Hungarian National Corpus II. – an attempt to gigacorpus

Maďarský národný korpus II. – pokus o gigabajtový korpus

Tibor Pintér (Jazykovedný ústav MAV)
8. Jún 2012, Bratislava
Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete
(10 rokov Slovenského národného korpusu)

MAĎARSKÝ NÁRODNÝ KORPUS
Hungarian National Corpus

- Bol vytvorený medzi rokmi
  - 1998-2002 – verzia s textmi z Maďarska (texty zo zahraničia len ojedineje)
  - 2002-2005 – verzia doplnená s textami z pohraničných štátoch/regiónoch
    - maďarský jazyk na Slovensku
    - maďarský jazyk v Rumúnsku
    - maďarský jazyk v Ukrajine
    - maďarský jazyk v Srbsku
- Cieľ: 100 millión tokenov, reprezentovať úzus maďarského jazyka

- Was created between
  - 1998-2002 – with texts from Hungary (Hungarian texts outside Hungary rarely)
  - 2002-2005 – version updated with Hungarian texts from outside Hungary
    - Hungarian from Slovakia
    - Hungarian from Romania
    - Hungarian from Ukraine
    - Hungarian from Sebia
- Aim: 1 billion tokens representing the real usage of the Hungarian language

KORPUS MAĎARSKÉHO JAZYKA
Hungarian National Corpus

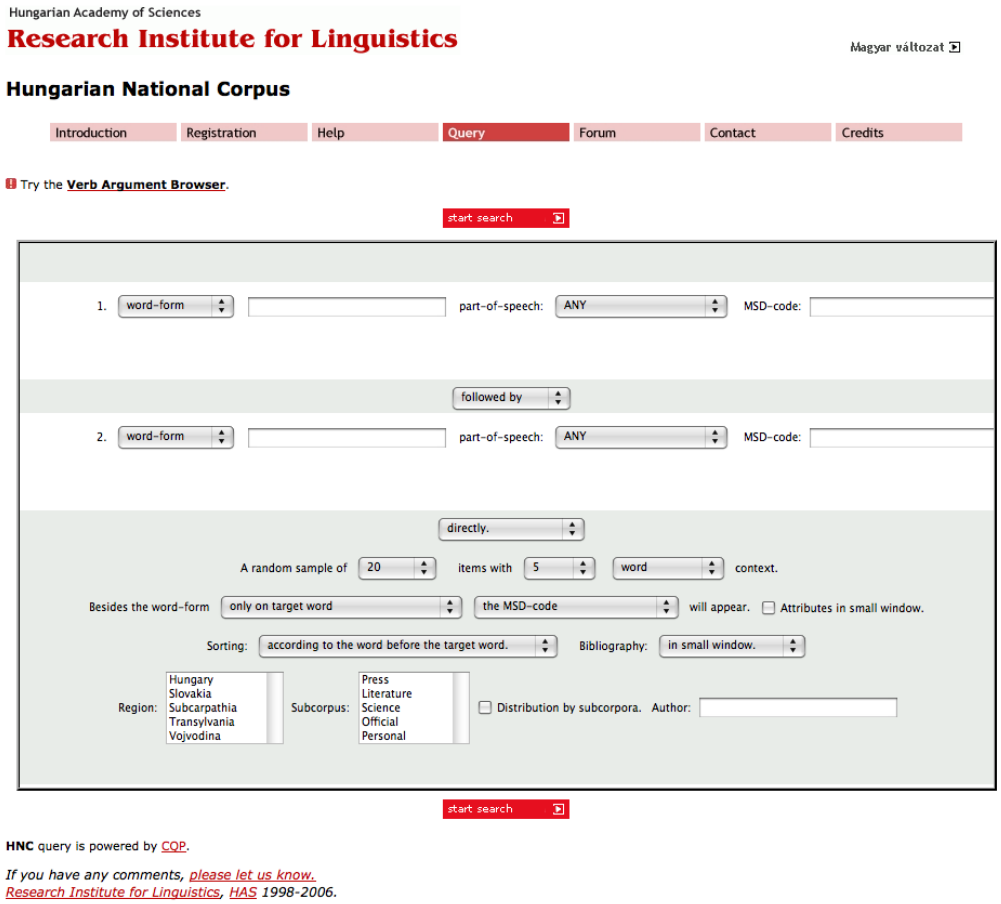
	HU	SK	RO	UK	SR	TOTAL
Tlač *Press*	71,0	5,7	0,7	5,5	1,5	**84,5**
Beletria *Literature*	35,5	1,4	0,4	0,8	0,2	**38,2**
Vedecký *Science*	20,5	2,3	0,7	1,6	0,3	**25,5**
Úradný *Official*	19,9	0,2	0,3	0,6	0,1	**20,9**
Hovorený/ osobný *Personal*	17,8	—	0,4	0,4	0,1	**18,6**
SPOLU TOTAL	**164,7**	**9,5**	**2,5**	**8,9**	**2,0**	**187,6**

