Bevezetés
az `e-magyar` programcsomag használatába

Vadász Noémi

2019. május 2.

MTA Nyelvtudományi Intézet
vadasz.noemi@nytud.mta.hu

Az előadás felépítése

1. szövegelemzés számítógéppel
   - elemzési lépések
   - elemzőláncok
   - az `e-magyar` működése
2. az `e-magyar` használata

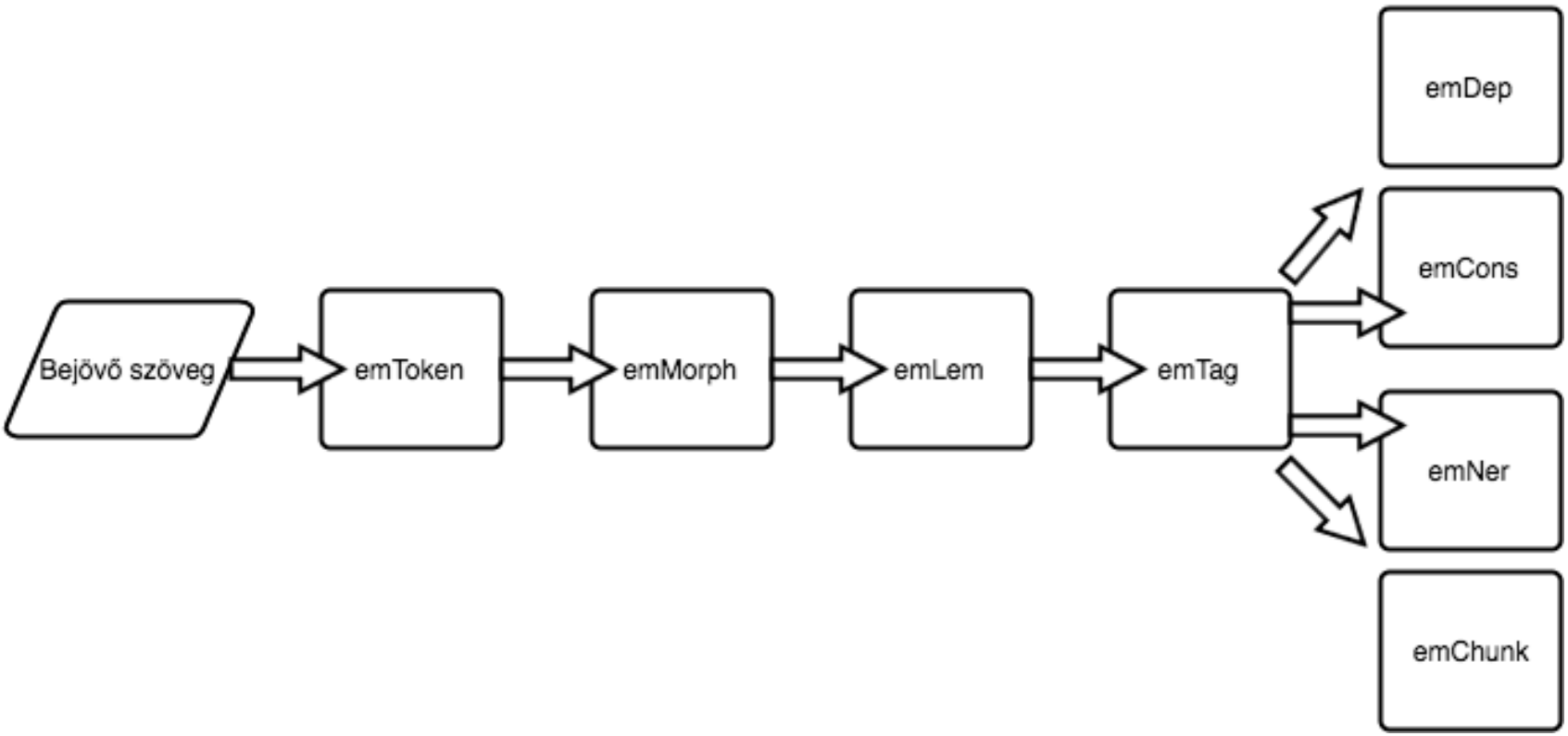
Szövegfeldolgozó lépések

1. szöveg felbontása kezelhető egységekre: **mondatszegmentálás** és **tokenizálás**
2. a szavak elemzése: **morfológiai elemzés**
3. az elemzések **egyértelműsítése** a mondatban
4. **főnévi csoportok**, **tulajdonnevek** meghatározása
5. **függőségi** és **összetevős** mondatelemzés

A lépéseket **lánc**ba köthetjük, ahol az alsóbb szintű elemzési lépés **kimenete** a felsőbb szintű elemzés **bemenet**ét képezi.

