ELTE | FACULTY OF HUMANITIES

Affixation, back-formation, or both?

A corpus-based study on the multiple motivation of Hungarian compound verbs

László Palágyi, Ágnes Kalivoda

ELTE Eötvös Loránd University, ELTE Research Centre for Linguistics

International Morphology Meeting

2026, June 18–21

Budapest

Introduction

Hungarian compound verbs

- *teszt-vezet-és* ‘test drive’ > *teszt-vezet* ‘test drive’
- test-drive-NLZR test-drive

vezetés$_N$: vezet$_V$ :: teszt-vezetés$_N$: teszt-vezet$_V$

- *ülő-sztrájk* ‘sit-in’ > *ülő-sztrájk-ol* ‘sit in’
- sitting-strike sitting-strike-VBLZ

sztrájk$_N$: sztrájkol$_V$ :: ülő-sztrájk$_N$: ülő-sztrájkol$_V$

- *párt-polit-ika* ‘party politics’ > *párt-polit-izál* ‘engage in party politics’
- party-polit-NLZR party-polit-VBLZ

politika$_N$: politizál$_V$ :: párt-politika$_N$: párt-politizál$_V$

- *szoknya-vadász* ‘skirt-chaser’ *szoknya-vadász-ik* ‘skirt-chase’
- skirt-hunter skirt-hunter-3SG

vadász$_N$: vadász(ik)$_V$ :: szoknya-vadász$_N$: szoknya-vadász(ik)$_V$

The issue: uncertainty of bases

mozgás-érzékel-ő ‘motion detector’ >
motion-detect-NML

mozgás-érzékel ‘detect motion’
motion-detect

mozgás-érzékel-és ‘motion detection’ >
motion-detect-NMLZ

jel-tolmács-ol-ás ‘sign language interpreting’ >
sign-interpreter-VBLZ-NMLZ

jel-tolmács-ol ‘interpret sign language’
sign-interpreter-VBLZ

jel-tolmács ‘sign language interpreter’ >
sign-interpreter

The issue: uncertainty of bases

mozgás-érzékel-ő ‘motion detector’ > ?‘use a motion detector’
motion-detect-NML

mozgás-érzékel ‘detect motion’
motion-detect

mozgás-érzékel-és ‘motion detection’ > ?‘do motion detection’
motion-detect-NMLZ

jel-tolmács-ol-ás ‘sign language interpreting’ > ?‘do sign language interpreting’
sign-interpreter-VBLZ-NMLZ

jel-tolmács-ol ‘interpret sign language’
sign-interpreter-VBLZ

jel-tolmács ‘sign language interpreter’ > ?‘act as a sign language interpreter’
sign-interpreter

Theoretical prerequisites

- In a network-based modell, distinct base words or even back-formation and affixation are not mutually exclusive.

- The new derivate might be based on a larger analogical pattern

$(S)N_1 : (S)N_2 : (S)V :: (C)N_1 : (C)N_2 : (C)V$
$érzékelő_N : érzékelés_N : érzékel_V :: mozgás\text{-}érzékelő_N : mozgás\text{-}érzékelés_V : mozgás\text{-}érzékel_V$
$tolmács_N : tolmácsolás_N\ tolmácsol_V :: jel\text{-}tolmács_N : jel\text{-}tolmácsolás_N\ jel\text{-}tolmácsol_V$

- Frequency as an indicator: if there is no significant difference between the frequencies of possible nominal compounds, there is no need to make an either-or decision, the structure of words may be multiply motivated

Booij & Audring 2018; Booij 2023; Hathout & Namer 2019

Question and goal

- How typical/frequent back-formation and multiple motivation is for compound verbs with *-l* verbalizer?
    - *ülő-sztrájk-ol* ‘sit in’, *jel-tolmács-ol* ‘to interpret sign language’

