AZ E-MAGYAR UD-KIMENETÉNEK CÍMKEKÉSZLETE + SEGÉDLET AZ ANNOTÁCIÓK JAVÍTÁSÁHOZ

VERB: ige

Aspect:

Freq: -gat/-get képzős igék (pl. *ütöget*)

Definite:

Ind: határozatlan (pl. *üt*)

Def: határozott (pl. *üti*)

2 (pl. *ütlek*)

Mood:

Ind: kijelentő (pl. *üt*)

Imp: felszólító (pl. *üss*)

Cnd: feltételes (pl. *ütne*)

Pot: ható igék (pl. *üthet*)

Cnd,Pot: feltételes + ható (pl. *üthetne*)

Imp,Pot: felszólító + ható (pl. *üthessen*)

Number:

Sing: egyes szám (pl. *üt*)

Plur: többes szám (pl. *ütnek*)

Person:

1: első személy (pl. *ütök*)

2: második személy (pl. *ütsz*)

3: harmadik személy (pl. *üt*)

Tense:

Pres: jelen idő (pl. *üt*)

Past: múlt idő (pl. *ütött*)

VerbForm:

Fin: véges ige (pl. *üt*)

Inf: főnévi igenév (pl. *ütni*)

Voice:

Act: aktív (pl. *üt*)

Cau: műveltető (pl. *üttet*)

AUX: segédige

- Az UD specifikációjában segédige: *volna*, *fog*, *talál*, *szokott*
- A magyarlanc 3.0 segédigeként annotálja: *fog*, *volna*

NOUN: főnév

Case: eset

Nom: nominative (pl. *ember, kormány*, *év*, *cég*, *forint*, *elnök*)

Acc: accusative (pl. *embert, százalékát, részt, törvényt, forintot)*
Dat: dative (pl. *embernek, lapunknak, parlamentnek, kormánynak*)
Loc: locative (pl. *Győrött, helyütt*)
Ins: instrumental (pl. *emberrel, százalékkal, évvel, alkalommal)*
Dis: distributive (pl. *emberenként, tonnánként, félóránként, hordónként*)
Ess: essive (pl. *emberként, ráadásul, vendégül, hírül, rabul, összegzésül*)
Tra: translative (pl. *emberré, közzé, társasággá, aduvá, autópályává*)
Ine: inessive (pl. *emberben, évben, kapcsolatban, napokban, esetben*)
Ill: illative (pl. *emberbe, őrizetbe, kórházba, hivatalba, tulajdonába*)
Ela: elative (pl. *emberből, szempontjából, bevételből, hibából*)
Ade: adessive (pl. *embernél, adóhatóságnál, bankautomatáknál*)
All: allative (pl. emberhez, ellátáshoz, elnyeréséhez, fűtőolajhoz, inflációhoz)
Abl: ablative (pl. *embertől, 1-jétől, elejétől, külvilágtól, várostól*)
Sup: superessive (pl. *emberen, héten, szerdán, elején, alapján*)
Sub: sublative (pl. *emberre, élére, forintra, színpadra, napirendre*)
Del: delative (pl. *emberről, posztjáról, kérdésekről, megszüntetéséről*)
Tem: temporal (pl. *éjfélkor, órakor, Letartóztatásakor, elrontásakor*)
Ter: termonative (pl. *emberig, végéig, napig, évig, 30-áig, ideig*)
Cau: causative (pl. *emberért, dollárért, BL-szereplésért, címért*)
Abs: absolutive (példaként, tagjaként, feladatként, Rendezőként) – ez valószínűleg bekerült az esszívuszba

Number: szám
Sing: egyes szám
Plur: többes szám

Number[psor]: birtokos száma
Sing: egyes szám (pl. *kutyája*)
Plur: többe szám (pl. *kutyájuk*)

Person[psor]: birtokos személye
1: első személy (pl. *kutyám*)
2: második személy (pl. *kutyád*)
3: harmadik személy (pl. *kutyája*)

Number[psed]: birtok száma
Sing: egyes szám (pl. *kutyáé*)
Plural: többes szám (pl. *kutyáéi*)

PROPN: tulajdonnév
- ugyanaz, mint a főneveknél

ADJ: melléknév

Case: eset (ugyanaz, mint a főneveknél)
- Az -an/-en/-án/-én ragos melléknevek ragja Ess (esszívusz)
Degree: fok
Pos: alapfok (pl. *szép*)
Cmp: középfok (pl. *szebb*)
Sup: felsőfok (pl. *legszebb*)
Abs: szuperlatívusz (pl**.** *legeslegszebb)*
Number: szám
Sing: egyes szám
Plur: többes szám
NumType:
Ord: ordinális szám (pl. *első*)
Number[psor]: birtokos száma
Sing: egyes szám
Plur: többe szám
Person[psor]: birtokos személye
1: első személy
2: második személy
3: harmadik személy
Number[psed]: birtok száma
Sing: egyes szám
Plural: többes szám
VerbForm:
PartPast: befejezett melléknévi igenév (pl. *épült*)
PartPres: folyamatos melléknévi igenév (pl. *épülő*)
PartFut: beálló melléknévi igenév (pl. *épülendő*)

