# KOMPYUTER LINGVISTIKASIDA SUN’IY INTELLEKT VA MASHINA TARJIMASI

## MASHINA TARJIMASIDA MATNNI MOSLASHTIRISH USULLARI

**Xamroyeva Shahlo,**
f.f.d., dotsent shaxlo.xamroyeva@navoiy-uni.uz
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
**Matyakubova Noila,**
doktorant nailya89mm@mail.ru
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

**Annotatsiya.** Matnni moslashtirish turli xil mashina tarjimasi tizimlarining muhim jarayonidir. Bu vazifa dastlabki matnning so‘zlari, jumlalari yoki paragraflari va ularning tarjimasi (parallel korpus) o‘rtasidagi yozishmalarni aniqlashdan iborat. Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: statistik usullar va leksikaga asoslangan usullar.

**Kalit so‘zlar:** *Matnni moslashtirish, mashina tarjimasi, statistik usullar, leksik usullar, tabiiy tilni qayta ishlash.*

**Abstract.** The text alignment is an important process of different Machine Translation systems. This task consists in identifying correspondences between words, sentences or paragraphs of a source text and their translation (parallel corpus). There are two main approaches to perform parallel corpus alignment: the statistical-based methods and lexical-based methods.

**Keywords:** *Text alignment, machine translation, statistical methods, lexical methods, natural language processing.*

**Аннотация.** Выравнивание текста — важный процесс различных систем машинного перевода. Эта задача заключается в выявлении соответствий между словами, предложениями или абзацами исходного текста и их переводом (параллельный корпус). Существует два основных подхода к параллельному выравниванию корпусов: статистические методы и методы на основе лексики.

**Ключевые слова:** *Выравнивание текста, машинный перевод, статистические методы, лексические методы, обработка естественного языка.*

**Kirish.** Tabiiy tilni qayta ishlash (NLP) - bu odamlar muloqot qilish uchun foydalanadigan tilni yaratish va tushunishga qaratilgan kompyuter fanining sohasi hisoblanib u juda ko'p muhim vazifalar va ilovalarga ega; ulardan biri matnni avtomatik ravishda bir tildan ikkinchi tilga tarjima qilishni maqsad qo‘llaniladigan mashina tarjimasi(MT)dir. MTni amalga oshirish uchun lug‘atlar, statistika yoki misollarga asoslangan bir nechta usullar mavjud. Ushbu usullarning o‘ziga xos

afzalliklari va kamchiliklari bo‘lsa ham, ular matnni moslashtirish jarayoni kabi ba’zi usullarni keng foydalanadi. Matnni moslashtirish (TA) jarayoni paragraflar, jumlalar, manba matnlarning so‘zlari va ularning tarjimalari o‘rtasida muvofiqlikni o‘rnatish uchun parallel korpusni tashkil qilishdan iborat. Parallel korpus turli tillardagi ikkita matn to‘plami sifatida belgilanishi mumkin, bu to‘plamlardan biri manba matni, ikkinchisi esa ularning tarjimasi hisoblanadi.

**Asosiy qism.** Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: statistik usullar va leksikaga asoslangan usullar. Leksikaga asoslangan yondashuvlar mavjud leksik bilimlarga tayanadi, masalan, antonim va sinonimlar, so‘zning tarjimalari va boshqalar; Statistik yondashuvlar esa leksik bo‘lmagan ma’lumotlarga tayanadi, masalan, jumla uzunligi, gapning o‘rni, takrorlanish chastotasi, ikki tilda gap uzunligi nisbati kabi holatlarga.

