Hellenic National Corpus: the current state

Maria Gavriilidou[1], **Nikos Sidiropoulos**[1]

Institute for Language and Speech Processing, Athena Research Center
Athens, Greece
(maria, nsidir)@athenarc.gr

Abstract

The Hellenic National Corpus (HNC) is an integrated online environment offering access to standard Modern Greek language material and to related analysis tools. The HNC corpus has been developed in two main phases, and currently comprises over 97 million words exclusively of written language, sourced from printed resources or scraped from the internet. The material has been automatically lemmatized and morphologically annotated, while a subset of 100,000 words has been further manually corrected, in order to produce a freely downloadable error-free corpus. Through the dedicated platform, the users have access to concordances, morphological analysis of words and statistical information (frequency) at word, lemma, part of speech and n-gram levels. Future steps include the expansion of the material in both historical and coverage dimensions: the inclusion of material from older phases of the language is foreseen, as well as the addition of dialectal material besides standards language.

Keywords: Greek corpus, access environment, analysis tools

1. Introduction

The Hellenic National Corpus (HNC)[1] is an integrated online environment offering access to Modern Greek language material and to related analysis tools. The HNC corpus currently comprises over 97 million words exclusively of written language, sourced from printed resources or scraped from the Internet. The corpus material has been automatically lemmatized and morphologically annotated. Through the dedicated platform, the users have access to KWIC-type or full sentence concordances, statistical information (frequency) and collocations at word, lemma and part of speech levels. Recent improvements of the HNC concern both the language material and the platform: quantitative and qualitative enrichment of the language material; addition of new text types and genres from the internet (digital press, social media, blogs, etc.); development of a new backend (document uploading platform and metadata editor); improvement of existing and addition of new search functionalities; redesign of the HNC user interface; improved visualizations of results; and, finally, creation of the Golden Part of Speech Tagged corpus, an automatically annotated and manually corrected small corpus subset, freely available through the CLARIN:EL National Infrastructure for Language Resources and Technologies[2] (Gavriilidou et al. 2024).

The current paper is structured as follows: Section 2 presents an overview of the historical evolution of the HNC; Section 3 describes the criteria and methodology for corpus collection, classification and processing across phases; Section 4 elaborates on the HNC Platform and its current functionalities; Section 5 focuses on the Golden Part of Speech Tagged corpus; and finally Section 6 concludes with future steps.

2. Historical Evolution of HNC

HNC was developed in two broad phases: the first phase (starting in 1992) set the objectives of the endeavor, specified the collection criteria and the methodology, and proceeded to collect the material and develop the platform for user access (Hatzigeorgiu et al., 2000, Gavrilidou, 2002). This phase resulted in the first version of HNC, which contained 42 million words. The largest portion of the material originated from Newspapers (Table 1), which were exclusively printed at that time, while the internet was critically under-represented. Publishing houses contributed literary and scientific works, which represented almost 10% of the material. These proportions, which were due to the availability (or scarcity) of the respective sources, rendered the corpus unbalanced.

Publication medium	
Book	9,4%
Internet	0,3%
Newspaper	61,3%
Periodicals	5,9%
Other	23,1%
Total	100,0%

Table 1: First phase proportions by Publication medium

Every single text that formed part of the HNC was accompanied by the appropriate license agreement. The license agreements which were signed with the publishers provided the material with strict restrictions, namely: HNC was allowed to include excepts but not the entire text provided

[1] https://hnc.ilsp.gr/

[2] https://www.clarin.gr/en

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 57–62. Palma, Mallorca. May 11, 2026. ©ELRA Language Resources Association, 2026

by the sources, to offer to the users very restricted amount of text results, i.e., the sentence containing the user query term, plus the previous and the next sentence; provision of the whole paragraph or even more of the whole text was forbidden due to Intellectual Property Rights restrictions.

The HNC Platform hosting the corpus and providing access to the language material was also designed and developed during the first phase. Its functionalities included filtering the material to select specific texts according to criteria (e.g., only newspapers, texts of a specific author, publisher or date, etc.), in order to construct sub-corpora on which the search was performed. Content search focused on word, lemma and/or part of speech, and up to 3 combinations thereof; additionally, HNC offered word and lemma frequencies. In order to ensure lawful use of the material granted by the publishing houses, the platform was designed to allow only online access but no downloading of the material.

