



Korpuszok létrehozása



Alapfogalmak

- **Korpusz: speciális célokra létrehozott, (gyakran tematikus) adatbázis – „szöveggyűjtemény”**
- **Annotáció: a szövegek nyelvi információval történő kézi jelölése (és kézi ellenőrzése)**
- **Gold standard (etalon) vs. silver standard: kézi vagy gépi jelölés**

Korpusztípusok

- **Egynyelvű**
- **Többnyelvű – párhuzamos korpusz: ugyanazok az adatok egynél több nyelven**
- **Beszédkorpusz: hanganyagok**
- **Írott nyelvi korpusz: szövegek**



Korpuszépítés

- **Mi a cél?**
- **Milyen szövegek kerüljenek bele?**
 - **Tematika (jog, irodalom...)**
 - **Nyelvi regiszterek (hivatalos, köznyelv, internetes nyelvhasználat...)**
 - **Homogén/heterogén**
 - **Milyen egyéb (meta)adatok? (idő, szerző...)**
- **Méret**
- **Nyelv**
- **Hozzáférhetőség (szerzői jogok, anonimizálás)**

Annotáció

- **Szöveg/dokumentum szintje**
 - Levél spam/nem spam
- **Mondat szintje**
 - Bizonytalan/tényszerű információt tartalmaz-e?
- **Szó/frázis szintje**
 - Morfológiai elemzés
 - Tulajdonnevek
- **Annotáció nélkül**
 - Szógyakoriság
 - Együtt előfordulás





Az annotáció típusa

- kézi
- félig automatikus: gépi úton bejelölt annotáció kézi javítása
- automatikus
- egyszeres: egy szövegen egy annotátor megy végig
 - olcsóbb
 - gyorsabb
- többszörös: egyazon szövegen több annotátor is teljes egészében végigmegy, egymástól függetlenül
 - időigényesebb
 - drágább
 - egyetértési arány mérése

Egyetértési arány

- az annotátorok mennyire értettek egyet (=mennyire jelöltek ugyanúgy) adott metrika szerint
 - Pontosság (accuracy)
 - F-mérték (pontosság – precision, fedés – recall)
 - Kappa
- az annotátorok által egyformán jelölt esetek arányát a gépi alkalmazások által elérhető felső határnak szokták tekinteni
- a feladat nehézségi fokának jelzése
- feladatfüggő!



Az annotáció formái

- Egy fájlban a szöveg és a jelölés (általában XML)
- Külön fájlban a szöveg és a jelölés (standoff/standalone)
- Előnyök/hátrányok:
 - Eredeti szöveg visszanyerése
 - Új szövegek hozzáadása
 - Szövegek törlése



<s id="Nepszava.24.2.1">Rövidtávú— féléves— kilátásait illetően a cégek egész évben októberben voltak a legoptimistábbak.

