TIL KORPUSLARINI LINGVISTIK TEGLASH BOSQICHLARI

Elov Botir Boltayevich
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi mudiri
Xusainova Zilola Yuldashevna
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU tayanch doktoranti

Annotatsiya. Til korpusi elektron shaklda taqdim etiladigan *katta hajmli* va *strukturlangan matnlar to‘plami* sifatida qaraladi. Til korpusi *yozma* yoki *og‘zaki* materialni ifodalab, NLP tizimini mavjud resurslarni o‘rganishi uchun lingvistik tahlilni amalga oshirish lozim. Bugungi kunda zamonaviy NLP modellarini yaratish lingvistik bilimlarning akustika va prosodika, fonetika, orfografiya, morfologiya, leksikologiya, sintaksis, semantika, pragmatika va diskurs kabi spetsifikatsiyasini talab qiladi. NLP bilan bog‘liq lingvistik bilimlar leksik, sintaktik, semantik va pragmatik xususiyatlarni o‘z ichiga oladi. Ushbu maqolada lingvistik teglashning asosiy amallari haqida fikrlar keltiriladi. Maqolada orfografiya bilan bog‘liq teglash – tokenizatsiya va gap chegaralarini aniqlash, token yoki tokenlar to‘plamining so‘z turkumini belgilash (PoS teglash), sintaktik, semantik teglash, korpus matnlaridagi koreferensiyani aniqlash masalalari va ularni amalga oshirish usullari haqida umumiy ma’lumot berilgan.

Abstract. A linguistic corpus is considered a large and structured collection of texts presented in electronic form. In order for the language corpus to represent written or spoken material, it is necessary to perform a linguistic analysis in order for the NLP system to learn the available resources. Today, the creation of modern NLP models requires the specification of linguistic knowledge such as acoustics and prosody, phonetics, orthography, morphology, lexicology, syntax, semantics, pragmatics and discourse. Linguistic knowledge related to NLP includes lexical, syntactic, semantic and pragmatic features. This article provides an overview of the basic operations of linguistic tagging. The article deals with orthography-related tagging – tokenization and sentence boundary determination, token or set of tokens tagging (PoS tagging), syntactic, semantic tagging, coreference detection in corpus texts, and general information about their implementation methods is given.

Аннотация. Лингвистический корпус – это большая и структурированная коллекция текстов, представленная в электронном виде. Чтобы языковой корпус представлял письменный или устный материал, необходимо провести лингвистический анализ, чтобы система НЛП могла изучить доступные ресурсы. Сегодня создание современных моделей НЛП требует уточнения таких лингвистических знаний, как акустика и просодия,

фонетика, орфография, морфология, лексикология, синтаксис, семантика, прагматика и дискурс. Лингвистические знания, связанные с НЛП, включают лексические, синтаксические, семантические и прагматические особенности. В этой статье представлен обзор основных операций лингвистического тегирования. В статье представлен обзор тегирования, связанного с орфографией – токенизация и определение границ предложения, токенизация или набор токенов (PoS-тегирование), синтаксическое, семантическое тегирование, обнаружение кореференции в корпусных текстах и методы их реализации.

Kalit so‘zlar: *til korpusi, lingvistik teglash, morfologik tahlil, tokenlash, lemmalash, POS teglash, sintaktik tahlil, semantik tahlil, NER, teglangan korpuslar*.

Kirish

Tabiiy tilni kompyuter yordamida tadqiq qilish va qayta ishlash, odatda, to‘rt jihatni: formallashtirish, algoritmlash, dasturlash va amaliyotga qo‘llashni o‘z ichiga oladi.

Birinchidan, tilshunoslik nuqtayi nazaridan o‘rganiladigan muammolarni **formallashtirish**, tilning ma’lum bir matematik shaklda qat’iy va muntazam ravishda ifodalanishi uchun formal modelni ishlab chiqish kerak.

Ikkinchidan, ushbu qat’iy va muntazam matematik shaklni algoritm shaklida ifodalash talab etiladi: bu jarayon **algoritmlash** deb ataladi.

