The new Gigaword Version of the Hungarian National Corpus

Csaba Oravecz

oravecz@nytud.hu

Language Technology Research Group,
Research Institute for Linguistics,
Hungarian Academy of Sciences

21 November 2014

Overview

- Origins
- Motivation
- Objectives
- Design considerations
- 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 $\rightarrow$ 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
    $\rightarrow$ the more data the better results
  - better language processing tools
    $\rightarrow$ higher quality and finer level of analysis and annotation
  - preservation of representativity
    $\rightarrow$ 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- HNC has become outdated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.

Design considerations

- Representativity?
  Illusionary goal on this scale $\rightarrow$ balancedness
- Collection of data by web crawling or acquiring large amounts of newswire text?
  Difficult to produce a solid, balanced resource with sufficient metadata
  $\rightarrow$ controlled, targeted resource collection, appropriate for each type of source. Easily processable source format is preferred
  $\rightarrow$ no pdf, no OCR.

Design considerations

- Clean IPR issues?
  Difficult, sometimes even impossible, to collect appropriate licenses
  $\rightarrow$ different availability options are offered for various sections of the HGC.

Preparation

Text collection

- (try to) clean up IPR issues
- collect data with extensive metadata

Preprocessing, normalization

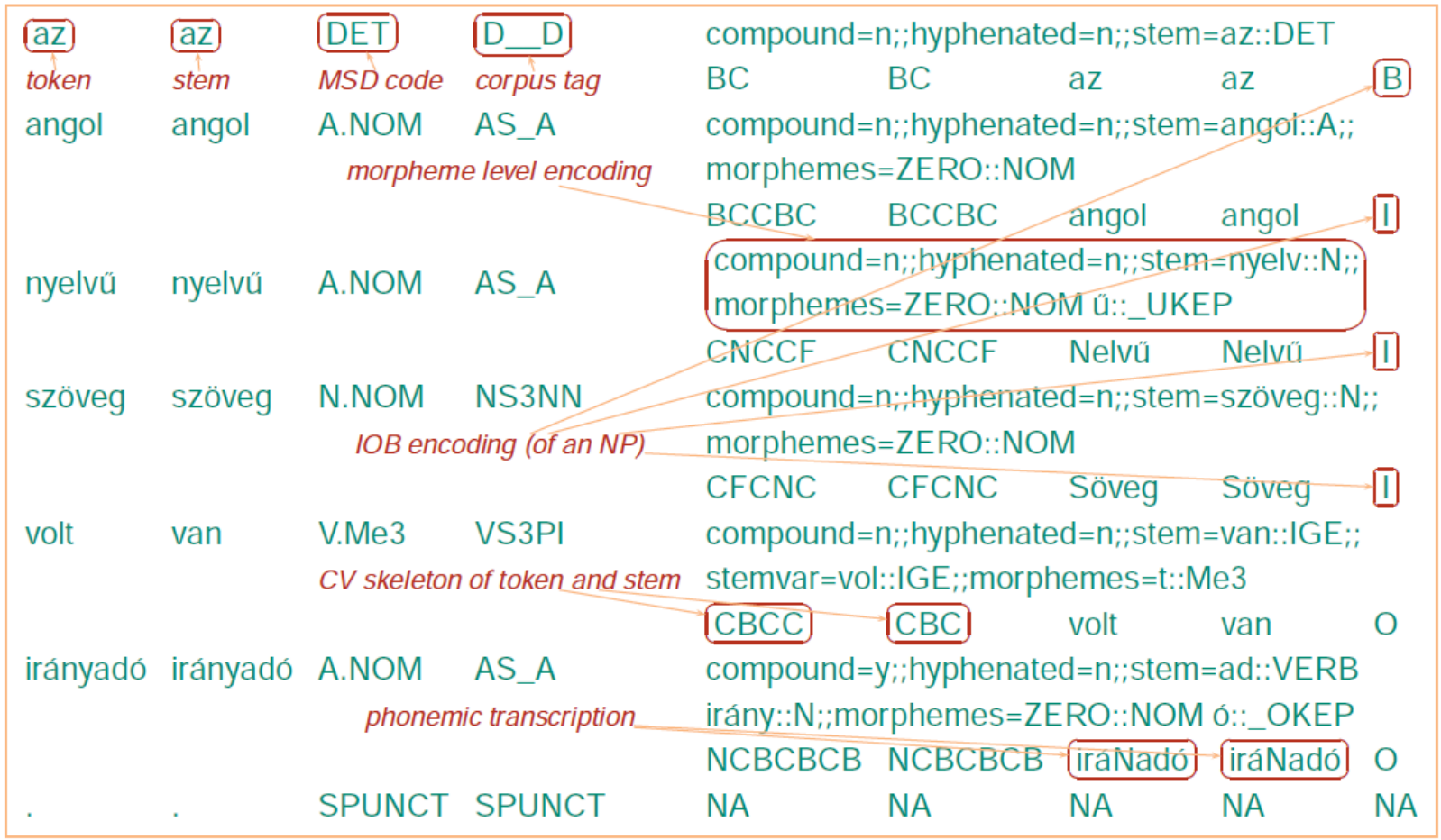
- identify textual content and basic document structure
- filter out (near-)duplicates and non-Hungarian sections

