

KOMPYUTER LINGVISTIKASIDA KORPUS VA PARSING MUAMMOLARI

Choriyeva Hilola Murod qizi,

Toshkent davlat Sharqshunoslik universiteti magistranti

Email: hilolachoriyeva@gmail.com

DOI: <https://doi.org/10.5281/zenodo.18785708>

Annotatsiya. Mazkur maqolada kompyuter lingvistikasining muhim yo'nalishlaridan bo'lgan lingvistik korpuslar va sintaktik parsing jarayonlari, ularning nazariy asoslari hamda amaliy muammolari yoritiladi. Korpus yaratishdagi muammolar - ma'lumotlar tanlovi, belgilash (annotatsiya), muvozanat va reprezentativlik masalalari tahlil qilinadi. Shuningdek, parsing jarayonida uchraydigan sintaktik noaniqliklar, morfologik murakkabliklar va tabiiy tillarning strukturaviy xilma-xilligi muammolari ko'rib chiqiladi. Maqolada o'zbek tili misolida mavjud muammolar va ularni bartaraf etish yo'llari ham taklif etiladi.

Kalit so'zlar: kompyuter lingvistikasi, lingvistik korpus, parsing, sintaktik tahlil, annotatsiya, tabiiy tilni qayta ishlash.

Аннотация. В данной статье рассматриваются лингвистические корпуса и процессы синтаксического парсинга как одни из важнейших направлений компьютерной лингвистики, их теоретические основы и практические проблемы. Анализируются трудности создания корпуса - отбор данных, разметка (аннотация), вопросы сбалансированности и репрезентативности. Также рассматриваются проблемы синтаксической неоднозначности, морфологической сложности и структурного разнообразия естественных языков, возникающие в процессе парсинга. В статье на примере узбекского языка предлагаются возможные пути решения существующих проблем.

Ключевые слова: компьютерная лингвистика, лингвистический корпус, парсинг, синтаксический анализ, аннотация, обработка естественного языка.

Annotation. This article examines linguistic corpora and syntactic parsing as key components of computational linguistics. It analyzes major challenges in corpus construction, including data selection, annotation, representativeness, and balance. Particular attention is paid to parsing problems such as syntactic ambiguity, morphological complexity, and structural diversity of natural languages. The paper also highlights current issues and future prospects of corpus linguistics and parsing, with a focus on the Uzbek language.

Keywords: computational linguistics, linguistic corpus, parsing, syntactic analysis, annotation, natural language processing.

Kirish. Kompyuter lingvistikasi tabiiy tillarni kompyuter yordamida modellashtirish, tahlil qilish va qayta ishlash bilan shug'ullanuvchi fan sohasi hisoblanadi. Bugungi kunda mashina tarjiması, nutqni tanish, matnni avtomatik tahlil qilish hamda sun'iy intellekt tizimlarining rivojida lingvistik korpuslar va parsing texnologiyalari muhim ahamiyat kasb etmoqda.

Lingvistik korpus - ma'lum mezonlar asosida tanlangan, elektron shaklda saqlanadigan va maxsus annotatsiyalangan matnlar majmuasidir. Parsing esa kompyuter yordamida gapning sintaktik tuzilishini avtomatik aniqlash jarayonidir.

Mazkur tadqiqotning maqsadi kompyuter lingvistikasida korpus va parsing bilan bog'liq nazariy hamda amaliy muammolarni tahlil qilish va o'zbek tili uchun istiqbolli yechimlarni taklif etishdan iborat.

Adabiyotlar tahlili. Kompyuter lingvistikasi va tabiiy tilni qayta ishlash bo'yicha fundamental tadqiqotlar orasida Daniel Jurafsky va James H. Martin tomonidan yaratilgan "*Speech and Language Processing*" asari alohida o'rin tutib, bu manbada NLP tizimlarining nazariy asoslari, parsing modellari va statistik yondashuvlar keng yoritilgan.

Korpus lingvistikasi metodologiyasi Tony McEnery va Andrew Hardiening "*Corpus Linguistics: Method, Theory and Practice*" asarida tizimli bayon etilgan. Statistik NLP asoslari esa Christopher D. Manning va Hinrich Schützening "*Foundations of Statistical Natural Language Processing* kitobida chuqur tahlil qilingan".

Shuningdek, o'zbek tilshunosligi grammatikasi G'. Abdurahmonov tomonidan yoritilgan bo'lib, bu tadqiqotlar o'zbek tilida avtomatik tahlil tizimlarini yaratishda nazariy asos bo'lib xizmat qiladi.

Tadqiqot metodologiyasi. Tadqiqot davomida ilmiy manbalar tahlil qilindi, korpus lingvistikasi va parsing modellari o'rganildi hamda mavjud muammolar tizimlashtirildi.

