UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
MTA Mesterséges Intelligencia kutatócsoport

# Információkinyerés magyar nyelvű szövegekből: megoldások és kihívások

Vincze Veronika

vinczev@inf.u-szeged.hu

# Bevezetés

- **Nyelv- és beszédtechnológia:**
  - **írott és a hangzó nyelv feldolgozása**
  - **nyelvi produktumok előállítása**
- **Célok:**
  - **az ember-ember, az ember-gép kommunikáció hatékonyabbá tétele**
  - **az emberi munkavégzés megkönnyítése újszerű, számítógépes termékek és szolgáltatások biztosításával**
  - **hátrányos helyzetű csoportok (siketek, gyengénlátók, baleset következtében beszédkészségüket elvesztők, idegen nyelveket nem tudók) életminőségének javítása**

# Határterületek

- **Nyelvészet**
- **Lexikográfia**
- **Szoftvertechnológia**
- **Pszichológia**
- **Matematika**
- **Informatika**
- **Fizika**
- **Fiziológia**
- **Neurológia**
- **Biológia**
- **…**

# Információfeldolgozás

- **Óriási mennyiségű szöveges adat**
  - **web**
  - **archívumok**
  - **adatbázisok**
- **Emberi feldolgozás lehetetlen vagy nagyon időigényes és költséges**
- **NLP-eszközök segíthetnek**

# Nyelvi feldolgozási lánc

- **Cél: szöveghalmazból hasznos információ kigyűjtése automatikus eszközökkel**
- **Bemenet: nyers szöveg**
- **Elemzési szintek:**
  - **szegmentálás**
  - **morfológia**
  - **szintaxis**
  - **szemantika**
- **Alkalmazások (pl.):**
  - **információkinyerés**
  - **véleménykinyerés**
- **Kimenet (pl.): adott dologhoz kapcsolt vélemények, események…**

# Meglévő eszközök

- **magyarlanc: szövegek szegmentálása + szófaji egyértelműsítése + függőségi elemzése**
- **Tulajdonnév-felismerés: tulajdonnevek felismerése és osztályba sorolása**
- **Sztenderd magyar szövegek feldolgozására lettek felkészítve (hírek, újságcikkek, irodalmi szövegek stb.)**
- **Ezeken nemzetközi szinten is jónak mondható eredményeket érünk el**
- **Hogyan teljesítenek nem sztenderd szövegeken?**

# Webes szövegek


- **Közösségi média**
- **Blogok**
- **Tweetek**
- **Facebook-státuszok**
- **Hozzászólások**
- **Üzenetek**
- **Skype-chat**
- **…**

# Példák - hozzászólások

belorusz, ukran (elmult 200-300 ev); szlovak; kettevalt a szerb-horvat (ez kifejezetten friss)..
es ez csak az, amit itt a kornyekunkon tudok.
ja, jut eszembe egy meg "aprobb" pelda: britt-amerikai angol szetvalasa. ez eppen most folyik...
es mond, ha mar en valaszoltam: irnal nehany peldat, ahol kihalt mostanaban egy-egy nyelv? ☺

@avmanster: ez egy keresõ, keresni lehet vele! Ez eddig egy összetett mondat **:D**

Egyébként csak bedobtam, mert itt már összegyült nyelvésztársadalom színe-java, akkor uccu neki csináljunk valamit, ne csak a szánkat tépjük! Mehááá ha csak

beszélünk a melórúú, akkó Sztahanov apánk mérgös lesz!

# Példák - facebook

***Zsuzsi Molnár Te voltál az 500. ☺ Kérjük vedd fel velünk a kapcsolatot, hiszen nyertéééééééél ☺***

***A francokat. S2 re tegnap tettek fel a 4.1.2 es androidot. Igy sem az S3,sem az S4 nem kell.***

***kiraktam a 7-ik képet 10 másodperc alatt..... szállguldok a célfelé..... meg is osztottam a nyereményjátékot! :D :D :D***