After a long period of maintenance of the HNC and user support but no addition of new material, the second phase (2020-2021) undertook the quantitative and qualitative enrichment of the corpus and the improvement of platform functionalities. The necessity for the enrichment of the corpus material was due firstly, to its small size for a national corpus, secondly, to the need to keep up with the technical evolution of the field, and thirdly, to cater for the new language production modes and language use established through the internet: in order to reflect digital-era language use, new genres and text types needed to be included, mainly born digital material (instead of digitized), social media material, etc. The text collection criteria and methodology were revised. A target size of 100 million tokens was set and new material was collected, processed, annotated and added to the HNC. The platform was redesigned as regards both backend and frontend, and it was enriched with additional search and analysis functionalities. These steps are detailed in the following sections.

3. Corpus Collection and Processing

3.1 Selection Criteria

HNC aimed to be a representative corpus of the current Greek language; therefore, the terminus post quem adopted was 1976, the year of the establishment of the standard Modern Greek language as the official language of education and public administration, which put an end to decades of diglossia. In order to focus on the current use of the language, most of the texts included have been produced from 1990 onwards, while a special exemption was made in the case of literature, where older significant and influential literary texts have also been included.

The issue of balance and representativeness has long concerned corpus linguists, in combination with the conflict between adherence to strict predefined design principles and the actual availability of sources. In the case of HNC, a practical approach was adopted rather than strict ratios between genders: the objective was to cover as many aspects of current language use as possible, through the inclusion of a large variety of genres, text types and topics. Given the focus on the standard language, dialectal material has not been included (geographical dialects as well as sociolects), as diverging from the standard. Readability was also used as text selection criterion: texts with high readability (high circulation newspapers, best-selling books) were preferred due to their influence in the evolution of the language.

In the initial phase, the collection strategy mainly consisted of requests to publishing houses and news agencies, frequently striving to overcome their reservations and skepticism, and convince them to sign provision agreements governing the lawful inclusion of their material in the HNC and the subsequent access provision for research purposes.

During the second phase, as mentioned in Section 2, expansion focused on including online-native content. For this, topic-focused web crawling techniques were used. Seed websites were selected after assessing their contribution towards both quantity and balance of content, based on the original selection criteria (standard, current, non-dialectal language, with a variety of genres and topics). The seed websites were fed to the ILSP Focused Crawler (Papavassiliou et al. 2013), a tool developed to automatically locate monolingual and bilingual texts of specific topics on the web. Initial identification of candidate texts for inclusion was carried out by this tool. The automatically gathered material was further screened for license: texts needed to be openly available for research purposes, without usage restrictions. The appropriate license was at least CC BY-NC-SA 4.0 (Creative Commons Attribution –NonCommercial–ShareAlike 4.0), which allows sharing for research purposes and adaptation with proper attribution. The detection of such licenses was first done automatically by the tool (via website disclaimers) and then manually reviewed. Entries linking to third-party content without appropriate licenses were excluded.

Duplicate entries (about 16%, mainly due to news agencies reposting each other) were automatically identified and only one version was included in the HNC.

Given that the second phase ended in 2021, the issue of synthetic data generated by LLMs (currently a hot issue for corpus creation), at the time had not yet appeared.

3.2 Text Classification

Every text included in the HNC corpus is accompanied by metadata providing bibliographic information (title, author, publisher, publication date), Publication Medium, Genre, and Topic. The typologies adopted in the first phase for Medium, Genre and Topic adhered to the specifications of EU project PAROLE (PAROLE, 1995), based on which many national corpora were documented in the years 1990-2000. One of these corpora was HNC, which, according to these specifications was an adequately sized corpus, especially for an under-resourced language as Greek at that time.

HNC was maintained throughout the following years, although no enrichments were possible. The second phase of the enrichment of HNC (2020-2021) aimed to benefit from technological advancements (greater computational capacity and storage), and to respond to wider social evolutions (increasing production of digital content, augmented use of social media, etc.). Consequently, it was considered necessary to focus on digital content sourced from the internet, which was not satisfactorily represented in HNC until then. Criteria for topic selection also needed updating, to reflect the digital linguistic production.

During the second phase of quantitative and qualitative enrichment, more than 65,000 new texts with almost 50 million words were added, sourced exclusively from the internet. This was dictated by the enrichment principles but was also enabled by the ease of access of digital textual material and clear licensing schemes. After the enrichment phase, digitally born material constitutes more than half of the material of HNC (55.74%), while digitized, originating as printed material totals 44.26%.

The new texts were selected to reflect a great variety of Genres (Table 2) and Topics (Table 3).