<choice>

<sic>

<w>Rövidtávú

<ana>

<humor><lemma>Rövidtávú</lemma><mscat>[X]</mscat></humor>

<msd><lemma>Rövidtávú</lemma><mscat>[X]</mscat></msd>

</ana>

<anav>

<humor><lemma>Rövidtávú</lemma><mscat>[X]</mscat></humor>

<msd><lemma>Rövidtávú</lemma><mscat>[X]</mscat></msd>

</anav>

</w>

</sic>

<corr>

<w>rövid

<ana>

<humor><lemma>rövid</lemma><mscat>[Afp-sn]</mscat></humor>

<msd><lemma>rövid</lemma><mscat>[Afp-sn]</mscat></msd>

</ana>

<anav>

<humor><lemma>rövid</lemma><mscat>[Afp-sn]</mscat></humor>

<msd><lemma>rövid</lemma><mscat>[Afp-sn]</mscat></msd>

</anav>

<anav>

<humor><lemma>rövid</lemma><mscat>[Nc-sn]</mscat></humor>

<msd><lemma>rövid</lemma><mscat>[Nc-sn]</mscat></msd>

</anav>

</w>

<w>távú

<ana>

<humor><lemma>távú</lemma><mscat>[Afp-sn]</mscat></humor>

<msd><lemma>távú</lemma><mscat>[Afp-sn]</mscat></msd>

</ana>

<anav>

<humor><lemma>távú</lemma><mscat>[Afp-sn]</mscat></humor>

<msd><lemma>távú</lemma><mscat>[Afp-sn]</mscat></msd>

</anav>

</w>

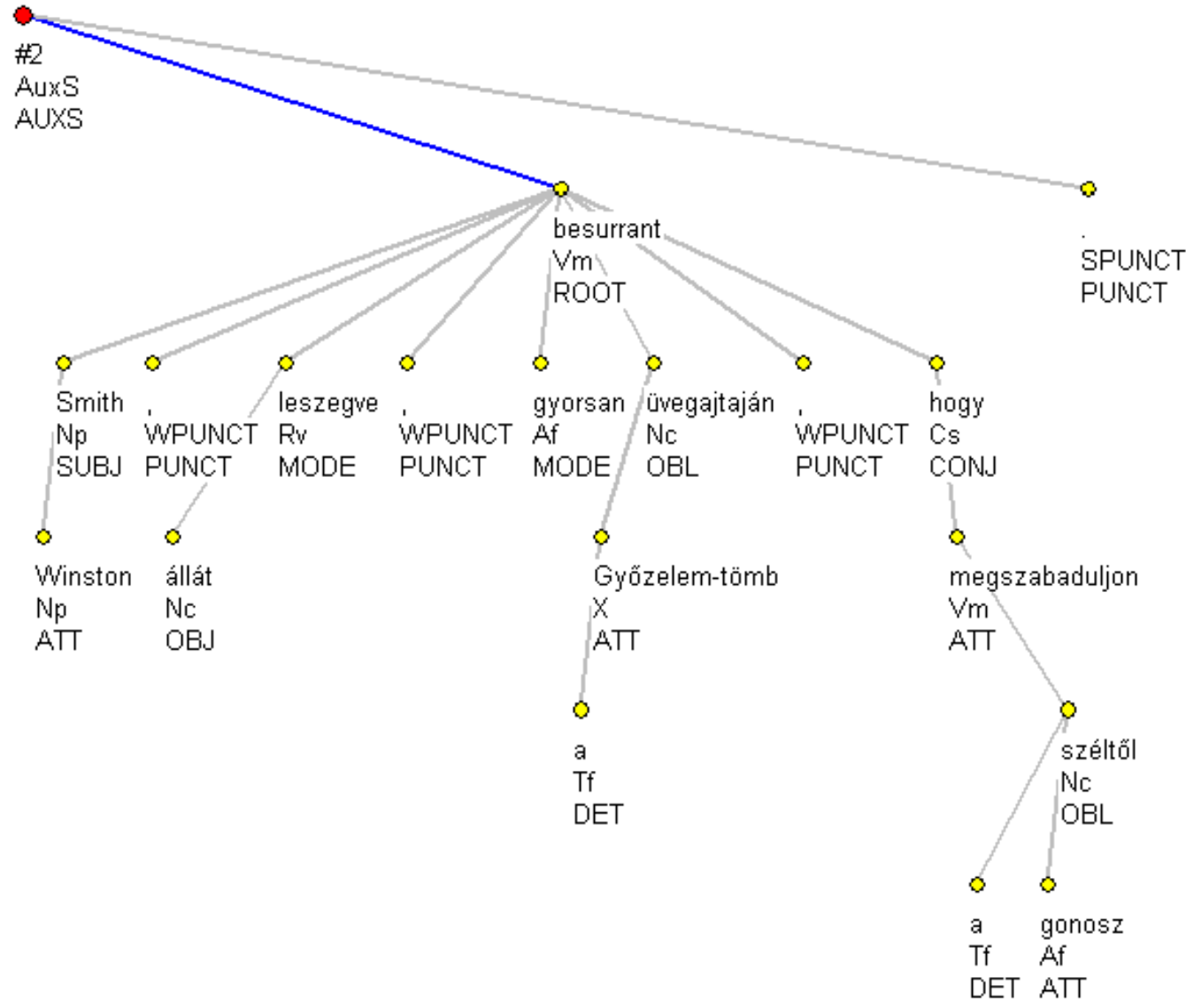
</corr>

</choice>



1	–	ROOT	ROOT	ELL	ELL	–	–	0	0
2	Japánban	Japán	Japán	N	N				
		SubPOS=p Num=s Cas=2 NumP=none PerP=none NumPd=none							
		SubPOS=p Num=s Cas=2 NumP=none PerP=none NumPd=none				1	1	OBL	
		OBL							
3	,	PUNCT	PUNCT	,	,	–	–	1	1
4	ahol	ahol	R	R		SubPOS=r Deg=none Num=none Per=none			
		SubPOS=r Deg=none Num=none Per=none	9	9		TLOCY	TLOCY		
5	1960-ban	1960	M	M					
		SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none							
		SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none				9	9	OBL	
		OBL							
6	közel	közel	R	R		SubPOS=x Deg=none Num=none Per=none			
		SubPOS=x Deg=none Num=none Per=none	7	7		MODE	MODE		
7	félmillió	félmillió	M	M					
		SubPOS=c Num=s Cas=n Form=l NumP=none PerP=none NumPd=none							
		SubPOS=c Num=s Cas=n Form=l NumP=none PerP=none NumPd=none				8	8	ATT	
		ATT							
8	válást	válás	N	N					
		SubPOS=c Num=s Cas=a NumP=none PerP=none NumPd=none							
		SubPOS=c Num=s Cas=a NumP=none PerP=none NumPd=none				9	9	OBJ	
		OBJ							
9	mondtak	mond	V	V		SubPOS=m Mood=i Tense=s Per=3 Num=p Def=n			
		SubPOS=m Mood=i Tense=s Per=3 Num=p Def=n				1	1	ATT	ATT
10	ki	ki	R	R		SubPOS=p Deg=none Num=none Per=none			
		SubPOS=p Deg=none Num=none Per=none	9	9		PREVERB	PREVERB		
11	,	PUNCT	PUNCT	,	,	–	–	9	9
12	1990-ben	1990	M	M					
		SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none							
		SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none				1	1	OBL	
		OBL							
13	már	már	R	R		SubPOS=x Deg=none Num=none Per=none			
		SubPOS=x Deg=none Num=none Per=none	15	15		MODE	MODE		
14	2,6	2,6	M	M					
		SubPOS=f Num=s Cas=n Form=d NumP=none PerP=none NumPd=none							
		SubPOS=f Num=s Cas=n Form=d NumP=none PerP=none NumPd=none				15	15	NUM	
