The Hungarian Gigaword Corpus

The Hungarian Language in the Digital Age

Budapest, Jan 18, 2013

Csaba Oravecz
oravecz.csaba@nytud.mta.hu
Research Institute for Linguistics
Hungarian Academy of Sciences

Overview

- Origins
- Motivation
- Objectives
- Preparation
- Outcome

Overview

- Origins
- Motivation
- Objectives
- Preparation
- Outcome

Overview

- Origins
- Motivation
- Objectives
- Preparation
- Outcome

Overview

- Origins
- Motivation
- Objectives
- Preparation
- Outcome

Overview

- Origins
- Motivation
- Objectives
- Preparation
- Outcome

Origins

The Hungarian National Corpus (HNC)

- developed between 1998 and 2001
- representative sample of the language use of the second half of the 90s → empirical evidence for status of language and data for theoretical analysis and language technology
- first major annotated Hungarian corpus, available freely through a search interface
- 187 million words, covering language variants from beyond the border of Hungary → Hungarian Minority Language Corpus
- more than 7000 registered users, dozens of research papers based on HNC data

Origins

The Hungarian National Corpus (HNC)

- developed between 1998 and 2001
- representative sample of the language use of the second half of the 90s → empirical evidence for status of language and data for theoretical analysis and language technology
- first major annotated Hungarian corpus, available freely through a search interface
- 187 million words, covering language variants from beyond the border of Hungary → Hungarian Minority Language Corpus
- more than 7000 registered users, dozens of research papers based on HNC data

Origins

The Hungarian National Corpus (HNC)

- developed between 1998 and 2001
- representative sample of the language use of the second half of the 90s $\rightarrow$ empirical evidence for status of language and data for theoretical analysis and language technology
- first major annotated Hungarian corpus, available freely through a search interface
- 187 million words, covering language variants from beyond the border of Hungary $\rightarrow$ Hungarian Minority Language Corpus
- more than 7000 registered users, dozens of research papers based on HNC data

Origins

The Hungarian National Corpus (HNC)

- developed between 1998 and 2001
- representative sample of the language use of the second half of the 90s $\rightarrow$ empirical evidence for status of language and data for theoretical analysis and language technology
- first major annotated Hungarian corpus, available freely through a search interface
- 187 million words, covering language variants from beyond the border of Hungary $\rightarrow$ Hungarian Minority Language Corpus
- more than 7000 registered users, dozens of research papers based on HNC data

Origins

The Hungarian National Corpus (HNC)

- developed between 1998 and 2001
- representative sample of the language use of the second half of the 90s → empirical evidence for status of language and data for theoretical analysis and language technology
- first major annotated Hungarian corpus, available freely through a search interface
- 187 million words, covering language variants from beyond the border of Hungary → Hungarian Minority Language Corpus
- more than 7000 registered users, dozens of research papers based on HNC data

Origins

The Hungarian National Corpus (HNC)

- developed between 1998 and 2001
- representative sample of the language use of the second half of the 90s $\rightarrow$ empirical evidence for status of language and data for theoretical analysis and language technology
- first major annotated Hungarian corpus, available freely through a search interface
- 187 million words, covering language variants from beyond the border of Hungary $\rightarrow$ Hungarian Minority Language Corpus
- more than 7000 registered users, dozens of research papers based on HNC data

Motivation

15 years after ...

- requirements against language resources have changed significantly
  - dominance of data oriented methods and applications in NLP
    → the more data the better results
  - better language processing tools
    → higher quality and finer level of analysis and annotation
  - preservation of representativity
    → subsequent sampling from language use needed
- HNC has become outdated

Motivation

15 years after ...

- requirements against language resources have changed significantly
  - dominance of data oriented methods and applications in NLP
    $\longrightarrow$ the more data the better results
  - better language processing tools
    $\longrightarrow$ higher quality and finer level of analysis and annotation
  - preservation of representativity
    $\longrightarrow$ subsequent sampling from language use needed
- HNC has become outdated

Motivation

15 years after ...

