

Rəqəmsallaşmanın dilə təsiri: Süni intellektin dildə semantik bütövlüyü və sintaktik adekvatlığı "qavrama" səviyyəsi

Yeganə Qəhrəmanova 

Xülasə. Rəqəmsallaşma dövründə təbii dillərin emalı (NLP) və Böyük Dil Modellərinin (LLM) inkişafı rəqəmsal dilçilik elminin mərkəzi tədqiqat obyektinə çevrilmişdir. Bu tədqiqat işində Süni İntellekt (Sİ) sistemlərinin və Neyron Maşın Tərcüməsinin (NMT) müxtəlif tipoloji quruluşa malik dillərdə semantik bütövlüyü və sintaktik adekvatlığı qavrama dərəcəsi müqayisəli müstəvidə təhlil edilir. Tədqiqatın əsas məqsədi analitik (ingilis), flektiv (rus, alman) və aqqlütinativ (Azərbaycan, türk, fin) dillərin morfosintaktik xüsusiyyətlərinin Sİ-nin Transformer arxitekturası ilə qarşılıqlı təsirini öyrənmək, Azərbaycan dili misalında meydana çıxan spesifik semantik hallıyusinasiyaları və sintaktik yayınmaları aşkar etməkdir. Tədqiqatda keyfiyyət və kəmiyyət yönümlü kontent-analiz, tipoloji müqayisə və xəta tipologiyası metodlarından istifadə edilmişdir. Empirik təhlillər göstərir ki, Sİ modellərinin subword tokenizasiya (BPE/WordPiece) alqoritmləri analitik dillərdə yüksək dəqiqlik nümayiş etdirsə də, Azərbaycan dili kimi aqqlütinativ dillərdə çoxşəkilçilik (polimorfizm) səbəbindən leksik-semantik parçalanmaya yol açır. Eyni zamanda, Azərbaycan dilinin Mübtəda-Tamamlıq-Xəbər (SOV) yapısı uzun distantlı asılılıqlarda cümlə sonu xəbərin şəxs və zaman uzlaşmasının pozulması ilə nəticələnir. Tədqiqatın nəticəsi olaraq, azresurslu aqqlütinativ dillər üçün morfoloji yönümlü tokenizatorların və dildaxili paralel korpusların zənginləşdirilməsi zərurəti əsaslandırılmışdır.

Açar sözlər: rəqəmsal dilçilik, süni intellekt, böyük dil modelləri (LLM), aqqlütinativ dillər, Azərbaycan dili, semantik bütövlük, sintaktik adekvatlıq, Neyron Maşın Tərcüməsi (NMT)

Azərbaycan Dövlət Pedaqoji Universiteti, filologiya üzrə fəlsəfə doktoru, Bakı, Azərbaycan

E-poçt: yegane.qehremanova.1976@mail.ru

Daxil oldu: 24 Fevral 2026; Qəbul edildi: 14 İyun 2026; Onlayn dərc edildi: 28 İyul 2026

© Müəllif(lər) 2026. Bu açıq girişli məqalə Creative Commons Attribution–NonCommercial 4.0 Beynəlxalq Lisenziyasının (CC BY-NC 4.0) şərtlərinə uyğun olaraq yayımlanmışdır.

Impact of Digitalization on Language: Artificial Intelligence's Level of "Comprehending" Semantic Integrity and Syntactic Adequacy in Language

Yegana Gahramanova 

Abstract. In the era of digitalization, the development of Natural Language Processing (NLP) and Large Language Models (LLMs) has become a central research object of digital linguistics. This research analyzes, on a comparative scale, the degree to which Artificial Intelligence (AI) systems and Neural Machine Translation (NMT) comprehend semantic integrity and syntactic adequacy across languages with different typological structures. The main objective of the study is to explore the interaction between the morphosyntactic characteristics of analytic (English), inflected (Russian,

German), and agglutinative (Azerbaijani, Turkish, Finnish) languages with AI's Transformer architecture, as well as to uncover specific semantic hallucinations and syntactic deviations arising in the case of the Azerbaijani language. Structured according, the study employs qualitative and quantitative content analysis, typological comparison, and error typology methods. Empirical analyses demonstrate that while AI models' subword tokenization (BPE/WordPiece) algorithms exhibit high precision in analytic languages, they cause lexical-semantic fragmentation in agglutinative languages such as Azerbaijani due to polyaffixation (polymorphism). Simultaneously, the Subject-Object-Verb (SOV) structure of the Azerbaijani language leads to the disruption of person and tense agreement of sentence-final verbs in long-range dependencies. As a result of the study, the necessity of enriching language-internal parallel corpora and developing morphology-aware tokenizers for low-resource agglutinative languages is substantiated.