Shuningdek, korpus yaratish bosqichlari, annotatsiya jarayoni, sintaktik tahlil mexanizmlari muammolari ilmiy jihatdan tahlil qilindi.

Tahlillar va natijalar. Bugungi kunda korpus lingvistikasi tillarni chuqur o'rganishda muhim ahamiyat kasb etmoqda. Korpuslar til birliklarining real qo'llanilishini tahlil qilish, til qonuniyatlarini aniqlash hamda avtomatik tizimlarni o'qitishda asosiy manba sifatida xizmat qiladi. Shu bois, korpus yaratish va undan foydalanish jarayoni kompyuter lingvistikasida dolzarb ilmiy yo'nalishlardan biri hisoblanadi. Korpuslarning asosiy turlari quyidagilardan iborat:

umumiy (milliy) korpuslar;

mualliflik korpuslari;

parallel korpuslar;

ta'limiy korpuslar;

lingvistik korpuslar.

Biroq korpus yaratish jarayoni murakkab va ko'p bosqichli bo'lib, bir qator muammolarni keltirib chiqaradi: korpus tilning barcha uslub va janrlarini yetarli darajada qamrab olishi kerak, aks holda, tadqiqot natijalari biryoqlama bo'lib qoladi. Annotatsiya muammolari - morfologik, sintaktik va semantik belgilash jarayonlari ko'p vaqt va resurs talab qiladi. Annotatsiyaning avtomatlashtirilishi esa xatoliklar ehtimolini oshiradi. Tilga xos murakkabliklar - agglyutinativ tillar, jumladan o'zbek tili uchun so'z shakllarining ko'pligi korpus yaratishni yanada qiyinlashtiradi.

Kompyuter lingvistikasida lingvistik korpuslar va parsing muammolari hozirgi kunda dolzarb muammolardan biri, chunki bunday korpuslarda uchraydigan orfografik beqarorlik, OCR (optik belgilarni tanish) xatolari, gap chegaralarini noto'g'ri aniqlash va morfologik teglashdagi noaniqliklar sintaktik tahlil natijalariga bevosita ta'sir ko'rsatadi. Ayniqsa, katta hajmli tarixiy korpuslar, masalan, 1665–1869-yillar oralig'idagi ilmiy matnlarni o'z ichiga olgan Royal Society Corpus asosida olib borilgan tadqiqotlar shuni

ko'rsatadiki, zamonaviy ingliz tili uchun mo'ljallangan parserlar bunday matnlarni past aniqlikda tahlil qiladi. Tarixiy ilmiy matnlarni parsing qilishda aniqlangan xatolarning sezilarli qismi bevosita sintaktik model kamchiliklari bilan emas, balki avvalgi bosqichlarda yuzaga kelgan texnik va lingvistik xatolar - imlo variantlarining ko'pligi, noto'g'ri tokenizatsiya, gap oxirini aniqlashdagi xatoliklar hamda POS-teglashdagi noaniqliklar bilan bog'liq bo'lib, ularning barchasi to'g'ri sintaktik daraxtlarning qurilishiga jiddiy to'sqinlik qiladi. Shu sababli, tarixiy va zamonaviy korpuslar asosida olib boriladigan kompyuter-lingvistik tadqiqotlarda parsing, anafora va koreferensiya muammolarini kompleks yondashuv asosida hal qilish, ya'ni orfografik normalizatsiya, xatolarni avtomatik tuzatish, davrga mos parserlarni qayta o'qitish hamda sifatli annotatsiya standartlarini ishlab chiqish muhim ilmiy vazifa sifatida namoyon bo'ladi.

Kompyuter lingvistikasi doirasida parsing - ya'ni gapning sintaktik tuzilishini avtomatik aniqlash jarayoni - tabiiy tilni qayta ishlashning eng murakkab bosqichlaridan biri hisoblanadi. Ayniqsa, o'zbek tili kabi agglyutinativ va erkin so'z tartibiga ega tillarda parsing jarayoni bir qator nazariy hamda amaliy muammolar bilan tavsiflanadi. Parsingdagi asosiy muammolar quyidagilardan iborat:

1. Sintaktik ko'pma'nolilik (ambiguity) muammosi. Tabiiy tilda bir gap bir nechta sintaktik talqinga ega bo'lishi mumkin. Bunday holat strukturaviy ko'pma'nolilik deb ataladi. Inson kontekst asosida to'g'ri talqinni tanlay oladi, biroq avtomatik parser kontekst yetishmasa ikkilanadi. Bu muammo, ayniqsa, statistik modellarda sezilarli darajada namoyon bo'ladi.