KORPUS – ŠALAMÚN
Corpus - Solomon

- Za 10 rokov existencie bolo použitý ako pomôcka pri posúdení jazykových javov/variantov.
  - je akýmsi sudcom (i etalónom) pre jazykovedcov
- Jedinečný v mnohých častiach medzi korpusmi maďarského jazyka:
  - rôzne žánre a registre
    - podľa textového typu
    - podľa geografického položenia
  - je vyvážený v mnohých častiach
- je ale starý

- Within the 10 years of its existance was used for help to decide the language varieties in use.
  - it plays a role of judge for linguists
- It is a unique corpus for Hungarian in a range of its characters:
  - a variety of genres/registers
    - by text types
    - by geographical localization of the texts
  - it is well balanced in a couple of proportions
- but it conciders to be now old

GUI


Hungarian Academy of Sciences
Research Institute for Linguistics
Magyar változat
Hungarian National Corpus
Introduction
Registration
Help
Query
Forum
Contact
Credits
Try the Verb Argument Browser.
start search
1.
word-form
part-of-speech:
ANY
MSD-code:
followed by
2.
word-form
part-of-speech:
ANY
MSD-code:
directly.
A random sample of
20
items with
5
word
context.
Besides the word-form
only on target word
the MSD-code
will appear.
Attributes in small window.
Sorting:
according to the word before the target word.
Bibliography:
in small window.
Region:
Hungary
Slovakia
Subcarpathia
Transylvania
Vojvodina
Subcorpus:
Press
Literature
Science
Official
Personal
Distribution by subcorpora.
Author:
start search
HNC query is powered by CQP.
If you have any comments, please let us know.
Research Institute for Linguistics, HAS 1998-2006.

MNK V POUŽÍVANÍ
HNC IN USE

- Zaregistrovaní používatelia:
  - 7 387
- Vyhľadávania za posledné roky
  - 2011: 33339
  - 2010: 59 727
  - 2009: 30 000
  - 2008: 10 414

- Registered users:
  - 7 387
- Queries for the last 4 years
  - 2011: 33339
  - 2010: 59 727
  - 2009: 30 000
  - 2008: 10 414

MAĎARSKÝ NÁRODNÝ KORPUS II.
HNC II.

- Korpus maďarského jazyka minimálne na úrovni jeho predku.
- Stará báza v „novom hábe“
  - niektoré vlastnosti dedí od predku
  - niektoré vlastnosti sú nové – teraz sme už schopní urobiť
- Cieľ: reprezentovať maďarský jazyk (za a mimo hranicami) v rôznych doménach písaného a hovoreného jazyka
  - vysnívaný cieľ: korpus 1000 millión slov s rôznymi typmi annotácií

- The new corpus should be at leas on the level of the previous one
- New vawes on old basis
  - Some chacarterisics will remain
  - Some are new – now we are able to do them
- Aim is to representthe Hungarian language inside and outside of the borders in certain domains of usage
  - Dreamed aim is to reach 1 0000 billion of tokens with a range of annotations