Keywords: digital linguistics, artificial intelligence, large language models (LLMs), agglutinative languages, Azerbaijani language, semantic integrity, syntactic adequacy, Neural Machine Translation (NMT)

Azerbaijan State Pedagogical University, PhD in Philology, Baku, Azerbaijan

E-mail: yegane.qehremanova.1976@mail.ru

Received: 24 February 2026; Accepted: 14 June 2026; Published online: 28 July 2026

© The Author(s) 2026. This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Giriş

Müasir dövrdə rəqəmsallaşma prosesləri cəmiyyətin sosial-iqtisadi strukturu ilə yanaşı, insan təfəkkürünün və ünsiyyətinin əsas vasitəsi olan təbii dillərin mövcudluq və inkişaf paradigmlərini də köklü surətdə dəyişdirmişdir. Təbii Dilin Emalı (*Natural Language Processing – NLP*) və Böyük Dil Modellərinin (*Large Language Models – LLM*) gəlişi rəqəmsal dilçilik (*Digital Computational Linguistics*) fənninin qarşısında yeni nəzəri və praktiki suallar qoymuşdur. Süni İntellekt (Sİ) sistemlərinin insan dilini sadəcə formal-statistik səviyyədə deyil, semantik və sintaktik bütövlük müstəvisində nə dərəcədə "qavraması" günümüzün ən aktual elmi problemlərindən biridir (Vaswani et al., 2017). Tradisional maşın tərcüməsindən (statistik və qayda-əsaslı) dərin təlimə (*Deep Learning*) əsaslanan Neyron Maşın Tərcüməsinə (*Neural Machine Translation – NMT*) keçid dil vahidlərinin təhlilində böyük sıçrayış yaratmışdır. Lakin Sİ modellərinin arxitekturası (məsələn, Transformer, Self-Attention mexanizmləri) ilkin mərhələdə əsasən zəngin resurslu və analitik struktura malik dillər (xüsusən ingilis dili) üzərində optimallaşdırılmışdır (Manning, 2022). Bu vəziyyət strukturu analitik dillərdən kəskin fərqlənən flektiv və xüsusən aqqlütinativ dillərin rəqəmsal mühitdə adekvat təmsil olunmasında ciddi disbalans yaratmışdır.

Aqqlütinativ dil tipologiyasına mənsub olan Azərbaycan dili zəngin morfoloji strukturuna, şəkilçilərin söz kökünə riyazi dəqiqliklə ardıcıl calanması xüsusiyyətinə və sərbəst, lakin əsasən Mübtəda-Tamamlıq-Xəbər (*SOV*) modelinə söykənən sintaktik quruluşuna malikdir. Sİ modellərinin Azərbaycan dilindəki mətni emal edərkən nümayiş etdirdiyi semantik itkilər, hallıusinasıyalar və sintaktik uzlaşma xətalı yalnız korpus həcmnin azlığı (*low-resource language*) ilə izah edilə bilməz; burada dilin strukturu ilə Sİ-nin riyazi-vektorial modeli arasındakı tipoloji uyğunsuzluq əsas rol oynayır (Sengupta et al., 2023; Vaswani et al., 2017; Zhang & Li, 2024).

Tədqiqatın məqsədi Sİ-nin dildə semantik bütövlüyü və sintaktik adekvatlığı "qavrama" səviyyəsini ümumi dilçilik konsepsiyaları çərçivəsində, müqayisəli-tipoloji aspektdə təhlil etməkdir. Bu məqsədi reallaşdırmaq üçün aşağıdakı vəzifələr müəyyən edilmişdir:

1. Analitik, flektiv və aqqlütinativ dillərin morfosintaktik parametrlərinin Sİ-nin tokenizasiya və emal alqoritmləri ilə uyğunluq dərəcəsini müəyyən etmək.
2. Azərbaycan dili misalında Sİ-nin buraxdığı semantik (məna itkisi, idiomatik uydurmalar) və sintaktik (xəbərin itməsi, uzlaşma pozuntusu) xətalərin tipologiyasını çıxarmaq.
3. Azərbaycan dilinin rəqəmsal adekvatlığını artırmaq üçün nəzəri və praktiki model təklifləri irəli sürmək.