Uchinchidan, NLPning turli amaliy tizimlarini shakllantirish uchun algoritmga asoslangan kompyuter dasturini yozish, uni kompyuterda amalga oshirish kerak: bu jarayonni **dasturlash** deb atash mumkin.

To‘rtinchidan, o‘rnatilgan NLP tizimini foydalanuvchining ehtiyojini qondirish, sifat va ish faoliyatini doimiy ravishda yaxshilab borish uchun baholash lozim; bu jarayonni **amaliylashtirish** deb atash mumkin.

Korpus matnlarini lingvistik teglash dastlab lingvistik nazariyalarni yoki bugungi kunda ma’lum bo‘lganidek, til korpuslarni ishlab chiqish va tahlil qilish uchun ma’lumot berish uchun amalga oshirildi [Hovy, Lavid, 2010; Arista, 2022]. Matnlarni “qo‘lda” teglash uchun ko‘p vaqt va kuch talab qilinadi. Biroq so‘nggi o‘ttiz yil ichida hisoblash texnikalarning quvvati va katta hajmdagu ma’lumotlarni saqlash imkoniyati hosil bo‘lgach, katta hajmdagi til kopruslarini ishlab chiqish va avtomatik teglash sohasidagi bir qator yutuqlarga erishildi [Xusainova, 2022; Bi, 2018]. Matnlardan iborat katta hajmli til korpuslari ustida turli lingvistik va amaliy tadqiqotlar olib borildi [Xusainova, 2023]. Tabiiy tilni yangi texnologiyalar asosda tadqiq qilish va, eng muhimi, ishonchli statistik modellarni ishlab chiqish uchun lingvistik teglangan matn va til korpusiga asoslangan tabiiy tilni qayta ishlash (NLP) sohasiga muhim xizmat qiladi.

Lingvistik teglash tavsiflovchi yoki analitik belgilarni til ma’lumotlari bilan bog‘lashni o‘z ichiga oladi. Strukturlanmagan (qayta ishlanmagan) ma’lumotlar matnli yoki har qanday manba yoki janrdan olingan yoki vaqt funksiyalari (audio,

video va/yoki fiziologik yozuvlar) shaklida bo‘lishi mumkin. Teglar barcha turdagi transkripsiyalarni (fonetik xususiyatdan nutqiy qurilmagacha), POS teglash, ma’no belgilari, sintaktik tahlil, NER obyektlar, semantik rol belgilari, vaqt va hodisalarni aniqlash, so‘zlarning sintaktik zanjirlarini, nutq darajasini o‘z ichiga olishi mumkin [Xusainova, 2022; Bi, 2018; Elov, Hamroyeva, Axmedova, 2023]. Lingvistik teglashning eng muhim komponenti bu tegishli teg birligi (masalan, tovush, token yoki so‘z, birikma, fragment, hujjat) bilan bog‘lanishi kerak bo‘lgan teglar va tegishli xususiyatlarni belgilaydigan teglash sxemasi hisoblanadi.

Bugungi kunda korpus matnlarini lingvistik teglashni qo‘lda yoki yarim avtomatik tarzda amalga oshirish mumkin. NLP bilan bog‘liq lingvistik bilimlarning jihatlari haqida turli xil qarashlar mavjud. Umuman olganda, NLP tadqiqotchilarining aksariyati NLP bilan bog‘liq lingvistik bilimlar, hech bo‘lmaganda, **leksik, sintaktik, semantik** va **pragmatik** xususiyatlarni o‘z ichiga olishi kerak, deb hisoblashadi.

Har bir xususiyat lingvistik ma’lumotni farqli yo‘llar bilan uzatadi. Masalan, leksik xususiyat so‘z darajasidagi asosiy tarkibiy qismlar (masalan, morfema) va uning flektiv shakllari haqidagi bilimlarni qamrab olishi mumkin; sintaktik xususiyat ma’lum bir tildagi so‘z yoki so‘z birikmalarining gap hosil qilishini o‘z ichiga oladi; semantika ma’lum so‘zlar yoki gaplarga qanday ma’no berishni; pragmatika suhbatda nutq markazidagi o‘zgarishlarni, berilgan kontekstdagi gap ma’nosini izohlash haqida bilimlarni o‘z ichiga oladi.