- Investigation of frequency patterns

    - The potential base words with *-Ás* and *-Ó* nominalizers (from which the verb can be explained by back-formation)
        - *jel-tolmács-ol-ás* ‘sign language interpreting’ > *jel-tolmács-ol* ‘to interpret sign language’

    - N-N compounds without nominalizer (in which case we can assume affixation)
        - *jel-tolmács* ‘sign language interpreter’ > *jel-tolmács-ol* ‘to interpret sign language’

Data and method

Data

- Corpus: a dependency-parsed, cleaned version of the **Hungarian Gigaword Corpus** (Oravecz et al. 2014)

- Target of the study: compound verbs containing the derivational suffix ***-Ol***
  (where O = *o, e, ö, á, é*)

- The problem: these verbs are **not explicitly annotated** in the corpus
  → identifying them required an iterative, heuristic search

Identifying the compound verbs

- A set of heuristics to extract candidate *-Ol* verbs:
  ≥ 8 chars; ends in -Ol; POS = VERB; no detailed emMorph analysis (Novák et al. 2016)

- Strip the -Ol ending to recover the potential **base form**,
  e.g. *lábjegyzetel* ‘write footnotes’ → *lábjegyzet* ‘footnote’

- Correct truncated stems,
  e.g. *fájlcserél* ‘exchange files’ → *fájlcsere*, not **fájlcser*

- Analyse base forms with **emMorph**; keep only **nominal compounds**
  (✗ *felragacs* lit. ‘up-goo’, ✓ *vendégszerep* ‘guest role’)

- Result: **536 items**, small enough to be **checked manually**

Building the dataset

- For each compound verb, manually add its corresponding ***-Ás*** and ***-Ó*** derivatives, giving the full paradigm:
- *díszcsomag* ‘gift box’ (N)
  díszcsomagol ‘gift-wrap’ (V)
  díszcsomagolás ‘gift wrapping’ (-Ás)
  díszcsomagoló ‘gift-wrapper’ (-Ó)

- Retrieve the **token frequency** of these lemmas from the corpus

- **Result**: a frequency-annotated dataset of compound verbs and their nominal counterparts
  → the input for the frequency analysis that follows