Elemzőláncok magyarra: magyarlanc és `e-magyar`

Az e-magyar elemzőlánc felépítése



Tokenizálás

Két feladat:

1. mondatszegmentálás
   - rövidítések, római számok, sorszámok (pont van a végén)
   - mondatok idézetben vagy zárójelben
2. tokenizálás
   - detokenizálhatóság
   - páros írásjelek (zárójel, idézőjel, aposztróf) kezelése
   - kérdőpartikula kezelése

Miért kell tokenizálni?

- a szintaktikai elemző mondatokon dolgozik
- a morfológiai elemző tokeneken dolgozik

emToken

- szabályalapú
- a leggyakoribb rövidítéseket ismeri (kb. 450 rövidítés)
- a kimenetben `xml`-tagek jelölik a szó- és mondathatárokat

A kutya váratlanul ugatni kezdett. Ettől úgy megijedt dr. Thorotzkay Alfréd, hogy hanyatt esett az aszfalton. Felesége, aki egyébként a BKV Zrt.-nél dolgozik, egyből rohant hozzá, amint ezt megtudta.

<s> ... </s>	mondat
<w> ... </w>	szó
<c> ... </c>	írásjel
<ws> ... </ws>	whitespace

emToken

<s><w>A</w><ws> </ws><w>kutya</w><ws></ws><w>váratlanul</w><ws>
</ws><w>ugatni</w><ws></ws><w>kezdett</w><c>.</c></s><ws> </ws>
<s><w>Ettől</w><ws> </ws><w>úgy</w><ws> </ws><w>megijedt</w><ws>
</ws><w>dr.</w><ws> </ws><w>Thorotzkay</w><ws>
</ws><w>Alfréd</w><c>,</c><ws> </ws><w>hogy</w><ws>
</ws><w>hanyatt</w><ws> </ws><w>esett</w><ws> </ws><w>az</w><ws>
</ws><w>aszfalton</w><c>.</c></s><ws>
</ws><s><w>Felesége</w><c>,</c><ws> </ws><w>aki</w><ws>
</ws><w>egyébként</w><ws> </ws><w>a</w><ws>
</ws><w>BKV</w><ws></ws><w>Zrt.-nél</w><ws>
</ws><w>dolgozik</w><c>,</c><ws></ws><w>egyből</w><ws>
</ws><w>rohant</w><ws> </ws><w>hozzá</w><c>,</c><ws>
</ws><w>amint</w><ws> </ws><w>ezt</w><ws>
</ws><w>megtudta</w><c>.</c></s>

emToken

A
kutya
váratlanul
ugatni
kezdett
.

Ettől
úgy
megijedt
dr.
Thorotzkay
Alfréd
,
hogy
hanyatt
esett

- `tsv` fájl
- minden sor egy token
- mondatok között üres sor

Morfológiai elemzés és tövesítés

- a tokenhez hozzárendeljük **az összes lehetséges elemzés**t
- különböző címkekészletek léteznek
- a címkékben lehetnek:
  - morfoszintaktikai információk
  - a jelentésre vonatkozó információk
  - a hangalakra vonatkozó információk (allomorfia)
  - szófajkód
  - lemma
  - morfológiai szegmentumok

emMorph és emLem

- tőtár, toldaléktár és nyelvtan
- szabályalapú morfológiai elemző (véges állapotú transzdúcer)
- szabályalapú tövesítés a tőalkotó morfémák figyelembevételével

szemét

1. szem[/N]=szem+e[Poss.3Sg]=é+t[Acc]=t
2. szem[/N]=szem+é[AnP]=é+t[Acc]=t
3. szemét[/N]=szemét+[Nom]=

Két morfológiai címkekészlet

emMorph

- nem tükrözi a morfológiai jelöltséget
- derivációt, szóösszetételt is kódol
- tartalmazza a lemmát, a morfológiai szegmentumokat, allomorfot
- `adtad ad[/V]=ad+tad[Pst.Def.2Sg]=tad`

UD (Universal Dependencies)

- külön szófajkód
- linearizált jegy–érték struktúra
- nem tükrözi a morfológiai jelöltséget
- képzést, derivációt nem kódol, csak inflexiót
- nem tartalmazza a lemmát, a morfológiai szegmentumokat
- `adtad ad VERB Definite=Def|Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin|Voice=Act`

Morfológiai egyértelműsítés

- minden tokenhez megvan az összes lehetséges elemzése
- el kell dönteni, hogy **az aktuális mondatban** melyik elemzést válasszuk
- tehát az egyértelműsítő már mondatokat néz
- a lehetséges elemzések legvalószínűbb láncát keresi

Péter Marira vár. Péter a vár előtt áll.