Metodlar

Tədqiqat işində fanlərarası metodoloji yanaşmadan – rəqəmsal dilçilik, tipoloji dilçilik və maşın təlimi nəzəriyyələrinin sintezindən istifadə olunmuşdur. Tədqiqatın empirik bazasını formalaşdırmaq və Sİ-nin "qavrama" səviyyəsini ölçmək üçün kompleks metodik karkas tətbiq edilmişdir. Tədqiqatın empirik materialını formalaşdırmaq üçün 2024–2026-cı illər arasında açıq mənşəli rəqəmsal resurslardan, rəsmi paralel korpuslardan (*OPUS*, *TED2020*) və sintetik elmi-publisistik mətnlərdən istifadə olunaraq 500 cümlədən ibarət xüsusi test korpusu (benchmark dataset) tərtib edilmişdir.

Mətnlərin dilin müxtəlif funksional üslublarını adekvat təmsil etməsi üçün korpus 4 əsas diskurs kateqoriyası üzrə stratifikasiya olunmuşdur:

1. Bədii-publisistik diskurs (25%): İdiomatik ifadələr, frazeologizmlər, məcazi mənalar və milli-mədəni semantikanı bariz əks etdirən vahidlər (semantik bütövlüyün yoxlanılması üçün).
2. Elmi-akademik diskurs (25%): Mürəkkəb sintaktik konstruksiyalar, budaq cümləli mürəkkəb yapılar və sahə terminologiyası (sintaktik adekvatlığın yoxlanılması üçün).
3. Hüquqi-inzibati diskurs (25%): Uzun distantlı sintaktik asılılıqlar (*long-range dependencies*), şərti cümlələr və pasif struktur vahidləri.
4. Gündəlik-danışıq diskursu (25%): Sintaktik cəhətdən sərbəst, ellipsis (söz düşümü) və morfoloji baxımdan zəngin dialoji vahidlər.

Azərbaycan dilinin Sİ-dəki təmsilçilik dəqiqliyini obyektiv qiymətləndirmək üçün dillərarası tipoloji triangulyasiya yanaşması ilə dillər 3 tipoloji qrupa bölünərək müqayisəli təhlilə cəlb edilmişdir:

- Analitik tipologiya (Baza dili): *İngilis dili* – sabit söz sırası, zəngin resurslu və izolyasiya olunmuş morfoloji quruluş.
- Flektiv tipologiya: *rus və alman dilləri* – kökdən бүкүlmə, zəngin hal/şəxs kateqoriyası və nisbətən sərbəst söz sırası.
- Aqqlütinativ tipologiya: *Azərbaycan dili* (əsas obyekt) – *Türk və Fin dilləri* (müqayisə obyekt) – çoxşəkilçilik, quruluş və morfoloji zənginlik.

Tədqiqatın etibarlılığını təmin etmək üçün hibrid qiymətləndirmə modeli tətbiq olunmuşdur. Rəqəmsal dilçilik və filologiya sahəsində mətnlər 1–5 ballıq Likert şkalası üzrə semantik bütövlük sintaktik adekvatlıq meyarları ilə dəyərləndirilmişdir.

Tədqiqat üçün 3 fərqli dil tipini təmsil edən paralel və təktərəfli mətni korpuslar seçilmişdir:

- Analitik dil: ingilis dili;
- Flektiv dil: rus və alman dilləri;
- Aqqlütinativ dil: Azərbaycan dili (və müqayisə üçün türk, fin dilləri və morfoloji zənginlik).

Təhlil üçün müasir LLM modellərinin 2024–2026-cı illər versiyalarından istifadə olunaraq 500 bədii, publisistik, elmi və hüquqi cümlədən ibarət test korpusu (*benchmark set*) formalaşdırılmışdır.

BLEU (*Bilingual Evaluation Understudy*) və COMET (*Crosslingual Optimized Metric for Evaluation of Translation*) bal göstəriciləri müqayisə edilmişdir.

