

O'ZBEK TILI KORPUSI ASOSIDA FRAZEOLOGIK BIRLIKLAR MA'LUMOTLAR BAZASINI SHAKLLANTIRISH BOSQICHLARI VA ER SXEMASI

Umid Yodgorov

Alisher Navoiy nomidagi Toshkent davlat
o'zbek tili va adabiyoti universiteti,
Kompyuter lingvistikasi mutaxassisligi 1-bosqich magistranti

Annotatsiya. Ushbu maqolada o'zbek tili korpusi asosida frazeologik birliklar ma'lumotlar bazasini shakllantirish bosqichlari va ER sxemasi haqida fikr yuritiladi. Lingvistik ma'lumotlar (til birliklari)ning ma'lumotlar bazasini yaratish bir necha bosqichda amalga oshiriladi. Frazeologik ma'lumotlar bazasi leksikografik manbalarni boyitish, til o'qitish va lingvistik tadqiqotlar uchun xizmat qiladi. Lingvistik ma'lumotlar bazasining bu ko'rinishini tuzishda ayrim o'ziga xos tomonlarni e'tiborga olish lozim.

Kalit so'zlar: o'zbek tili korpusi, frazeologizm, ma'lumotlar bazasi, ER sxemasi

Abstract. This article discusses the stages of forming a database of phraseological units based on the Uzbek language corpus and the ER scheme. The creation of a database of linguistic data (language units) is carried out in several stages. The phraseological database serves to enrich lexicographic resources, language teaching and linguistic research. Some specific aspects should be taken into account when creating this type of linguistic database.

Keywords: Uzbek language corpus, phrasema, database, ER scheme

Lingvistik ma'lumotlar bazasi so'z, ibora, grammatik shakl, idioma va boshqa lisoniy unsurlarni saqlaydigan ma'lumotlarning tartiblangan to'plamidir. Bunday ma'lumotlar bazalaridan turli maqsadlarda foydalanish mumkin, jumladan:

1. Leksikografiyada: lug‘at va ma’lumotnomalar yaratish.
2. Tabiiy tilni qayta ishlash: avtomatik tarjima, nutqni aniqlash va matn generatsiyasi bilan bog‘liq texnologiyalarni qo‘llab-quvvatlash.
3. Lingvistik tadqiqotlar: til strukturasi, til birliklarining o‘zgarishi va qo‘llanilishini tahlil qilish.
4. Ta’limda: til o‘rganish uchun o‘quv materiallari va vositalarini yaratish.

Ma’lumotlar bazasi ma’lum bir maydonlarga ega jadvallardan tuzilishi mumkin, bu yerda har bir til birligining ta’rifi, gap bo‘laklari, illyustrativ misollar va boshqa metama’lumotlarni qamrab oladi.

Lingvistik ma’lumotlar (til birliklari)ning ma’lumotlar bazasini yaratish bir necha bosqichda amalga oshiriladi. Bu kabi ma’lumotlar bazasini yaratishning bosqichma-bosqich qadamlarini quyidagicha izohlaymiz:

1. Maqsad va talab belgilanadi. Ma’lumotlar bazasini tuzishdan oldin ma’lumotlar bazasi nima uchun ishlatilishi aniqlanadi:

- MB qaysi til birliklarini (so‘z, ibora, grammatik qurilmalar) o‘z ichiga olishi belgilab olinadi;
- har bir birlik haqida qanday ma’lumotlar saqlanadi (ta’rif, so‘z turkumi, misollar, sinonim, antonim va boshqalar);
- ma’lumotlarni qidirish va filtrlash mezonlari aniqlanadi.

2. Platformani tanlash. Ma’lumotlar bazasini saqlash uchun platforma tanlanadi va belgilanadi:

- relyatsion ma’lumotlar bazalari (masalan, MySQL, PostgreSQL, SQLite).
- NoSQL ma’lumotlar bazalari (masalan, MongoDB).

- kichik loyihalar uchun elektron jadval muharrirlari (masalan, Microsoft Excel, Google Sheets).

3. Ma'lumotlar bazasi strukturasini loyihalash. Ma'lumotlar bazasi uchun jadval strukturasini yaratiladi. Asosiy jadvallar quyidagilardan iborat bo'lishi mumkin:

- til birliklari jadvali: har bir birlik haqidagi asosiy ma'lumotlarni o'z ichiga oladi;
- izoh(ma'no)lar jadvali: ta'rif, izoh va qo'llanish holatlari misollari tavsiflanadi;
- grammatik belgilar jadvali: so'z turkumi, shaxs, son va boshqalar grammatik belgilar haqida ma'lumot.