Statisztikai alapú egyértelműsítés:

- sok, már elemzett és emberek által egyértelműsített mondatot megnézett a program
- ebből megtanulja, hogy mik a gyakori mintázatok a szövegben
- nem egyszerűen azt nézi, hogy a vár gyakrabban ige, mint főnév
- hanem azt, hogy egy névelőt sokkal gyakrabban követ főnév, mint ige
- és azt, hogy a -rA esetragú főnév után valószínűbb az ige, mint a főnév

A kastély nem vár.

A	a	[/Det|art.Def]
kastély	kastély	[/N][Nom]
nem	nem	[/Adv]
vár	vár	[/N][Nom]
.	.	[/PUNCT]

A kastély nem vár senkire.

A	a	[/Det|art.Def]
kastély	kastély	[/N][Nom]
nem	nem	[/Adv]
vár	vár	[/V][Prs.NDef.3Sg]
senkire	senki	[/N|Pro][Subl]
.	.	[/PUNCT]

Függőségi mondatelemzés

- bemenete a mondat egyértelműsített tokenekkel
- kimenete a függőségi fa
- függőség: a szavak közötti reláció
- a fa minden csomópontja egy szó
- a csomópontok közötti él a gyerek (függő) csomóponttól mutat a szülő felé
- az élcímkék a függőségi viszony típusai

Az `e-magyarba` épített függőségi mondatelemző statisztikai megközelítésű.

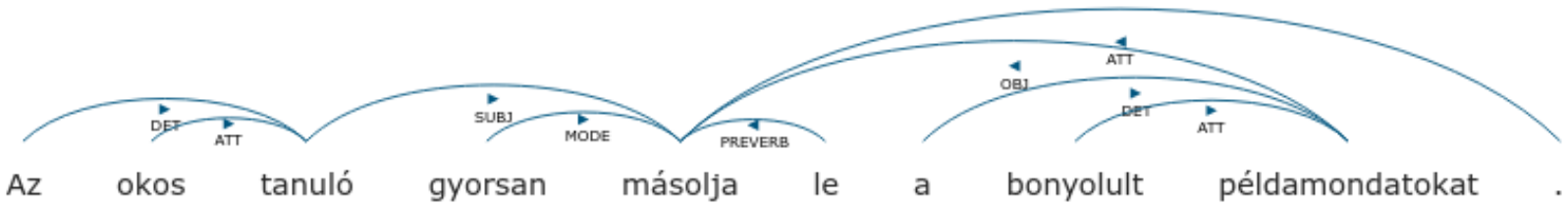
emDep

Az exkatonát kórházba szállították, ahol két műtétet is végrehajtottak rajta.

1	Az	2	DET
2	exkatonát	4	OBJ
3	kórházba	4	OBL
4	szállították	0	ROOT
5	,	4	PUNCT
6	ahol	10	LOCY
7	két	8	ATT
8	műtétet	10	OBJ
9	is	8	CONJ
10	végrehajtottak	4	ATT
11	rajta	10	OBL
12	.	0	PUNCT

emDep

Az okos tanuló gyorsan másolja le a bonyolult példamondatokat.



Összetevős mondatelemzés

- bemenete a mondat egyértelműsített tokenekkel
- kimenete az összetevős fa
- a mondatot frázisokra bontja
- a fa levelei a mondat szavai (terminálisok)
- a fa nem-levél csomópontjai a frázisok (nem-terminálisok)
- az élek címkézetlenek