Yuqorida keltirilgan to‘qqizta xususiyat, asosan, lingvistik bilimlarni o‘z ichiga oladi, chunki NLP, asosan, lingvistik muammo deb hisoblanadi. Biroq lingvistik bilimga qo‘shimcha ravishda NLP *axborot texnologiyalari*, *matematika*, *psixologiya*, *falsafa*, *statistika* va *biologiya* sohalardagi boshqa bilimlarni ham o‘z ichiga olishi mumkin.

Tilshunoslikda lisoniy hodisalarni tahlil qilish va tavsiflash, odatda, bir necha alohida bosqichlarda amalga oshiriladi. Tilning turli tovushlari **fonetika**da tavsiflansa, yozuv tizimi **orfografiya**da o‘rganiladi. **Morfologiya** so‘zning shakllanishi, o‘zgarishini, **sintaksis** so‘zlarning joylashuvi, ularning so‘z birikmasi va gapga aylanishini aniqlaydi. **Semantika** so‘z *(leksik semantika)*, birikma hamda gapning ma’nosini *(kompozitsion semantika)* tahlil qiladi. So‘z va so‘z birikmalarining turli sharoit va vaziyatda o‘ziga xos ma’no anglatishini **pragmatik tahlil** natijasida o‘rganish mumkin. Shaxs va narsalar qanday qilib mavzu sifatida kiritilishi va keyinchalik qanday tilga olinishi **diskurs tahlil**ning predmeti sanaladi.

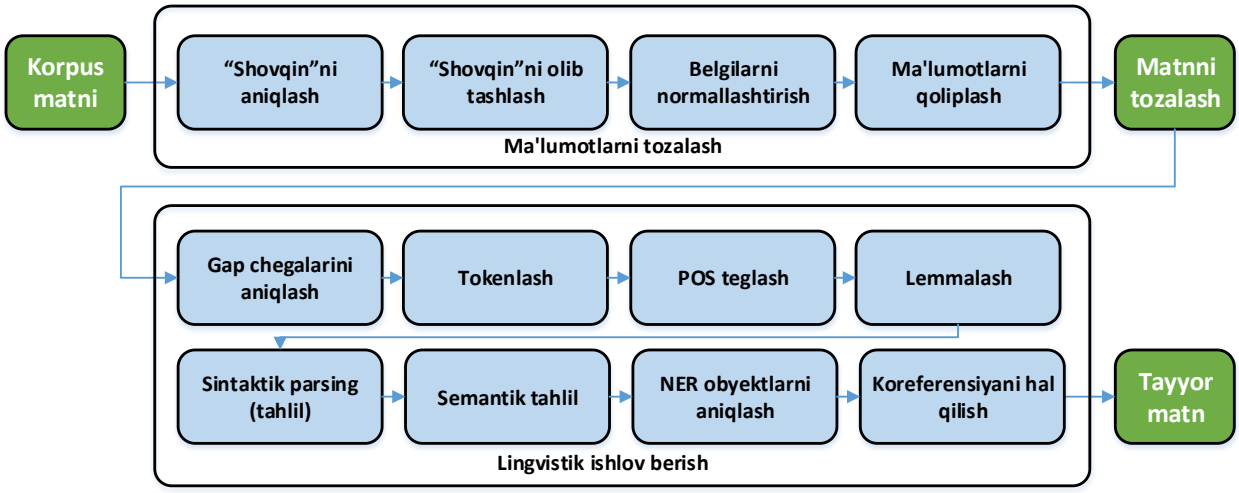
Fonetika va **orfografiya** eng kichik birliklar (alohida tovush va harf) bilan shug‘ullanadi. **Morfologiya, sintaksis** va **semantika** o‘rta hajmdagi birlik(so‘z, so‘z birikmasi va gap)larni o‘rganadi. **Diskurs** va **pragmatika**ning predmeti yuuqorida keltirilganidek eng katta birliklar (*paragraf* va *dialog*)dir.

Bugungi kunda tilshunoslikda matnni teglashning zamonaviy usullari, shuningdek, turli xil teglash vazifalarini quyidagi – 1-jadval va 1-rasmda

ko‘rsatilganidek, o‘xshash qatlamlar to‘plamiga joylashtirish mumkin bo‘lgan turli bosqichlarga ajratish mumkin.

