Managing Growth in a National Corpus: The Hungarian National Corpus 3.0 (MNSZ3)

Noémi Ligeti-Nagy[1], **Enikő Héja**[1], **Ágnes Bánfi**[1], **Flóra Földesi**[1], **Bence Sárossy**[1], **Boglárka Skrabák**[2], **Tamás Váradi**[1], **Gábor Prószéky**[1]
[1] ELTE Research Centre for Linguistics, Budapest, Hungary
[2] ELTE Faculty of Informatics, Budapest, Hungary
ligeti-nagy.noemi@nytud.hu, heja.eniko@nytud.hu, banfi.agnes@nytud.hu, foldesi.flora@nytud.hu, sarossy.bence@nytud.hu, skrabakbogi@gmail.com, varadi.tamas@nytud.hu, proszeky.gabor@nytud.hu

Abstract

The third generation of the Hungarian National Corpus (MNSZ3) aims to provide a large-scale, curated, and well-described corpus resource needed for the sustainable digital presence of Hungarian. Building on the domain structure and proportions of MNSZ2 (v2.0.5; 1.04 billion running words), the project targets a substantial increase in scale while also strengthening the coverage and metadata description of Hungarian language use outside Hungary. MNSZ3 retains the six traditional domains of the earlier corpus—press, fiction, scientific, official, personal, and transcribed spoken language—and is planned to reach approximately 10 billion tokens. This paper presents the motivation and design principles of the project, outlines the practical decisions and procedures used in data collection and cleaning, and discusses the annotation strategy developed for large-scale processing. In planning the linguistic analysis, we build on the complementary strengths of HuSpaCy and e-magyar: HuSpaCy provides the unified and efficient UD-oriented processing backbone, while e-magyar (emMorph) is preserved as an explicit additional layer for morphology and lemmatisation.

Keywords: Hungarian, national corpus, parsing

1. Introduction

[1]

Large, searchable, and metadata-rich corpora are basic infrastructure for modern linguistic research. They provide an empirical basis for theoretical investigations, descriptive work such as dictionary and grammar writing, and the development and evaluation of language technology applications. Over the past decade, expectations concerning corpora have partly shifted: alongside carefully sampled reference corpora, continuously expanding, multi-billion-token monitor corpora have become increasingly important. At the same time, scaling up also makes methodological risks more visible, including quality assurance, duplication, access and copyright constraints, and different kinds of bias.

For Hungarian, a sustainable digital presence requires large-scale, curated corpora that are well described with metadata and available to both researchers and developers. A key resource in this area has been the Hungarian Gigaword Corpus (MNSZ2, Oravecz et al., 2014), which was preceded by the first Hungarian National Corpus (MNSZ1, Váradi, 2002). MNSZ2 contains approximately 1.04 billion running words distributed across six domains and includes not only language use from Hungary but also Hungarian texts from neighbouring countries.

The aim of MNSZ3 is to expand this resource by an order of magnitude while preserving the domain structure and balance of MNSZ2. The project is driven by three closely related goals: increasing scale in a controlled way, improving regional coverage – especially for Hungarian used outside Hungary – and establishing a reproducible, maintainable processing workflow for collection, cleaning, annotation, and access.

1.1. International context

Internationally, the label “national corpus” covers several partly different practices: classical balanced reference corpora, continuously updated monitor corpora, and corpora built from multiple subcorpora that cover different time periods and registers. One well-known example of the reference approach is the British National Corpus (BNC), a roughly 100-million-word collection of spoken and written English (BNC Consortium, 2007). Its modern counterpart, BNC2014, was built on a similar scale with a focus on British English of the 2010s (Brezina et al., 2021). At the other end of the spectrum, the German DeReKo contains more than 42 billion units and continues to grow (Kupietz et al., 2018). Czech corpus practice combines a large versioned written corpus (SYN v13) with smaller balanced reference corpora such as SYN2020 (Hnátková et al., 2014; Jelínek et al., 2021). Comparable large national resources also exist for Polish, Spanish, and Russian (Narodowy

[1]This paper is a slightly extended version of Ligeti-Nagy et al. (2025).

