The Hungarian National Corpus

Tamás Váradi

Research Institute for Linguistics
Hungarian Academy of Sciences, Budapest
varadi@nytud.hu

Abstract

The paper reports on the development of the Hungarian National Corpus, which was completed at the end of 2001 after four years' effort. The HNC is designed to be a balanced reference corpus of current written Hungarian consisting of 150 million words. The paper first discusses basic design issues concerning the composition of the corpus. The HNC adopts a fairly pragmatic approach, focusing on five major text types. The second half of the paper contains details of the annotation and tagging system used.

1. Introduction

Work on the Hungarian National Corpus began early 1998 at the Research Institute of the Hungarian Academy of Sciences. The project grew out of previous work in the compilation of the Hungarian Historical Corpus, a 20 million word corpus of texts from the 18th to the late 20th century designed to serve the purposes of the unabridged Academy dictionary of Hungarian. Valuable corpus linguistic expertese and tools were also accumulated in the MULTEXT-EAST project.

The initial objective was to create a balanced reference corpus of present-day Hungarian consisting of 100 million words. The project was funded by Országos Tudományos Kutatási Alap (National Fund of Scientific Research) for a period of four years. The HNC project started with very modest resources amounting to no more than two full-time and and one part-time researchers, in fact. Hence it was imperative to fully exploit computational methods wherever possible.

1.1. Design considerations

1.1.1. Spoken vs. written language

From the beginning, work on the compilation of the HNC faced a number of theoretical and practical constraints. First of all, we had to rule out capturing spoken language because it was deemed far too labour intensive for our resources available. As far as spoken Hungarian is concerned, the transcripts of the Budapest Sociolinguistic Interview (Kontra and Váradi, 1997), roughly 600 hours of taped interviews conducted with a representative sample of 250 speakers from Budapest, will be a valuable component of the HNC. Transcripts of the 50 informants involved in Phase 2 of the BSI is expected to be completed by the end of 2003.

1.1.2. Printed vs. electornic sources

Shortage of manpower also forced us to consider as source material only texts already available in electronic media. It appeared a rather stringent limitation at the time. However, it proved a decision that was justified by the phenomenal rate of expansion of the availability of Hungarian documents on the web. This decision more or less defined the timescale within which the data originated. The overwhelming majority of the texts date from no earlier than the mid-nineties. Among the theoretical issues we still had to face the weighty problem of compiling a representative sample of present-day Hungarian.

1.1.3. Geographical coverage

Hungarian is spoken by approximately 14 million native speakers, 4 million of whom live outside Hungary (Kontra, 1999). The HNC cannot aspire to provide comprehensive cover for the varieties of Hungarian spoken around the globe. However, the Hungarians living in the neighbouring countries in areas formerly belonging to Hungary constitute a speacial case. Their geographical position in Hungarian culture is conceived as a kind of semi-distance, *határontúl* 'across the border' intermediate between *belföld* 'inland' and *külföld* 'abroad'. They are considered as Hungarian nationals who are citizens of foreign countries. The status of their language variety is a matter of hot debate among linguists. In recognition of their special position in Hungarian culture, the HNC does contain a sample of newspapers from Romania, Slovakia and Novi Sad (Serbia and Macedonia).

1.1.4. Temporal coverage

The HNC aims to cover the present-day state of the Hungarian language. The overwhelming majority of the texts in the whole corpus indeed date from no earlier than the mid-nineties. There are two possible exceptions. The *belles-lettres* subcorpus consists of the writings of a group of authors who are considered living classics. As the complete *oevres* is included from each author, it is the case that texts dating from much earlier times can be found in this subcomponent. The same is true of the scientific literature subcorpus, although there we had some freedom to influence the choice of selection along the dimension of time.

1.1.5. Representativeness

After an extensive review of the literature we have come to the conclusion that representativeness is an ideal that cannot be attained in principle. Biber (1993), one of the most influential papers on this issue, sets out the inherent difficulties in this objective with exceptional clarity. However, he proposes to overcome these difficulties by blithely declaring the conventional notion of representativeness irrelevant to corpus linguistics and proposing to replace it with one that equates representativeness with maximal variety of text types. As argued elsewhere (Váradi, 2001), to eliminate the difficulty with a concept by redefining it altogether is a rather facile "solution". The notion of representativeness is well understood by the general public and is inextricably linked to the notion of proportionality, which Biber also rejects. The practice of flouting fellow disciplines and the general public alike and bending such a well understood notion to suit one's own purposes is not only unwise but, in our opinion, even threatens the integrity of the field.