PRÍPRAVA
PREPARATION PHASE

- 5 fáz prípravy korpusu (ohniská problémov)
  1. Zber textov
  2. Analýza a anotácia
  3. Systém spravovania korpusu – korpusový manažér (engine)
  4. GUI
  5. Prostriedky odvodené u korpusu

- 5 points of preparational phase
  1. Text collecion
  2. Alnalyse and annotation
  3. Engine
  4. GUI
  5. Other/minor resources derived from the HNC2

PROBLÉMY A RIEŠENIA
PROBLEMS AND THEIR SOLVING

- Problémy sa vyskytnú na každej rovine
- Problémy nie sú len technického rázu
  - napr. pri zbere materiálov – aby čo najviac textov bolo legálnych
  - napr. pri zbere materiálov – aby sme texty dostali i s metadátami
  - napr. štatút textov v korpuse – keďže ponúkame 3 možnosti prístupu ku korpusu
    - pseudokorpus – korpus na prihľadávanie
    - korpus – na sťahovanie 1 (texty miešané)
    - korpus – na sťahovanie 2 (texty v origináli)

- Problems occure on each level
- Problems are not just of technical character
  - eg. In case of collection – we should focus on cleared IPR issues
  - eg. we should be provided with text acconpanished with metadata
  - this has an inpact on the follwong – we are offering 3 types of corpus access
    - pseudocorpus for queries
    - ídownloadable corpus 1 (randomised texts)
    - downloadable corpus 2 (NOT randomised texts)

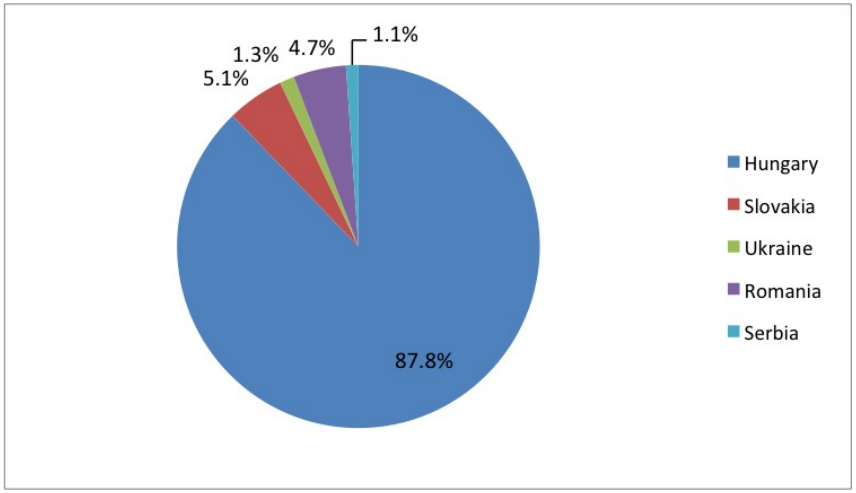
ZBER TEXTOV

1. žánrová pestrosť
2. zbierka materiálov
3. právna stránka - IPR
4. prepracovanie zozbieraných materiálov
   - kódovanie znakov
   - filtrovanie textov
   - filtrovanie maďarských textov