2. Agglyutinativ morfologiya murakkabligi. O'zbek tilida grammatik ma'nolar qo'shimchalar orqali ifodalanadi. Natijada bir so'z tarkibida bir nechta morfologik shakl mujassam bo'ladi. Masalan: kelmaganligingizdan

Mazkur birlik quyidagi grammatik elementlarni o'z ichiga oladi: fe'l asosi (kel), inkor shakli, sifatdash qo'shimchasi, egalik qo'shimchasi, kelishik qo'shimchasi. Parser bu birlikni to'g'ri segmentatsiya qilishi, grammatik xususiyatlarini aniqlashi, gap ichidagi sintaktik rolini belgilashi zarur. Morfologik tahlil bosqichidagi xatolik keyingi sintaktik tahlil natijasiga bevosita ta'sir ko'rsatadi.

3. Erkin so'z tartibi. O'zbek tilida so'z tartibi nisbatan erkin hisoblanadi. Masalan:

- Men kitobni o'qidim.
- Kitobni men o'qidim.
- O'qidim men kitobni.

Mazmun o'zgarmaydi, biroq daraxt struktura shakllanishi farqlanadi. Qoidaviy (rule-based) parserlar uchun bu o'zgaruvchanlik murakkablik tug'diradi, chunki ular qat'iy tuzilma andozalariga tayangan holda ishlaydi.

4. Murakkab va qo'shma gaplar tahlili. Ergash gapli qo'shma gaplar sintaktik daraxtni murakkablashtiradi. Masalan: U kelgach, biz ishni boshladik. Bu yerda u kelgach ergash gap, biz ishni boshladik bosh gap hisoblanadi.

Ergash va bosh gap o'rtasidagi bog'lanish turini aniqlash parserdan yuqori darajadagi strukturaviy aniqlik talab qiladi.

5. Avtomatik tahlil (kompyuter orqali tahlil) muammosi. Kompyuter tilni tushunish uchun grammatik qoidalarni ishlatadi. Agar bir gap bir nechta ma'no yoki grammatik tuzilishga ega bo'lsa, kompyuter qaysi ma'noni tanlashni bilmay qoladi. Bu orqali kompyuter noto'g'ri sintaktik daraxt tuzishi mumkin. Natijada fe'l, ot yoki gap qismlari noto'g'ri bog'lanadi. Xulosa chiqarishda xatolik yuz beradi.

Natijalar shuni ko'rsatadiki, korpus sifati va parsing aniqligi o'zaro uzviy bog'liq. Past sifatli tokenizatsiya yoki morfologik teglash sintaktik daraxtning noto'g'ri shakllanishiga olib keladi. Zamonaviy neyron yondashuvlar katta hajmdagi annotatsiyalangan ma'lumot talab qiladi. Shu bois o'zbek tilida kompyuter lingvistikasini rivojlantirish uchun milliy korpusni kengaytirish, sifatli annotatsiya standartlarini ishlab chiqish, zamonaviy parsing modellari yaratish, sun'iy intellekt asosidagi yondashuvlarni joriy etish muhim hisoblanadi.

Xulosa. Xulosa qilib aytganda, lingvistik korpuslar va parsing muammolari kompyuter lingvistikasining markaziy masalalaridan biri hisoblanadi. Ularni hal etish tabiiy tilni qayta ishlash tizimlarining sifatini oshirishga xizmat qiladi. O'zbek tili misolida bu sohada amalga oshiriladigan ilmiy va amaliy ishlar milliy til texnologiyalarining rivojiga katta hissa qo'shadi. Tadqiqotlar shuni ko'rsatadiki, yuqori sifatli korpuslar va aniq sintaktik parsing natijalari nafaqat morfologik va sintaktik tahlil, balki anafora va koreferensiya kabi yuqori darajadagi diskursiv hodisalarni aniqlash uchun ham muhim asos bo'lib xizmat qiladi. Shu bilan birga, zamonaviy tillar uchun ishlab chiqilgan parserlar tarixiy matnlarning lingvistik xususiyatlarini to'liq aks ettira olmasligi sababli ularni mos davr materiallari asosida qayta o'qitish va boyitish zarurati yuzaga keladi. Shuning uchun kompyuter lingvistikasi sohasida korpus yaratish, annotatsiya standartlarini takomillashtirish, parsing aniqligini oshirish hamda xatolarni avtomatik aniqlash va tuzatish mexanizmlarini ishlab chiqish kompleks yondashuv asosida amalga oshirilishi lozim.

Foydalanilgan adabiyotlar ro'yxati:

1. Jurafsky D., Martin J. *Speech and Language Processing*. Pearson, 2023.
2. McEnery T., Hardie A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012.
3. Manning C., Schütze H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
4. Abdurahmonov G'. O'zbek tilining grammatikasi. Toshkent, 2018.
5. Nivre J. *Dependency Parsing*. Morgan & Claypool, 2006.
6. O'zbek tili milliy korpusi materiallari.