# Példák - facebook

*Igen a nagy ötletet én is használnám! 18 hónapja szivat a vodafon a nokiával együtt, egy rokkantat,akire rágyujtották a házát? megáll-ták gy.hibás a tel. a nokia 5230 de 5-ször hozták vitték, rosszabbnál rosszabb cserekész-kel szenvedtem, 500 mg a tel.beáll-ra ment el, söt a futárszolg.saját szml-ra fiz-tem a semmit nem tudó ügyfélsz.utasitására, 3-szor volt a nokia szak-ba mivel ennek anyagi vonzata sem kevés, jelenleg sincs lakásom és ezzel szenvedtettek,mig ingyen tul adtam rajta, mivel hüségnyil-om van, fizetni kell, most samsung SG-G700-as tel-nal küz-ök, irtam 5 old.lev. a válasz bizt-ták a szolg-t,még jó mikor fiz-tem, ezt a rossz tel-val nincs vége az ügyf.sz. ebben sem tud seg-ni 18 hónapja nincs lev.r. a márkaképv.Szolnok nemhogy beáll-ták még azt sem tudta,hol van rajta a GALÉRIA minden beut. 3 e ft, irtam ujra még az édesanyám sz.dátumát is kérik hogy kivizsg-ják, ennyit a vodafonról! ELNÉZËST A KELLEMETLENSÉGÉRT EZT HALL. Az ügyfél.sz, itt nem kellemetl-ről van szó, hanem arról hogy semmilyen segitséget nem vehettem igénybe, MINDEN TISZTELETEM A BEHAJTÓ CSOPORTNAK MINDEN SEGITSÉGET A KERETEIK KÖZÖTT MEGADTAK. 8 éve vodafonos vagyok MEGÉRTE Bocs hogy itt irtam gondjaimról.*

# Példák - skype

(wave)

rogton (Wave)

oks

no vagyok

nő?

:D

:D

az is :)

hát akkor így nézne ki

elsőre jó

reduce is változik

vagyis NEM

:)

mert az elhagyogattuk

nem, az marad :)

hanem

hanem hanem hanem

nos nos nos nos

KRtoMSD változik

# Példák - skype

ejjj
van 1 perced?
persze
kellett csinálnom egy
1 órás
olasz zenei
összeállítást
de
valaki azt mondja most, hogy ez nem olyasz
olasz :)
nem vicces
:D
és most végig kell hallgatnom az 1 órát? :D
lehet h elküldtem valami román izát
(worry)

# Webes szövegek jellemzői

- **Valós idejű (online) kommunikáció**
- **Noha írásban zajlik, mégis beszélt nyelvi sajátosságokat mutat**
- **Gyorsaságra való törekvés (elírások, rövidítések)**
- **Érzelmek kifejezése (emotikonok, nagybetűs írásmód)**
- **Hezitáció megjelenése**

# Nyelvi jellemzők

- **Ékezetek hiánya (*lehet egy hulye kerdesem?*)**
- **Központozás hiánya**
- **Nagybetűk hiánya, ill. csupa nagybetűs írásmód**
- **Félregépelések (*mindegyekinek*)**
- **Sajátos rövidítések (*sztem, lécci, h*)**
- **Angolból átvett szavak/rövidítések (*cool, wtf, pls*)**
- **Emotikonok (írásjelek új funkcióban!) ☺ :D**
- **Különírás-egybeírás hibái (*hibás/népies szavakat kijavítóennek a neve?*)**
- **Írásjelek halmozása (*lehet ezekkel dolgozni???*)**
- **Betűk halmozása (*pontosannn*)**
- **Hezitáció (*öööööö*)**

# A szövegek nyelvi elemzése

- **Magyarlanccal elemeztük a fenti szövegtípusokat + FB-szövegeken tulajdonnév-felismerést hajtottunk végre**
- **sztenderd szövegekre készült elemző hogyan teljesít nem sztenderd szövegeken?**
- **Tipikus hibák feltérképezése**