Genre proportions	
Opinion	26,2%
Information	62,8%
Official	1,3%
Scientific/Educational	2,1%
Private	0,2%
Literature	1,4%
Instruction	0,9%
Proceedings	0,1%
Discussion	5,0%
Miscellaneous	0,1%
TOTAL	100,0%

Table 2: Current Proportions by Genre

Topic proportions	
Society	49,6%
Economy	12,4%
Leisure	9,3%
Arts	8,7%
International issues	5,7%
Politics	5,6%
Health	4,3%
Sciences	3,7%
Culture	0,6%
Miscellaneous	0,3%
TOTAL	100,0%

Table 3: Current proportions by Topic

Genre and Topic classification in the second phase was performed semi-automatically: the initial classification was performed by the ILSP Focused Crawler based on relevant information identified in the text, followed by manual correction.

3.3 Text Processing

Before being added to the HNC, all texts went through three main stages: normalization, annotation (structural and linguistic), and metadata addition. Normalization removed non-textual elements and converted files into standard XML format. Structural annotation (segmentation and tokenization) identified structural elements such as paragraphs, sentences, and tokens (words, abbreviations, numbers, dates). Linguistic annotation included lemmatization and morphological annotation (part of speech tagging and morphological analysis). These processing stages were conducted using the ILSP Feature-based multi-tiered POS Tagger[3] (available through the CLARIN:EL infrastructure for Language Resources and Technologies); the tool's accuracy is 96.28% (Papageorgiou et al., 2000). Manual correction of the morphological annotation results has been performed exclusively for the Golden Part of Speech Tagged Corpus (see Section 5); manual correction of approximately 100 million words was considered not feasible. It must be noted that the process of normalization and annotation was applied also on the already existing material of the initial phase, to secure conformity.

4. The HNC Platform

The corpus is accessible to the users through a dedicated platform, allowing users to search for texts via filters based on the metadata described.

[3] http://hdl.handle.net/11500/ATHENA-0000-0000-23E8-3

Thus, users can select specific texts or define sub-corpora based on date of publication, author, topic or any of the metadata. Texts are accessed through the specially designed user interface, but are not available for downloading, due to the agreements signed with the copyright holders (publishing houses, institutions, etc.), that allow access to but not free distribution of their material.

The HNC platform consists of two web applications: the frontend interface which provides user access, and the backend interface which provides platform administration. Both applications were developed in PHP scripting language, and they are hosted in a Internet Information Server (IIS).

4.1 Technical Description: the Backend

The core of the HNC platform is the SQL Server Database, where data are stored. Queries to the database are performed via SQL Query language, the response is handled by PHP and presented through the frontend.

The backend interface caters for the preparation of the documents to be inserted to HNC. A web-based management environment was implemented in PHP and connected to the corpus database. Mass document insertion was achieved via a PHP script which loaded the contents of the XML files into the SQL Server Database that hosts the HNC.

The backend platform also included a metadata editor for single XML file editing, classification, and insertion. Through this editor, annotators inspect, validate or correct (if needed) the automatically added metadata identified by the crawler during the collection process; they can also add the appropriate metadata in missing cases. Each document's metadata can be edited through the respective form (Figure 1), while word annotation editing is also available (Figure 2).

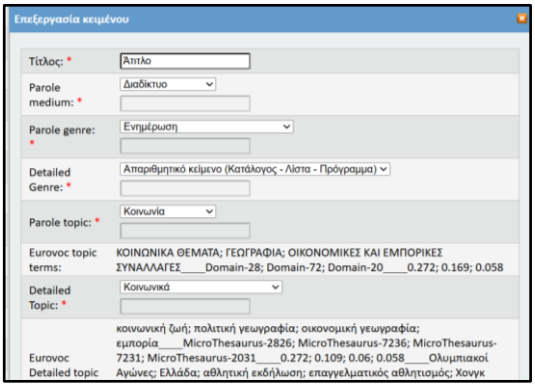

Figure 1: Document metadata editing

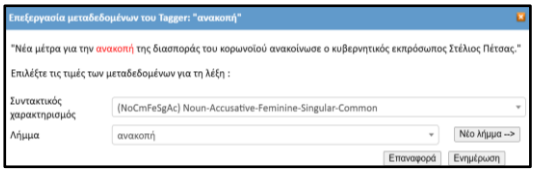

Figure 2: Word annotation editing

4.2 The Frontend: Search Functionalities

The frontend interface was designed to provide user-friendly interaction and easy customization in query building. Although the HNC frontend is open to use without registration, its main functionalities are available only to registered users. Guest users have limited results to their queries (50 sentences) while registered users have a significantly higher sentence limit (5,000 sentences). Furthermore, registered users can create and save sub-corpora to use in their searches; they also have access to the *Analysis* and *Correlation* functionalities (see Section 4.2), whereas these features are not available to non-registered users. This approach safeguards the performance of the HNC by protecting it from simultaneous multiple queries by large numbers of non-registered users.