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 (in prep.)
- XML annotation compatible with international standards

Annotation format



Annotation problems

Multiple derivational paths

in possibilities:

```
lehetőség[N]+ek[PL]+ben[INE]
lehető[A]+ség[_PROP]+ek[PL]+ben[INE]
*lesz[V]=le+hető[_HATO]+ség[_PROP]+ek[PL]+ben[INE]
=>lehetőség[N][PL][INE]/NP3N2
```

→ select the analysis with the highest number of morphemes

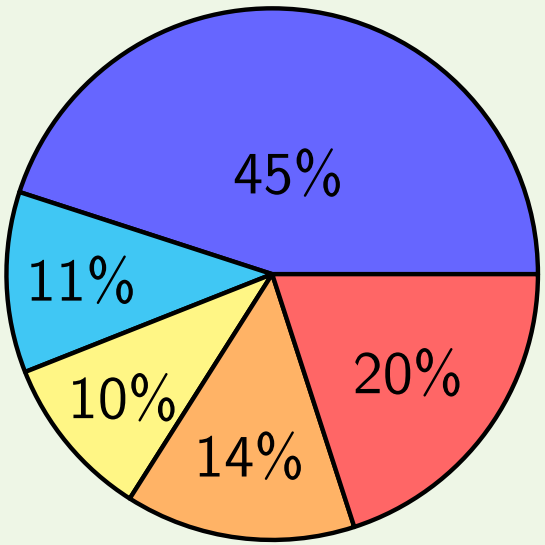
Compound overgeneration

lázadó (“rebel”) = láz (“fever”) + adó (“tax”)

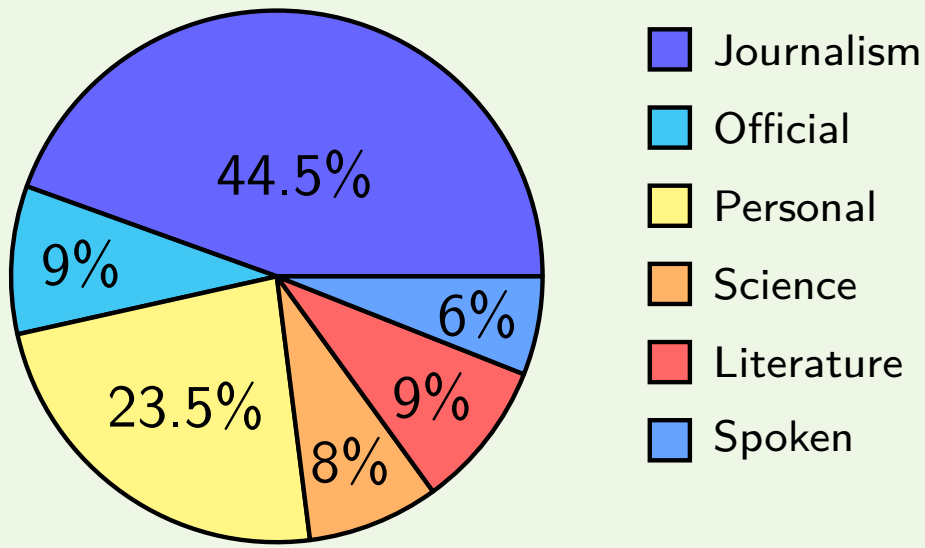
→ filter rules with regular expressions

Outcome

HNC: 187 m.



HGC (+HNC): 1437 m.