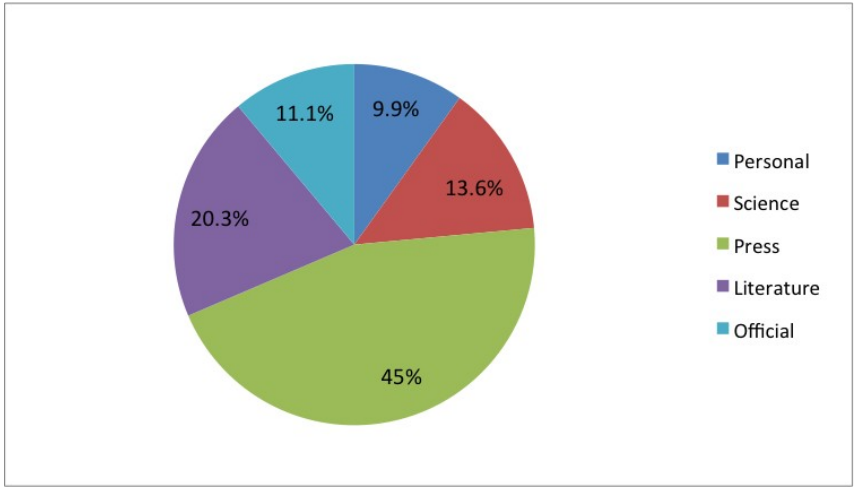
- Parts of the issue:
  1. full of genres (the more is the better)
  2. Collection
  3. IPR issues
  4. Preprocessing of the collected materials
     1. character encoding
     2. text filtring (only texts)
     3. text filtring (only HUNGARIAN texts)