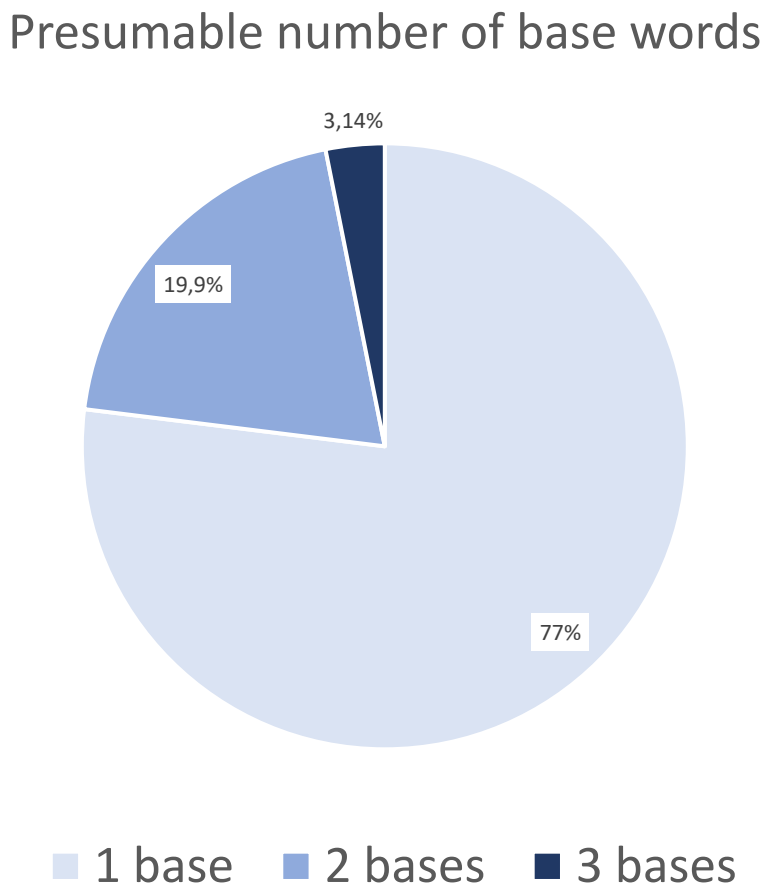
Results

Presumable number of bases

Noun 1 without an affix	Verb with the *-l* suffix	Noun 2 with the *-Ás* nominalizer	Noun 3 with the *-Ó* nominalizer
1. ND 2. (relatively) few data 3. (relatively) large amount of data = base!	191 types 580 tokens 121 hapaxes	1. ND 2. (relatively) few data 3. (relatively) large amount of data = base!	1. ND 2. (relatively) few data 3. (relatively) large amount of data = base!

- An average of 1.24 base words in the sample

Multiple motivation is neither common nor exceptional for the compounds with *-l* verbalizer