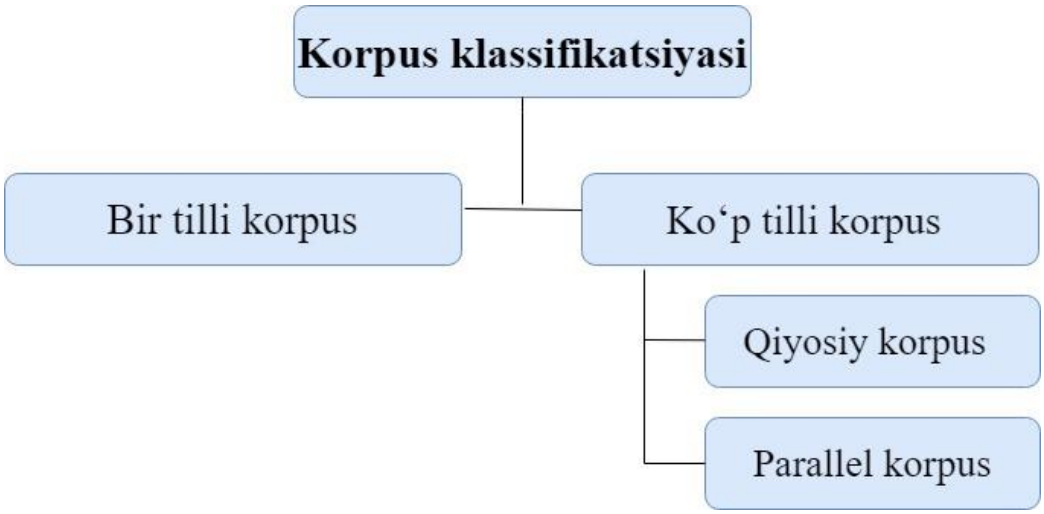
TA jarayonini amalga oshirish uchun biz yozma korpusdan foydalanishimiz kerak. Buni tadqiqot uchun, ayniqsa tarjima dasturlarini ishlab chiqish va tabiiy tilni qayta ishlash uchun foydalaniladigan matnlar to‘plami yoki katta hajmli matn sifatida aniqlash mumki. Korpuslar nomlangan yoki nomlanmagan bo‘lishi mumkin. Nomlangan korpuslar turli atributlarni yoki lingvistik ma’lumotlarni aniqlash uchun izohlanadi, masalan, korpus tarkibidagi hujjatlarning mavzulari yoki so‘zlarning nutq qismi va boshqalarni. Masalan, "atirgullar" so‘zi uchun korpusda belgilangan atributlar ot, ko‘plik va boshqalar bo‘lishi mumkin. Nomlanishi mumkin bo‘lgan lingvistik ma’lumotlar uning lemmasi, yani ma’lum bir lug‘at bo‘yicha so‘zning to‘g‘ri ma’nosi va boshqalar bo‘lishi mumkin. Rus va ingliz kabi tillarda, otning jinsini ifodalovchi belgilar qo‘shilishi mumkin. Nomlangan korpus **/sentence/word/lemma/pos/id** ni aniqlaydigan maxsus belgi bilan belgilanadi *[*Moore, R. C. 2005,. 1-8b*]*. Ammo, nomlanmagan korpuslar lingvistik ma’lumotlarga ega emas va aniq tuzilishga ham ega emas. Bu ko‘pincha elektron pochta yoki tezkor xabarlar, hujjatlar yoki ijtimoiy media xabarlari kabi foydalanuvchilar tomonidan yaratilgan ma’lumotlardir.

Korpusning har xil turlari mavjud bo‘lib mashina tarjimasi sohasida bir tilli va ko‘p tilli korpusning tasnifi muhim ahamiyatga egahisoblanadi. Ko‘p tilli korpuslar bir nechta tillardagi matnlar bo‘lib, ularni quyidagi kichik toifalarga bo'lish mumkin:

1. Parallel korpusni turli tillardagi ikkita matn to'plami sifatida aniqlash mumkin, bu to‘plamlardan biri manba matnlari, ikkinchisi esa ularning tarjimalari. Ushbu matnlarning har biri bitextlar deb nomlanishi mumkin [Harris,B. 8-10]. Parallel korpus bir yo‘nalishli, ikki tomonlama yoki ko‘p yo‘nalishli bo‘lishi mumkin. Masalan, Navoiyning “Xamsa” dostonlari va ularning turli tillardagi nusxalarini parallel korpus deb hisoblash mumkin.

