Irányelvek tárgyszavazáshoz

KDK-SZTAKI MILAB PROJEKT

Tárgyszókészlet:

- Jelenleg van egy hierarchiába rendezett, 220 tárgyszót tartalmazó listánk, amit első körben az ELSST tezaurusz-fordításból, aztán az abból kiválogatott 600+ szavas listából, majd az abból szűkített 242-es szószedetből állítottunk elő, olyan kiegészítésekkel, amik nem szerepeltek a bővebb szószedetekben
- A hierarchiában minden szónak pontosan megjelölt helye van, ÉS minden szó csak egy helyen szerepel, legyen az alsó-, középső- vagy felsőszintű fogalom
- Ami a 600+-as listából nem került be a rövidített listába, azt megfeleltettük valaminek a 220-as listán
- A szókészlet elkészítésénél igyekeztünk áttekintően eljárni, néhány, a gyűjteményeink szempontjából hangsúlyos részt kiemelve. pl. holokauszthoz kapcsolódó kifejezések (de általánosságban elmondható, hogy egy ilyen hosszúságú listába csak alapszintű fogalmak férnek bele)

A tárgyszavazás céljának tisztázása

- mi a tárgyszavazás célja?
    - az, hogy olyan tárgyszavakat társítsunk a szakaszokhoz, amik jól leírják az azokban megjelenő, társadalomtudományos szempontból fontos tartalmakat - itt nem elsődleges cél a gép tanítása
        - a kiválasztott tárgyszavakhoz pedig olyan hívószavakat jelöljünk ki, amik jól utalnak a tárgyszavakra - ez a gép tanítása
    - minden szakaszt kezeljünk úgy, mintha az egy teljesen új, önálló szöveg lenne (bármilyen furcsán is hangzik) - technikailag így tudja értelmezni a gép
        - azaz, ha a családról már sokszor volt szó, és egy adott szakaszban újra szó van róla, akkor ott is tárgyszavazzuk (ez egyébként azért is fontos, mert a végén az, hogy milyen kiemelt tárgyszavakat ad a gép a szövegeknek, lehet, hogy épp azon is múlik majd, hogy egy adott tárgyszó hány szakaszhoz rendelődött hozzá - tehát nem mindegy, hogy valaki egyszer beszél a családjáról vagy folyamatosan)

Label Studio:

- **Feladat: az interjúk egyes szakaszaihoz tárgyszavakat rendelni, utóbbiakhoz a szövegből hívószavakat választani**
    - **Szakasz**: az interjúk szakaszokra bontva olvashatóak és tárgyszavazandóak
    - **Tárgyszó**: a fix tárgyszólista (KDK-Tezaurusz) egy eleme, ami leírja az adott szakasz egy-egy fontos tartalmi egységét

    - **Hívószó**: olyan, az adott szakaszban konkrétan (szó szerint) szereplő, a kódoló által kiválasztott kifejezés, ami jól utal egy-egy fontos tartalmi egységre, és kapcsolható a kiválasztott tárgyszóhoz

– A ***cél a hangsúlyos és társadalomtudományos szempontból releváns tartalmak tárgyszavazása*** (így pl. ha valaki egyszer említ valamit, de nem tűnik fontosnak, akkor ne tárgyszavazzuk, akkor sem, ha amúgy “fontos téma”, pl a holokausztról olvasott egy könyvet, és ezt megemlíti, de amúgy másról beszél)
– Minden esetben **a lehető legpontosabb szót/kifejezést** válassza mindenki, mind a tárgyszavak, mind a hívószavak közül:
    - ez adott esetben jelentheti azt, hogy ha a szövegben előrehaladva találunk jobb kapcsolatot, érdemes az előzőt kicserélni vagy az újat is hozzáadni
    - az értelmező megfeleltetéseket kerüljük, ami a gyakorlatban azt jelenti, hogy ha pl a megszökést/menekülést emlegeti az interjúalany, és a kontextusból kiderül, hogy börtönből megszökésről van szó, akkor meg kell keresni a legfontosabb, legpontosabb kifejezést, amihez nem értelmezés jelleggel adjuk a tárgyszót, pl az említett példa esetében a *börtön* lenne ideális, nem a *megszökés* (hiszen megszökni sok helyről/helyzetből lehet)

Irányelvek:

– egy szakaszhoz maximum 5 tárgyszó társuljon (lehet 0 is)
– egy tárgyszóhoz maximum 4 hívószó társuljon, de ha indokolt, társulhat több is
– egy hívószó egy konkrét szó, vagy egy konkrét szókapcsolat (2, maximum 3 egymás utáni szó) lehet, fél mondatok ne legyenek kijelölve, azzal az MI nem tud dolgozni
– ha nagyon nincs jó hívószó, de a szakasz tartalma miatt fontos lenne egy adott tárgyszó megadása, akkor nem kell hozzá hívószó
– ha a bejelölni akart hívószó több elemből áll (pl elváló ige, többszavas szókapcsolat...) és ezek az elemek nem közvetlenül egymás után állnak, akkor vagy ne jelöljünk hívószót csak adjuk meg a tárgyszót, vagy külön-külön jelöljük be a hívószó elemeit, és adjuk hozzájuk egyesével ugyanazt a tárgyszót, de leginkább kerüljük az ilyen eseteket, próbáljunk másik hívószót találni az adott tárgyszóhoz
– egy hívószó kapcsolódhat több tárgyszóhoz is, ha így logikus
– érdemes a tárgyszóval teljesen azonos hívószavakat is a tárgyszóhoz kapcsolni (tárgyszó=asszimiláció / hívószó=asszimiláció - ez így oké)

A kódolás menetére egy módszer – lépések:

1. a szakasz olvasása közben a fontosnak tűnő tartalmakhoz hívószavak kijelölése
2. menet közben / a szakasz végén azon hívószavak törlése, amik nem voltak igazán jók és lett helyettük jobb
3. a szakasz végén a hívószavak átnézése, azon hívószavak törlése, amelyek mégsem fontos hívószavak (mégsem fontos tartalmak)

4. a szakasz tartalmainak átgondolása (mik a fontos tartalmak) -> ki kell-e venni hívószót / van-e olyan tartalom, amihez nincs még hívószó (ha van, akkor keresni hozzá, ha nem lehet jót találni, akkor csak tárgyszó lesz hozzá, hívószó nem)
5. tárgyszavak kiválasztása a listából, amelyek jól leírják a megállapított tartalmakat
6. ha 5-nél több a tárgyszó, akkor annak eldöntése, hogy indokolt-e a több tárgyszó (indokolt, ha az átlagnál hosszabb a szakasz) -> ha indokolt, akkor lehet több, de akkor a számot a kódolótárssal közölni kell / ha nem indokolt, akkor az 5 legfontosabb tárgyszó kiválasztása
7. tárgyszavak véglegesítése
8. hívószavak és tárgyszavak párosítása
9. ellenőrzés – biztos minden párosítás megtörtént-e
10. ha van extremitás (pl.: van olyan fontos tartalom, amihez van tárgyszó, de nincs jó hívószó, stb, stb), akkor annak az ezt gyűjtő doksiban való jelzése

Technikai megjegyzések a kódoláshoz:

- **kódolás közben a megfelelő tárgyszó megtalálásához érdemes lehet az eredeti excelt is használni néha, abban kibontva látszik a hierarchia összes eleme, könnyebben átlátható, lehet benne keresni... - hasznos tud lenni!**
- gyűjtsük külön doksiban a konkrét extrém eseteket, pl.: adott tárgyszóhoz nincs jó hívószó, vagy messze vannak egymástól a hívószó elemei, pedig csak ezt lehetne a kiválasztott tárgyszóhoz hozzákapcsolni... - hogy lássuk, milyen ilyen esetek vannak, esetleg kitaláljunk erre jobb megoldást a mostaninál
- ha hiányzik nagyon egy tárgyszó a listából, akkor gyűjtjük a javaslatokat (a tárgyszólista esetleges későbbi javításához).
- submit gomb:
    - akkor érdemes véglegesíteni egy adott szakaszt a submit gombbal, ha az készen van
    - ugyan az update gombbal még lehet változtatni a kódoláson, de ha rányom az ember a submit-ra, akkor szabad kezet ad a SZTAKI-nak, hogy a szakaszhoz tartozó tárgy- és hívószavakat exportálja és dolgozni kezdjen velük
- ha a két kódoló elkészült a közös interjújával, akkor:
    - szólni a SZTAKI-nak (e-mail), akik exportálnak, összehasonlítanak, és visszajeleznek, hogy kell-e harmadik kódoló
    - az összehasonlítás csak a tárgyszavak alapján történik
    - ha nem kell harmadik kódoló, akkor a két eredeti kódoló készíti el a végleges kódolást (arany standard), amin a gép tanul majd
    - ha kell harmadik kódoló, akkor kijelölünk egy harmadik kódolót, és ő dönt az arany standardról

Technikai megjegyzések a Label Studiohoz:

- ha rákattint az ember egy, már kiválasztott hívószóra, akkor közvetlenül a szöveg fölött megjelenik egy zöldkeretes boxban a hozzá kapcsolódó tárgyszó - itt talán jobban látszik, hogy mihez mit rendeltünk, mint a jobboldali sávban
- ha egy alacsonyabb szintű tárgyszót választunk ki, akkor csak az fog megjelenni a kiválasztottak között, az exportált verzióban azonban (amivel a SZTAKI dolgozik majd), ott lesznek az adott szó fölé tartozó tárgyszavak is, mint az adott szakaszhoz tartozó tárgyszavak