The frontend interface allows users to perform filter-based selection of texts, to formulate queries to perform content search, to request morphological analysis of a word, or statistics of single or multiple items. Specifically, content search is achieved through word, lemma, or part of speech search, and combinations thereof (e.g. lemma X followed by lemma Y, lemma X followed by a preposition and then by a noun, etc.). The system retrieves sentences which comply with the search criteria and provides KWIC-type concordances (Figure 3) or full sentences (Figure 4) containing the requested query terms.

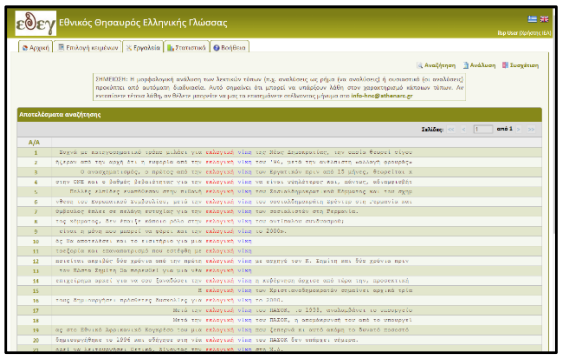

Figure 3: HNC concordance of a two-item query, each one marked in different color

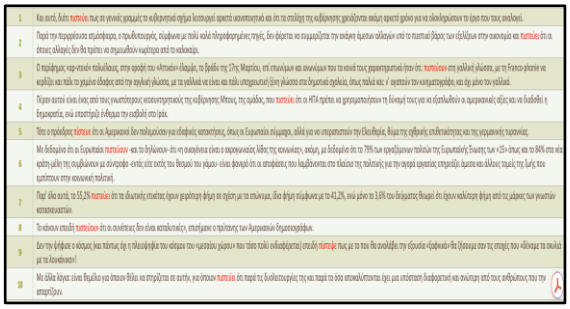

Figure 4: Full-sentence concordance

The users can also obtain statistical information, i.e., frequency of words, lemmas or parts of speech, n-gram frequencies, and lists with the most frequent words and lemmas in HNC.

The *Analysis* functionality offers the 'linguistic profile' of an item (word or lemma), namely, morphological analysis, frequent n-grams it participates in, and the yearly distribution of the item in HNC, i.e., its frequency on the time scale, based on the date of publication of the texts that contain the specific item (Figure 5).

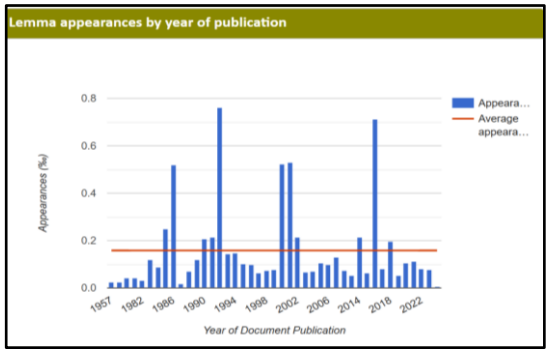

Figure 5: Yearly distribution of a lemma

Collocations a word participates in can be studied in detail through the most frequent words preceding or following a specific word (Figure 6).

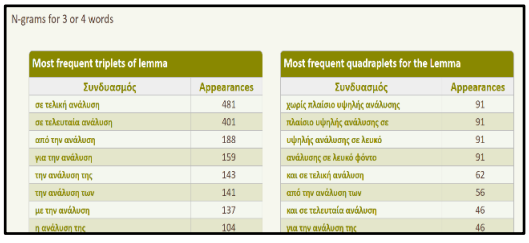

Figure 6: Most frequent n-grams for a lemma

The relation between two words is analyzed through the *Correlation* functionality, which provides information on their joint appearance in HNC: if and how frequently they appear together and how far apart (Figure 7) and their comparative frequency through the years (Figure 8).

