# The Hungarian Gigaword Corpus

**Csaba Oravecz, Tamás Váradi, Bálint Sass**
Research Institute for Linguistics, Hungarian Academy of Sciences
Benczúr u. 33, H-1068 Budapest
{oravecz.csaba,varadi.tamas,sass.balint@nytud.mta.hu}

**Abstract**

The paper reports on the development of the Hungarian Gigaword Corpus, an extended new edition of the Hungarian National Corpus, with upgraded and redesigned linguistic annotation and an increased size of 1.5 billion tokens. Issues concerning the standard steps of corpus collection and preparation are discussed with special emphasis on linguistic analysis and annotation due to Hungarian having some challenging characteristics with respect to computational processing.

**Keywords:** gigaword corpora, corpus annotation, morphological analysis

## 1. Introduction

Corpora have a significant impact on the success of research and applications in NLP along two critical dimensions, corpus quality and quantity. Recent initiatives have focused on achieving a substantial increase not only in the former, providing more and more detailed, "deep" annotation, but also in the latter, resulting in billion word corpora for several languages, many of them available from the LDC (see eg. Wei-yun and Huang (2006), Parker et al. (2011), Halácsy et al. (2008) or Ferraresi et al. (2008)). The paper describes the process of the creation of such a resource for Hungarian, a language with some challenging characteristics for computational processing. As the Hungarian Gigaword Corpus (HGC) is designed to serve as a resource for a wide range of linguistic research as well as for the interested public, a number of issues had to be resolved which were raised by trying to find a balance between the above two application areas.

## 2. Origins

The HGC has its roots in the Hungarian National Corpus (HNC) (Váradi, 2002) developed between 1998 and 2001 as a representative (balanced) sample of the language use of the second half of the 90s, providing valuable empirical evidence for the status of the Hungarian language for theoretical analysis and language technology alike. It was the first major annotated Hungarian corpus of about 187 million words, covering language variants inside the country and from neighbouring countries as well (yielding the Hungarian Minority Language Corpus as a subset of the HNC). The HNC has been a fairly popular language resource with more than 8000 registered users at the web search interface and dozens of research papers based on its data.

In the last 10-15 years, however, expectations for language resources have changed dramatically, especially in the following 3 areas:

- *size*: the dominance of data oriented methods and applications in NLP has led to the need for more and more data to achieve better and better performance,
- *quality*: the quality of language processing tools used for corpus processing has improved, calling for higher precision and finer levels of analysis and annotation in corpora,
- *coverage*: the need for the preservation of representativity has demanded subsequent samplings from language use including registers that are not yet covered by the HNC.

As a natural consequence of the above requirements, the HNC has become severely outdated in many respects and badly in need of major revision. To remedy these problems the following main objectives have been defined for the development of the HGC, focusing on the pivotal concept of *increase* in:

- *size*: extending the corpus to minimum 1 billion words,
- *quality*: using new technology for development and analysis,
- *coverage* and *representativity*: taking new samples of language use and including further variants (transcribed spoken language data and user generated content (social media) from the internet in particular).

If these objectives are fulfilled the HGC will be an up-to-date language resource that will service the current needs of the research community as well as the interested public.

## 3. Collection

### 3.1. Design considerations

Compiling corpora faces a number of theoretical and practical constraints, many of which have not changed much in recent years. Therefore, the reader is referred to the discussion in Váradi (2002) with respect to general issues about the design of the corpus, and, in particular, with respect to the concept of representativity, which, in its strict original sense has proved impossible to achieve and therefore has been replaced by the notion of balancedness. In this section, we focus only on problems originating from idiomatic features of the HGC, most prominently its size.

The predominant method for the collection of data on this scale has been either crawling the web or acquiring large amounts of newswire text. However, if used exclusively, both methods have their obvious weaknesses to produce a solid, balanced resource with sufficient metadata . The pros and cons of using the web as corpus are discussed in some detail in, for example, Baroni and Ueyama (2006), and their findings are valid for the HGC as well. This fairly opportunistic approach often results in very noisy data, which although can be filtered with various methods, frequently lacks even the basic metadata without which linguistic research cannot use the text reliably, and so could be applied only for specific text domains. A newswire corpus can have serious deficiencies with respect to representativity. Therefore, significant effort had to be put into acquiring the source data through controlled, targeted resource collection, appropriate for each type of source: crawling for user generated social media content, negotiating with publishers to have access to archives of news agencies and other digital text collections. Resources already in some structured (semi)standard and easily processable format have been given preference over ad hoc collections of files in various formats. The scale of the enterprise ruled out scanning documents and using OCR from the beginning due to the lack of necessary labour resources.