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 84–90. Palma, Mallorca. May 11, 2026. ©ELRA Language Resources Association, 2026

Korpus Języka Polskiego (NKJP) / Porowski, S., Bańko, M., et al., 2024; Real Academia Española, 2025; Savchuk et al., 2024).

These examples show that there is no single model for a national corpus. MNSZ3 is positioned between the tradition of balanced reference corpora and the present-day need for substantially larger resources.

1.2. Hungarian background

The direct predecessors of MNSZ3 are the first Hungarian National Corpus (MNSZ1) and the Hungarian Gigaword Corpus (MNSZ2). The common principle behind both was a balanced, domain-based corpus model that aims not at strict statistical representativeness but at a professionally motivated and interpretable balance of text types and sources.

MNSZ1 was designed as a balanced reference corpus of contemporary written Hungarian. Because of practical constraints, spoken language could not yet be included in a comprehensive way, and the available sources were mainly electronically accessible written texts. Even at this stage, however, the corpus design already took the geographical dimension into account and included samples from Hungarian communities in neighbouring countries.

MNSZ2 was developed as a larger and reannotated successor to MNSZ1. Its development was motivated by three main considerations: scale, to support data-driven methods and the study of rare phenomena; quality, through better analysers and finer annotation; and coverage, through resampling and the inclusion of previously underrepresented registers.

1.3. Objectives

The aim of this paper is twofold. First, it presents the motivation and design principles of the MNSZ3 corpus-building programme in the context of earlier Hungarian corpus work. Second, it documents the practical decisions and procedures needed to build a multi-billion-token corpus that remains searchable, reusable, and sustainable.

The project is organised around three goals:

1. **Scaling up with a balanced structure.** The size of MNSZ2 is increased by an order of magnitude while the domain structure and proportions remain controlled.

2. **Strengthening regional coverage.** Hungarian texts from outside Hungary are included in much greater quantities and with richer metadata, especially in those region–domain combinations that were missing or marginal in MNSZ2.

3. **Quality and sustainability.** The project gives greater weight to curated sources, explicitly handles legal and access constraints, and establishes reproducible workflows for collection, cleaning, and versioning.

2. Corpus design principles and structure

This section summarises the main design principles of MNSZ3: the domain system and its definitions, regional and temporal coverage, the metadata model, and the basic requirements of the planned query and access framework.

2.1. Design principles

For national corpora, strict statistical representativeness is not achievable in practice. Hungarian corpus building has therefore traditionally relied on the notion of *balance*. In this approach, the corpus does not claim to mirror reality directly; rather, it models text types and source groups in a proportionate and professionally motivated way, with an explicit domain and partly regional structure (Váradi, 2002; Oravecz et al., 2014). MNSZ3 follows this principle: it preserves the inherited domain structure and proportions of MNSZ2 and scales up within that framework.

A second guiding principle is scalability and reproducibility. The corpus relies on processing schemes that can be applied across large volumes of data and rerun in a controlled manner when the corpus is updated. This requires versioned outputs, stable identifiers, and documented filtering and cleaning steps.

2.2. Domains

The domain system of MNSZ3 follows that of MNSZ2 and consists of six major domains: press, fiction, scientific and popular scientific texts, personal texts, official texts, and transcribed spoken language. The purpose of this system is twofold: it provides broad register coverage and at the same time preserves a modular structure that can be queried at subcorpus level, separately or in combination.

In operationalising the domains, special care is needed for boundary cases, such as the relation between press and opinion writing, or the different genres grouped under official texts. To preserve queryability, domain labels need to be complemented by richer document-type and source metadata.

Table 2 shows the regional distribution of MNSZ2 together with the planned targets for MNSZ3. The

Table 1: Main domains of MNSZ3.

Domain	Definition / typical document types
Press	Edited news and background material: articles, interviews, reports, opinion pieces, lifestyle texts; publisher and source metadata, topical labels.
Fiction	Primarily fictional prose: novels, novellas, short stories; temporal control; translations marked separately.
Scientific / popular scientific	Scientific and popular scientific texts; genre and source metadata; thematic labelling.
Personal	User-generated and personal written communication, such as blogs, forums, and social platforms.
Official	Laws, public and institutional documents, minutes, and related materials; finer sublabelling by document type.
Transcribed spoken language	Transcribed spoken texts, such as conversations, interviews, and broadcasts, with metadata on source and date.