If the notion of representativeness in its original sense of a proportional sample cannot be achieved in principle in corpus linguistics, it is more honest to abandon it rather than bend it to suit our purposes. The notion of a balanced corpus seems to be a suitable concept as it is free from the strict statistical commitments involved in the term "representative".

1.2. A modular corpus

In the end, the HNC was designed to cover five major text varieties, which constitute five sub-corpora. They can be queried individually or in any combination. Justification for this design strategy is based on two points. Unfortunately, there does not exist for Hungarian the body of careful text typological studies such as those pioneered by Biber in numerous publications. Hence we do not have the criteria, whatever they are worth, that are used in English to establish a set of text types on the basis of their linguistic characteristics. Second, (Biber et al., 1999), a major corpus-based grammar of English, analyses language use in similarly broad categories. Apparently, a more detailed scheme would have been intractable when it comes to presenting a broad picture of language use.

1.3. Sources

Table 1 contains a breakdown of the internal composition of the HNC.

Register	Words	Source
Journalism	75	Daily/weekly newpapers
Belles-letters	15	Digital Literary Academy
(Popular) science	20	Hung. Electronic Library
Official	20	Web sites of public admin.
Personal	20	Internet forums
Total	150	

Table 1: The composition of the HNC

The size of the literary component is expected to grow to approximately 40 million words when all the data targeted for it is available for the HNC.

1.3.1. Newspapers

Newspaper texts make up half of the corpus in its present release. It is customary to mention newspapers in a somewhat derisory manner probably because of their ready availability. However, the issue of accessibility has nothing to do with their relevance for corpus linguistic purposes. In fact, newspapers represent a very broad mix of language varieties both in terms of horizontal and vertical stratification of language use. On reflection, the mix of newspapers represented in the HNC are slightly slanted in that quality dailies are probably overrepresented. However, at the early phase when text collection began, tabloid papers were not available on the Internet.

1.3.2. *Belles-lettres*

The literary component of the HNC is particularly valuable. It consists of the complete material of the Digital Literary Academy, which is a Government sponsored major project to publish the total oeuvres of 52 living Hungarian authors on the Internet. The distinguished writers are paid a monthly fee in remuneration for their copyright. Under an agreement with the Neumann-ház, which was entrusted with the technological implementation of publishing the literary archives, the HNC obtained the right to incorporate the DIA material in its entirety. The digitization and coding of the texts are still in progress. The whole of the targeted collection in DIA is estimated to reach 40 million words, which will be all taken over into the HNC on completion.

1.3.3. Scientific texts

The third component, (popular) scientific works, also come from a well-established archive, the Hungarian Electronic Library, which grew from the voluntary effort of a few enthusiastic individuals to be a vast collection of a broad range of texts.

1.3.4. Official language

The fourth subcorpus is aimed to cover language use in official contexts. The texts include legislation, regulations, by-laws, transcripts of parliamentary debates, all sorts of documents that one might informally characterise as officialese.

1.3.5. Personal

The subcorpus branded personal communication consists of discussions in internet forums operated by index.hu, one of the oldest and biggest Internet portals in Hungary. This language variety is an interesting specimen as it is possibly the closest that one can get in written form to spontaneous communication. In some cases, the lively and rapid exchanges come very close to spoken interaction.

2. Corpus annotation

2.1. Preprocessing

Figure 2 contains an overview of the first phase of the corpus annotation process.

Raw text
(.txt/HTML)
Conversion
Validation
Validated
SGML

Figure 1: SGML conversion

The majority of the input files were formatted in HTML. The HTML encoding proved fairly useful in allowing us to implement some heuristic algorithm to identify the structure of the text and to infer some key bibliographical information such as the author, title, date etc. Unfortunately, this procedure was entirely dependent on the in-house conventions used by the various publishing houses, which not only