Correlation & distance	
Joined appearances	498
Sequential appearances (% of joined ones)	0 (0 %)
Appearances in distance between 2 and 5 words (% of joined ones)	232 (46.59 %)
Appearances in distance between 6 and 10 λέξεις (% of joined ones)	72 (14.46 %)
Appearances in distances over 10 words (% of joined ones)	194 (38.96 %)
Minimum distance in sentences	2
Maximum distance in sentences	84
Average distance in sentences	11

Figure 7: Correlation & distance of two words

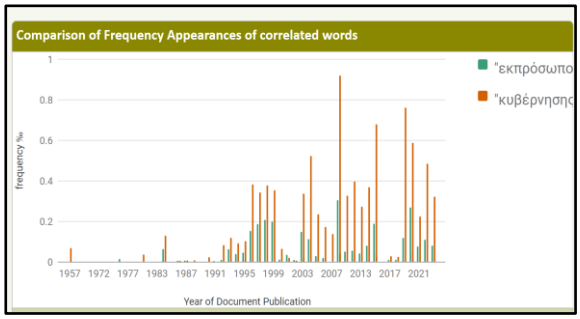

Figure 8: Joint yearly appearance of two words

Currently, registered users can store the sub-corpora they create. Query results are not stored, but search history is preserved along with its parameters so it can be replicated. Statistics, and specifically the top-frequency words/lemmas are also provided to registered users in CSV format for download.

5. The Golden Part of Speech Tagged Corpus

The Golden Part of Speech Tagged Corpus is a small subset of HNC (100,000 words), which consists of texts selected from a variety of sources covering various domains. Given that the material was planned from its conception to be freely distributed, the texts were selected based on their license (either CC0 4.0 or CC BY 4.0). The texts have been crawled from the web and underwent cleaning and removal of boilerplate material, manual correction of typos and spelling mistakes, automatic lemmatization and part-of-speech tagging for each word using the ILSP Feature-based multi-tiered POS Tagger, and manual correction of the results by linguists, in order to provide error-free material. The Golden Part of Speech Tagged Corpus is freely downloadable via CLARIN:EL as a single XML file[4].

[4] http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C

6. Future Steps

Although maintenance of HNC, as well as user support are taken care of, further enrichments or improvements have not been possible after the second phase, i.e. since 2021, due to the lack of funding.

With the proviso of funding availability, future steps concern the enrichment of the content with older material (diachronic expansion) and the addition of dialectal material (geographical and social dialects). Diachronic expansion will proceed stepwise, from the most recent to the older versions of the language, as this has also repercussions on the accompanying processing tools which need to be updated to successfully deal with older nominal and verbal inflection systems. As regards annotation tools, experimentation with LLMs is planned, focusing on the use of existing models for the processing of the existing material, with the aim to test their performance and to provide a new lemmatized and annotated version if appropriate. Additional steps include the finetuning of the existing models for the processing of the dialectal material which will be part of the corpus.

Issues to be tackled concern the identification and treatment of synthetic data and other machine-generated data such as translationese.

Finally, freely available material of the HNC is foreseen to be made available through the CLARIN:EL infrastructure, whose platform allows the downloading of both material and processing results.

7. Bibliographical References

Gavriilidou, M., Piperidis, S., Galanis, D., Pouli, K., Labropoulou, P., Bakagianni, J., Tsiouli, I., Deligiannis, M., Kolovou, A., Gkoumas, D., Voukoutis, L., and Gkirtzou, K. 2024. The CLARIN:EL infrastructure: Platform, Portal, K-Centre. In Lindén, K., Kontino, T. and Niemi, J (eds.) *Selected papers from the CLARIN Annual Conference 2023,* Linköping Electronic Conference Proceedings 210. ISBN: 978-91-8075-740-9. DOI: https://doi.org/10.3384/ecp210005

Gavrilidou, M. 2002. The Hellenic National Corpus on-line. In*: Revue belge de philologie et d'histoire.* Tome 80 fasc. 3, 2002. Langues et littératures modernes - Moderne taal en litterkunde. pp. 1003-1015. https://www.persee.fr/doc/rbph_0035-0818_2002_num_80_3_4652

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H., and Demiros, I. 2000. Design and implementation of the online ILSP Greek Corpus. In *Proceedings of Language Resources and Evaluation Conference* (LREC-2000). Athens, Greece. European Language Resources Association (ELRA) https://aclanthology.org/L00-1250/

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. 2000. A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC'00), Athens, Greece. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/181.pdf

Papavassiliou, V., Prokopidis, P., and Thurmair, P. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria. Association for Computational Linguistics, https://aclanthology.org/W13-2506/, pp. 43-51.

PAROLE. Design and composition of reusable harmonized written language reference corpora for European languages. 1995. Technical report, PAROLE Consortium. MLAP: 63-386

8. Language Resource References

Institute for Language and Speech Processing - Athena Research Center (2021). Golden Part of Speech Tagged Corpus. Version 1. [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C

Institute for Language and Speech Processing - Athena Research Center (2015). ILSP Feature-based multi-tiered POS Tagger. Version 1. [Software (Tool/Service)]. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-23E8-3