Table 2: Regional distribution of MNSZ2 by domain (million running words) and planned targets for MNSZ3 (million tokens).

Domain	MNSZ2 region (M words) HU	SK	UA	RO	RS	MNSZ2 total	MNSZ3 target	Main focus of expansion
Press	350.5	11.6	0.7	0.6	1.5	364.8	3918	proportional growth; targeted expansion outside Hungary
Fiction	77.0	2.3	0.4	0.8	0.2	80.6	161	moderate growth; better metadata for translations and originals
Scientific	112.0	3.3	0.7	1.6	0.3	117.9	1300	proportional growth; targeted inclusion of institutional sources
Official	98.0	0.2	0.3	0.6	0.03	99.0	1171	targeted expansion of official texts from neighbouring countries; finer document-type labels
Personal	300.3	–	0.4	0.4	0.03	301.1	3011	filling missing Slovakian material; expansion in all regions
Spoken	76.2	–	–	–	–	76.2	836	inclusion of spoken-language material outside Hungary; unified transcription and metadata framework
Total	1013.9	17.3	2.5	3.9	2.0	1039.7	10397	approximately tenfold overall expansion

Note: HU = Hungary, SK = Slovakia, UA = Transcarpathia, RO = Transylvania, RS = Vojvodina. MNSZ2 figures are given in million running words; MNSZ3 targets are in million tokens.

table makes it clear which region–domain combinations were missing or only marginally represented in MNSZ2; their targeted inclusion is one of the priorities of the expansion.

3. Data collection and sources

The expansion of MNSZ3 is not conceived as the creation of an unrestricted web corpus. Instead, it is organised as domain-based data collection that preserves the domain structure and proportions established in MNSZ2 and increases scale within that framework.

Throughout the project, we have aimed to ensure that both source selection and processing follow explicit procedures rather than ad hoc decisions. We fixed temporal coverage domain by domain: for press texts, publication date is the primary reference point; for official texts, the relevant date depends on the document type; for fiction, the date of first publication is used, and for translations, the year of translation. In several domains, 1990 serves as a practical lower boundary, although in fiction we go back to 1945 when necessary to secure enough material under genre constraints.

At document level, we record at least the source, date, and domain, and, where possible, also additional fields such as author, section or topic, and document type. We also anticipated large-scale repetition from the start: identical or near-identical content appears frequently in web and institutional collections, so document-level deduplication is typically combined with boilerplate removal and further domain-specific filtering.

3.1. Press

Press texts were collected through several complementary channels. In web-archive-based harvesting, especially from Common Crawl, we used pre-filtering and content extraction to target article-like pages. We then deduplicated at document level and handled typical cases of near-duplication, such as the same article under different URLs or pages with repeated recommendation blocks. In targeted portal-level harvesting, we aimed at broad coverage of each news site's article inventory and adapted metadata extraction – title, section or topic, author, publication date – to the structure of each

source.

Quality filtering in the press domain consistently excluded index and listing pages, excessively short or incomplete texts, non-Hungarian content, and pages that did not match the genre profile of press articles. The main source groups are the following:

- web-archive material, especially Common Crawl, based on curated press-domain lists;
- targeted harvesting from major Hungarian news sites such as `24.hu`, `hvg.hu`, `nemzetisport.hu`, `blikk.hu`, `vg.hu`, `nlc.hu`, `telex.hu`, `444.hu`, `femina.hu`, and `vezess.hu`;
- inherited or earlier collections already used in MNSZ2;
- press sources from neighbouring countries, for example `hirek.sk`, `gutaonline.sk`, `bulvar.parameter.sk`, `amikassa.sk`, `maszol.ro`, and `hirmondo.ro`.

The current press material collected from Hungarian sources amounts to roughly 2.6 billion tokens from around 100 news domains after deduplication, language filtering, and boilerplate removal. Where topic metadata is incomplete or inconsistent, automatic topic detection is used to assign corpus-wide thematic labels in a consistent way (Osváth and Héja, 2025).