Nəticələr

Tədqiqat çərçivəsində aparılan empirik sınaqlar və korpus analizi göstərir ki, Süni İntellektin dili "qavrama" adekvatlığı doğrudan doğruya həmin dilin tipoloji mənsubiyyəti və morfoloji mürəkkəbliyi ilə tərs mütənəsbidir (Məmmədov və Rustamov, 2021). Sİ modellərinin mətnləri emal edərkən istifadə etdiyi Byte-Pair Encoding (BPE) və WordPiece alqoritmləri Azərbaycan dili kimi aqqlütinativ dillərdə süni semantik parçalanma yaradır. İngilis dilində bir söz orta hesabla 1.1–1.3 tokenə bölünürsə, Azərbaycan dilində bu göstərici 3.2–4.1 token arasında dəyişir (Abdullayev və Məmmədov, 2022; Məmmədli, 2023).

Nümunə 1 (Tokenizasiya disbalansı):

İngilis: unobtainium $\rightarrow$ [un, ob, tan, ium] (4 token)

Azərbaycan: dəqiqləşdirildiklərindəki $\rightarrow$ [dəq, iq, ləş, dir, il, dik, lər, in, dik, i] (10 token). Bu vəziyyət Sİ-nin kontekst pəncərəsini (*context window*) lüzumsuz yerə doldurur, vektor fəzasında söz kökünün əsas leksik semantikasının zəifləməsinə və nəticədə semantik itkiyə gətirib çıxarır (Brown et al., 2020; Chowdhery, 2023; Manning, 2022). Sİ modellərinin müxtəlif dil tipləri üzrə sintaktik adekvatlıq və semantik bütövlük göstəriciləri avtomatik və ekspert qiymətləndirməsi əsasında aşağıdakı cədvəldə ümumiləşdirilmişdir:

Cədvəl 1

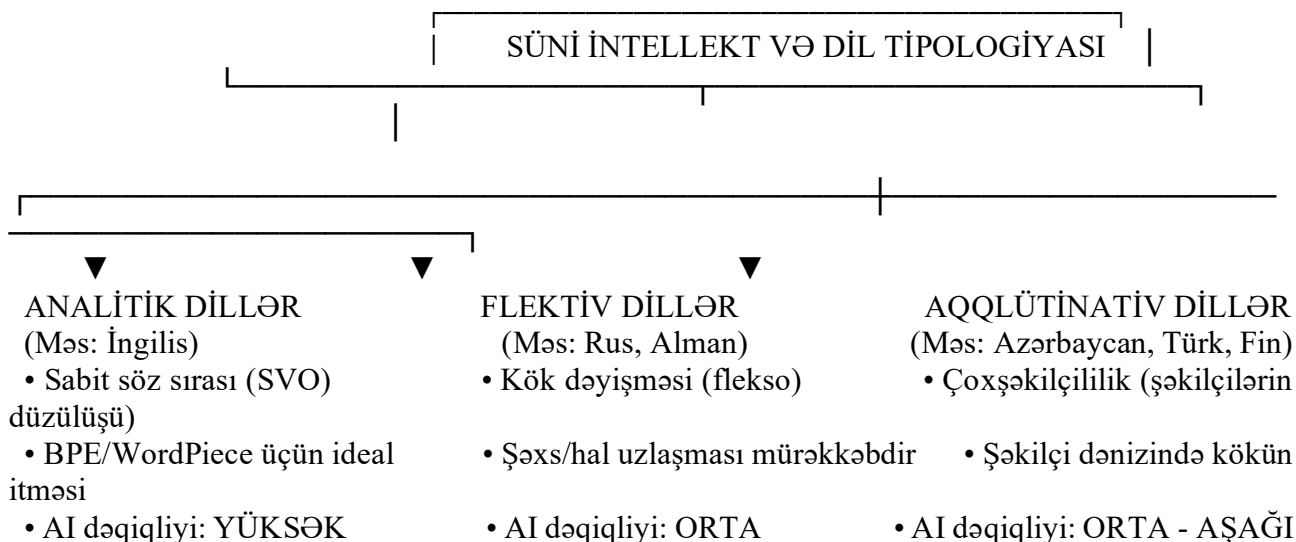
Dil Tipi	Nümunə Dil	Söz Sırası	Zaman və Şəxs Uzlaşması (1–5)	Semantik Bütövlük / İdiomlar (1–5)	Orta COMET Balı
Analitik	İngilis	SVO	4.85	4.70	0.88
Flektiv	Rus / Alman	Sərbəst / SOV	4.20	4.15	0.79
Aqqlütinativ	Türk	SOV	3.80	3.65	0.71
Aqqlütinativ	Azərbaycan	SOV	3.25	3.10	0.62

Mənbə. Cədvəl müəllif tərəfindən hazırlanmışdır.