# Elemzési példák - szegmentálás

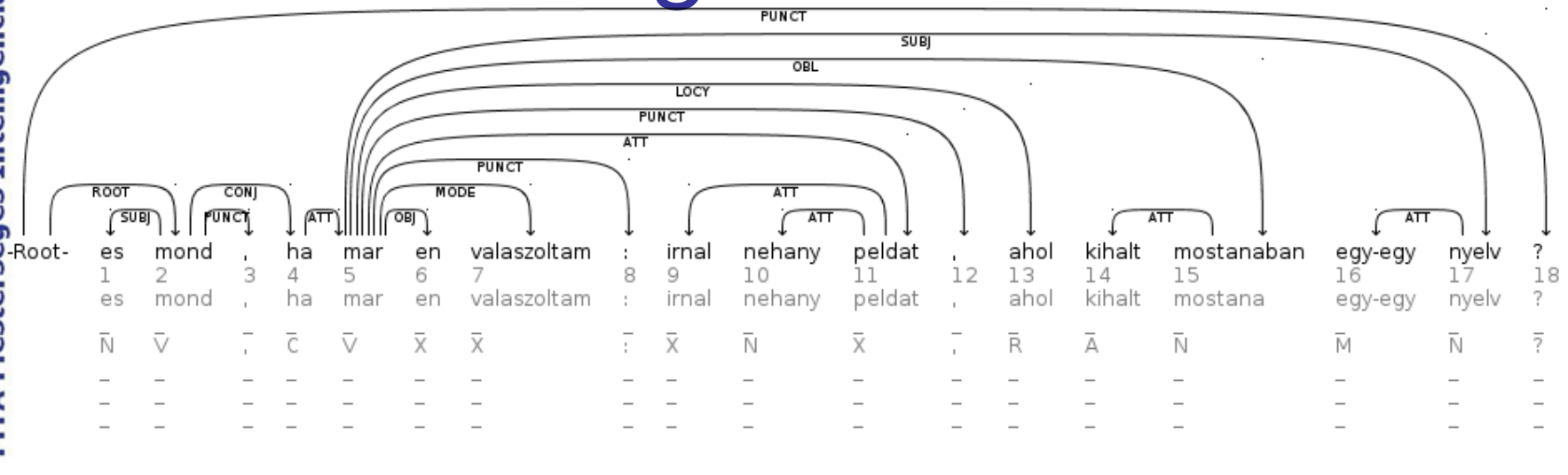
| | | |
|---|---|---|
| ( | ( | K |
| wave | wave | Nn-sn |
| ) | ) | K |
| rogton | rogton | X |
| ( | ( | K |
| Wave | wave | Nn-sn |
| ) | ) | K |
| oks | oks | X |
| no | no | Io |
| vagyok | van | Vmip1s---n |
| nő | nő | Nn-sn |
| ? | ? | ? |
| : | : | : |
| D | D | Nn-sn |
| : | : | : |
| D | D | Nn-sn |
| az | az | Pd3-sn |
| is | is | Ccsp |
| : | : | : |
| ) | ) | K |

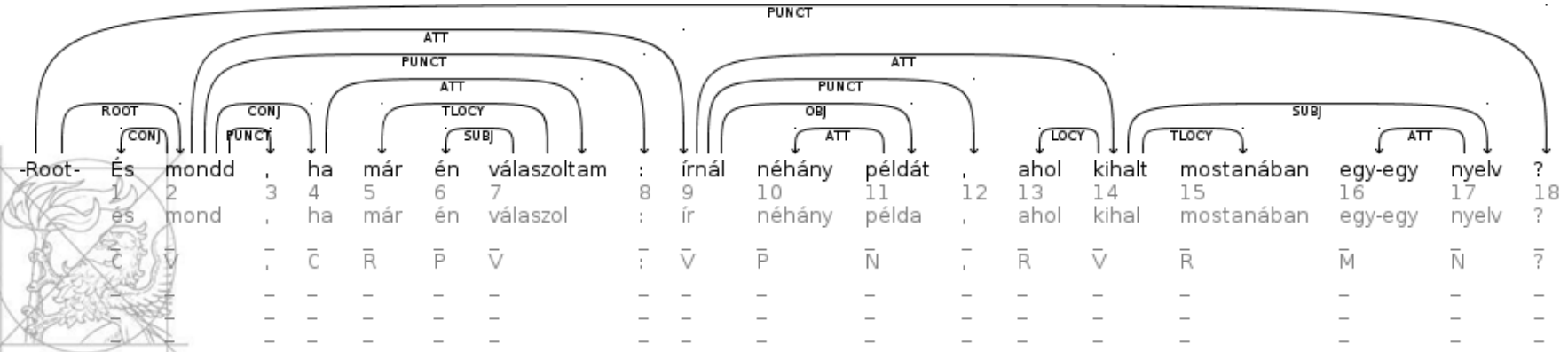
| | | |
|---|---|---|
| ( | ( | K |
| wave | wave | Nn-sn |
| ) | ) | K |
| rogton | rogton | X |
| ( | ( | K |
| Wave | wave | Nn-sn |
| ) | ) | K |
| oks | oks | X |
| no | no | Io |
| vagyok | van | Vmip1s---n |
| nő | nő | Nn-sn |
| ? | ? | ? |
| : | : | : |
| D | D | Nn-sn |
| : | : | : |
| D | D | Nn-sn |
| az | az | Pd3-sn |
| is | is | Ccsp |
| : | : | : |
| ) | ) | K |