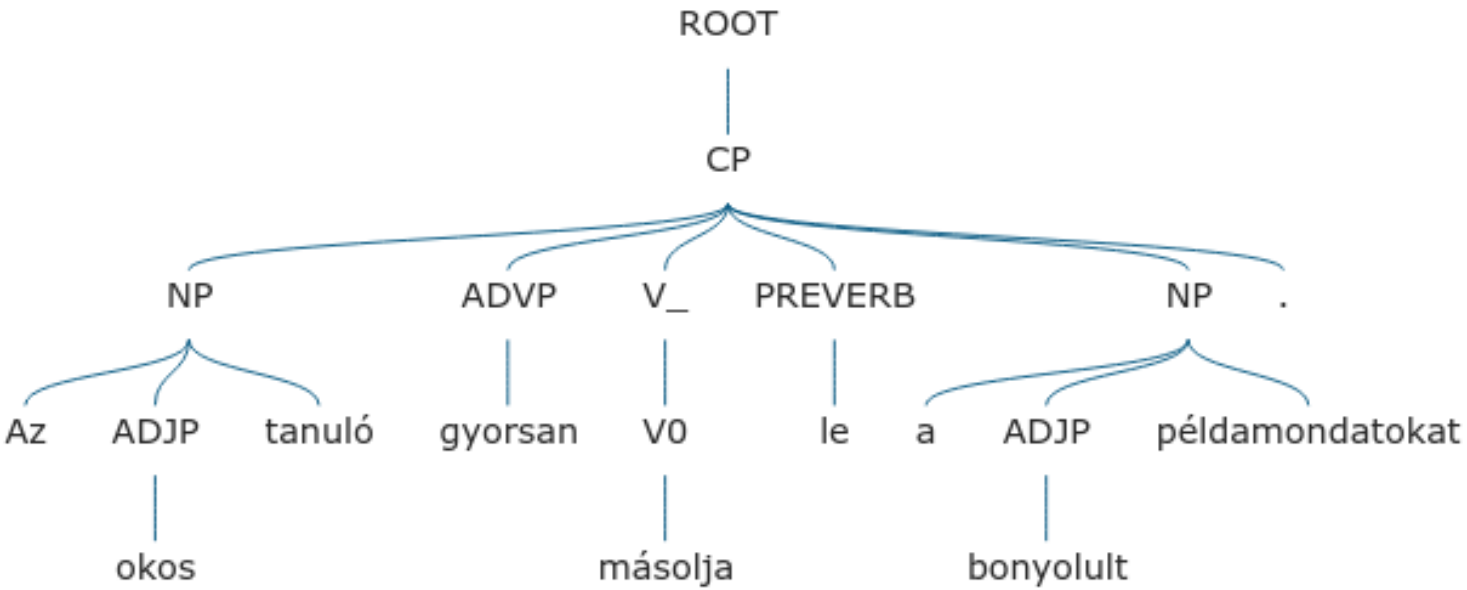
Az `e-magyarba` épített összetevős mondatelemző statisztikai megközelítésű.

emCons

Az exkatonát kórházba szállították, ahol két műtétet is végrehajtottak rajta.

1	Az	(ROOT(CP(NP*
2	exkatonát	*)
3	kórházba	(NP*)
4	szállították	(V_(V0*))
5	,	*
6	ahol	(ADVP*)
7	két	(NP*
8	műtétet	*)
9	is	(C0*)
10	végrehajtottak	(V_(V0*))
11	rajta	(NP*)
12	.	*))

Az okos tanuló gyorsan másolja le a bonyolult példamondatokat.



A főnévi csoportok és a tulajdonnevek felismerése

- ún. szekvenciális címkézési feladatok
- főnévi csoport keresése: **maximális NP** (olyan főnévi csoport, amely nem része főnévi csoportnak)
- **tulajdonnevek** keresése és névkategóriákba sorolása (személynév, intézménynév, földrajzi név, egyéb)
- statisztikai gépi tanuláson alapul, a HunTag3 program végzi
- sok szöveget látott, amelyben be vannak jelölve az NP-k/tulajdonnevek
- és hogy milyen tulajdonságokra figyeljen a tanuláskor (nagybetűvel kezdődött, mondat elején van, szófaj)
- összefüggéseket talál a tulajdonságok és a címkék között, amelyből valószínűségi modellt épít

emChunk

A szállásunk egy Balaton melletti kis üdülőfaluban, Zamárdiban volt.

A	B-NP
szállásunk	E-NP
egy	B-NP
Balaton	I-NP
melletti	I-NP
kis	I-NP
üdülőfaluban	I-NP
,	I-NP
Zamárdiban	E-NP
volt	O
.	O

[...] közölte Wolf László, az OTP Bank vezérigazgató-helyettese az MTI érdeklődésére.

közölte	0
Wolf	B-PER
László	E-PER
,	0
az	0
OTP	B-ORG
Bank	E-ORG
vezérigazgató-helyettese	0
az	0
MTI	1-ORG
érdeklődésére	0
.	0