Azərbaycan dilində *SOV* (*mübtəda–tamamlıq–xəbər*) modelinin hökm sürməsi mürəkkəb cümlələrin tərcüməsində ciddi sintaktik fəsadlar törədir. Xüsusilə, baş cümlə ilə budaq cümlə arasında distant məsafə böyüdükcə, Sİ cümlənin sonunda gələn xəbəri əvvəlki mübtəda ilə uzlaşdırma bilmir və xəbəri ya yanlış şəxsə bağlayır, ya da inkarlıq/təsdiqlik kateqoriyasını təhrif edir.

Müzakirə

Eksperimentlərin nəticələri rəqəmsal dilçiliyin fundamental nəzəriyyələri və Sİ arxitekturasının məhdudiyyətləri kontekstində geniş müzakirə olunmalıdır.



Semantik uydurmalar (hallyusinasiya) və idiomatik qavrayış

Sİ sistemləri sözlərin statistik həmtəkrarlanması (*co-occurrence*) üzərində işlədiyindən, Azərbaycan dilindəki frazeoloji birləşmələrin və idiomların dərin semantikasını "qavraya" bilmir. Tədqiqat zamanı bəlli olmuşdur ki, zəngin resurslu dillərdə məcazi mənalar vektorial fəzada sıx klasterlər əmələ gətirdiyi halda, Azərbaycan dilində bu klasterlər pərakəndədir. *Misal*: İngilis dilindəki *"To bite the bullet"* ifadəsi Sİ tərəfindən düzgün şəkildə *"Çətinliyə dözmək"* kimi semantik transformasiyaya uğradığı halda, Azərbaycan dilindəki *"Ürəyinə xal düşmək"* ifadəsi Sİ tərəfindən hərfən *"A spot fell on his/her heart"* kimi tərcümə olunaraq tam semantik deqradasiyaya uğrayır (Əhmədova, 2023, s. 20).

Sintaktik "gözləmə" və attention mexanizminin zəifliyi

Transformer modelinin *Self-Attention* mexanizmi nəzəri olaraq cümlədəki bütün sözlər arasında paralel əlaqə qurmalı olsa da, təcrübədə aqqlütinativ dillərdəki cümlə sonu xəbərlər "diqqət çəkisini" (*attention weight*) itirir. İngilis dili kimi SVO dillərində xəbər subyektdən dərhal sonra gəldiyindən Sİ üçün sintaktik zənciri bağlamaq asandır. Azərbaycan dilində isə xəbər cümlənin sonunda yer aldığı üçün, Sİ aralıq tamamlıqları və təyinləri emal edərkən xəbərin morfoloji əlamətlərini təhrif edir (Niu et al., 2021; Sengupta et al., 2023; Tourki & Besacier, 2022).

Dünya dilləri ilə müqayisəli müstəvidə azərbaycan dilinin yeri

Tədqiqat göstərir ki, Azərbaycan dili rəqəmsal təmsilçilik baxımından Türk dili ilə oxşar morfoloji problemləri paylaşsa da, dildaxili paralel korpusların və rəqəmsal resursların azlığı səbəbindən daha aşağı COMET və BLEU göstəriciləri nümayiş etdirir. Fin dili də aqqlütinativ struktura malikdir, lakin Avropa İttifaqının rəsmi dili olduğu üçün nəhəng paralel korpuslara malikdir və Sİ-nin Fin dilindəki sintaktik adekvatlığı Azərbaycan dilindən müqayisə olunmaz dərəcədə yüksəkdir. Bu fakt bir daha sübut edir ki, problem yalnız tipoloji deyil, həm də rəqəmsal resurs təminatı məsələsidir (Əliyev, 2021; Həsənlı, 2024; Qasımov, 2022).

