* (‘seize’), *kitör* (‘erupt’), *környékez* (‘approach’), and *bizsereg* (‘tingle’). A smaller cluster encodes the suppression or fading of the urge: *elfojt* (‘suppress’), *visszafog* (‘hold back’), *elmúlik* (‘pass’), *elpárolog* (‘evaporate’), *erőt vesz* (‘overcome’). This supports the interpretation of the construction as profiling an internal, temporally bounded impulse.

6 Conclusion

The impulsative construction illustrates how expressive language and more specifically constructions that encode internal states, affect, or evaluations, can contribute to constructional productivity and reorganization (Gyselinck, 2018). In the case of the impulsative, expressive meaning appears to facilitate both semantic extension and the construction’s integration into larger patterns of clause-level organization.

This case contributes to ongoing work on how constructions can be characterized and compared in distributional terms, and raises questions about how semantic coherence is maintained as constructions become more schematic and productive.

Acknowledgements

Ágnes Kalivoda’s research has been supported by the OTKA PD project No. 142317 funded by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the PD 22 funding scheme.

References

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. HPLT: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 517–518, Tampere, Finland. European Association for Machine Translation.

MaryEllen Cathcart. 2011. *Impulsatives: The syntax and semantics of involuntary desire*. Ph.D. thesis, University of Delaware.

Emmeline Gyselinck. 2018. *The role of expressivity and productivity in (re)shaping the constructional network*. Ph.D. thesis, Ghent University.

Martin Hilpert. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory*, 2(2):243–256.

Ágnes Kalivoda, Farrell Ackerman, and Robert Malouf. 2026. Morphological change as systemically motivated bricolage: Hungarian impulsative constructions. *Morphology*, 36(4):1–33.

Robert Malouf. 2025. Treesearch: Pattern matching for dependency treebanks. https://github.com/rmalouf/treesearch.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

György Orosz, Gergő Szabó, Péter Berkecz, Zsolt Szántó, and Richárd Farkas. 2023. Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In *Text, Speech, and Dialogue*, pages 58–69, Cham. Springer Nature Switzerland.

Florent Perek. 2015. *Argument Structure in Usage-Based Construction Grammar*. John Benjamins, Amsterdam/Philadelphia.

Hans-Jörg Schmid and Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction. *Cognitive Linguistics*, 24(3):531–577.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.