Copyright with respect to the development of language resources is a sensitive and hot topic, perhaps even more so than it was 10 years ago (Clercq and Perez, 2010; Reynaert et al., 2010). We found that it was substantially more difficult, sometimes even impossible, to collect appropriate licenses again for more data from all text providers whose data was already included in the corpus. Despite every reasonable effort there remained fair amounts of texts in the corpus which could not be covered with a license for full-scale access. The only possibility under this situation remained for us to offer different availability options for various sections of the HGC (see Section 6. below).

### 3.2. Sources

The composition of the HGC again basically follows that of the HNC and is discussed in detail in Váradi (2002). The distribution of tokens are illustrated in Table 1.[1] At first blush the table shows an increased dominance of the press genre, but it should be noted that the size of all other subcorpora has grown significantly in absolute terms, and there is a completely new genre, the (transcribed) spoken language as well.

Unfortunately, despite our expectation that after more than 10 years of the compilation of the HNC it would be an easy routine to collect electronic texts from data sources since document management and storage must have sufficiently advanced to follow established standards, we were faced with a large number of issues regarding accessibility, format and metadata of the source texts. This problem was further aggravated by the unforeseen and surprising reluctance by some text owners to issue licenses for their resources to be used even for strictly research purposes. As a result, recent materials from some news sources already present in the HNC and published continuously ever since are painfully missing from the HGC.

It is important to note that the HGC is not a faithful archive of the sources collected but primarily a language resource, a collection of linguistic data. It is not uncommon therefore for sections of very noisy source data to be removed from the corpus. The amount of this text, however, is insignificant compared to the full available data of the specific source and so have no influence on the result of investigations, experiments or the operation of applications (to be) based on the HGC.

## 4. Corpus preparation

The development of corpora of this magnitude is often influenced by practical constraints (such as the availability of human resources), nevertheless the standard steps in corpus preparation (preprocessing, normalization, up-translation, annotation) are usually followed, as it was done in the HGC, too. In the preprocessing and normalization phase textual content and basic document structure are identified in the raw data, and (near-)duplicates and non-Hungarian sections are filtered out. Language identification is carried out with near perfect precision/recall at the level of identifiable paragraph-like units longer than a specified threshold of characters using the algorithm of Lui and Baldwin (2012). Detecting duplicates proved to be a more complex issue excluding the use of standard methods developed for large scale web corpora (Pomikalek, 2011). The wide spectrum of sources (ranging from social media through official, legal documents and newswire to literature) required customized processing that is primarily based on the Kupietz (2005) toolkit, but the default detection has to be followed by manual post-editing identifying typical types of duplicates which have to be removed or, on the contrary, preserved, as the case may be. There are near-identical textual segments whose unalienable feature is their repetitiveness in language use, and therefore removing them would lead to data distortion. Typical examples are weather report sections in newspaper data, which use a language so constrained that automatic detection is prone to identify them as (near-)duplicates, but they have to be preserved since this kind of repetition is entirely deliberate.

To facilitate linguistic analysis an extensive normalization is carried out at the character level, in which various renderings[2] of characters are mapped to the (near-)equivalent characters of the Hungarian alphabet or some appropriate other character. These may include ligatures, normal text rendered by calligraphic unicode symbols, some fancy punctuation and the like.

The output of the preparation step is the input for the linguistic analysis in the form of a clean XML file for each type of source, level 1 encoded according to a slightly modified DTD based on the Corpus Encoding Standard (Ide, 1998), with all major structural units marked up down to the paragraph level. Metadata is encoded in TEI conformant headers.

## 5. Analysis and annotation

### 5.1. The processing pipeline

All tools used for analysis at all levels of processing have been updated to produce a more precise, detailed and reli-

[1] It is of course subject to some change as the corpus is growing due to further development.