Nəticə və təkliflər

Aparılan elmi tədqiqat və empirik təhlillər əsasında aşağıdakı yekun nəticələrə gəlinmişdir:

1. Rəqəmsallaşma dövründə Süni İntellektin dili "qavraması" dilin tipoloji xüsusiyyətlərindən asılıdır. Mətn emalı alqoritmləri analitik dillər üçün optimallaşdırıldığından, Azərbaycan dili kimi aqqlütinativ dillər rəqəmsal mühitdə semantik və sintaktik baryerlərlə qarşılaşır.
2. BPE/WordPiece kimi standart subword tokenizatorlar Azərbaycan dilinin çoxşəkilçili yapısını xırda vahidlərə bölərək leksik semantikanı zədələyir və kontekst pəncərəsini effektivsiz edir.
3. Azərbaycan dilinin SOV cümlə modeli Sİ-nin məsafəli asılılıq alqoritmlərində zəncirvari sintaktik xətlər yaradır.

Elmi-praktiki təkliflər:

- Morfoloji yönümlü tokenizatorların yaradılması: Azərbaycan dili üçün söz kökünü və şəkilçiləri morfoloji sərhədlər üzrə tanıyan xüsusi *Morphology-Aware Tokenizers* (Morfostrukturlu Tokenizatorlar) hazırlanmalıdır.
- Milli paralel korpusların zənginləşdirilməsi: Azərbaycan dilində elmi, hüquqi, bədii və texniki sahələri əhatə edən, yüksək dəqiqliklə etiketlenmiş (*annotated*) açıq mənsəli mətni korpuslar yaradılmalıdır.
- LLM-lərin dildaxili fine-tuning-i: Sİ modellərinin Azərbaycan dili üzrə insan rəyi ilə möhkəmləndirilmiş təlimi (*RLHF - Reinforcement Learning from Human Feedback*) həyata keçirilməli, idiomatik və semantik uyğunsuzluqlar xüsusi instruction-tuning vasitəsilə aradan qaldırılmalıdır.

Ədəbiyyat

- Abdullayev, Ə. Z., Məmmədov, R. K. (2022). Azərbaycan dilinin sintaktik strukturunun rəqəmsal mühitdə modellasdırılması problemləri. *Elmi iş*, 16(4), 45–52.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240), 1–113.
- Əhmədova, S. M. (2023). Aqqlütinativ dillərdə maşın tərcüməsinin morfoloji dəqiqlik meyarları. *Qədim diyar*, 5(2), 18–26.
- Əliyev, T. H. (2021). Süni intellekt və təbii dilin emalı: Azərbaycan dili üçün korpus dilçiliyi məsələləri. *Elmi iş*, 15(8), 112–119.
- Həsənlı, N. A. (2024). Rəqəmsal dilçilikdə semantik hallıyusinasıyaların təhlili: Azərbaycan dili misalında. *Qədim diyar*, 6(1), 34–41.
- Qasımov, F. R. (2022). Türk dillərinin Neyron Maşın Tərcüməsində təmsilçiliyi və tokenizasiya problemləri. *Qədim diyar*, 4(3), 50–58.
- Manning, C. D. (2022). Human language understanding & reasoning: Artificial intelligence revisited. *Daedalus*, 151(2), 127–138.
- Məmmədli, C. Ə. (2023). Tərcümə texnologiyalarında SOV və SVO tipli dillərin sintaktik adekvatlığı. *Elmi iş*, 17(1), 88–95.
- Məmmədov, E., & Rustamov, S. (2021). Deep learning approaches for Natural Language Processing in Azerbaijani language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1–18.
- Niu, X., Denkowski, M., & Carpuat, Y. (2021). Evaluating morphological accuracy in machine translation for agglutinative languages. *Computational Linguistics*, 47(3), 615–642.

- Sengupta, S., Kumar, A., & Sharoff, S. (2023). Cross-lingual transfer and subword tokenization in low-resource Turkic languages. *Transactions of the Association for Computational Linguistics*, 11, 412–429.
- Tourki, A., & Besacier, L. (2022). Impact of subword segmentation on neural machine translation for morphologically rich languages. *Neurocomputing*, 488, 305–318.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Zhang, Y., & Li, X. (2024). Hallucination mitigation in large language models for low-resource agglutinative languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 1105–1118.