A nyelvi elemző korlátai

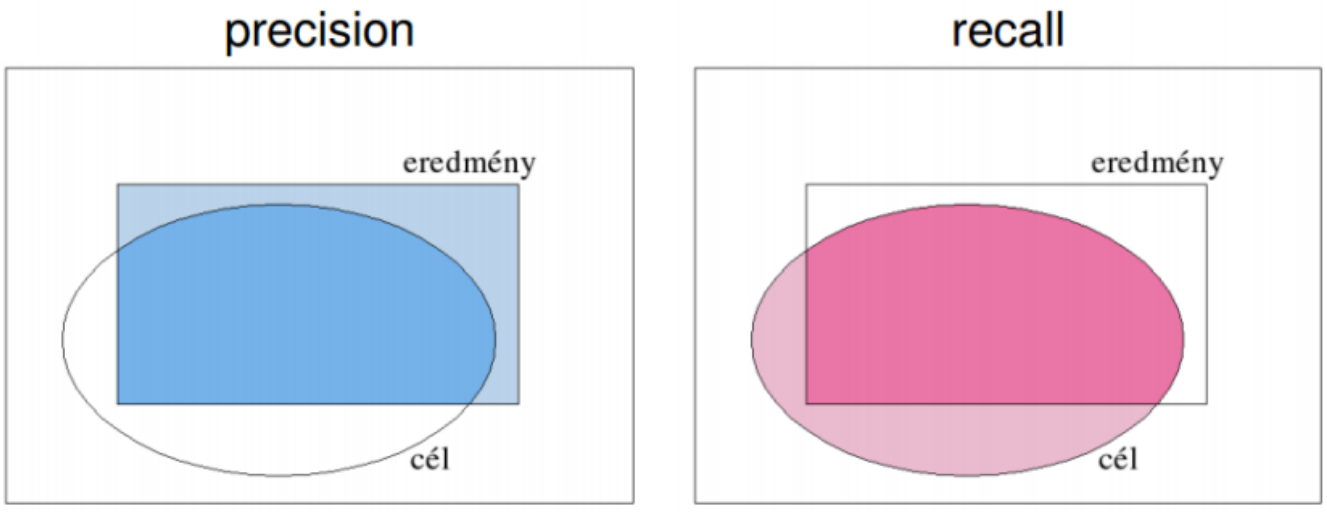
gold standard: kézzel címkézett anyag, amihez hasonlítunk

	cél pozitív	cél negatív
eredmény pozitív	True positive (TP)	False positive (FP)
eredmény negatív	False negative (FN)	True negative (TN)

Egy példa: egy kép gépi feldolgozása

- egy gép eldönti 1 000 képről, hogy van-e rajta kismacska
- az 1 000 képet ember is kiértékeli (ez lesz a gold standard)
- összevetjük az ítéleteket
- összeszámoljuk a TP, TN, FN találatokat
- a számok segítségével értékeljük a gép teljesítményét