2. Qiyosiy korpus - bu bir xil asosiy mavzuni yorituvchi, lekin uni ko‘rib chiqish uslubida farq qiluvchi turli tillardagi matnlar to‘plamidir. Bu shuni anglatadiki, qiyosiy korpus manba matni va ularning tarjimasi hisoblanmaydi. Misol tariqasida jurnallar yoki yangiliklarni uzatish tizimlaridan olingan yangi maqolalar

to‘plamlari olish mumkin, chunki ular turli tillarda bir xil voqeaga ishora qiladi va shuning uchun qiyosiy korpus deb hisoblanishi mumkin [Simões, A. 2004. 5-8b].




**1-jadval Korpus klassifikatsiyasi**

Korpusni moslashtirish - bu manba matnlarning paragraflari, jumlalari yoki so‘zlari va ularning tarjimalari o‘rtasida muvofiqlikni o‘rnatish uchun parallel korpusni tashkil qilishdan iborat. Shunga qaramay, parallel korpuslarni avtomatik ravishda moslashtirish ba’zi til juftliklari uchun katta ahamiyatga ega vazifa hisoblanadi.

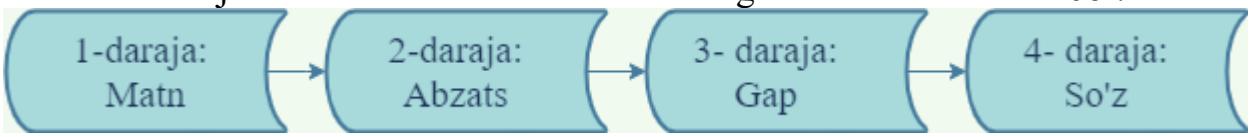
Kanadalik tilshunos Elliot Maklovich MTda qo’llaniladigan moslashtirish usullarini 4 darajaga bo‘lib, ularni quyidagicha tasniflagan [Macklovitch, E., & Hannan, M.L.1998. 41-57b.]:

1-darajali moslashtirish: matn yetarlicha uzun bo‘lmaganda butun matnni moslashtiradi.

2-darajali moslashtirish: bu daraja abzatslarni moslashtirishni takrorlaydi.

3-darajali moslashtirish: jumlalarni moslashtirishni tavsiflaydi.

4-darajali moslashtirish: bitextlar orasidagi so‘zlarni moslashtiradi.