- requirements against language resources have changed significantly
  - dominance of data oriented methods and applications in NLP
    $\longrightarrow$ the more data the better results
  - better language processing tools
    $\longrightarrow$ higher quality and finer level of analysis and annotation
  - preservation of representativity
    $\longrightarrow$ subsequent sampling from language use needed
- HNC has become outdated

Motivation

15 years after ...

- requirements against language resources have changed significantly
  - dominance of data oriented methods and applications in NLP
    $\longrightarrow$ the more data the better results
  - better language processing tools
    $\longrightarrow$ higher quality and finer level of analysis and annotation
  - preservation of representativity
    $\longrightarrow$ subsequent sampling from language use needed
- HNC has become outdated

Motivation

15 years after ...

- requirements against language resources have changed significantly
  - dominance of data oriented methods and applications in NLP
    $\longrightarrow$ the more data the better results
  - better language processing tools
    $\longrightarrow$ higher quality and finer level of analysis and annotation
  - preservation of representativity
    $\longrightarrow$ subsequent sampling from language use needed
- HNC has become outdated

Motivation

15 years after ...

- requirements against language resources have changed significantly
  - dominance of data oriented methods and applications in NLP
    $\longrightarrow$ the more data the better results
  - better language processing tools
    $\longrightarrow$ higher quality and finer level of analysis and annotation
  - preservation of representativity
    $\longrightarrow$ subsequent sampling from language use needed
- HNC has become outdated

Objectives

Increase ...

- *quality.* Use new technology for development and analysis.
- *size.* Extend the corpus to 1 Gw.
- *coverage and representativity.* Take new samples of languge use and include further variants (transcribed spoken language data in particular).

HGC: Develop an up-to-date language resource that will service the research community as well as the interested public.

Objectives

Increase ...

- *quality.* Use new technology for development and analysis.
- *size.* Extend the corpus to 1 Gw.
- *coverage and representativity.* Take new samples of languge use and include further variants (transcribed spoken language data in particular).

HGC: Develop an up-to-date language resource that will service the research community as well as the interested public.

Objectives

Increase ...

- *quality.* Use new technology for development and analysis.
- *size.* Extend the corpus to 1 Gw.
- *coverage and representativity.* Take new samples of languge use and include further variants (transcribed spoken language data in particular).

HGC: Develop an up-to-date language resource that will service the research community as well as the interested public.

Objectives

Increase ...

- *quality.* Use new technology for development and analysis.
- *size.* Extend the corpus to 1 Gw.
- *coverage and representativity.* Take new samples of languge use and include further variants (transcribed spoken language data in particular).

HGC: Develop an up-to-date language resource that will service the research community as well as the interested public.

Objectives

Increase ...

- *quality.* Use new technology for development and analysis.
- *size.* Extend the corpus to 1 Gw.
- *coverage and representativity.* Take new samples of languge use and include further variants (transcribed spoken language data in particular).

HGC: Develop an up-to-date language resource that will service the research community as well as the interested public.

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing → no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing → no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing → no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing $\rightarrow$ no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing $\rightarrow$ no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing $\rightarrow$ no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Text collection

- clean IPR issues
- extensive metadata (simple webcrawling not sufficient)
- ease of processing $\rightarrow$ no pdf, no OCR

Preprocessing, normalization

- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

Preparation

Analysis and annotation

- detailed morphosyntactic analysis and disambiguation with updated processing toolchain
  (information on stem, each morph and compounding)
- NP chunking, Named Entity recognition
- XML annotation compatible with international standards