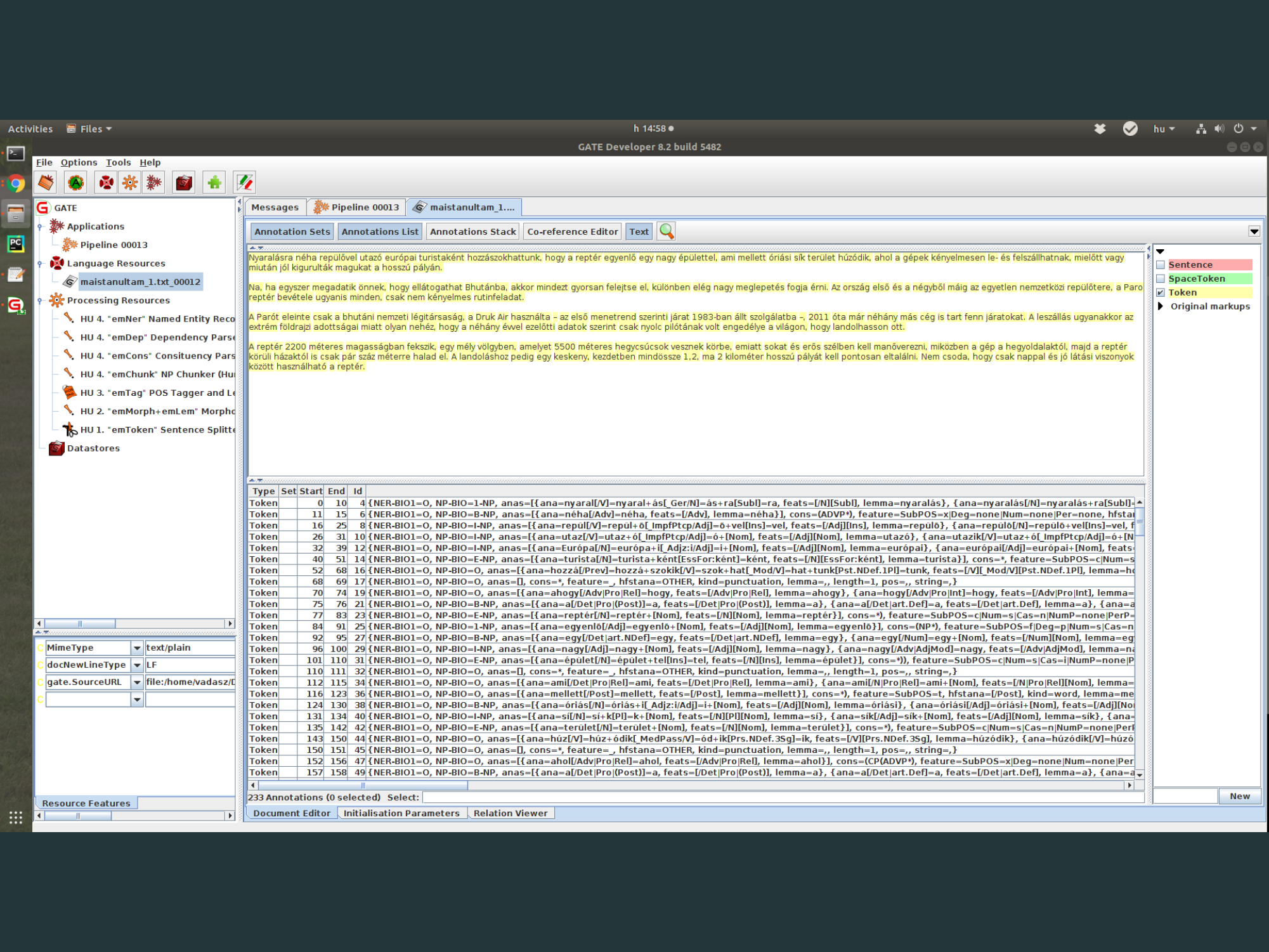
A nyelvi elemző korlátai



- **pontosság** (precision): TP/(TP+FP)
  A macskásnak ítélt képek közül hány tartalmaz az ember szerint is kismacskát?
- **fedés** (recall): TP/(TP+FN)
  Megtalálta-e a gép az összes macskás felvételt?
- **F-mérték**: a pontosság és a fedés harmonikus közepe

Az e-magyar háromféle használata

1. GATE
   - nyelvfeldolgozó keretrendszer
   - grafikus felület
   - platformfüggetlen
   - egységes annotációs modell
   - más elemzőeszközök is beilleszthetők a láncba
   - jól dokumentált
   - soklépéses, viszonylag bonyolult telepítés
2. emTSV
   - terminálból működtethető
   - egységes adatformátum a modularitásért
   - az elemzőlánc lépései között ki- és beléphetünk
   - a kimenet egy könnyen feldolgozható fejléces `tsv`
   - kétféle morfológiai címkekészlet: emMorph és UD
3. honlap
   - a modulok és a lánc kipróbálására
   - max. 6 000 karakter hosszú szövegek elemzésére
   - kétféle formátumban letölthető kimenet: `xml` és `tsv`