**2-rasm. MTda qo‘llanadigan moslashtirish darajalari**

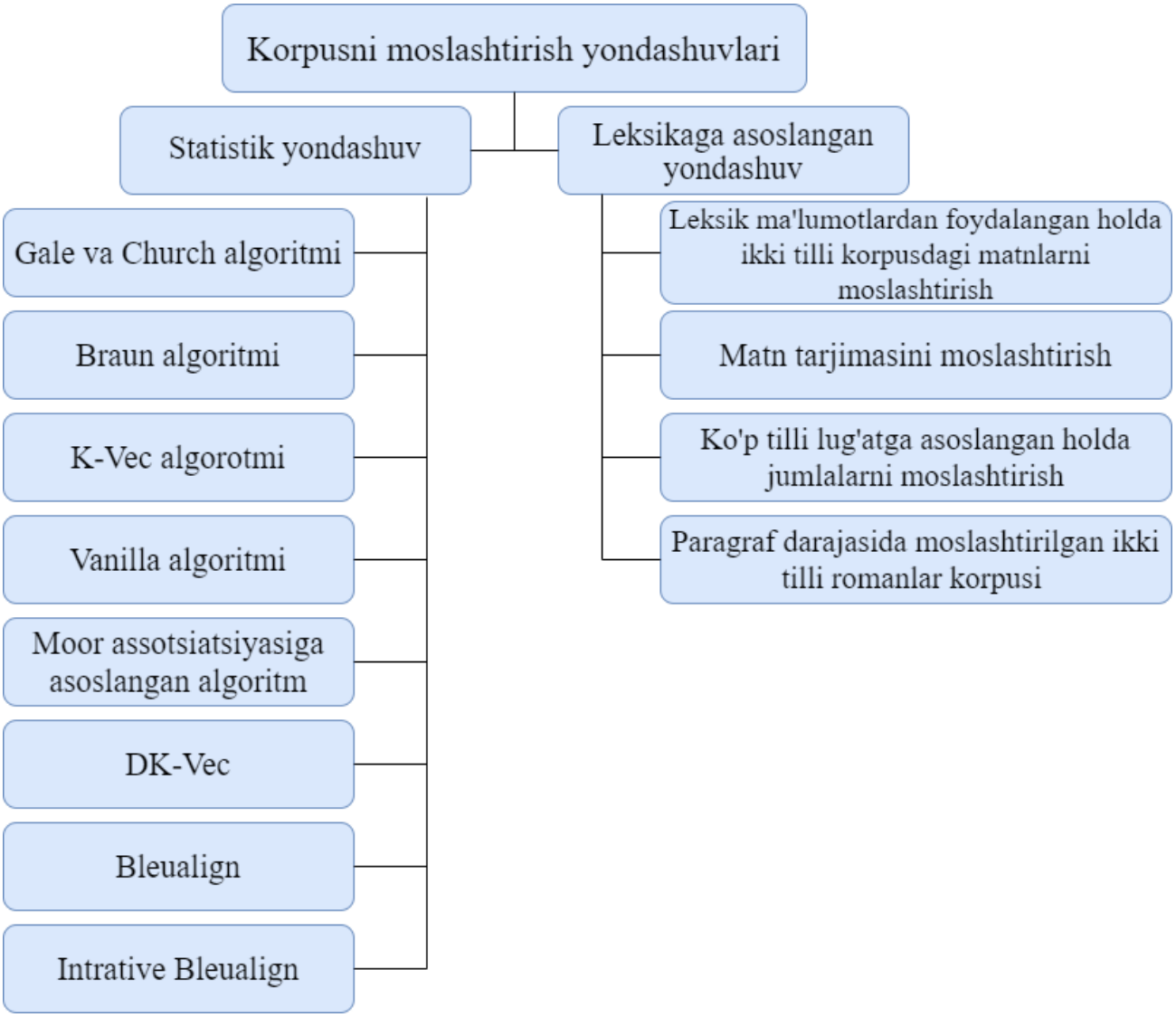
**Korpusni moslashtirish yondashuvlari**

Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: birinchi yondashuv statistik ma’lumotlarga asoslanadi, ikkinchisi esa qo‘shimcha lingvistik bilimlarni qo‘llaydi. Ushbu farqning asosi qayta ishlash usullaridan mustaqil ravishda qayta ishlanadigan ma’lumotlar turi bilan bog‘liq [Gelbukh, A., Sidorov, G., & Vera-Félix, J. Á.2006. 16-23b).]. Ushbu yondashuvlar asosida bir

nechta texnikalar ishlab chiqilgan bo‘lib, ularning har biri o‘zining afzalliklari va kamchiliklariga ega.

Leksikaga asoslangan yondashuvlar mavjud leksik resurslarga, masalan, katta hajmdagi ikki tilli lug‘atlar, tillar haqida ma’lumot olish uchun maxsus atamalar ro‘yxatiga, antonim va sinonimlar, so‘zning tarjimalari va boshqalarga tayanadi. Ushbu usullar statistik ma’lumotlarga asoslangan usullardan ko‘ra sekinroq va tilga bog‘liq bo‘ladi. Ushbu usullarning asosiy kamchiligi shundaki, ularning ishlashi ko‘p jihatdan moslashtirish jarayonida qo‘llaniladigan leksik ma’lumotlarga bog‘liq. Biroq, bu usullarning ko‘pchiligi statistik usullardan ko‘ra yaxshiroq natijalarga erishish mumkinligi uchun ishlab chiqilmoqda.