[2] Usually in the form of a unicode symbol.

| Register | HNC | HGC | | Source |
|---|---|---|---|---|
| Journalism | 84,500,000 | 643,257,776 | (42%) | Daily/weekly newspapers |
| Literature | 38,200,000 | 221,731,436 | (14.5%) | Digital Literary Academy |
| (Popular) science | 25,500,000 | 110,903,157 | (7.2%) | Hungarian Electronic Library |
| Personal | 18,600,000 | 338,600,000 | (22.1%) | Social media |
| Official | 20,900,000 | 135,401,305 | (8.8%) | Documents from public admin. |
| (Transcribed) spoken | – | 83,040,104 | (5.4%) | Radio programs |
| | **187,000,000** | **1,532,933,778** | | |

Table 1: The composition of the HGC in number of tokens

able annotation than the one in the HNC. The toolset works as a pipeline consisting of separate modules for the main processing stages: a tokenizer/segmenter, a morphological analyzer, a POS tagger and tools for higher level processing.

Tokenization and sentence segmentation is carried out with an extended and updated version of the Huntoken tool[3], highly customized to cope with erroneous and noisy input (such as social media downloaded from the web, in particular). With Hungarian being a highly inflectional language, reliable morphological analysis is exceptionally important. The Humor morphological analyzer tool (Prószéky and Tihanyi, 1996) has undergone a major update to give extended information on stems, each morph and compounding. The representation of each morph in the annotation presented two new challenges not yet handled in Hungarian:

- Any usable morphological analyzer for Hungarian will produce significant structural ambiguity in many cases with respect to the possible combinations of stems and derivational suffixes (see Váradi and Oravecz (1999) for a detailed illustration). According to normal procedure, since the rightmost derivational suffix determines the part of speech of the word, all derivational details are eliminated, the stem is taken as including the rightmost derivational suffix and the resulting wordform that is input to the POS tagger is composed of the stem and only the inflectional suffixes.[4] This is a necessary step to make POS tagging tractable. However, if all derivational details are to be preserved for the sake of annotation (but still not for tagging), this added ambiguity must be taken care of. A simple heuristic that was applied is to select the analysis with the highest number of morphemes, this is always the most informative method about the internal structure of the token. If there are more than one analyses with the same highest number of morphemes all analyses are preserved. This represents a derivational ambiguity extremely difficult if not impossible to resolve automatically.
- Compounding is very productive in Hungarian and to ensure acceptable coverage, the morphological analyzer has to allow a wide scale of combinations of stems, which inevitably leads to overgeneration, not necessarily permitting bad compounds but rather unusual combinations with unnatural, far-fetched interpretations, bringing another layer of unwanted ambiguity to the analysis.[5] For a corpus designed to serve human research user experience is critical, so eliminating errors is not only driven by frequency but also by quality or the language user's sensitivity. Unusual compounds fall into this "sensitive" category, consequently, if possible, they must be completely eliminated regardless of frequency. The current solution is to manually produce filter rules with regular expressions to get rid of the unwanted analyses.[6]

The disambiguation framework based on Oravecz and Dienes (2002) and Halácsy et al. (2006) has been retrained with a 1 million word manually tagged training corpus yielding high precision output (near 98%), and new layers of analysis have been added in the form of NP chunking, and named entity recognition (Varga and Simon, 2007). Initial results for these higher level annotations are not very convincing, and some further work is needed to fine-tune the tools to produce higher quality output.

### 5.2. Annotation format

The hub of the corpus encoding for linguistic analysis is the output of sentence splitting and tokenization. Each token is on a separate line with empty lines marking sentence boundaries. All further annotation is added as tab separated columns similarly to the WaCky format (Baroni et al., 2009), resulting in a flexible and easy to process output, which can be readily converted to XML (and validated) at any stage of the processing pipeline. This format is illustrated with a small extract for the phrase "the English language text [is] the primary" in Figure 1.[7] The first nine columns stand for the token, stem, morphosyntactic description (as output from the morphological analyzer), corpus tag (for pos tagging), morpheme level encoding with compounding information, syllable structure for the token

[3] https://lrt.clarin.eu/tools/huntoken-tokenizer-and-sentence-splitter

[4] This is the level of analysis that is encoded for example in the MULTEXT-East specifications (Erjavec, 2004). Since the HGC analysis is a lot more detailed, the application of this standard is ruled out.

