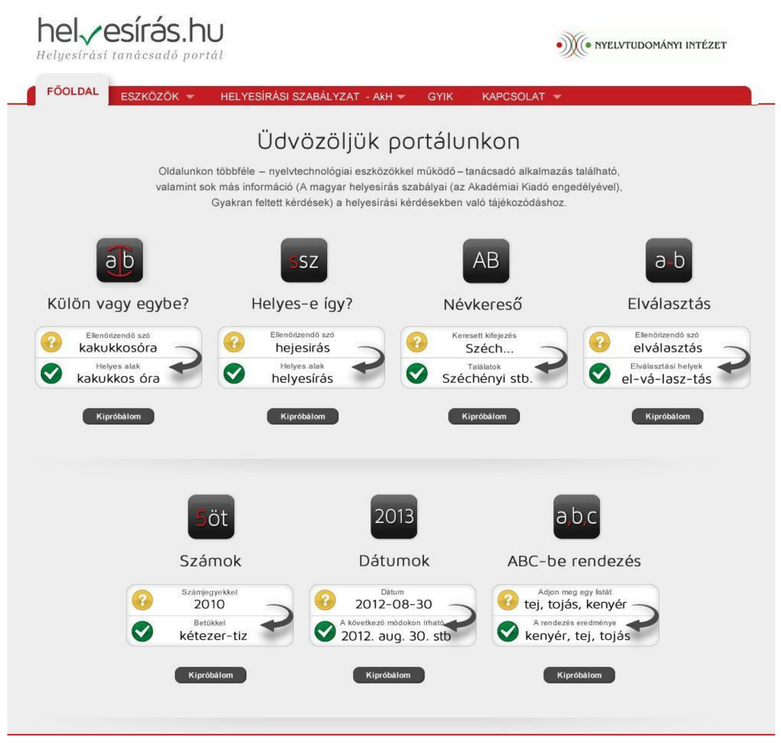
Helyesírás.hu – Nyelvtechnológiai megoldások automatikus helyesírási tanácsadó rendszerben

Miháltz Márton, Ludányi Zsófia
MTA Nyelvtudományi Intézet, Nyelvtechnológiai Kutatócsoport
{mihaltz.marton, ludanyi.zsofia}@nytud.mta.hu

0. Bevezetés

- Felhasználói kérdések hatékony kezelhetősége
  - kompromisszum: egyszerű felhasználói felület ↔ várható inputról a lehető legtöbb információt begyűjtő rendszer
  - tetszőleges szöveg bevitele, automatikus feldolgozása nem oldható meg
  - megoldás: problémakategóriák egyértelmű meghatározása → kötelező típusválasztó
- Bemeneti főkategóriák
  - felhasználói választás → megfelelő modul hozzárendelése



- Modultípusok – milyen bemenetet várnak:
  - speciális, megszorított bemenet (pl. számnevek, keltezés)
  - szűrt szabad szöveges bemenet (külön-egybe, helyesírás-ajánló)
- Erőforrások:
  - rendszerhez csatolt erőforrások (morfológiai elemző, lexikális adatbázis)
  - felhasználótól kért információ

1. Számnevek

- Feladat: számjegyekkel írt számok betűkkel való átírása, kommentekkel
- Egész számok, sorszámnevek, (tizedes) törtek
- Példák:
  - *1200* → *ezerkétszáz* v. *ezerkettőszáz*
  - *3.* → *harmadik*
  - *3/4* → *háromnegyed v. három negyed*
  - *3,14* → *három egész tizennégy század*

2. Keltezés

- Feladat: éééé-hh-nn formátumú dátumok AkH. által elfogadott írásmódjainak visszaadása + leggyakoribb toldalékolt formák
- Példák:
  - *1582. október 10.*
  - *1582. október 10-én*
  - *1582. október 10-e óta*
  - *1582 októberében*

3. Betűrendbe sorolás

- Bemenet: külön sorba írt szavak
- Klasszikus rendezési algoritmus: 2 sztring összehasonlítása karakterenként, az első különböző karakterpár összehasonlítása adja az eredményt

4. Tulajdonnevek helyesírása

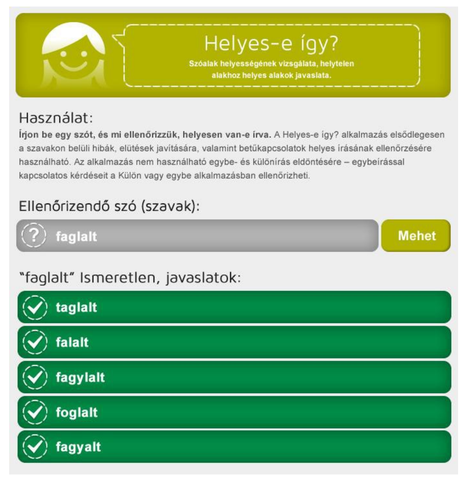
- Begépelés közben prediktív megjelenítés
- Tulajdonnevek azonosítása: földrajzi név, személynév, egyéb jegyek (keresztnév, településnév stb.)

5. Elválasztás

- Felhasznált modulok:
  - huhyphn: OpenOffice/LibreOffice elválasztómodulja
  - HUMor morfológiai elemző
- Probléma: egy szóalak → többféle elválasztás (*me-gint* ~ *meg-int*) → a hunhyphen csak az egyik megoldást adja vissza
- Megoldás: HUMor → összetétel-e?, összetételi határok megjelölése | jellel

6. Helyesírás-ajánló

- Önálló szóalakok vizsgálata: helyes / helytelen + javaslatok
- Hunspell 1.3.2 + Humor



7. Különírás-egybeírás

- Környezetfüggetlen, jegystruktúrás kifejezésnyelvtan
- Szófaj, morfológiai jegyek, szótagok/összetételi tagok száma
- Lexikális jegyek: színnevek, foglalkozások, rangok, keresztnevek, népek és nyelvek nevei, rövidítések...

Példa egy szabályra:

```
id: M_EK_ANYAGNEV_1_1_1
rule: N(sem="Material", ncomparts=1) +
  N(match=".*,NOM", ncomparts=1) == N(sep='')
comment: "Az anyagnévi jelzőt, ha egyszerű szó,
  egybeírjuk a nem összetett főnevekkel."
refs: AKH-115, OH-117
ex: bőr + kabát = bőrkabát, selyem + ing =
  selyeming
kill: M_EK_JELOLETLEN_BIRTOKOS
```

Példa egy generált elemzési fára:

```
Input: bőr + kabát
N(sep=[''], ncomparts="2") : M_EK_ANYAGNEV_1_1_1
  N(wordform="bőr", stem="bőr", match="FN,NOM",
    sem=['Material'], ncomparts="1", nsylls="1") :
    0.
  N(wordform="kabát", stem="kabát", match="FN,NOM",
    sem=[], ncomparts="1", nsylls="2") : 1.
```

A kimenet:

```
Javasolt alak: „bőrkabát”
Magyarázat:
A "bőr" főnevet és a "kabát" főnevet egybeírjuk az
alábbi szabály alapján:
"Az anyagnévi jelzőt, ha egyszerű szó, egybeírjuk
a nem összetett főnevekkel." (AKH-115, OH-117)
```

8. Összefoglalás

- Moduláris, nyelvtechnológiai eszközökkel támogatott automatikus helyesírás-támogató rendszer
- Elemzők és lexikai adatbázisok szükségszerű használata
- Felhasználói elégedettség kritikus → állandó, célzott fejlesztés igénye