2-jadval. Lingvistik teglash bosqichlari

№	Teglash usuli	Tavsif
1.	Gap chegalarini aniqlash (sentence boundaries)	matnni gaplarga ajratish
2.	Tokenlash (tokenization)	matnni so‘zlarga ajratish
3.	Lemmalash	so‘zshaklni lemma (lug‘atdagi shakli)ga keltirish
4.	POS teglash (part-of-speech tagging)	so‘zlarning turkumlari bilan aniqlash va belgilash
5.	Sintaktik parsing (tahlil)	gapdagi tarkibiy so‘z birikmalarini tahlil qilish
6.	Semantik parsing (tahlil)	predikat-argument munosabatlarini belgilash (labeling predicate-argument relations)
7.	NER obyektlarini aniqlash (named entity recognition)	atoqli otlarni aniqlash va belgilash
8.	Koreferensiyani hal qilish (coreference resolution)	matndagi bir xil obyektlarga havolalarni bog‘lash



1-rasm. Korpus matnlarini teglashning zamonaviy usullari

Biroq til korpuslarini teglashda bajariladigan vazifalar bilan tilshunoslik nazariyasidagi tavsif bosqichlari o‘rtasida faqat umumiy muvofiqlik mavjud. Ushbu maqolada matnlarni teglashning zamonaviy yondashuvlari ko‘rib chiqiladi.

Orfografiya bilan bog‘liq teglash vazifalari **tokenlash** va gap chegaralarini aniqlashdir. Ushbu vazifalar matnni alohida **so‘z (belgi)lar** va alohida **gaplarga**

ajratishda qo‘llanadi. Odatda, bu ikki vazifaning qaysi biri oldin bajarilishi muhim emas, lekin ikkala vazifa ham keyingi bosqichdagi vazifalarni bajarishdan oldin bajarilishi lozim.

Matn alohida tokenlarga ajratilgandan so‘ng har bir token yoki tokenlar to‘plamini **so‘z turkumi** (ot, fe’l, olmosh va boshqalar) bilan belgilash mumkin. Bu vazifa NLPda **so‘z turkumini aniqlash va teglash** yoki **PoS teglash**dir. PoS teglash morfologik tahlil sanaladi, chunki o‘zakka qo‘shilgan grammatik shakl orqali so‘z turkumini aniqlashga yordam beradi. Misol uchun, morfologik teglash jarayoni *-gan, -di* qo‘shimchalarining fe’llar bilan qo‘llaniladigan morfologik birlik ekanligini anglatadi (istisnolar ham mavjud).

Matn gaplarga bo‘lingandan so‘ng har bir gapdan *otli birikma* va *fe’lli birikma* kabi tarkibiy so‘z birikmalari aniqlanadi. Bu tahlil sintaktik tahlil bo‘lib, **lingvistik daraja (sintaksis)** va **annotatsiya vazifasi (sintaktik parsing)** o‘rtasida aniq muvofiqlik mavjud. Bugungi kunda lingvistik tahlil bosqichi bo‘yicha ko‘plab lingvistik nazariyalar mavjud bo‘lib, o‘zbek tili uchun hozirda keng qo‘llaniladigan **so‘z birikmalari modeli** yondashuvidan foydalanish maqsadga muvofiq.

Lingvistik tavsifning yuqori darajalari hisoblangan **semantika, pragmatika** va **diskursda** turli nazariyalar mavjud. Bugungi kunda o‘zbek tili korpusini ushbu darajalar asosida lingvistik teglashdan NLPning amaliy vazifalarida foydalanish uchun aniq konsensus topish qiyin. Shu sababli til korpusining semantik yoki pragmatik tahlili haqida o‘zbek tilidagi ilmiy tadqiqotlar deyarli amalga oshirilmagan. Lingvistik tavsifning diskurs darajasiga mos keladigan lingvistik teglash vazifalari haqida ayrim ishlar mavjud. Til korpusi matnlaridan **NER (Named Entity Recognition) obyektlarini aniqlash** – bu matndagi atoqli otlarni aniqlash va ularni teglash vazifasi (shaxs, tashkilot, joy nomlari va boshqalar) hisoblanadi. **Koreferensiyani hal qilish** – matndagi aynan bir obyektni atovchi yoki ularga bir xil obyektlarga ishora qilishini, tegishli ekanligini aniqlash vazifasi.