Activities
Files
h 14:58
hu
GATE Developer 8.2 build 5482
File Options Tools Help
GATE
Applications
Pipeline 00013
Language Resources
maistanultam_1.txt_00012
Processing Resources
HU 4. "emNer" Named Entity Reco
HU 4. "emDep" Dependency Parse
HU 4. "emCons" Consituency Pars
HU 4. "emChunk" NP Chunker (Hu
HU 3. "emTag" POS Tagger and Le
HU 2. "emMorph+emLem" Morpho
HU 1. "emToken" Sentence Splitte
Datastores
MimeType text/plain
docNewLineType LF
gate.SourceURL file:/home/vadasz/D
Resource Features
Messages
Pipeline 00013
maistanultam_1....
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
Nyaralásra néha repülővel utazó európai turistaként hozzászokhattunk, hogy a reptér egyenlő egy nagy épülettel, ami mellett óriási sík terület húzódik, ahol a gépek kényelmesen le- és felszállhatnak, mielőtt vagy miután jól kigurulták magukat a hosszú pályán.
Na, ha egyszer megadatik önnek, hogy ellátogathat Bhutánba, akkor mindezt gyorsan felejtse el, különben elég nagy meglepetés fogja érni. Az ország első és a négyből máig az egyetlen nemzetközi repülőtere, a Paro reptér bevétele ugyanis minden, csak nem kényelmes rutinfeladat.
A Parót eleinte csak a bhutáni nemzeti légitársaság, a Druk Air használta – az első menetrend szerinti járat 1983-ban állt szolgálatba –, 2011 óta már néhány más cég is tart fenn járatokat. A leszállás ugyanakkor az extrém földrajzi adottságai miatt olyan nehéz, hogy a néhány évvel ezelőtti adatok szerint csak nyolc pilótának volt engedélye a világon, hogy landolhasson ott.
A reptér 2200 méteres magasságban fekszik, egy mély völgyben, amelyet 5500 méteres hegycsúcsok vesznek körbe, emiatt sokat és erős szélben kell manőverezni, miközben a gép a hegyoldalaktól, majd a reptér körüli házaktól is csak pár száz méterre halad el. A landoláshoz pedig egy keskeny, kezdetben mindössze 1,2, ma 2 kilométer hosszú pályát kell pontosan eltalálni. Nem csoda, hogy csak nappal és jó látási viszonyok között használható a reptér.
Sentence
SpaceToken
Token
Original markups
Type Set Start End Id
Token 0 10 4 {NER-BIO1=O, NP-BIO=1-NP, anas=[{ana=nyaral[/V]=nyaral+ás[_Ger/N]=ás+ra[Subl]=ra, feats=[/N][Subl], lemma=nyaralás}, {ana=nyaralás[/N]=nyaralás+ra[Subl]=
Token 11 15 6 {NER-BIO1=O, NP-BIO=B-NP, anas=[{ana=néha[/Adv]=néha, feats=[/Adv], lemma=néha}], cons=(ADVP*), feature=SubPOS=x|Deg=none|Num=none|Per=none, hfstan
Token 16 25 8 {NER-BIO1=O, NP-BIO=I-NP, anas=[{ana=repül[/V]=repül+ő[_ImpfPtcp/Adj]=ő+vel[Ins]=vel, feats=[/Adj][Ins], lemma=repülő}, {ana=repülő[/N]=repülő+vel[Ins]=vel, f
Token 26 31 10 {NER-BIO1=O, NP-BIO=I-NP, anas=[{ana=utaz[/V]=utaz+ó[_ImpfPtcp/Adj]=ó+[Nom], feats=[/Adj][Nom], lemma=utazó}, {ana=utazik[/V]=utaz+ó[_ImpfPtcp/Adj]=ó+[N
Token 32 39 12 {NER-BIO1=O, NP-BIO=I-NP, anas=[{ana=Európa[/N]=európa+i[_Adjz:i/Adj]=i+[Nom], feats=[/Adj][Nom], lemma=európai}, {ana=európai[/Adj]=európai+[Nom], feats=
Token 40 51 14 {NER-BIO1=O, NP-BIO=E-NP, anas=[{ana=turista[/N]=turista+ként[EssFor:ként]=ként, feats=[/N][EssFor:ként], lemma=turista}], cons=*, feature=SubPOS=c|Num=s
Token 52 68 16 {NER-BIO1=O, NP-BIO=O, anas=[{ana=hozzá[/Prev]=hozzá+szokik[/V]=szok+hat[_Mod/V]=hat+tunk[Pst.NDef.1Pl]=tunk, feats=[/V][_Mod/V][Pst.NDef.1Pl], lemma=h
Token 68 69 17 {NER-BIO1=O, NP-BIO=O, anas=[], cons=*, feature=_, hfstana=OTHER, kind=punctuation, lemma=,, length=1, pos=,, string=,}