3.2. Official texts

The official domain is built primarily from large structured collections of legal and administrative documents. Extraction and cleaning had to be adapted to several different web and markup environments, including HTML-based sources, structured exports, and TEI-like parallel files. Particular attention was paid to isolating the linguistically relevant text core and excluding headers, footers, repeated navigation elements, attachments, tables, and metadata blocks.

According to the current project records, the official domain reaches close to 0.9 billion words. The main source groups include the following:

- the National Law Repository (approximately 83 million words);
- the Repository of Municipal Decrees (approximately 290 million words);
- anonymised court decisions (approximately 431 million words);
- the Hungarian part of JRC-Acquis (approximately 44 million words);
- the Hungarian side of Europarl (approximately 12 million words);
- the Hungarian subset of DCEP (approximately 35 million words);
- supplementary parliamentary material from Hungary, such as documents and minutes, with internal sublabelling inside the official domain (approximately 79 million words);
- institutional and municipal sources from neighbouring countries.

3.3. Fiction

The fiction domain was planned as a controlled expansion from the start. Here, accessibility, copyright, and quality assurance—especially OCR errors and heterogeneous metadata—set the main limits. For this reason, the collection prioritises prose works with clear dates and verifiable origin, and ambiguous cases such as borderline genres or uncertain translation status are handled in a separate control procedure.

The main sources are the Hungarian Electronic Library (MEK), the Digital Literary Academy (DIA), and the inherited fiction component of MNSZ2. At the current stage, the fiction subcorpus contains 884 documents with a total size of 50,345,981 tokens. Within this material, metadata is maintained both for origin (including texts from outside Hungary and translated works) and for source-side genre labels.

The domain is also one of the most difficult to scale. In practice, the main problems are OCR errors, untidy metadata, encoding issues, and the textual heterogeneity of literary works. Because of these constraints, fiction cannot realistically be expanded at the same rate as the other major domains.

3.4. Scientific texts

In the scientific domain, the most important source group is the REAL repository. These materials were not obtained through direct web harvesting, but were integrated from a parallel project and adapted to the requirements of MNSZ3. The main task here is quality assurance, because a large share of the material is OCR-based and therefore contains character substitutions, line-break errors, hyphenation artefacts, and intrusive headers, footers, and page numbers.

Cleaning is carried out in several steps: character encoding and basic typography are normalised, repeated non-content elements are removed, rule-based corrections are applied to common OCR patterns, and the corrected texts are added to the corpus after quality control. At present, the scientific domain of MNSZ3 contains approximately 1.2 billion words.

3.5. Personal texts

Since MNSZ2, access conditions to social-media platforms have changed substantially. This affects one of the key source types for the personal domain: large platforms have become much harder to harvest because of stricter usage terms and data-handling practices. For this reason, the personal domain relies mainly on publicly accessible and stably citable textual sources where document-level processing and minimal metadata recording are feasible.

In practice, blogs are the main source group. A central source is `blog.hu`, harvested in a targeted way and then cleaned by removing navigation, comment, and recommendation blocks, followed by language filtering and deduplication. Where possible, the corpus distinguishes between post-level and comment-level material. In addition to blogs, we also include Reddit, which is still technically accessible for this purpose. The current Reddit-based personal material is on the order of 10 million words.

3.6. Transcribed spoken language

The spoken-language domain is currently being expanded mainly through podcast-like recordings with longer stretches of continuous speech. The main source at present is the podcast collection of the National Széchényi Library.

For transcription, we use an uploader and controller tool attached to the BEAST2 system (Kádár et al., 2023). The program uploads the audio and stores the machine transcripts in a documented directory structure, from which they are reintegrated into the corpus-building pipeline. Automatic post-correction is then carried out with a dedicated correction component designed to address typical transcription problems such as mishearing, punctuation, sentence boundaries, and unstable spelling of proper names.

3.7. Hungarian texts from neighbouring countries

Regional extension started from web-archive-based domain lists that contained Hungarian-language content. From these lists, we selected domains associated with neighbouring countries and assigned them manually to the MNSZ3 domains. Because the sources are highly heterogeneous, project-specific harvesting and extraction scripts were developed. Text and metadata were stored separately, then filtered for language, deduplicated, and cleaned of boilerplate content.