Xulosa

NLPga asoslangan ilovalar(axborot tizimlari)ni ishlab chiqish uchun NLP tizimni mavjud ma’lumotlarni o‘rganishini ta’minlash kerak. Ushbu amalni til korpusi vositasida amalga oshirish mumkin. Til korpusi elektron shaklda taqdim etiladigan *katta hajmli* va *strukturlangan matnlar to‘plami* sifatida qaraladi. Til korpusi *yozma* yoki *og‘zaki* materialni ifodalab, NLP tizimini mavjud resurslarni o‘rganishi uchun lingvistik tahlilni amalga oshirish lozim. Korpus NLP tizimlarining asosidir. Ular AI va mashinali o‘rgatish tizimlarini o‘rgatish uchun ishlatiladi. Ular axborot hayotiy davrlarini modellashtirish, bashorat qilish uchun keng va xilma-xil ma’lumotlar to‘plamini taqdim etadi. Korpus NLP tizimini tabiiy tilni boshqarish va sharhlash uchun tayyorlaydi, bu esa odamlar bilan tabiiy tilda oson muloqot qilish imkonini beradi.

Til korpusi – bu tilning haqiqiy foydalanuvchilari tomonidan ishlab chiqarilgan, so‘z, ibora va, umuman, til qanday ishlatilishini tahlil qilish uchun ishlatiladigan juda katta matnlar to‘plami. U tilshunos, leksikograf, ijtimoiy olim,

gumanitar fanlar, tabiiy tillarni qayta ishlash bo‘yicha mutaxassislar va boshqa ko‘plab sohalarda qo‘llaniladi. Korpus, shuningdek, dasturiy ta’minotni ishlab chiqishda ishlatiladigan turli til ma’lumotlar bazalarini yaratish uchun ishlatiladi, masalan, *bashoratli klaviatura, imloni tekshirish va tuzatish, matn/nutqni tushunish tizimlari, matndan nutqqa modullar, mashina tarjimasi tizimlari* va boshqalar.

Til korpusi foydalanuvchilar uchun to‘liq foydali bo‘lishi uchun uni teglash kerak. Bugungi kunda dunyodagi mashur korpuslar turli mezonlar bo‘yicha teglangan.

Lingvistik teglash – bu qaror qabul qilish maqsadida kompyuterda o‘qiladigan ma’lumotlarni uning ma’nosiga bog‘lash jarayoni. Texnik jihatdan, u tildagi murakkab naqshlarni aniqlash uchun hissiyotlarni tahlil qilish yoki NLP ilovalari tomonidan ishlatilishi mumkin bo‘lgan lingvistik metama’lumotlarga ega matnga izoh berishni o‘z ichiga oladi.

So‘nggi bir necha yil ichida lingvistik tahlillar onlayn korpuslar orqali amalga oshirilmoqda. Masalan, Sketch Engine 38 milliard so‘zni o‘z ichiga olgan English Web 2020 (enTenTen20) kabi eng mashhur ochiq korpuslar xizmatini taqdim etadi. Shunisi e’tiborga loyiqki, ijtimoiy tarmoqlar va axborot tizimlaridagi katta hajmdagi strukturlanmagan ma’lumotlar faqat matn emas, balki turli formatlarda (audio, video va boshqalar) ham bo‘lishi mumkin. Natijada, lingvistik teglash matn tahlili bilan chegaralanib qolmaydi. Teglar turli formatlarga qo‘llanilishi mumkin: *transkripsiya, vaqt belgisi, nutq sharhi, ma’no teglari* va boshqalar.

Tabiiy tilni tushunishga (natural language understanding, NLU) asoslangan ko‘plab NLP ilovalarini ishlab chiqish uchun, til korpuslarini shakllantirish va ularni teglash lozim. Shu sababli mualliflar tomindan o‘zbek tili korpusni lingvistik teglash turlari hisoblangan, *fonetik, morfologik, sintaktik* va *semantik teglar*ning o‘zbek tiliga mos to‘plami shakllantirildi hamda korpusga qo‘llandi.