Statistik yondashuvlar leksik bo‘lmagan ma’lumotlarga tayanadi, masalan, jumla uzunligi, jumlaning o‘rni, birgalikda sodir bo‘lish chastotasi, ikki tildagi jumla uzunligi nisbati. Bu usullar moslashtirish jarayonini tezlashtiradi, biroq, ushbu usullarning asosiy kamchiligi shundaki, ularning ishlashi maqsadli matn va bitextlarning manba matni o‘rtasidagi tarkibiy o‘xshashlikka bog‘liq [Gelbukh, A., Sidorov, G., & Vera-Félix, J. Á.2006. 16-23b]. 3- rasmda bugungi kunda korpus moslashuvida keng qo‘llanilayotgan usullar va algoritmlar keltirilgan.




**3- rasm. Korpusni moslashtirish yondashuvlari**

**Gale va Church algoritmi** - ushbu algoritm "bir tildagi uzunroq jumlalar boshqa tildagi uzunroq jumlalarga tarjima qilinadi va qisqaroq jumlalar qisqaroq jumlalarga tarjima qilinadi" g‘oyasi asosida ishlaydi va moslashtirish tamoyila aynan shu g‘oyaga asoslangan.[ Gale, W. A., & Church, K. W. 1993. 75-102b] Ushbu algoritm faoliyati uchun paragraflarda allaqachon moslashtirilgan parallel korpus talab qilinadi. Ushbu algoritm matnlarni moslashtirish uchun jumlalarning uzunligini belgilarda hisobga oladi. Bular har bir jumlalar juftligi (manba matnlardan biri va maqsadli matnlardan biri) uchun masofa o‘lchovi deb ataladigan qiymatni hisoblash uchun ishlatiladi. Masofa o‘lchovi qanchalik past bo‘lsa, jumlalarning o‘zaro mos kelishi ehtimoli shunchalik yuqori bo‘ladi.

**Braun algoritmi** - "Braun klastering" deb ham nomlanadi, chunki u dastlab 1990-yillarda Braun universitetida ishlab chiqilgan. Gale va Church algoritmiga o‘xshab, bu algoritm matnlarni moslashtirish uchun jumlalar uzunligini hisobga oladi, ammo farqi shundaki, o‘lchovda jumla uzunligi so'zlarda hisobga olinadi.

Unda Kanada Hansard korpusining TEX formatiga kiritilgan teglardan moslashish jarayonida foydalaniladi. Algoritm asosiy va kichik bog‘lanish nuqtalarini ko‘rib chiqadi va moslashtirishni ayni ushbu ikki bosqichda amalga oshiradi.[ Brown, P. F., Lai, J. C., & Mercer, R. L. 1991. 169-176b]

**K-Vec algoritmi**- bu algoritm Paskal Fung va Kennet Uord Cherxi [7] tomonidan ishlab chiqilgan sentence aligner vositasi hisoblanadi. K-vec juda muhim xususiyatga ega; u jumla chegaralariga bog‘liq emas. Ushbu xususiyatdan foydalanib, algoritm o‘xshash bo‘lmagan tillarni moslashtirishni maqsad qiladi; masalan, ingliz va yapon, yevropa tillari kabi. Ushbu algoritm "agar ikkita so‘z bir-birining tarjimasi bo‘lsa, ular ikkita bo‘lmagan so‘zlarga qaraganda bir xil segmentlarda paydo bo‘lish ehtimoli ko‘proq" tamoyili asosida yaratilgan[Fung, P., & Church, K. W. 1994.1096-1102b]. Bittekstdagi so‘z boshqa so‘zning tarjimasi ekanligini aniqlash uchun ushbu algoritm ularning tegishli matnda taqsimlanishining o‘xshashligiga e’tibor qaratadi.