[5] An example can be the simple noun *lázadó* ("rebel") as composed of *láz* ("fever") + *adó* ("tax"), the compound reading being extremely improbable.

[6] This issue opens up a whole new domain of possible future research of trying to algorithmically solve the problem.

[7] For presentation purposes, the annotation layout is edited.

```
az       az       DET     D__D    compound=n;;hyphenated=n;;stem=az::DET
                                  BC      BC      az      az      B
angol    angol    A.NOM   AS_A    compound=n;;hyphenated=n;;stem=angol::A;;
                                  morphemes=ZERO::NOM
                                  BCCBC   BCCBC   angol   angol   I
nyelvű   nyelvű   A.NOM   AS_A    compound=n;;hyphenated=n;;stem=nyelv::N;;
                                  morphemes=ZERO::NOM ű::_UKEP
                                  CNCCF   CNCCF   Nelvű   Nelvű   I
szöveg   szöveg   N.NOM   NS3NN   compound=n;;hyphenated=n;;stem=szöveg::N;;
                                  morphemes=ZERO::NOM
                                  CFCNC   CFCNC   Söveg   Söveg   I
az       az       DET     D__D    compound=n;;hyphenated=n;;stem=az::DET
                                  BC      BC      az      az      O
irányadó irányadó A.NOM   AS_A    compound=y;;hyphenated=n;;stem=ad::VERB
                                  irány::N;;morphemes=ZERO::NOM ó::_OKEP
                                  NCBCBCB NCBCBCB iráNadó iráNadó O
.        .        SPUNCT  __SPUNCT__   __NA__  __NA__  __NA__  __NA__  __NA__
```

Figure 1: Sample of the raw IOB format

and the stem[8], and pseudo-phonemic transcription for the token and the stem, respectively. For annotations spanning over several tokens the standard IOB (Inside, Outside, Beginning) encoding scheme (Ramshaw and Marcus, 1995) is used. In Figure 1 the tenth column uses this format to encode noun phrases.

The higher level XML encoding of document structure is kept separately as standoff annotation, which can be merged with the linguistic annotation to produce a unified output.

## 6. Implementation and distribution

The corpus engine selected for the implementation of the HGC is the Manatee/Bonito corpus management system (Rychlý, 2007), the open source part of the engine behind the Sketch Engine (Kilgarriff et al., 2004). This is a mature toolkit, very fast both in indexing and querying and able to handle several billion tokens. The skeleton of the HGC search interface is based on the Bonito application, so the standard built-in services of this package are readily accessible. However, the interface has been substantially extended to allow for complex searches on all layers of the detailed (morphonological) annotation (syllable structure, CV skeleton, morpheme types, compounding etc.) providing user-friendly access and supporting linguists in doing extensive qualitative and quantitative research based on the HGC. Figure 2 illustrates the level of details of the annotation and also the extensive possibilities of the query interface. In this example, when selecting verb as part of speech, a roll down menu of all derivational and inflectional properties of Hungarian verbs is displayed. Here we search for verb forms which contain the '-hAt' derivational suffix (meaning "able"), and the '-lAk' inflectional ending encoding first person singular subject and second person direct object at the same time, in declarative mood.

With respect to the accessibility of the corpus, the full version is currently available only through the web search interface due to copyright restrictions. Sources for which the licenses make it possible will be freely accessible in full text version as well.

## 7. Future work

The development of the HGC has benefited from previous experience gained during the creation of its predecessor but also from user feedback. A fair amount of work has been invested at all stages of the process to produce a language resource unprecedented for Hungarian not only in quantity but also in quality in this magnitude. This makes the corpus an ideal base to derive further resources by utilizing appropriate post-processing algorithms. These resources might include frequency dictionaries, collocation lists, verb subcategorization frame lexica etc.

A framework is being developed to make periodic update and extension of the corpus viable from continuously monitored data sources providing repeated sampling of language use, and to immediately update the quality of analysis when any of the processing tools receive an upgrade, correcting errors in the linguistic analysis.