Outcome

Composition (m. words)

Register	HNC	HGC	Source
Journalism	84,5	639,2	Daily/weekly newspapers
Literature	38,2	130,7	Digital Literary Academy
(Popular) science	25,5	110,9	Hungarian Electronic Library
Personal	18,6	338,6	Social media
Official	20,9	135,4	Documents from public admin.
(Transcribed) spoken	–	83,0	Radio programs
	187,7	**1,437,8**	

Search interface

Corpus query engine

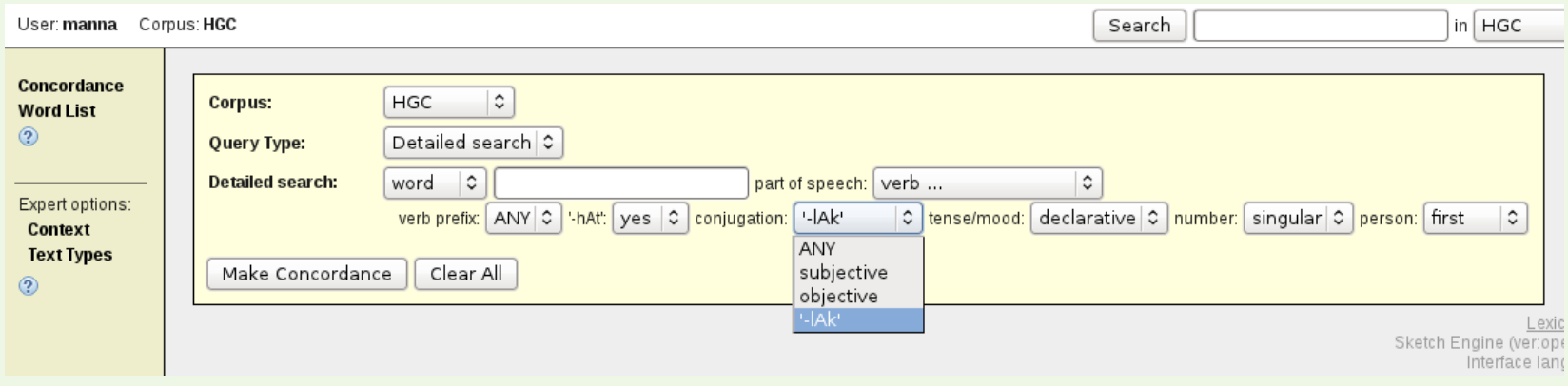
- robust, able to handle several gigawords
- quick response (depending on complexity of queries)

"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)

Queries

Searching for a specific verb form



Results

Concordance list for *-hAtlAk* verbs

Concordance
Word List

Save
View options
KWIC/Sentence
Sort
Left | Right
Node
References
Shuffle
Sample
Filter
Frequency
Node tags
Node forms
Doc IDs
Collocations
ConcDesc