Corpus query engine

- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

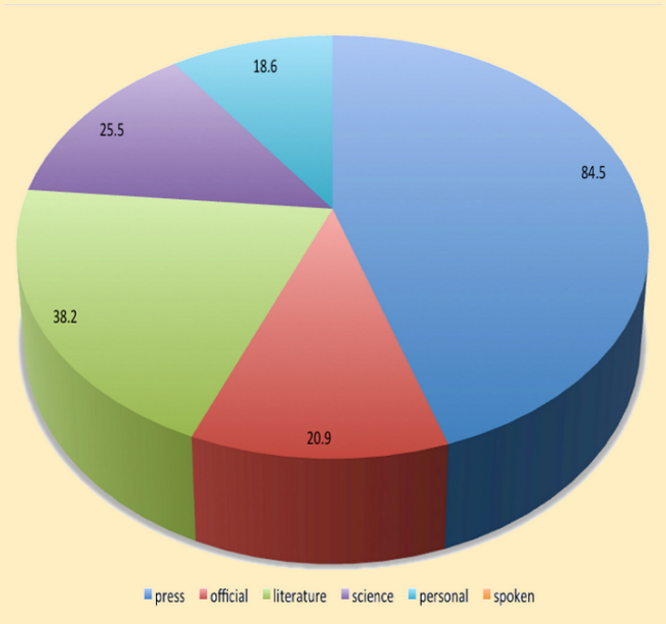
Outcome

HNC: 187 m.

HGC (+HNC): 1091 m.

Outcome

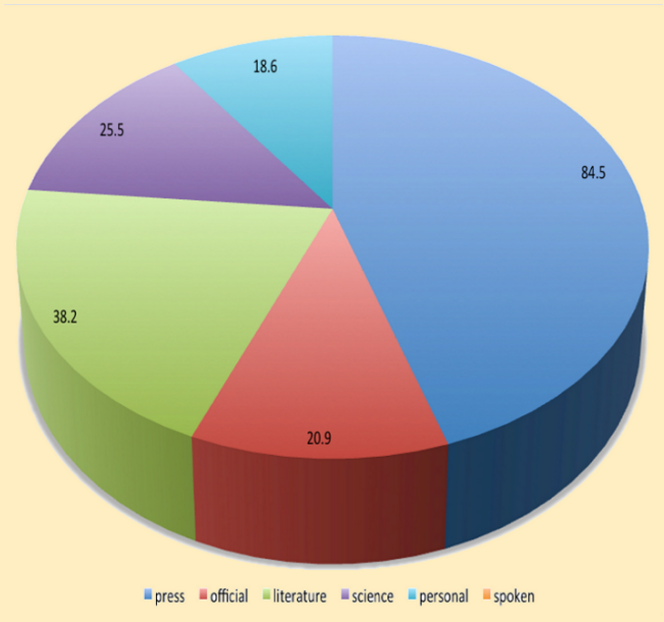
HNC: 187 m.



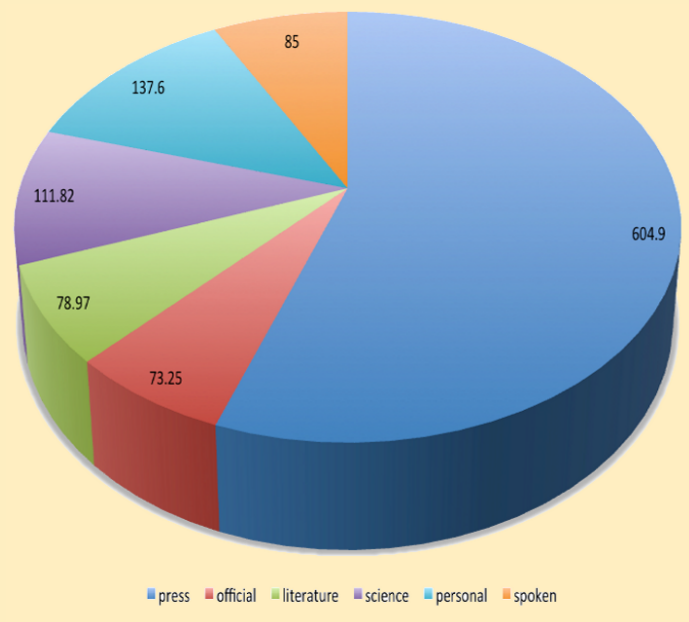
HGC (+HNC): 1091 m.