NUM: számnév
Case: eset (ugyanaz, mint a főneveknél)
NumType:
Card: kardinális (pl. *három*)
Frac: törtszám (pl. *háromnegyed*)
Dist: disztributív szám (pl. *három-három*)
Number:
Sing: egyes szám
Plur: többes szám
Number[psor]: birtokos száma
Sing: egyes szám
Plur: többe szám
Person[psor]: birtokos személye
1: első személy
2: második személy
3: harmadik személy
Number[psed]: birtok száma

Sing: egyes szám
Plural: többes szám

PRON: névmás
Case: eset (ugyanaz, mint a főneveknél)
Number: szám
Sing: egyes szám
Plur: többes szám
Person:
1: első személy
2: második személy
3: harmadik személy
Number[psor]: birtokos száma
Sing: egyes szám
Plur: többes szám
Person[psor]: birtokos személye
1: első személy
2: második személy
3: harmadik személy
Number[psed]: birtok száma
Sing: egyes szám
Plural: többes szám
Poss: birtokos névmás
Yes: igen (pl. *tiéd*)
PronType: névmásfajta
Prs: személyes vagy birtokos személyes névmás (pl. *te*)
Rcp: kölcsönös névmás (pl. *egymás*)
Int: kérdő névmás (pl. mi, milyen, mit, Ki, mire, hány, kik, kivel, melyik, mennyit)
Rel: vonatkozó névmás (pl. *aki*)
Dem: mutató névmás (pl. azt, az, ez, annak, arra, ezt, ennek, arról, ezzel, azzal)
Tot: általános névmás (pl. minden, mindenki, senki, valamennyi, bárki, semmi)
Ind: határozatlan névmás (pl. valaki, egyik, néhány, más, másik, valaki, mások)
Reflex: visszaható névmás
Yes: igen (pl. *magad*)

ADV: határozószó
PronType:
Neg: tagadószó
? Art: névelő

Int: kérdő névmás (-e kérdőpartikulánál kell használni)

Degree: fok

Pos: alapfok (pl. *későn*)

Cmp: középfok (pl. *később*)

Sup: felsőfok (pl. *legkésőbb*)

Abs: szuperlatívusz (pl**.** *legeslegkésőbb)*

PronType: ?

VerbForm:

Trans: határozói igenév (e-magyar)

(Conv: határozói igenév - az újabb verzióban így hívják)

DET: névelő

Definite:

Ind: határozatlan (pl. *egy*)

Def: határozott (pl. *a*)

PronType: névmásfajta

Art: névelő

ADP: névutó
PART: partikula („PARTs are those words in Hungarian that only function as preverbs and are not used elsewhere.”; ez csak a *meg* és az *utol* igekötő)
CONJ: mellérendelő kötőszó
SCONJ: alárendelő kötőszó
INTJ: indulatszó
PUNCT: központozás
SYM: szimbólumok
X: idegen szavak

Néhány dolog, amire érdemes figyelni:

- A **lemmákat** mindig a sztenderd helyesírásnak megfelelően adjuk meg. Ha például a szövegben *sziv, szivek* stb. szerepel, akkor a lemma *szív*.
- Ha hibásan **tulajdonnévként** (PROPN) lett annotálva egy szó, akkor a lemma nagy kezdőbetűjét is kicsire kell javítani.
- Az **igekötők** mindig ADV szófajúak, kivéve a *meg* és az *utol* igekötőt, ami mindig PART.
- Itt a **partikula** (PART) szófaji kategóriája teljesen máshogy értelmeződik, mint a magyar leíró nyelvtanban. PART címkét kizárólag a *meg* és az *utol* igekötő kap. Ennek a rendszernek a szemeteszsákja az ADV szófaji címke (a leíró nyelvtan partikulái nagyrészt ide kerülnek).