ŽÁNROVÁ PESTROSŤ

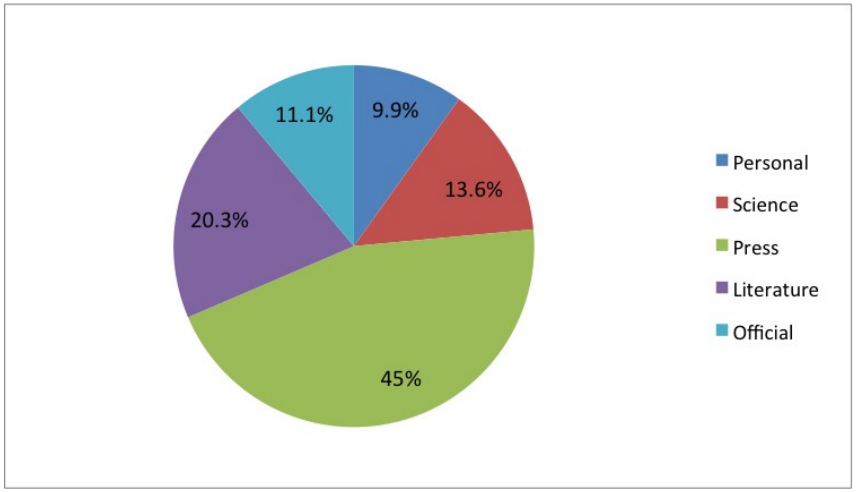
MNK 1 – HNC 1



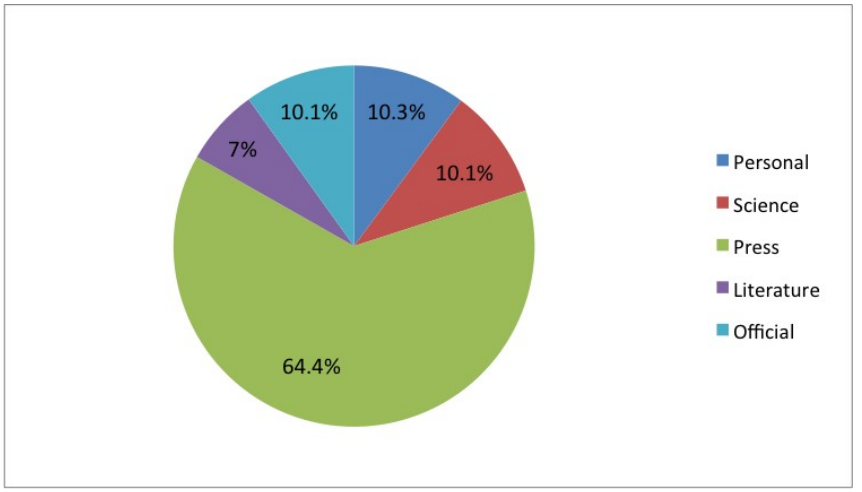
MNK 1 – HNC 1



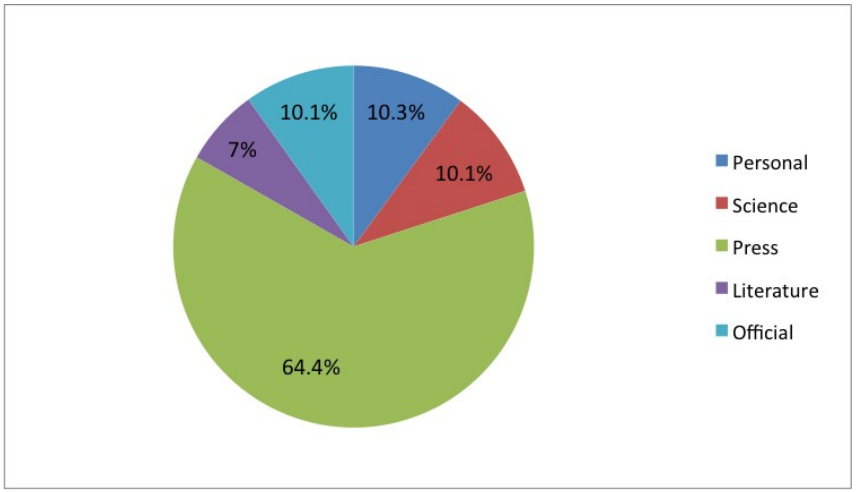
MNK 1 – HNC 1



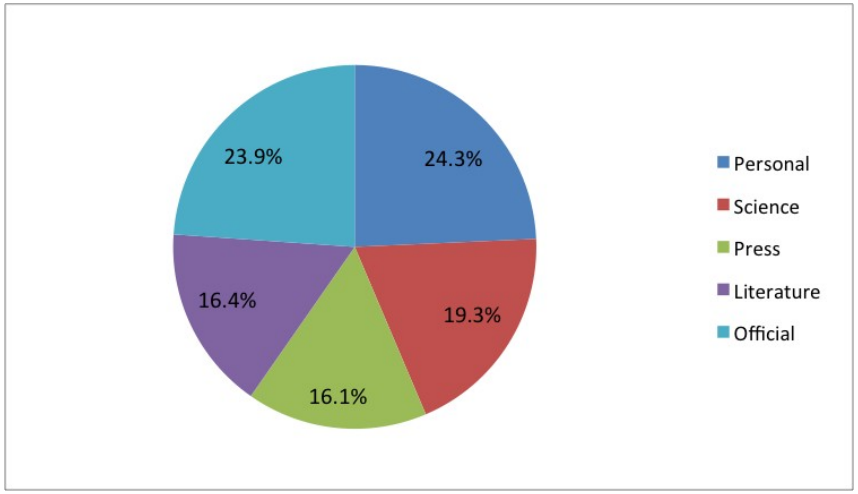
MNK 2 – HNC 1



publicistika dominuje
press in prevail



publicistika nedominuje
press not in prevail



ZBER TEXTOV - ČO
COLLECTING TEXTS - WHAT

- okolo 1000 miliónov textových jednotiek
  1. tlač
     - čo najširšie reprezentovať škálu časopisov a novín (tematicky, politicky atď.)
  1. vedecký
     - čo najpestrejšiu škálu vedeckých textov
     - beletriadigitalizované texty súčasných maďarských autorov
  1. odborný
     - odborbné texty z Maďarskej elektronickej knižnice
  1. hovorený/osobný
     - najmä blogy, prepisy televýznych
     - a rozhlasových relácií

- cca. 1 G text units
- press
- science
- literature
- official
- personal

ZBER TEXTOV - AKO
COLLECTING TEXTS - HOW

- Snaha o zber textov „v textovej podobe”
  - vyhnúť sa digitlizácie, OCR
- Snaha o legálne texty – texty nesťahujeme, ale dostaneme od majiteľov
  - nielen texty, ale i štruktúru a metadáta
  - sťahujeme texty iba ak nie je iná možnosť
- Licenčné práva sa snažíme brať legálne
- Texty nielen z Maďarska a z okolitých regiónov
  - zámorie a z čo najširšieho okruhu
- Len maďarský jazyk (preklad i originál)

