Útmutató a magyar nyelvű COLA készítéséhez

1. A cél
    - A GLUE benchmark részét képező Corpus of Linguistic Acceptability előállítása magyar nyelvű anyaggal
    - A korpuszban magyar nyelvű mondatok szerepelnek, a mondatok címkéi a 0 (“nem elfogadható”) és az 1 (“elfogadható”)
2. A korpusz előállításának lépései
    2.1. Mondatgyűjtés

    Nyelvészeti munkákból elfogadhatósági ítéletes példák gyűjtése excel fájlba
        - Ahol valami csillagozott / kérdőjelezett mondat feltűnik, onnan az összes példamondatot kigyűjteni
        - Az információk, amiket gyűjtünk:
            - A mondat maga
            - Egy “x”, ha a példa nem egy teljes mondat (pl. **három kutyák*). Az első cellába kerüljön a “mondat” abban a formában, ahogy találtuk, majd következő lépésben formázzuk teljes mondattá
            - A jel: *, ?, ??, ?*, és ami még előfordul; illetve üres cella, ha elfogadhatónak van ítélve a mondat
            - A nyelvi jelenség, amelynek kapcsán előkerül
            - A forrás
            - Egyéb komment
        - Hogy gyűjtünk?
            - Csak alfanumerikus karaktereket gyűjtünk az excelbe, tehát az **Erzsi Imrétől a születésnapjára [$_{VP}$nyakláncot kapott], és Mari szintén [$_{VP}$nyakláncot kapott Imrétől a születésnapjára].* példa “Erzsi Imrétől a születésnapjára nyakláncot kapott… “ formában kerül az excelbe. Üres névmást jelentő zérókat és más dolgokat sem tüntetünk fel, azokat ignoráljuk.
            - Glosszázott példamondatokat is egyben, folyószövegként írunk le, tehát a *nyaklánc-ot* az “nyakláncot” lesz az excelben.
            - Ha zárójeles dolog, opció kerül elő, akkor két külön sort készítünk belőle. Pl. *Megnézzük (*a) Budapest hídjait* -> “Megnézzük Budapest hídjait” és “*Megnézzük a Budapest hídjait”
        - Amit nem gyűjtünk:
            - Ha egy mondat azért nem elfogadható, mert egy adott jelentést nem “jelenthet”, azt nem gyűjtjük, pl.: **Megver Péter.* ‘Péter megveri Pétert’ jelentésben (Str. M. Ny. Mondattan, 49. oldal)
            - Ha egy mondatban valamilyen nonszensz szó van.
            - Ha a fókusz helyzete miatt nem jó egy mondat.
            - Mondatpárokat. Ha az első mondat azért nem jó, mert a második az konkrétan az az adott mondat…

    2.2. Mondatjavítás

    2.2.1. Mondatkiegészítés
        - A legfontosabb alapelv: röviden és egyszerűen egészítünk

- Lehetőleg épp annyit teszünk hozzá a szerkezethez, hogy mondat legyen (legyen benne valamilyen predikátum).
    - Egyetlen igét
    - Vagy egy alanyt, ha az hiányzik.
- A semleges szórend megtartására törekszünk, ha lehetséges: nem teszünk be fókuszt, kérdést a mondatba, ha az eredeti szerkezetben sem volt.
- Ugyanannak a szerkezetnek különböző változatait (*annak a fiúnak a kalapja, a fiú kalapja, *a fiú a kalapja* stb.) ugyanúgy egészítjük ki.
- *Valaki*, *valami*: lecseréljük, ha muszáj (*eltörpül valami mellett*)