- A **segédigék** nagy része VERB szófaji címkét kap. Kizárólag a következő segédigék kapnak AUX címkét: *volna*, *fog*, *talál*, *szokott*
- **Igenevek:** A főnévi igenév szófaja VERB, lemmája pedig az igei alapalak. A határozói igenév szófaja ADV, lemmája pedig az igei alapalak, a melléknévi igenevek szófaja ADJ, lemmája pedig a melléknévi igenévi alak. Az igenévi jelleget a VerbForm jegy értékével jelezzük.
- A melléknevek esetében a **melléknévi igenevet** csak akkor kell feltüntetni a VerbForm jegy értékeként, ha az igéből képzett szóhoz kapcsolódik valamilyen, az igei használatból örökölt bővítmény vagy igemódosító (pl. *könyvet olvasó ember, a királyra felesküdött katona, szépen rajzoló gyerek*). Emellett folyamatos melléknévi igenevekként annotálhatók még azok az esetek, ahol az igéből képzett szó nem állandó, hanem egy entitás időleges/pillanatnyi tulajdonságára utal (pl. a meghaló ember).
- A **melléknév** + -an/-en/-án/-én ragos alakok (pl. *szépen, gyorsan*) melléknévnek számítanak, tehát ADJ szófaji címkét kapnak, lemmájuk a melléknév alapalakja. A ragot a Case=Ess (esszívusz) jelzi
- A **sorszámnevek** (pl. *harmadik*) ADJ címkét kapnak, minden más számnév NUM címkét kap. A sorszámnevek lemmája az alapalak (pl. első → egy, harmadik → három)
- A **névmásokat** az e-magyar hajlamos rossz névmástípusba sorolni. Ezeket bátran javítsátok.
- Az *alattam, alattad, előttünk* típusú **személyes határozószavakat** az e-magyar következetesen névutókként, ADP szófaji címkével annotálja, a morfológiai jegyek között pedig semmit nem ad meg. A szófaji címkét ne változtassátok meg, viszont a morfológiai jegyek közé rakjátok be a szám (Number) és a személy (Person) kategóriákat ebben a sorrendben (pl. Number=Sing|Person=3). E szavak lemmájaként a névutós alakot kell megadni (pl.: alatta → alatt, mellettem → mellett)
- Az **igék Definite** morfológiai jegye az igék határozott vs. határozatlan ragozására vonatkozik, régebbi terminológiával tárgyas vs. tárgyatlan ragozására. Csak akkor kerül ide Def érték, ha az igeragozás határozott/tárgyas. Minden más esetben Ind az érték.
- Ha egy magyar mondatban szerepel egy **idegen szó**, amely értelmesen illeszkedik a mondat magyar szintaxisába, akkor próbáljunk neki szófaji és morfológiai címkéket adni, akkor is, ha az idegen szó nincs ragozva valamilyen magyar ragozási paradigma alapján (főnév esetén például ebben az esetben case=Nom jeggyel látható el). Az X szófaji címkét lehetőleg akkor használjuk csak, ha több idegen szóból álló szövegrész (szókapcsolat, tagmondat) szavait kell felcímkézni.

- Az elsődleges jelentésükben kérdő, de az adott kontextusban **vonatkozó névmásokként** szereplő *mely*, *hol, mi, ki, miért* szavakat, *amely, ahol, ami, aki, amiért* alakokként lemmatizáljuk. Az egyéb eseteket vegyük fel a hibalistába, és beszéljünk róluk még.

Tulajdonság-érték (feature-value) párok feltüntetésének sorrendje:

- A tulajdonság-érték párokként feltüntetett morfoszintaktiai jellemzők listájában a tulajdonságok ábécé sorrendben követik egymást (kis- és nagybetű különbsége nem számít). Erre akkor kell figyelni, ha új tulajdonságot adunk hozzá a listához.

Egymásnak ellentmondó morfológia és szintaxis:

- Az olyan agrammatikus mondatoknál, ahol a szintaktikai szerkezet és a szó morfológiája ellentmond egymásnak, a morfológiai szerkezet alapján adjuk meg a morfoszintaktikai jellemzőket. (Pl. *András evett egy csoki* – a csoki esete Nom, nem pedig Acc)

Hibás tokenizálás:

Ha esetleg rossz a tokenizálás, és két token egy tokenkét jelenik meg, akkor az XML-be rakjatok be egy plusz <w> vagy <pc> elemet a megfelelő attribútumokkal. Amennyiben egy token hibásan két tokenként jelenik meg, akkor pedig töröljétek a felesleges <w> vagy <pc> elemet. A tokenizálás szempontjából javított <w> és <pc> elemek @type attribútumába írjátok be azt, hogy "token_problem". A bevezetett új tokenek @xml:id attribútumának értékeként adjatok egy olyan tetszőleges értéket, amely nem szerepel a fájlban (pl. w1000). Fontos, hogy két szónak vagy írásjelnek nem lehet ugyanaz az xml:id-je, mert az invalid XML-t eredményez, az viszont nem probléma, ha felborul a számozás sorrendje (pl. w101, w1000, w102), ezt utólag automatizáltan lehet javítani, ha szükséges.

Az e-magyar és az UDPiepe kimenetének különbségei (A Lírakorpuszhoz nem releváns):

- emagyar: határozói igenév Trans, UDPipe: Conv
- emagyar: csak Dat, UDPipe: Dat és Gen
- e-magyarnál bizonyos szavak ADV, míg UDpipe-nál PRON
- határozószói névmásoknál emagyar: PRON, UDPipe: ADV
- hol e-magyar: PRON vs udPipe: ADV

- meddig e-magyar PRON, UDPipe ADV
- igen
- -gat/-get képzős igék
- semmi
- volna
- …dik (melléknév)
- vala
- többi: e-magyar: melléknév
- e-magyar: a határozószó fokozása nem annotálódik