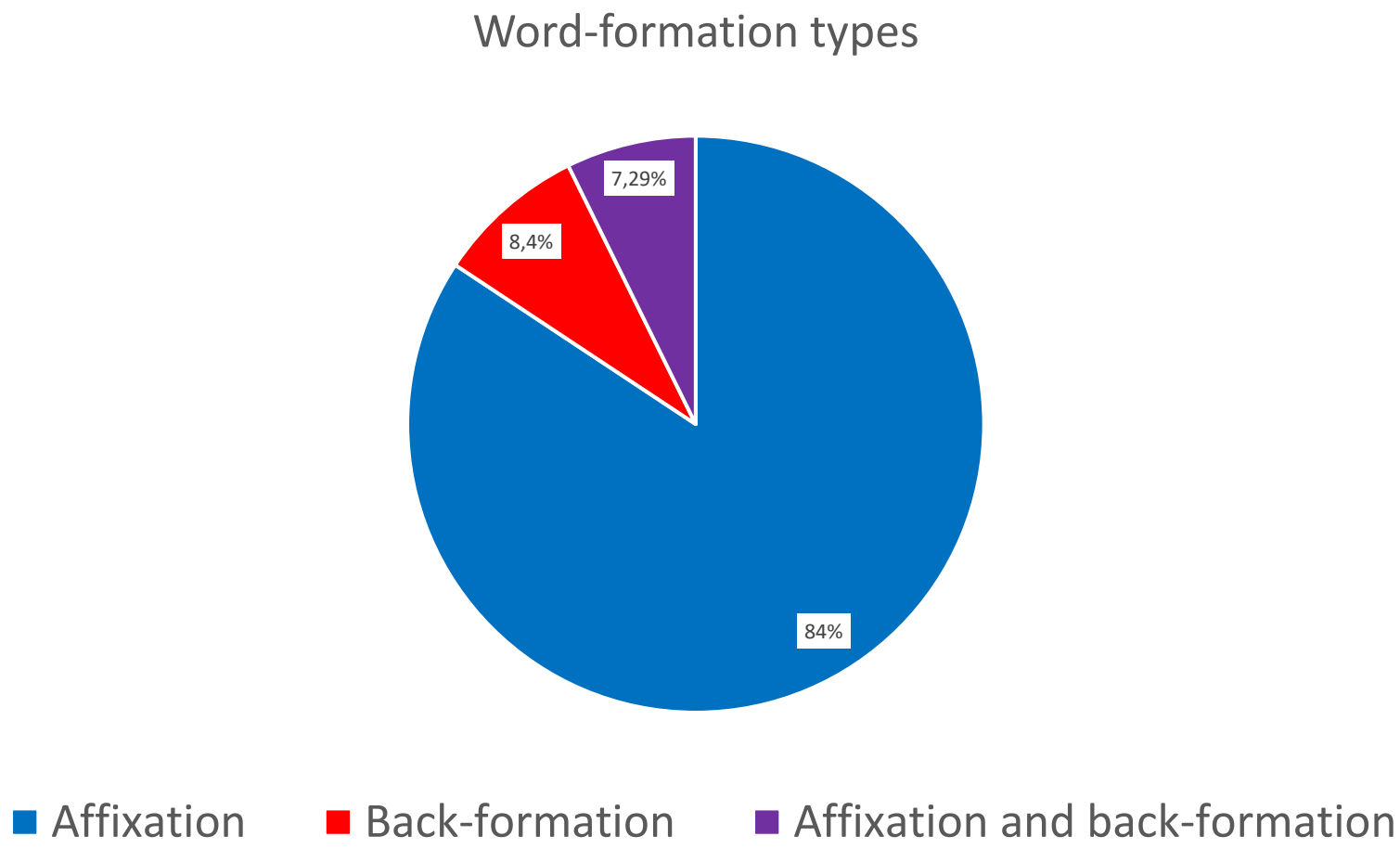


Distribution of presumable bases

Number of bases	Noun 1 without an affix	Verb with the -l suffix	Noun 2 with the -Ás nominalizer	Noun 3 with the -Ó nominalizer
1	Base	X		
1		X	Base	
1		X		Base
2	Base	X	Base	
2	Base	X		Base
2		X	Base	Base
3	Base	X	Base	Base

- All three nominal constructions occurred as single bases, and examples were found for all their combinations

Distribution of word-formation processes



Possible correlations

1. Token frequency of verbs ↔ Multiple motivation

2. The number of the attested corresponding nouns was higher for more frequent verbs
   - 121 verbal hapaxes – 145 nouns 1.198 noun/verb
   - 70 verbal non-hapax – 96 nouns 1.371 noun/verb
   - Mann-Whitney U test: U=3521, z=2.23, p=0.026, r=0.16

- More common compound verbs serve as bases themselves, thus increasing the number of their corresponding nouns.

jel-tolmács$_{331\ \text{tokens}}$ ‘sign language interpreter’ > *jel-tolmács-ol* ‘interpret sign language’
sign-interpreter sign-interpreter-VBLZ

jel-tolmács-ol ‘interpret sign language’ > *jel-tolmács-ol-ás*$_{18\ \text{tokens}}$ ‘sign language interpreting’
sign-interpreter-VBLZ sign-interpreter-VBLZ-NMLZ

> *jel-tolmács-ol-ó*$_{2\ \text{tokens}}$ ‘sign language interpreter’
sign-interpreter-VBLZ-NMLZ

Summary

- How typical/frequent back-formation and multiple motivation is for compound verbs with *-l* verbalizer?

- Multiple motivation is neither common nor exceptional for the Hungarian compounds with *-l* verbalizer
  - Some compound verbs had three presumable bases (equally conventional corresponding nouns)
  - In some cases, there was no reason to exclude either back-formation or affixation

- Uniformly viewed concatenative and non-concatenative morphological operations can be effectively examined within the same network of morphological relations.

References

- Booij, Geert (2023). Paradigmatic relations in Construction Morphology: The case of Dutch Noun+Verb compounds. *Journal of Word Formation* 7(2). 13-33.
- Booij, Geert – Jenny Audring (2017). Construction Morphology and the Parallel Architecture of Grammar. *Cognitive Science* 41 (S2) 277–302.
- Hathout, Nabil – Namer, Fiammetta (2019). Paradigms in word formation: what are we up to? Special issue, *Morphology* 29(2). 153-165
- Palágyi László (2024). Back-formation, Cross-formation, and Forward-formation in the same construction. Lingue e Linguaggio.
- Novák, Attila, Borbála Siklósi and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In N. Calzolari et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA). 1315–1322.
- Oravecz, Csaba, Tamás Váradi and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In N. Calzolari et al. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik: European Language Resources Association (ELRA). 1719–1723.

ELTE | FACULTY OF HUMANITIES

Thank you for your attention!