POS	Num	Pers	Stem [NAR] Mood/Tense [V]	Case [N] Def [V]	Owner's Num	Owner's Pers	Total
N	2 [PS]	3 [123]	5 [QAVNP]	21	2 [PS]	3 [123]	2058*
A			2 [AV]				2*
R			2 [RV]				2*
V	2 [PS]	3 [123]	5 [PRCSI]	3 [ID2]			79*
Invariant minor categories: Q, D, PRE, RP, C, Int, Y							7
							2148

N = Noun	A = Adjective	R = Adverb	V = Verb
Q = Numeral	D = Article	PRE = Verbal prefix	RP = Postposition
C = Conjunction	Y = Abbreviation	Int = Interjection	

Def = Agreement in definiteness with object (def, indef, 2nd person)
Owner's Num = sing. or plural owner   Owner's Pers = person marker of owner
* = not all combinations are possible, so not a simple product
[NAR][V][N] = POS categories to which the attribute apply

Table 2: The combinatorial scheme of Hungarian inflections

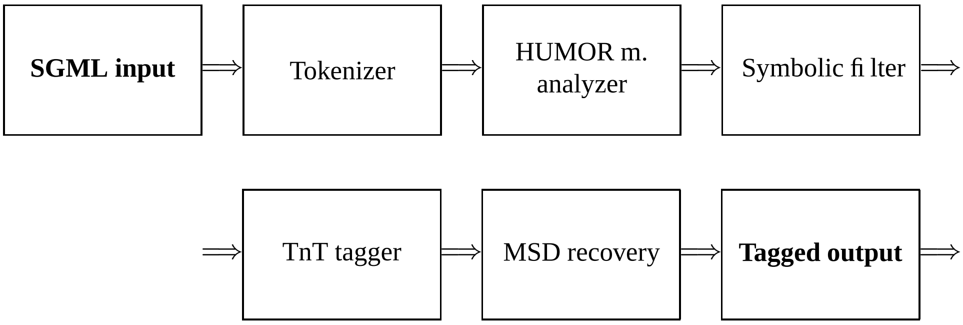

Figure 2: Components of the tagging system

varied from one to the other but also was subject to change without prior notice. It was therefore necessary to monitor the automatic downloading of file regularly and adjust the conversion algorithm wherever necessary.

The documents were encoded according to the Corpus Encoding Specification (Ide, 1998).

2.2. Tagging

2.2.1. The morphological profile of Hungarian

The morphosyntactic tagging of Hungarian presented a major challenge at the outset of the project. Not only was this to be a completely novel enterprise but the notriously complex morphological system of Hungarian destined it to be a highly difficult process. To illustrate the difficulties at hand, one recent estimate (Tihanyi, 1996), which only considered the combinatorial possibilities of inflection and derivation, produced the figure of slightly more than 4 million forms for a single verb. This would amount to close a 20 billion word forms for a medium sized vocabulary of 50,000 nouns and 9,400 verbs. Such an abundance of word forms clearly rules out the possibility of processing by lexical lookup from tables. While it is claimed (Elworthy, 1995) that prior morphological analysis of the corpus may take over a large part of the job of the tagger, such a high number of word forms made it extremely difficult to establish a tag set that does justice to the rich morphosyntactic information encoded within the words and at the same time remains computationally tractable.

Table 2 displays the list of combinatorial possibilities of Hungarian inflectional suffixes. It appears that even without compounding and derivation, both very productive in Hungarian, one has to contend with no less than 2148 forms. This may be considered the full set of corpus tags – which would present severe problems of data sparsity for a stochastic tagger.

2.2.2. Disambiguation

Figure 2.1. pesents an overview of the linguistic annotation of the HNC.

The tokenization was carried out with the MTSEG tool (Russell and Petitpierre, 1995), a freely available tool developed in the Multext project. It has been customized for Hungarian with auxiliary lexicons of abbreviation lists, multiword units and date format templates.

Morphological analysis of the data was done with the help of HUMOR, the morphological analyzer developed by MorphoLogic (Prószéky and Tihanyi, 1996). Some postprocessing was applied to the output of HUMOR to eliminate spurious multiple analyses and to establish a single lemma, which was taken to be the rightmost relative stem

Units	Unique	Ambiguous	Unknown	Total
Word forms	1,048,263 (60.6%)	228,105 (13.2%)	452,403 (26.2%)	1,728,771
Tokens	50,437,483 (68.1%)	20,542,442 (27.7%)	3,083,286 (4.2%)	74,063,211

Table 3: Summary figures of the morphological analysis (74m running words)

in case of compund and derived forms.