Corpus: **HGC**
Hits: **753**

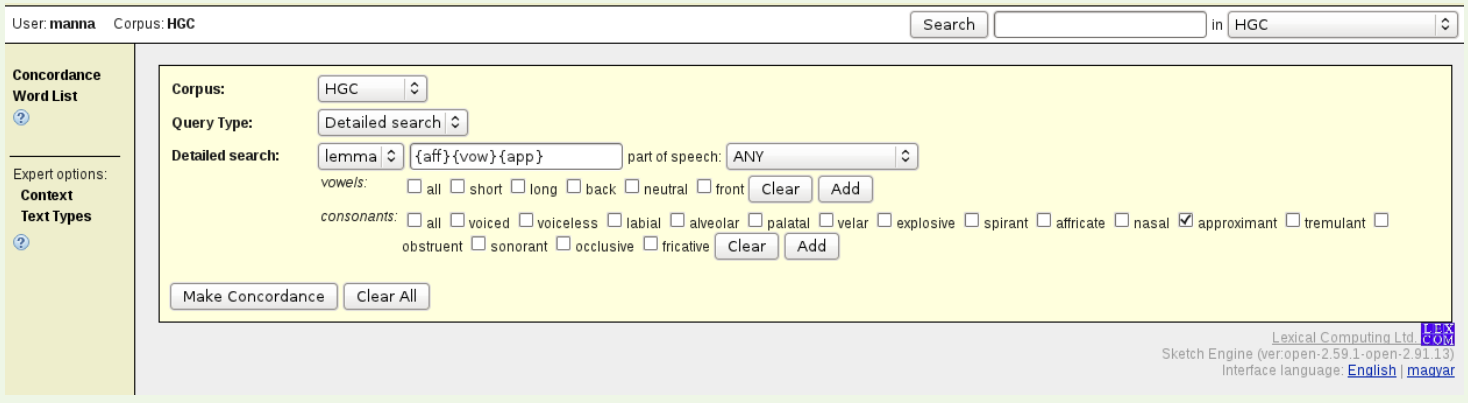
Page 1 of 38 Go Next | Last

doc#2	<!--btw_p_resz_begin:resz--> A. Milyen ízed van ?	Megcsókolhatlak	? Szeretsz ? Van -e öled ? Nőstényszagod
doc#2	szerelmet így tudta kifejezni : " nagyon	szerethetlek	, ha már a sálam is integet utánad " .
doc#4	éjszakára is . Olyan jó lett volna , ha úgy	hívhatlak	föl , hogy szabadon beszélhetsz velem .
doc#6	Gyere , gyere csak , foglalj helyet ! Mivel	kínálhatlak	meg ? Azért ne siess ! A királylány jól
doc#8	képre ? K. : Á ... Á . : S ha lekerülök ,	megkereshetlek	-e , mint az idősebb katonát , mutasd meg
doc#9	Jó is volna , oly hosszú idő után egyszer	kaphatlak	meg , és te máris menni akarsz . Hisz az
doc#15	talán , talán lehet még , hiszen hiszem -	lehetlek	, s én se vagyok még emlék , és élj-se
doc#18	talán , talán lehet még , hiszen hiszem -	lehetlek	, s én se vagyok még emlék , és élj-se
doc#18	vessen , nem táltosló , de táltosan szabad .	Elveszíthetlek	? Akár a száj a szót ? Szívdobogást a szív
doc#20	mosolygott . Túl udvariasan . Várt . -	Meghívhatlak	egy sörre ? - kérdezte Jocó . - Megyek
doc#20	- Meghívlak . Olyan a helyzet , hogy	meghívhatlak	. Megtaláltam az asszonyt ... Szerencsém
doc#25	ide . A vörös bor talán . Ebéd után ...	Megkínálhatlak	borral is ? - Ha nem édes .
doc#41	, kiről beszélsz . Na megvársz ? -	Megvárhatlak	- mondta Sivár . Nopritz Andrea bement
doc#45	befogadja . - Jó - mondta Andrej Bodor -,	biztosíthatlak	, mindezt pont így tudatni fogom . <head>
doc#46	, kiről beszélsz . Na megvársz ? -	Megvárhatlak	- mondta Sivár . Nopritz Andrea bement
doc#47	mindennapi szökdelést irigylitek ? Vagy hogy	fölfalhatlak	benneteket ? Véretek mocska ; szentkép
doc#47	után . Félálomban és dadogva hazámnak is	mondhatlak	téged . Visszatért fecskéid itt rajcsúroznak
doc#50	Baranyi Antal vagyok , ülj le , kérlek ,	megkínálhatlak	egy kis disznósajttal , a vágóhíd igazgatója
doc#50	</title> </head> Hagyjuk abba . Azt hittem ,	rászedhetlek	, hogy ellazítsd magadat legalább félórára
doc#50	alakja születik meg az ő verseiben , hogy :	lehetlek	. Azaz : én te lehetek majd , mert az szeretnék