- Collection of texts in text format
  - no pdf, no OCR
- Intentions: texts of cleared IPR – no downloading, to get from IPR holders
  - not just texts, but metadata too
  - download only if it is neccessary
- IPR issues are to be taken serious
- Texts not only from Hungary, but from countries where Hungarian is used
- Only Hungarian – no other language

ANALÝZA A ANOTÁCIA
ANALISYS AND ANNOTATION

- MNK 1/2 – inline anotácia, XML-štruktúra
  - structure-type: odsek, veta
  - token-type: lemma, postag
- Reprezentácia jednotlivých anotačných vrstiev
  - segmentácia (<s>, <w>, <c>)
  - morfologická anotácia a unifikácia (<w lemma="..." msd="...">...</w>)
  - vnútorná segmentácia viet (IOB atribúty)
  - NP (IOB)
  - NER (IOB)
  - verbálne prefixy

- Structure: inline annotation, XML structure
  - Structure-type: paragraph, sentence
  - token-type: lemma, postag
- Representatrion of annotation layers:
  - segmentation (<s>, <w>, <c>)
  - morphologic annotation and unification (<w lemma="..." msd="...">...</w>)
  - inner segmentation of sentences (IOB atributes)
  - NP (IOB)
  - NER (IOB)
  - verb prefixes

ANALÝZA A ANOTÁCIA – PRO & CONTRA
ANALISYS AND ANNOTATION

- jednoduchý
- rýchly
- je možné rozšíriť o nové vrstvy (čo potrebujeme pri MNK2)
- naraz inline a standoff

- niektoré textové typy sú ťažko kódovateľné
- problémy môžu byť pri viacslovných jednotkách

- Easy
- Fast
- It is possible to extend with new layers (of annotations) – what we need
- Inline and standoff in a moment

- BUT
- Some types of texts are hard to encode
- Problems with multiword units

ANALÝZA A ANOTÁCIA - ČO EŠTE
ANALISYS AND ANNOTATION – WHAT ELSE

- automatické pridelenie kľúčových slov k textom
- automatické kategorizovanie textov podľa tém

- nové možnosti pri vyhľadávaní

- authomatic key-word annotation
- authomatic cathegorisation by themes

- new possibilities in queries

KORPUSOVÝ MANAŽÉR
CORPUS ENGINE

- Otázky ešte otvorené:
  - NoSketch Engine: Manatee/Bonito
  - IMS Open Corpus Workbench (CWB) – aj s CQP
  - Emdros – možno…
  - Pravdepodobne NoSketch Engine

- Questions still open
  - NoSketch Engine: Manatee/Bonito
  - IMS Open Corpus Workbench (CWB) – aj s CQP
  - Emdros – maybe…
  - Probably NoSketch Engine