Outcome

HNC: 187 m.



HGC (+HNC): 1091 m.



Outcome

"Intelligent" corpus

- complex searches based on every piece of information in the annotation
  - morpho(phono)logical phenomena
  - multiword expressions: collocations, verbal arguments
- display settings: context, metadata
- distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)

Outcome

"Intelligent" corpus

- complex searches based on every piece of information in the annotation
  - morpho(phono)logical phenomena
  - multiword expressions: collocations, verbal arguments
- display settings: context, metadata
- distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)

Outcome

"Intelligent" corpus

- complex searches based on every piece of information in the annotation
  - morpho(phono)logical phenomena
  - multiword expressions: collocations, verbal arguments
- display settings: context, metadata
- distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)

Outcome

"Intelligent" corpus

- complex searches based on every piece of information in the annotation
  - morpho(phono)logical phenomena
  - multiword expressions: collocations, verbal arguments
- display settings: context, metadata
- distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)

Outcome

"Intelligent" corpus

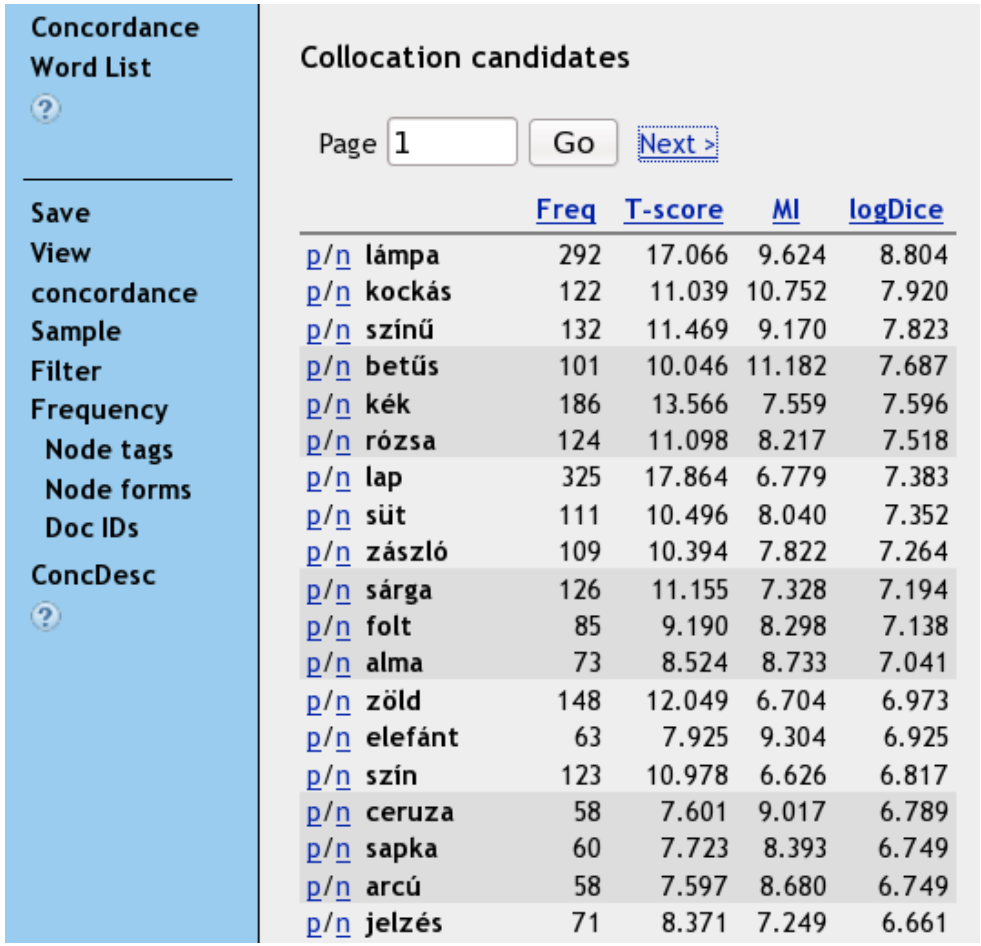
- complex searches based on every piece of information in the annotation
  - morpho(phono)logical phenomena
  - multiword expressions: collocations, verbal arguments
- display settings: context, metadata
- distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)