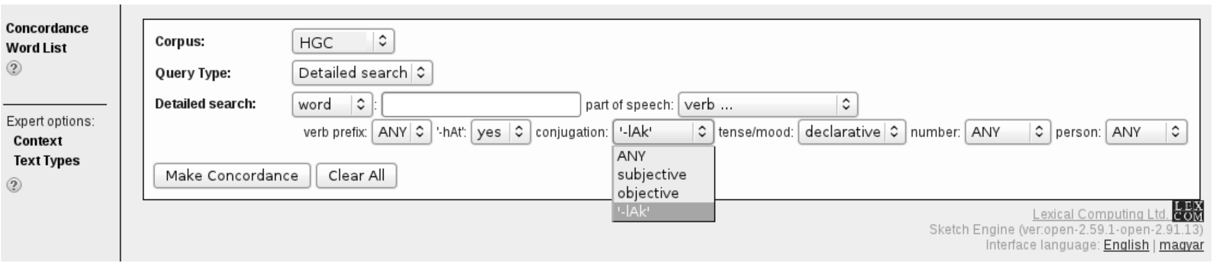
## 8. References

Baroni, M. and Ueyama, M. (2006). Building general- and special-purpose corpora by web crawling. In *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application.*, pages 31–40, Tokyo, Japan.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Clercq, O. D. and Perez, M. M. (2010). Data collection and IPR in multilingual parallel corpora: Dutch parallel corpus. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evalua-*

[8] F: front, B: back , N: neutral vowel; C: consonant.



Corpus: HGC

Query Type: Detailed search

Detailed search: word :

part of speech: verb ...

verb prefix: ANY '-hAt': yes conjugation: '-lAk' tense/mood: declarative number: ANY person: ANY

Expert options:

Sketch Engine (ver:open-2.59.1-open-2.91.13)

Interface language: English | magyar

Figure 2: A section of query interface of the HGC.

*tion (LREC'10)*, pages 3383–3388, Valletta, Malta. European Language Resources Association (ELRA).

Erjavec, T. (2004). MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *LREC-04*, pages 1535–1538. ELRA.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Halácsy, P., Kornai, A., Oravecz, Cs., Trón, V., and Varga, D. (2006). Using a morphological analyzer in high precision pos tagging of Hungarian. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2245–2248, Genoa.

Halácsy, P., Kornai, A., Németh, P., and Varga, D. (2008). Parallel creation of gigaword corpora for medium density languages – an interim report. In *In Proceedings of Language Resources and Evaluation Conference (LREC08)*.

Ide, N. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *Proceedings of the First International Language Resources and Evaluation Conference*, pages 463–470.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of Euralex*, pages 105–116.

Kupietz, M. (2005). Near-duplicate detection in the IDS corpora of written German. Technical Report IDS-KT-2006-01, Institut für Deutsche Sprache.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea. Demo Session.

Oravecz, Cs. and Dienes, P. (2002). Efficient stochastic part-of-speech tagging for Hungarian. In *LREC-02*, pages 710–717, Las Palmas, Canary Islands, Spain.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English Gigaword Fifth Edition. Linguistic Data Consortium.

Pomikalek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Masaryk University, Faculty of Informatics, Brno.

Prószéky, G. and Tihanyi, L. (1996). Humor – a morphological system for corpus analysis. In *Proceedings of the first TELRI seminar in Tihany*, pages 149–158, Budapest.

Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 82–94.

Reynaert, M., Oostdijk, N., Clercq, O. D., van den Heuvel, H., and de Jong, F. (2010). Balancing SoNaR: IPR versus processing issues in a 500-million-word written Dutch reference corpus. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Rychlý, P. (2007). Manatee/Bonito – a modular corpus manager. In *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno: Masaryk University.

Varga, D. and Simon, E. (2007). Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 18(2):293–301, February.

Váradi, T. and Oravecz, Cs. (1999). Morphosyntactic ambiguity and tagset design for Hungarian. In *Proceedings of the Workshop on Linguistically Interpreted Corpora*, EACL'99, pages 8–13, Bergen. Association for Computational Linguistics.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 385–389, Las Palmas.

Wei-yun, M. and Huang, C.-R. (2006). Uniform and effective tagging of a heterogeneous gigaword corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 2182–2185, Genoa, Italy.