Page 1 of 38 Go Next | Last

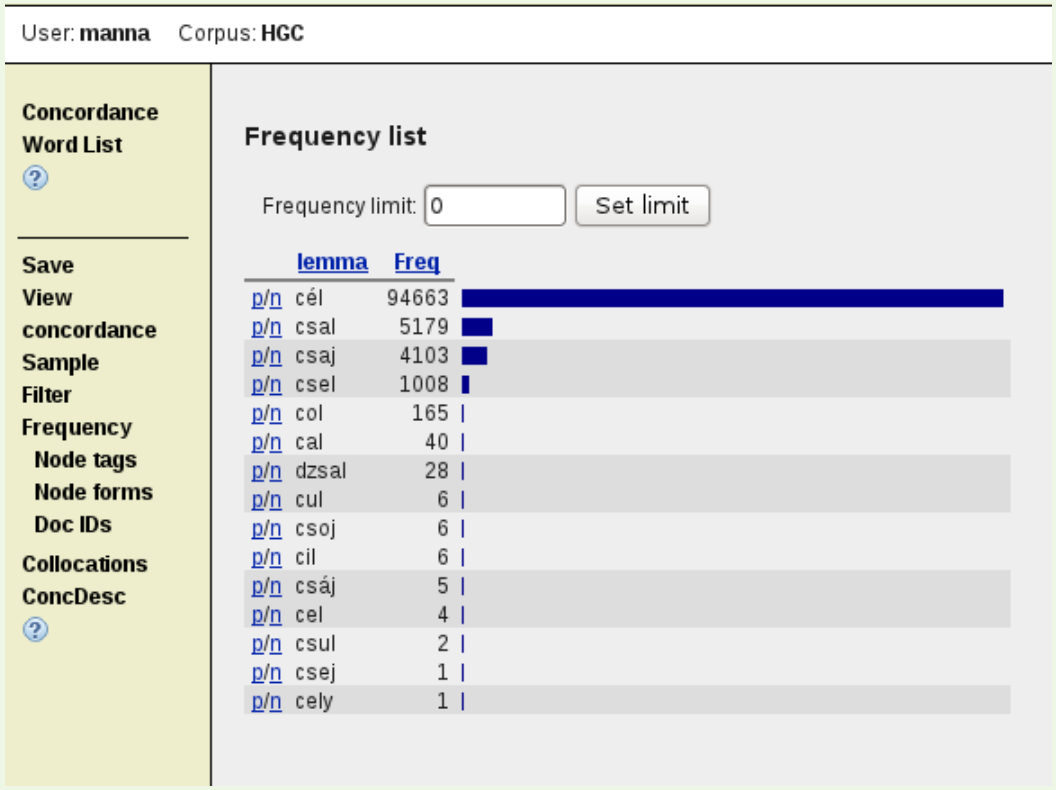
Queries

Searching for phonological phenomena

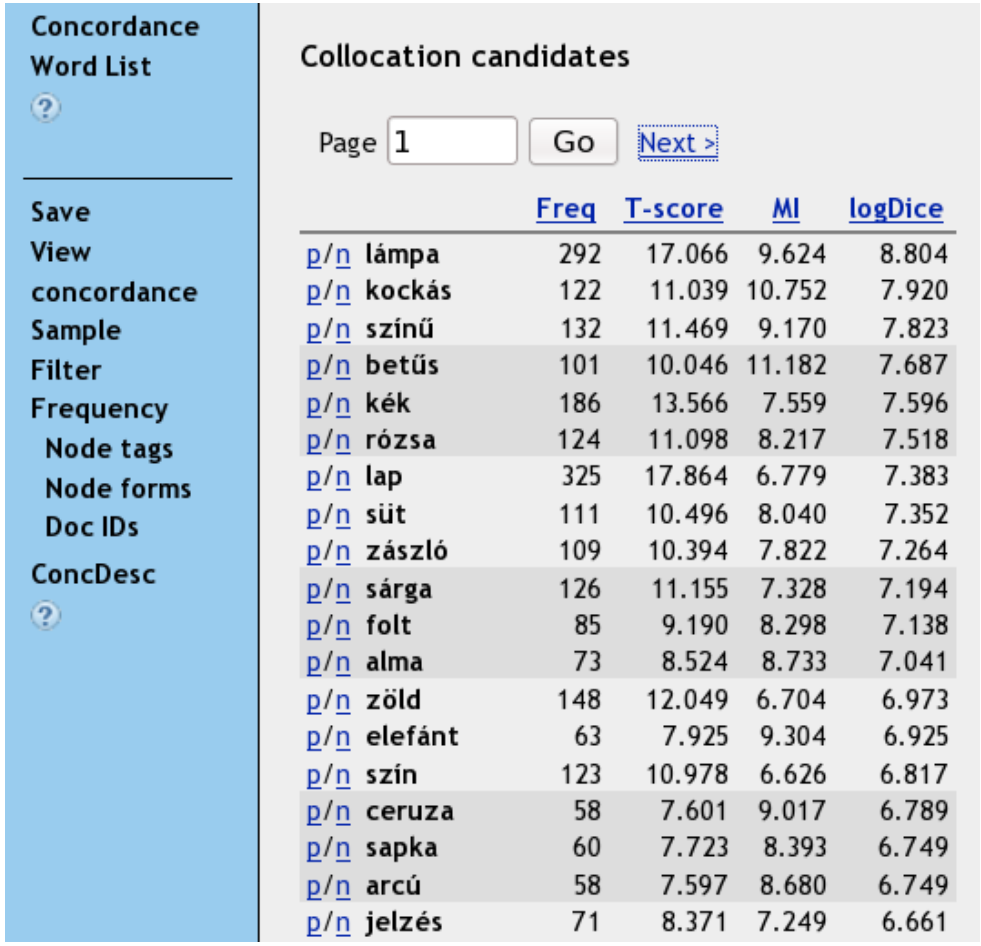


Results

Frequency list generated from search result



"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://hnc.nytud.hu`

The end

Thank you for your attention