**Vanilla Aligner** – 1997-yilda Pernilla Danielson va Daniel Ridings tomonidan taqdim etilgan va Gale va Church algoritmining takomillashtirilganidir. Xuddi o‘zidan oldingisi kabi, bu ham sentence aligner vositasi hisoblanib, jumlalar chegaralariga bog‘liq. Ushbu alignerning asosiy afzalligi SGML formatidagi bitextlar bilan ishlashga mosligidir [Danielsson, P., & Ridings, D.1997. 75-80b]. Bitextlarni SGML formatida ishlatishning afzalliklaridan biri shundaki, standart shakl yoki tuzilma o‘rnatilishi mumkin va u jumla chegaralarini osonroq aniqlashga yordam beradi.

**Moor assotsiatsiyasiga asoslangan algoritm** – ushbu algoritm asosiy ikkita tamoyilga asoslanib yaratilgan: birinchidan, ular sentence alignerlar phrase (iboraga asoslangan) alignerlarni yaratish uchun yaxshi debocha bo'lishi mumkinligiga qattiq ishonishgan; ikkinchidan, o'sha vaqtga qadar taqdim etilgan algoritmlar, masalan, Braun yuqori hisoblashda birmuncha murakkabliglar va kamchiliklarga ega, ammo past hisoblashda proportsional yaxshi aniqlik bilan ishlaydi. Moor so‘zlarni moslashtirishning uchta turli strategiyasini taqdim etadi [Moore, R. C. 2005. 1-8b).]:

1) bir so‘zni bitta ma’nosi bilan moslashtirish
2) bir so‘zni bir nechta ma’nolari bilan moslashtirish
3) Tokenlarni mosligini tanlash.

Har bir strategiyada bir nechta muammolarni yechish uchun ikki yoki undan ortiq usullar mavjud va ularning barchasi leksikon tarjimasini yaratishda qo‘llanilgan Log-Likelihood-Ratio (LLR) assotsiatsiyasi o‘lchoviga asoslangan.

**DK-vec algoritmi – (**Dinamik K-vec algoritmi) hozircha DK-vec o‘zining ajdodi K-vec algoritmiga asoslangan bo‘lib, u ikkita so‘z bir-birining tarjimasi bo‘lsa, bir segmentda paydo bo‘lish ehtimoli ko‘proq degan faraz ostida ishlaydi. Biroq, bu odatda o‘xshash bo‘lmagan tillarda sodir bo‘lmaydi. Bundan tashqari, k-vec algoritmi tilning eski ma’lumotlarini yoki uning ishlashini pasaytiradigan korpus xususiyatlarini hisobga olmaydi.

**Bleualign -** ushbu sentence aligner Riko Sennrich va Martin Volk tomonidan yaratilgan. Uning asosiy g'oyasi moslashtirish jarayonida yordam berish uchun MT tizimi va tarjima baholovchi, BLEUdan foydalanishdir [Sennrich, R., & Volk, M. 2010. 10-13b]. Alignerni yaxshiroq tushunish uchun BLEU haqida bilish kerak. BLEU - bu MT tarjima natijasi sifatini baholovchi algoritm. Ushbu sifatni o‘lchash uchun BLEU MT tarjima natijasini bir yoki bir nechta inson tarjimalari bilan mosligini taqqoslash orqali baxolanadi.

**Intrative Bleualign** - Ushbu sentence aligner algoritmi 2011 yilda Riko Sennrich va Martin Volk tomonidan moslashtirish jarayonida mashina tarjima qilish tizimidan foydalanishdagi kamchiliklarni chuqurroq tahlil qilish natijasida yaratilgan. Sennrich va Volk MT asosidagi alignerlar manba matnining to‘g‘ri tarjimasiga kuchli bog‘liqligini aniqladilar va MT tizimlari odatda moslangan matnlar bilan oziqlanganligini hisobga olsak, aylana bog‘liqligi mavjudligi aniq bo‘ladi. Ushbu qaramlikni bartaraf etish uchun ushbu algoritm moslashtirishni amalga oshirish uchun bootstrapping (yuklash) usulini taqdim etadi.