Token 70 74 19 {NER-BIO1=O, NP-BIO=O, anas=[{ana=ahogy[/Adv|Pro|Rel]=hogy, feats=[/Adv|Pro|Rel], lemma=ahogy}, {ana=hogy[/Adv|Pro|Int]=hogy, feats=[/Adv|Pro|Int], lemma=
Token 75 76 21 {NER-BIO1=O, NP-BIO=B-NP, anas=[{ana=a[/Det|Pro|(Post)]=a, feats=[/Det|Pro|(Post)], lemma=a}, {ana=a[/Det|art.Def]=a, feats=[/Det|art.Def], lemma=a}, {ana=a
Token 77 83 23 {NER-BIO1=O, NP-BIO=E-NP, anas=[{ana=reptér[/N]=reptér+[Nom], feats=[/N][Nom], lemma=reptér}], cons=*), feature=SubPOS=c|Num=s|Cas=n|NumP=none|PerP=
Token 84 91 25 {NER-BIO1=O, NP-BIO=1-NP, anas=[{ana=egyenlő[/Adj]=egyenlő+[Nom], feats=[/Adj][Nom], lemma=egyenlő}], cons=(NP*), feature=SubPOS=f|Deg=p|Num=s|Cas=n
Token 92 95 27 {NER-BIO1=O, NP-BIO=B-NP, anas=[{ana=egy[/Det|art.NDef]=egy, feats=[/Det|art.NDef], lemma=egy}, {ana=egy[/Num]=egy+[Nom], feats=[/Num][Nom], lemma=eg
Token 96 100 29 {NER-BIO1=O, NP-BIO=I-NP, anas=[{ana=nagy[/Adj]=nagy+[Nom], feats=[/Adj][Nom], lemma=nagy}, {ana=nagy[/Adv|AdjMod]=nagy, feats=[/Adv|AdjMod], lemma=na
Token 101 110 31 {NER-BIO1=O, NP-BIO=E-NP, anas=[{ana=épület[/N]=épület+tel[Ins]=tel, feats=[/N][Ins], lemma=épület}], cons=*)), feature=SubPOS=c|Num=s|Cas=i|NumP=none|P
Token 110 111 32 {NER-BIO1=O, NP-BIO=O, anas=[], cons=*, feature=_, hfstana=OTHER, kind=punctuation, lemma=,, length=1, pos=,, string=,}
Token 112 115 34 {NER-BIO1=O, NP-BIO=O, anas=[{ana=ami[/Det|Pro|Rel]=ami, feats=[/Det|Pro|Rel], lemma=ami}, {ana=ami[/N|Pro|Rel]=ami+[Nom], feats=[/N|Pro|Rel][Nom], lemma=
Token 116 123 36 {NER-BIO1=O, NP-BIO=O, anas=[{ana=mellett[/Post]=mellett, feats=[/Post], lemma=mellett}], cons=*), feature=SubPOS=t, hfstana=[/Post], kind=word, lemma=me
Token 124 130 38 {NER-BIO1=O, NP-BIO=B-NP, anas=[{ana=óriás[/N]=óriás+i[_Adjz:i/Adj]=i+[Nom], feats=[/Adj][Nom], lemma=óriási}, {ana=óriási[/Adj]=óriási+[Nom], feats=[/Adj][No
Token 131 134 40 {NER-BIO1=O, NP-BIO=I-NP, anas=[{ana=si[/N]=si+k[Pl]=k+[Nom], feats=[/N][Pl][Nom], lemma=si}, {ana=sík[/Adj]=sík+[Nom], feats=[/Adj][Nom], lemma=sík}, {ana=
Token 135 142 42 {NER-BIO1=O, NP-BIO=E-NP, anas=[{ana=terület[/N]=terület+[Nom], feats=[/N][Nom], lemma=terület}], cons=*), feature=SubPOS=c|Num=s|Cas=n|NumP=none|PerP
Token 143 150 44 {NER-BIO1=O, NP-BIO=O, anas=[{ana=húz[/V]=húz+ódik[_MedPass/V]=ód+ik[Prs.NDef.3Sg]=ik, feats=[/V][Prs.NDef.3Sg], lemma=húzódik}, {ana=húzódik[/V]=húzó
Token 150 151 45 {NER-BIO1=O, NP-BIO=O, anas=[], cons=*, feature=_, hfstana=OTHER, kind=punctuation, lemma=,, length=1, pos=,, string=,}
Token 152 156 47 {NER-BIO1=O, NP-BIO=O, anas=[{ana=ahol[/Adv|Pro|Rel]=ahol, feats=[/Adv|Pro|Rel], lemma=ahol}], cons=(CP(ADVP*), feature=SubPOS=x|Deg=none|Num=none|Per
Token 157 158 49 {NER-BIO1=O, NP-BIO=B-NP, anas=[{ana=a[/Det|Pro|(Post)]=a, feats=[/Det|Pro|(Post)], lemma=a}, {ana=a[/Det|art.Def]=a, feats=[/Det|art.Def], lemma=a}, {ana=a
233 Annotations (0 selected) Select:
New
Document Editor
Initialisation Parameters
Relation Viewer

Hasznos linkek

e-magyar honlap

`http://e-magyar.hu`

e-magyar GATE

`https://github.com/dlt-rilmta/hunlp-GATE`

emTSV

`https://github.com/dlt-rilmta/emtsv`

magyarlanc

`http://rgai.inf.u-szeged.hu/index.php?lang=en&page=magyarlanc`

morfológiai címkekészletek

`https://github.com/dlt-rilmta/panmorph`