		NUM							
15	milliót	millió	M	M					
		SubPOS=c Num=s Cas=a Form=l NumP=none PerP=none NumPd=none							
		SubPOS=c Num=s Cas=a Form=l NumP=none PerP=none NumPd=none				1	1	OBJ	
		OBJ							
16	.	PUNCT	PUNCT	.	.	–	–	0	0





Shadow_Riders.txt

The Shadow Riders, known as the in the original Japanese language version, are a fictional group of villains in the Yu-Gi-Oh! GX anime series, appearing between episodes 29-49.

Composed of seven duelists and their leader of varying origins and backgrounds who each have their own agendas, the Shadow Riders serve as the main antagonists of the series' first season, intent on resurrecting the Sacred Beasts.

However, one of them returns in the fourth and final season as the true mastermind behind the mysterious attacks that take place in Duel Academy and Domino City.

Shadow_Riders.txt.annotation

NE_ORG 4	17		
NE_MISC48	56		
NE_MISC116	128		
MWE_COMPOUND_NOUN	129	141	
SENT_BOUND	170	175	
NE_ORG 294	307		
NE_MISC394	407		
NE_MISC_SB	401	407	
MWE_LVC	527	537	
MWE_LVC_VERB	527	531	
MWE_LVC_NOUN	532	537	
NE_LOC 541	553		
NE_LOC 558	569		
NE_LOC_SB	565	569	
NE_ORG 576	589		
NE_PER 626	638		
NE_PER_SB	634	638	
NE_PER 691	702		
SENT_BOUND	794	803	
MWE_COMPOUND_NOUN	814	825	
MWE_COMPOUND_NOUN	855	872	
NE_MISC873	897		
SENT_BOUND	994	1002	

The **Shadow Riders**, known as the in the original **Japanese** language version, are a fictional group of villains in the **Yu-Gi-Oh! GX** anime series, appearing between episodes 29-49.

Composed of seven duelists and their leader of varying origins and backgrounds who each have their own agendas, the **Shadow Riders** serve as the main antagonists of the series' first season, intent on resurrecting the **Sacred Beasts**.

However, one of them returns in the fourth and final season as the true mastermind behind the mysterious attacks that **take place** in **Duel Academy** and **Domino City**.

Annotációs eszköz előnyei

- Grafikus kezelői felület
- Ember számára értelmezhetőbb
- Átláthatóbb
- Kisebbségi a hibázási arány



A korpuszépítés folyamata

- 1. Szövegek gyűjtése, gépi előkészítése**
- 2. Kézi annotálás**
 - kettős jelölés – egyetértés aránya
 - egyszeres jelölés
- 3. Az eltérések feloldása, ellenőrzés**
 - a kétféle annotáció közti eltérések egyértelműsítése
- 4. Záró munkálatok**
 - a korpusz végső formába hozása, formai hibák javítása, a korpusz publikálása



Kézi annotálás

- Nyelvi háttér kidolgozása
- Útmutató készítése
- Próbaannotáció végzése
(többszörösen / több emberrel)
- Jellemző hibák feltérképezése,
javítása
- Útmutató javítása
- Indulhat az annotáció...



A korpuszok felhasználhatósága

- Referencia
- Vizonyítási pont
- (Gépi tanuló) algoritmusok tanítása
- Algoritmusok tesztelése
- Nyelvészeti adatok gyűjtése