Outcome

"Intelligent" corpus

- complex searches based on every piece of information in the annotation
  - morpho(phono)logical phenomena
  - multiword expressions: collocations, verbal arguments
- display settings: context, metadata
- distributional analysis, built in post-processing (multilevel frequency lists, subsequent searches on previous results)

"piros ..."


Concordance
Word List
Save
View
concordance
Sample
Filter
Frequency
Node tags
Node forms
Doc IDs
ConcDesc
Collocation candidates
Page 1
Go
Next ›
Freq
T-score
MI
logDice
p/n lámpa 292 17.066 9.624 8.804
p/n kockás 122 11.039 10.752 7.920
p/n színű 132 11.469 9.170 7.823
p/n betűs 101 10.046 11.182 7.687
p/n kék 186 13.566 7.559 7.596
p/n rózsa 124 11.098 8.217 7.518
p/n lap 325 17.864 6.779 7.383
p/n süt 111 10.496 8.040 7.352
p/n zászló 109 10.394 7.822 7.264
p/n sárga 126 11.155 7.328 7.194
p/n folt 85 9.190 8.298 7.138
p/n alma 73 8.524 8.733 7.041
p/n zöld 148 12.049 6.704 6.973
p/n elefánt 63 7.925 9.304 6.925
p/n szín 123 10.978 6.626 6.817
p/n ceruza 58 7.601 9.017 6.789
p/n sapka 60 7.723 8.393 6.749
p/n arcú 58 7.597 8.680 6.749
p/n jelzés 71 8.371 7.249 6.661

"piros lámpa"

történt volna . Így eshetett meg , hogy	**piros**	***lámpánál*** az út közepén futottam , s éppen
autótolvajok új stratégiát követnek . Így például a	**piros**	***lámpánál*** hátulról belehajtanak a kiszemelt
Módszerük az , hogy az autópályáról levezető	**piros**	***lámpánál*** megálló gépkocsi mellé berobogó
Felelőtlenség a gyerekeknek rossz példát mutatni ,	**piros**	***lámpánál*** átszaladni , a szabályokat nem
Mindennapi tapasztalatot elevenít fel a	**piros**	***lámpánál*** megrekedt autósról szóló jelenet
össze a motoros rendőrökkel . Ha még a	**piros**	***lámpán*** is áthajtott , és a kerékpár felszereltsége
vezető minden alkalommal megjelent , hatalmas	**piros**	elemes ***lámpát*** tartott a kezében , és elmondta
arcom . De az biztos , ha a stúdióban ég a	**piros**	***lámpa*** , akkor mindenki egy kicsit összekapja
és kíséretét szállító konvoj minden egyes	**piros**	***lámpánál*** megállt Lahorban - a szemtanúk
gépkocsioszlop szabályosan megállt minden egyes	**piros**	***lámpánál*** , nem kis megdöbbenést okozva
beavatkozásokat eszközölni , ők csupán várakoznak a	**piros**	***lámpánál*** és adományokat gyűjtenek az autósoktól
" technika . Az előző módszernél a	**piros**	***lámpánál*** álló vezetőt kikényszerítik autójából
például annak , hogy amikor Pesten járt , és a	**piros**	***lámpánál*** egy autóbusz utolérte , a sofőr
kikapcsolása . (A három futó állapotát egy sor	**piros**	***lámpa*** is jelzi . Csakhogy Szaloniki felett

Web interface

http://mnsz.nytud.hu

The end

Thank you for your attention