Strukturlanmagan matnlarni tozalash NLP vazifasi orqali matn tahlil qilish sifati va aniqligini oshirish uchun muhim qadamdir. Matnni tozlash jarayoni: imlo va formatlashdagi nomuvofiqliklarni bartaraf etish orqali matnni kichik harflarga aylantirish bilan birga maxsus belgi, raqam va nomuhim so‘zlar kabi ahamiyatsiz (yoki ortiqcha) ma’lumotlarni olib tashlashni o‘z ichiga oladi. Matnni tozalash, shuningdek, imlo xatolarni qayta ishlash, so‘zlarni o‘zak shakliga keltirish (lemmatizatsiya) va matnni kodlash muammolarini hal qilish yechimlarini taklif qiladi. Ijtimoiy tarmoqlar, onlayn servislarda hosil qilinadigan katta hajmdagi ma’lumotlar, matn terishdagi xatolik, nomuvofiq so‘zlar, sheva unsurlarini misol sifatida keltirish mumkin. Tozalanmagan matnlar NLP modeli ish jarayoniga salbiy ta’sir o‘tkazadi.

Ushbu maqolda matnni teglashning zamonaviy usullari hisoblangan: gap chegalarini aniqlash, tokenlash, lemmalash, POS teglash, sintaktik tahlil, semantik tahlil, NER obyektlarini aniqlash va koreferensiyani hal qilish kabi NLP vazifalari yechimlari o‘zbek tili leksik birliklari misolida keltirildi. Maqolada keltirilgan teglash usullaridan turli NLP ilovalarida, masalan, *matnni tushunish, ma’lumot*

olish, hissiyotlarni tahlil qilish, imloni tekshirish, hujjatlarni umumlashtirish va *mashina tarjimasida* foydalanish mumkin.

Foydalanilgan adabiyotlar:

1. Hovy, E., & Lavid, J. (2010). Towards a ‘Science’ of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, *22*(1).

2. Arista, J. M. (2022). Toward the morpho-syntactic annotation of an old english corpus with universal dependencies. Revista de Linguistica y Lenguas Aplicadas, 17. https://doi.org/10.4995/rlyla.2022.16787

3. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi – Toshkent, 2022. №.1. – B.154-163.

4. Bi, P. (2018). Handbook of Linguistic Annotation. Journal of Quantitative Linguistics. https://doi.org/10.1080/09296174.2018.1424495

5. Xusainova Z.Y. BPE algoritmi asosida tokenizatsiya jarayonini amalga oshirish // O‘zbekiston milliy universiteti xabarlari jurnali. Toshkent, 2023. №1/3/1. – B.296-298.

6. Boltayevich, E. B., Mirdjonovna, H. S., & Ilxomovna, A. X. (2023). Methods for Creating a Morphological Analyzer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13741 LNCS. https://doi.org/10.1007/978-3-031-27199-1_4

7. Elov B., Hamroyeva Sh., Xudayberganov N., Yodgorov U., Yuldashev A. Pos tagging of Uzbek texts using hidden Markov models (HMM) and Viterbi algorithm. O‘zMU xabarlari. Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti ilmiy jurnali. 2023 yil Maxsys son.

8. Gries, S. Th., & Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. In Handbook of Linguistic Annotation. https://doi.org/10.1007/978-94-024-0881-2_15

9. Finlayson, M. A., & Erjavec, T. (2017). Overview of Annotation Creation: Processes and Tools. In Handbook of Linguistic Annotation. https://doi.org/10.1007/978-94-024-0881-2_5

10. Elov, B.B., Hamroyeva, Sh.M., Abdullayeva, O.X., Husainova, Z.Y., Xudoyberganov, N.U. 2023. “Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)”. O‘zbekiston: til va madaniyat 2: 6-39

11. Arista, J. M. (2022). Toward the morpho-syntactic annotation of an old english corpus with universal dependencies. Revista de linguistica y Lenguas Aplicadas, 17. https://doi.org/10.4995/rlyla.2022.16787