Table 2.2.1. shows the amount of ambiguity after the morphological analysis. The figures were established on roughly half of the data. It appears that about 28 % of the tokens need disambiguation. In fact, a fintuning of the MSD set and some post editing of the output of the morphological analizer enabled us to push this figure to around 23 %

We decided to adopt a stochastic method in the tagging of the HNC. For an alternative approach based on symbolic learning algorithm to induce disambiguation rules from a training corpus see (Alexin et al., 1999). The stochastic process was complemented with a simple symbolic prepocessor operating with environmentally conditioned deterministic rules (in effect, a small constraint grammar). As is noted in Section 5 of (Oravecz and Dienes, 2002), this device proved very efficient, resulting in about 10% reduction of errors.

The tagging was carried out with the *Trigrams 'n Tags* (TnT) system, an HMM based trigram tagger developed by Thorsten Brants (2000). Courtesy of the developer, we could use a slightly modified version of the tool in which each input token was associated with a list of possible tags (allowing the application of the ambiguity reduction to the output of the symbolic prepocessor). The tagger was trained on a manually disambuated corpus of 270,830 words.

In addition to the large set of morphosyntactic descriptions owing to the richness and productivity of morphology, which result in data sparseness and computational inefficiency mentioned in 2.2.1., what exacerbated the tagging further was the huge number of possible Hungarian word forms and the fact the HMM based models calculate lexical probabilities from a word form lexicon generated from the training corpus. Inevitably this results in a large number of unknown words for the model when new data is analyzed.

In the end, a solution was developed to tackle both problems. The tagging was based on the idea of tiered tagging pioneered by Tufiş (1999). This allowed us to use reduced tagsets down to the cardinality of 50-100 and still recover all the information contained in the Full tagset. Tests of tagging performance on a randomly chosen test set of 70 words resulted in 97.6% in terms of the standard *correctly tagged/all correct* performance measure.

Space constraints prevent an in-depth discussion of the technical details of the tagging process, which have been reported in a series of articles (Tufiş et al., 2000; Dienes and Oravecz, 2000; Oravecz and Dienes, 2002).

2.3. Implementation

The corpus engine selected for the implementation of the HNC was the Corpus Workbench System (Christ, 1994). As a matter of technical convenience the whole of the HNC was indexed as a single uniform body of texts. Acces to the various subcorpora individually or in any combination was implemented as a matter for the query system. The corpus itself can be accessed directly via the command line by those who have the proper authorication and the facility to handle the CQP query system. This is, for the moment, still the most efficient and flexible way to access the corpus data. It is admittedly not the easiest but at the same time the most rewarding way, in that it is the only way that data can be queried without limitation.

2.4. User interface

At the moment, the CQP engine is accessed via a CGI script written in Perl. The implements most of the functionality of the CQP system and adds a few relating to the annotation involved. Hence, it is possible to toggle subcorpora, to filter the search space in terms of author, genre, time and corpora.

The complex morphological information associated with each token (see Figure 3 for a sample) presents a major challenge for the user interface of the query system. The difficulty lies in devising a system which allows a detailed yet flexible specification of the target word or words and all this in a simple and user friendly manner.

At the moment a prototype user interface is available at $http : \backslash\backslash corpus.nytud.hu/test$, which still lacks quite a number of the functionalities of the envisioned system but allows specification of the linguistic characteristics of the search expression through the definition of the required MSD. As this presupposes familiarity with pretty arcane technical details far beyond the average user, this manner of search expression definition must obviously be replaced with some graphical, menu driven technique.

3. Future work

The current release is only the first public version of the HNC. The size of the corpus is expected to grow, if only because of the addition of more literary texts after data entry in the Digital Literary Academy project is finished. Preliminary tests of the tagging system has so far returned reasonably good results. Further enhancements can only come with the next release of the HUMOR system when its vocabulary has been overhauled. Obviously, the corpus annotation can be taken a stage further by implementing some shallow parsing. We are currently engaged with MorphoLogic on joint work to develop a syntactic analyser.