GUI

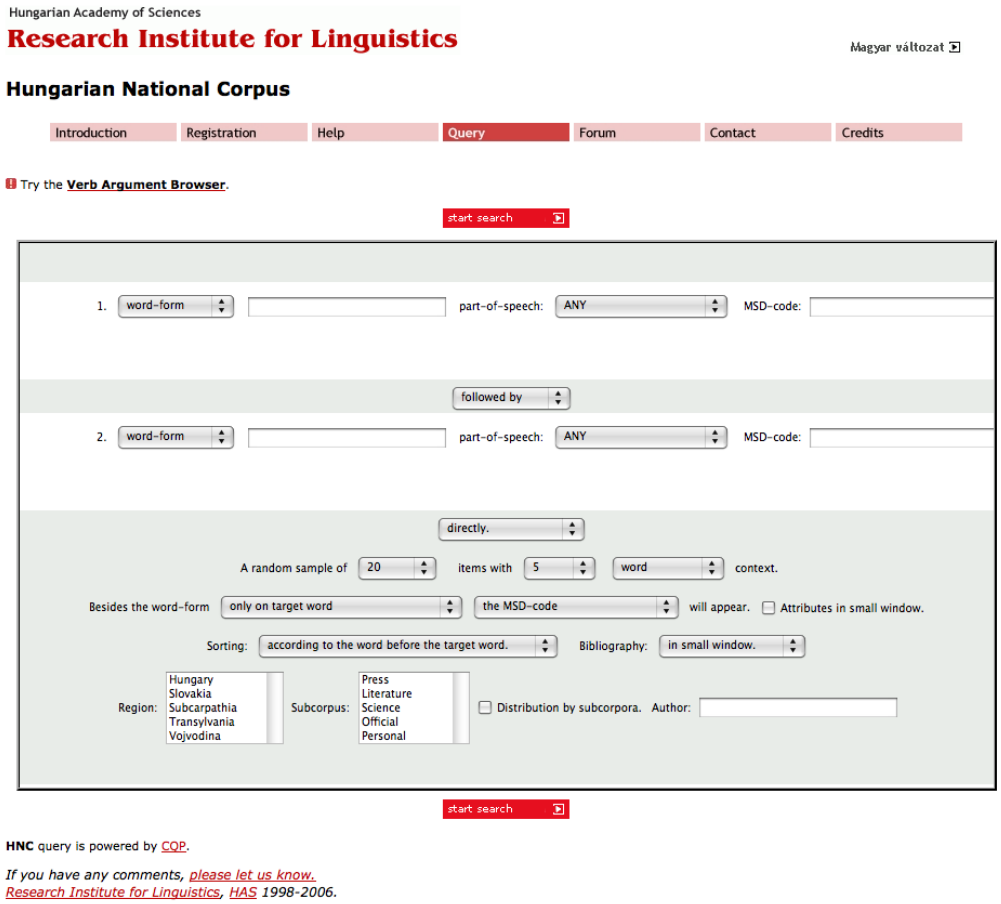
- Interaktívna plocha
- sofistikované vyhľadávanie
- viac informácií a možností než je dnes
- interaktívna ponuka odvodených prostriedkov
- možnosť generovania rôznych štatistík

- možno XAIRA (XML Aware Indexing and Retrieval Architecture)

- Interactive GUI
- Sophisticated queries
- More information on texts than is today at HNC1
- Interactive offer of other/minor resources derived from the HNC2
- Range of statistics offered

- maybe XAIRA (XML Aware Indexing and Retrieval Architecture)

GUI


Hungarian Academy of Sciences
Research Institute for Linguistics
Magyar változat
Hungarian National Corpus
Introduction
Registration
Help
Query
Forum
Contact
Credits
Try the Verb Argument Browser.
start search
1.
word-form
part-of-speech:
ANY
MSD-code:
followed by
2.
word-form
part-of-speech:
ANY
MSD-code:
directly.
A random sample of
20
items with
5
word
context.
Besides the word-form
only on target word
the MSD-code
will appear.
Attributes in small window.
Sorting:
according to the word before the target word.
Bibliography:
in small window.
Region:
Hungary
Slovakia
Subcarpathia
Transylvania
Vojvodina
Subcorpus:
Press
Literature
Science
Official
Personal
Distribution by subcorpora.
Author:
start search
HNC query is powered by CQP.
If you have any comments, please let us know.
Research Institute for Linguistics, HAS 1998-2006.

PROSTRIEDKY ODVODENÉ U KORPUSU
DERIVATES

- Viac možností:
  - frekvenčný slovník
  - časť korpusu na stiahnutie
    - s miešanými vetami
    - s vetami ako v originálach
  - API
  - n-gramy

- To be offered:
  - Word frequency list
  - A part of the corpus to be download
    - With sentencies in randomised order
    - With sentencies in original order
  - API
  - N-grams

ĎAKUJEM ZA POZORNOSŤ

Thank you for your attention!!!

MAGYAR TUDOMÁNYOS AKADÉMIA
NYELVTUDOMÁNYI INTÉZET