# Elemzési példák - szegmentálás

| | | | |
|---|---|---|---|
| http | http | X | |
| : | : | : | |
| / | / | K | |
| / | / | K | |
| cdn | cdn | X | |
| . | . | . | |
| grumpycats.com | grumpycats.com | X | |
| / | / | K | |
| wp-content | wp-content | X | |
| / | / | K | |
| uploads | uploads | X | |
| / | / | K | |
| 2013 | 2013 | Mc-snd | |
| / | / | K | |
| 04 | 04 | Mc-snd | |
| / | / | K | |
| IMG_2692-1-625x677 | IMG_2692-1-625x677 | X | |
| . | . | . | |
| jpg | jpg | X | |

# Elemzési példák – morfológia és szintaxis



# Tapasztalatok

- **A sztenderd szövegeken tanult elemzők nincsenek felkészítve a nem sztenderd szövegek sajátosságaira**
- **Nem egyformán nehezek a szövegek (pl. blogok, hivatalos oldalak facebook-posztjai könnyebbnek tűnnek, mint a tweetek vagy skype-chat…)**
- **A hibatípusok nagy része átfedést mutat szövegtípusok között – egységes megoldás adható rájuk, kisebb finomhangolással**

# Jellemző hibák - szegmentálás

- **Mondatok összevonása (mi egy mondat?)**
- **Emotikonok szétbontása írásjelre és betűkre (rengeteg írásjel)**
- **Webhelyek szétbontása**

# Jellemző hibák - morfológia

- **Ismeretlen szavak (X kód):**
  - **Nincsenek benne az elemző szótárában (*Kvantifikált*)**
  - **Elírások (*eljfelejtődni*)**
- **Rossz kód:**
  - **Létező szóalak, de a kontextusban nem jó (*mer*)**
- **Helyesen felismert alakok:**
  - **hááát hát Nn-sn**
  - **feljődést feljődés Nn-sa**
  - **elemzve elemzve Rv**

# Jellemző hibák – tulajdonnév-felismerés

- **Azonos tulajdonnév több osztályba sorolva:**
  - Meki/I-LOC
  - Meki/I-PER
  - Meki/I-ORG
- **Nagybetűs írásmód:**
  - Juhéééé/I-ORG
  - Sziasztok/I-PER
- **Nagybetűs formulák:**
  - Boldog Névnapot/I-PER
- **Csupa nagybetű:**
  - MERT TUDJÁTOK A NÉP/I-MISC
  - GIGALÁJK/I-ORG
- **Tulajdonnevet követő/megelőző nagybetűs szó:**
  - Erzsébet Garamszeginé Egyetértek/I-ORG
  - Kedves Norbi/I-PER
- **Egymást követő tulajdonnevek**
  - FittJóga Stúdióba , Kis Ernő/I-ORG

# Normalizálás

- **Standardszerűvé tétel**
- **Hibák jelentős része szegmentálási eredetű**
  - **Szóközhiány és -felesleg**
- **Egyéb**
  - **Ékezet**
  - **Halmozás**
- **Szabályok, reguláris kifejezések**

# Javítási lehetőségek - szegmentálás

- **Szabályok felvétele: 1 sor – 1 mondat / felhasználó 1 megnyilatkozása 1 mondat? (skype vs. FB)**
- **Nagybetűs mondatkezdet ne legyen elvárás**
- **Írásjelek nem mindig mondathatárt jelölnek**
- **Emotikonok beépítése a szótárba**
- **Webhelyek kezelése (reguláris kifejezések)**
- **Sajátos jelentéstartalommal bíró jelek (@ # …) kezelése**

# Javítási lehetőségek - morfológia

- **Ékezetesítés**
- **Szótár bővítése: új szavak, rövidítések felvétele**
- **Halmozott írásjelek, betűk egyszerűsítése (2-nél több azonos betű nem lehet egymás mellett a magyarban: *hozzáállás*): *jeeee jee jeeeeeeeee***
- **Hangulatjelek külön kezelése**
- **Elírások kezelése (helyesírás-ellenőrző, szóvégi extra karakterek): *KR-benű***
- **Felhasználói javítások figyelembevétele (*csak kirakták a honlapra a dyjakat díjakat*)**

# Javítási lehetőségek – tulajdonnév-felismerés

- **Nem minden nagybetű tulajdonnevet jelöl**
- **Nem minden tulajdonnév kezdődik nagybetűvel**
- **Osztályok megállapításához valószínűleg kézi annotáció szükséges**

# Összegzés

- **Webes magyar szövegek nyelvi elemzése**
- **Jellemző hibák feltérképezése**
- **Megoldási javaslatok nyújtása**
- **Ezek megvalósítása zajlik – magyarlanc 3.0**