**Ikki tilli korpusda jumlalarni moslashtirishning leksik yondashuvlari** turli tillardagi tegishli jumlalarni aniqlash uchun leksik ma’lumotlardan foydalanishni o‘z ichiga oladi. Ushbu yondashuv turli tillardagi tegishli jumlalar o‘xshash leksik tarkibga ega bo‘ladi degan taxminga asoslanadi. Leksik asosda moslashtirishning keng tarqalgan usullaridan biri ikki tilli lug‘atdan foydalanishga asoslangan. Ushbu yondashuv boshqa tilda to‘g‘ridan-to‘g‘ri tarjimalari bo‘lgan bir tildagi so‘zlarni aniqlashni va keyin tegishli jumlalarni moslashtirish uchun ushbu tarjimalardan foydalanishni o‘z ichiga oladi. Misol uchun, agar ingliz tilidagi "cat" so‘zining o‘zbek tilidagi "mushuk" ga to‘g‘ridan-to‘g‘ri tarjimasi bo‘lsa, unda bu so‘zlarni o‘z ichiga olgan jumlalarni moslashtirish mumkin.

Leksik yondashuvda moslashtirishning yana bir usuli statistik modellardan foydalanishni o‘z ichiga oladi. Bu modellar jumlalar orasidagi yozishmalarni birgalikdagi so‘zlarning chastotasiga qarab aniqlash uchun ehtimollik algoritmlaridan foydalanadi. Misol uchun, agar turli tillardagi ikkita jumlada ko‘plab umumiy so‘zlar mavjud bo‘lsa, ular mos kelishi ehtimoli ko'proq. Umuman olganda, leksikaga asoslangan yondashuvlar ikki tilli korpusdagi jumlalarni moslashtirish uchun samarali bo‘lishi mumkin, ayniqsa, birlashtirilayotgan tillar ko‘p qarindosh bo‘lsa (ya’ni, o‘xshash shakl va ma’nolarga ega so‘zlar). Biroq, bu usullar leksik tuzilmalari juda xilma-xil bo‘lgan tillar uchun yoki idiomatik iboralar bilan ishlashda unchalik samarali bo‘lmasligi mumkin.

**Paragraf darajasida moslashtirilgan ikki tilli romanlar korpusi** - bu ikki tildagi matnlar to‘plami bo'lib, ular paragraflari asosida tekislangan. Ushbu turdagi korpusda bir tildagi har bir paragraf boshqa tildagi tegishli paragraf bilan birlashtirilib, matnlarni oson taqqoslash va tahlil qilish imkonini beradi. Paragraf darajasida moslashtirilgan ikki tilli romanlar korpusini yaratish odatda bir necha bosqichlarni o‘z ichiga oladi. Birinchidan, har ikki tildagi romanlarni tanlab, raqamlashtirish kerak. Keyinchalik, har bir romandagi paragraflarni aniqlash va

ajratish kerak. Keyin bir tildagi paragraflar boshqa tildagi tegishli paragraflar bilan mos kelishi kerak. Bu moslashtirish qo‘lda yoki aligner dasturi yordamida amalga oshirilishi mumkin.

Moslash tugallangandan so‘ng, ikki tilli korpus turli maqsadlarda ishlatilishi mumkin, masalan:

1. Qiyosiy adabiy tahlil: Tadqiqotchilar turli tillardagi romanlarning uslublari, mavzulari va hikoya tuzilmalarini solishtirishlari mumkin.
2. Tilni o‘rgatish va o‘rganish: O‘qituvchilar korpusdan til o‘rganuvchilarning o‘qishni tushunish va tarjima qilish ko‘nikmalarini yaxshilash uchun materiallar ishlab chiqishda foydalanishlari mumkin.
3. Mashina tarjimasi: Korpusdan mashina tarjimasi modellarini ularning aniqligi va ravonligini oshirish uchun o‘rgatish uchun foydalanish mumkin.