In the press domain, the material currently collected from neighbouring countries contains 232,323,495 tokens and 1,206,612 documents. Most of this material comes from Romania and Slovakia, with smaller but still relevant collections from Austria, Croatia, Serbia, Slovenia, and Ukraine.

In the official domain, the corresponding material currently contains 36,094,964 tokens and 114,971 documents, again with the largest share coming from Romania and additional contributions from Slovakia, Ukraine, and Serbia.

In the personal domain, collection is still in progress. The available material comes mainly from open web sources such as blogs, personal homepages, service-oriented self-presentations, political personal pages, and some forum-like sites. Because the source base remains heterogeneous and uneven across countries, no final aggregate figures are given for this component at the current stage.

4. Linguistic annotation

One of the main commitments of MNSZ3 is that it will not only be large and balanced, but also linguistically analysed. In addition to tokenisation and lemmatisation, the corpus is planned to provide morphosyntactic information, dependency parsing, named entity recognition, and keyword extraction. The goal of annotation is twofold: to improve corpus usability for search and linguistic investigation, and to establish a technological basis for further components such as terminology extraction and domain-specific normalisation.

For large-scale processing, we compared two widely used Hungarian language-processing pipelines: e-magyar (Váradi et al., 2017; Indig et al., 2019) and HuSpaCy (Orosz et al., 2022, 2023). The two systems overlap in several functions, but differ in architecture and strengths. E-magyar is a modular, research-oriented framework whose most distinctive component is the rich emMorph-based morphological analysis (Novák et al., 2016). HuSpaCy, by contrast, is built on spaCy (Honnibal et al., 2020), provides a unified UD-oriented pipeline, and is well suited to efficient large-scale processing.

Because annotation in MNSZ3 means running the analysis on billions of words, individual component quality is not the only consideration. Output consistency, reproducibility, error handling, and stability across domains are equally important. To support the comparison, we built a unified evaluation framework (*Launcher*) that runs both analysers on the same input, applies minimal normalisation where necessary, compensates for tokenisation shifts heuristically, and returns differences in a form that supports both aggregation and manual inspection. This part of the work is based on Skrabák and Ligeti-Nagy (2025).

4.1. Test material and main findings

The comparison was carried out on a parliamentary-record sample. This material is both linguistically varied and formally heterogeneous, which makes it a suitable stress test for tokenisation, morphology, and syntax. In the sample, HuSpaCy identified 46,607 tokens and e-magyar 46,452 tokens, with the difference largely attributable to recurring tokenisation patterns.

The results can be summarised as follows.

- **Tokenisation.** HuSpaCy performed better overall. Weighted by frequency of occurrence, it gave the better solution in 89.9% of all differing cases and in 93.9% of the unambiguous ones.
- **Morphology and lemmatisation.** E-magyar provided the more reliable output. In the HuSpaCy–emMorph integration, 4,987 out of 46,635 tokens returned `None` at the morphological level. Even after corrections to the integration, substantial differences remained between the outputs of the two systems.
- **Part-of-speech tagging.** Many differences were driven by tokenisation and punctuation handling rather than purely by tag assignment.
- **Dependency parsing.** HuSpaCy produced the more consistent and reusable output. In e-magyar, some sentences lacked a ROOT relation, which is a serious problem for downstream processing.
- **Named entity recognition.** HuSpaCy tended to identify more named-entity tokens, while e-magyar was more conservative but more precise in type assignment on the manually checked sample.

4.2. Annotation strategy

The comparison suggests that the two pipelines should not be seen as a simple "better versus worse" contrast. Rather, they have complementary strengths. HuSpaCy is stronger in tokenisation, dependency parsing, and named entity recognition; e-magyar provides better morphology and lemmatisation.

The annotation strategy of MNSZ3 therefore combines them. HuSpaCy provides the unified and efficient backbone for large-scale UD-oriented processing, including tokenisation, POS/UD morphology, dependency parsing, and named entity recognition. At the same time, the emMorph-based output of e-magyar is preserved as an explicit additional layer for morphology and lemmatisation. In parallel, the HuSpaCy–emMorph integration remains an independent line of development, with the aim of bringing the morphological and lemmatisation behaviour of the HuSpaCy-based workflow closer to that of e-magyar.