The most urgent task, however, is the development of a user-friendly interface to enable as precise and complete access to the data as possible. This may mean a higher level of abstraction through the use of templates, which maximally utilize the advanced search facilities of the Corpus Workbench System.

```
<!-- HVG ./0116/0116009.htm --> <div type="article" column="unspec">
<opener> <dateline> <w lemma="HVG" msd="N.NOM" ctag="NS3NN">HVG</w>
<w lemma="2001/16" msd="DIG" ctag="Q">2001/16</w> <c lemma="."
msd="SPUNCT" ctag="SPUNCT">.</c> <w lemma="szám" msd="N.NOM"
ctag="NS3NN">szám</w> <date iso8601="04-21-2001"> <w
lemma="2001._április_21." msd="DATUM"
ctag="DATUM">2001._április_21.</w> </date> </dateline> </opener>
<head rend="IT" type="unspec"> <s> <w lemma="egészségügyi"
msd="A.NOM" ctag="AS_A">Egészségügyi</w> <w lemma="szigorítás"
msd="N.PL.NOM" ctag="NP3NN">szigorítások</w> </s> </head> <head> <s>
<w lemma="sok" msd="Num.NOM" ctag="Q">Sok</w> <w lemma="zseb"
msd="N.ELA" ctag="NS3NE">zsebből</w> <w lemma="vérzik" msd="V.e3"
ctag="VS3RI">vérzik</w> </s> </head> <head rend="BO" type="display">
<s> <w lemma="Alaposan" msd="Adv" ctag="R">Alaposan</w> <w
lemma="felkavar" msd="Pre.V.TMe3" ctag="@VS3PD">felkavarta</w> <w
lemma="a" msd="Det" ctag="D">a</w> <w lemma="kedély" msd="N.PL.ACC"
ctag="NP3NA">kedélyeket</w>
```

Figure 3: A sample of the corpus annotation

4. Acknowledgements

Work reported in this article was supported by grant number T026091 of the Országos Tudományos Kutatási Alap.

5. References

Zoltán Alexin, Tamás Váradi, Csaba Oravecz, Gábor Prószéky, János Csirik, and Tibor Gyimóthy. 1999. Fgt – a framework for generating rule-based taggers. In *ILP-99 Late-Breaking papers*, Bled, Slovenia.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4):243–257.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, Seattle, WA.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94*, pages 23–32, Budapest.

Péter Dienes and Csaba Oravecz. 2000. Bottom–up tagset design from maximally reduced tagset. In Anne Abeille, Thorsten Brants, and Hans Uszkoreit, editors, *Proceedings of the Workshop on Linguistically Interpreted Corpora*, COLING 2000, pages 42–47.

David Elworthy. 1995. Tagset design and inflected languages. In *Proceedings of the ACL-SIGDAT Workshop*, Dublin.

Nancy Ide. 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463–70, Granada. ELRA.

Miklós Kontra and Tamás Váradi. 1997. *The Budapest Sociolinguistic Interview: Version 3*. Working Papers in Hungarian Sociolinguistics No. 2. Linguistics Institute, Hungarian Academy of Sciences.

Miklós Kontra. 1999. The sociolinguistics of hungarian outside hungary. Technical report, Open Society Institute, March. http://e-lib.rss.cz/digilib/pdf/22.pdf.

Csaba Oravecz and Péter Dienes. 2002. Efficient stochastic part-of-speech tagging for hungarian. In *Third International Conference on Language Resources and Evaluation*. (this volume).

Gábor Prószéky and László Tihanyi. 1996. Humor – a Morphological System for Corpus Analysis. In *Proceedings of the first TELRI Seminar in Tihany*, pages 149–158, Budapest.

G. Russell and D. Petitpierre. 1995. Mmorph – the multext morphology program, version 2.3. Technical report, CNRS. MULTEXT deliverable report for task 2.3.1.

László Tihanyi, 1996. *MULTEXT-EAST Delivearable D1.2. Application to Hungarian. Appendix 2*, chapter Number of Hungarian Word Forms. , May.

Dan Tufiş, Péter Dienes, Csaba Oravecz, and Tamás Váradi. 2000. Principled hidden tagset design for tiered tagging of Hungarian. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens.

Dan Tufiş. 1999. Tiered tagging and combined language models classifiers. In F. Jelinek and E. Nöth, editors, *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 1692, pages 28–33. Springer.

Tamás Váradi. 2001. The linguistic relevance of corpus linguistics. In *Corpus Linguistics 2001*.