Umuman olganda, paragraf darajasida moslashtirilgan ikki tilli romanlar korpusi tadqiqotchilar, til o‘rganuvchilar va mashina tarjimasi tizimlarini ishlab chiquvchilar uchun qimmatli manba bo‘lishi mumkin.

**Xulosa.** Korpus tilshunosligida statistik va leksik moslashtirish so‘zlar yoki iboralarni parallel yoki taqqoslanadigan korpuslar o‘rtasida moslashtirish uchun ishlatiladigan ikkita usuldir. Ikkala texnikaning ham afzalliklari va kamchiliklari bor va ulardan foydalanish aniq tadqiqot savollari va tadqiqot maqsadlariga bog‘liq. Statistik moslashtirish katta hajmdagi ma’lumotlar bilan ishlashda va asosiy e’tiborni avtomatik qayta ishlashga qaratilayotganda foydali bo‘lsa, leksik moslashtirish muayyan so‘z yoki iboralarni semantik va sintaktik tahlil qilishga qaratilgan bo‘lsa ko‘proq mos keladi.

## Foydalanilgan adabiyotlar:

1. Brown, P. F., Lai, J. C., & Mercer, R. L. (1991, June). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics.(169-176b)
2. Danielsson, P., & Ridings, D. (1997, February). Practical presentation of a “vanilla” aligner. In *TELRI Workshop in alignment and exploitation of texts*, February.
3. Fung, P., & Church, K. W. (1994, August). K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics*-Volume 2. Association for Computational Linguistics. (1096-1102b)
4. Fung, P., & McKeown, K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. *arXiv preprint cmp-lg/9409011*.
5. Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, *19*(1), (75-102b).

6. Gelbukh, A., Sidorov, G., & Vera-Félix, J. Á. (2006). A bilingual corpus of novels aligned at paragraph level. In *Advances in Natural Language Processing*. Springer Berlin Heidelberg. (16-23 b)

7. Harris, B. (1988). Bi-text, a new concept in translation theory. *Language Monthly*, 54,(8-10b).

8. Kit, C., Webster, J. J., Sin, K. K., Pan, H., & Li, H. (2004). Clause alignment for Hong Kong legal texts: A lexical-based approach. *International Journal of Corpus Linguistics*,(29-51b).

9. McEnery, A., & Xiao, R. Z. (2008). Paralell and comparable corpora: what are they up to?. *Incorporating Corpora: Translation and the Linguist. Translating Europe.Clevendon:Multilingual Matters*.

10. Macklovitch, E., & Hannan, M. L. (1998). Line ‘em up: advances in alignment technology and their impact on translation support tools. *Machine Translation*, 13(1),(41-57b).

11. Meyers, A., Kosaka, M., & Grishman, R. (1998, October). A multilingual procedure for dictionary-based sentence alignment. In *Conference of the Association for Machine Translation in the Americas*. Springer Berlin Heidelberg. (187-198b)

12. Moore, R. C. (2005, June). Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics. (1-8b)

13. Sennrich, R., & Volk, M. (2010, November). MT-based sentence alignment for OCRgenerated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado.

14. Sennrich, R., & Volk, M. (2011, May). Iterative, MT-based sentence alignment of parallel texts. In *18th Nordic Conference of Computational Linguistics, NODALIDA*

15. Simões, A. (2004). *Parallel corpora word alignment and applications* (master thesis).Universidade do Minho, Braga.

16. Natural Language Toolkit (NLTK): https://www.nltk.org/

17. British National Corpus (BNC): https://www.natcorp.ox.ac.uk/

18. Text Encoding Initiative (TEI): https://tei-c.org/