5. Conclusion

MNSZ3 is intended to provide a qualitative step forward for the digital presence and research infrastructure of Hungarian. A multi-billion-token corpus that is curated, consistently described with metadata, and linguistically analysed will support both traditional corpus-based linguistic work and present-day language technology. It will allow more reliable investigation of rare phenomena, finer tracking of genre and diachronic differences, and broader inclusion of registers that were previously only marginally available.

At the same time, the project has direct practical value. A more balanced domain structure, stronger representation of Hungarian used outside Hungary, and the extension of spoken-language material all improve the range of linguistic variation available for study. Detailed metadata and a reproducible processing workflow improve transparency and reusability, and lower the practical threshold for corpus use among researchers, developers, and educators.

6. Bibliographical References

BNC Consortium. 2007. British National Corpus, XML Edition. Available at http://www.natcorp.ox.ac.uk/,.

V. Brezina, A. Hawtin, and T. McEnery. 2021. The written British National Corpus 2014 – design and comparability. *Text & Talk*, 41(5-6):595–615.

Milena Hnátková, Michal Křen, Pavel Procházka, and Hana Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Péter Kundráth, and Noémi Vadász. 2019. emtsv – Egy formátum mind felett. In *Magyar Számítógépes Nyelvészeti Konferencia*, Szeged. In Hungarian.

Tomáš Jelínek, Jan Křivan, Vladimír Petkevič, Hana Skoumalová, and Jitka Šindlerová. 2021. SYN2020: A new corpus of Czech with an innovated annotation. In *Text, Speech, and Dialogue*, Lecture Notes in Computer Science, pages 48–59. Springer.

Máté Soma Kádár, Gergely Dobsinszki, Katalin Mády, and Péter Mihajlik. 2023. "Feeding the BEAST" – A BEA Speech Transcriber továbbfejlesztése és integrálása neurális nyelvmodellel. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, volume 19, pages 135–143.

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German reference corpus DeReKo: New developments – new opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4353–4360, Miyazaki, Japan. European Language Resources Association (ELRA).

Noémi Ligeti-Nagy, Enikő Héja, Ágnes Bánfi, Flóra Földesi, Mariann Lengyel, Bence Sárossy, Boglárka Skrabák, Tamás Váradi, and Gábor Prószéky. 2025. Expanding the Hungarian Gigaword Corpus. In *CLARIN Annual Conference Proceedings 2025*, pages 188–192, Vienna. CLARIN ERIC.

Narodowy Korpus Języka Polskiego (NKJP) / Porowski, S., Bańko, M., et al. 2024. Narodowy korpus języka polskiego. Dostępne online na https://nkjp.pl/.

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)*.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723. 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.

György Orosz, Zsolt Szántó, Péter Berkecz, Gergő Szabó, and Richárd Farkas. 2022. HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 59–73.

Mátyás Osváth and Enikő Héja. 2025. Internetes hírek automatikus osztályozása. In *Magyar Számítógépes Nyelvészeti Konferencia (21.)*, volume 21, pages 29–39, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet. Konferenciaközlemény. Elérhető: http://acta.bibl.u-szeged.hu/id/eprint/88770.

Real Academia Española. 2025. CORPES XXI: Corpus del Español del Siglo XXI. Accessed: 22/01/2026.

S. O. Savchuk, T. Arkhangelskiy, A. A. Bonch-Osmolovskaya, O. V. Donina, Yu. N. Kuznetsova, O. N. Lyashevskaya, B. V. Orekhov, and M. V. Podryadchikova. 2024. Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2):7–34.

Boglárka Skrabák and Noémi Ligeti-Nagy. 2025. A huspacy és e-magyar elemzőláncok teljesítményének átfogó összehasonlítása országgyűlési szövegeken: a tokenizálástól a függőségi elemzésig. In *Magyar Számítógépes Nyelvészeti Konferencia*, volume 21, pages 171–184, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet.

Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 385–389.

Tamás Váradi, Eszter Simon, Bálint Sass, Mátyás Gerőcs, Iván Mittelholtz, Attila Novák, Balázs Indig, Gábor Prószéky, and Veronika Vincze. 2017. Az e-magyar digitális nyelvfeldolgozó rendszer. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, pages 49–60, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.