4. Ma'lumotlarni kiritish. Ma'lumotlarni yig'ish va jadvallarga kiritish jarayoni qo'lda bajaridishi mumkin yoki skriptlar yordamida avtomatlashtirilishi mumkin (masalan, matn ma'lumotlarini tahlil qilish uchun Python yordamida).

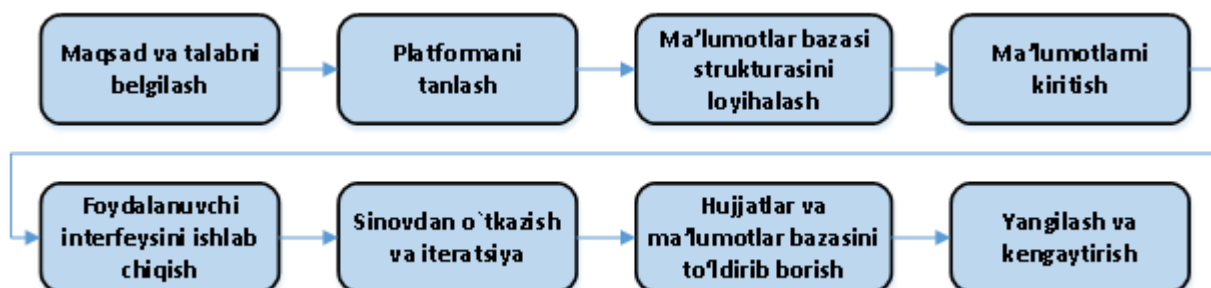
5. Kirish interfeysini ishlab chiqish. Agar ma'lumotlar bazasi boshqa foydalanuvchilar tomonidan foydalanilsa, qulay foydalanish uchun interfeys yaratish maqsadga muvofiq. Bunday interfeysni quyidagi ko'rinishda yaratish mumkin:

- ma'lumotlarni kiritish va izlash shakllariga ega veb-ilova;
- mobil ilova.

6. Sinov va iteratsiya. Ma'lumotlar bazasining maqsadga muvofiq ishlashiga ishonch hosil qilish uchun sinovdan o'tkazish talab etiladi. Foydalanuvchi(lar)ning fikr-mulohazalari asosida xatolar tuzatiladi.

7. Hujjatlar va ma'lumotlar bazasini to'ldirib borish. Ma'lumotlar bazasidan foydalanishda hujjatlar, yangi til birliklari va tuzatishlar qo'shish orqali yangilab turiladi.

8. Yangilash va kengaytirish. Yangi birliklar qo'shish, xatolarni tuzatish va foydalanuvchi ehtiyojlariga ko'ra strukturani yaxshilash orqali ma'lumotlar bazasi muntazam yangilab turiladi.



1.-rasm. Ma'lumotlar bazasini yaratishning bosqichlari

Yuqorida ko'rsatilgan bosqichlarni bajarish orqali funksional va samarali lingvistik ma'lumotlar bazasini yaratishga erishish mumkin.

Yuqorida lingvistik ma'lumotlar yaratishning umumiy tavsifi keltirildi. Quyida frazeologik ma'lumotlar bazasi haqida fikr yuritamiz.

Frazeologik ma'lumotlar bazasi leksikografik manbalarni boyitish, til o'qitish va lingvistik tadqiqotlar uchun xizmat qiladi. Lingvistik ma'lumotlar bazasining bu ko'rinishini tuzishda ayrim o'ziga xos tomonlarni e'tiborga olish lozim.

1. Ma'lumotlar bazasi tarkibiga qo'yilgan talablar:

1. Frazeologik birliklarning tipi (ibora, maqola, turg'un birikma)ni aniqlash.

2. Atributlarni aniqlash:

- frazeologizm (ibora);
- izohi;
- tarjimasi (agar ma'lumotlar bazasi boshqa til ma'lumotlarini ham qamrab olsa);
- misollar;
- mavzu doirasi, kategoriyasi: masalan, tabiat, oziq-ovqat va h.;
- qo'llanilish darajasi;
- sinonim va antonimlari (eagar mavjud bo'lsa).

3. Ma'lumotlar bazasi tarkibini loyihalash talab etiladi:

Ma'lumotlar modeli: ma'lumotlar bazasi tuzilishini shakllantirish. Masalan:

- “iboralar” jadvali (ID, frazeologizm, izohi, kategoriyasi)
- “misollar” jadvali (ID, frazeologizm ID, misol)
- “tarjimalar” jadvali (ID, frazeologizm ID, tarjima, tili)

Lingvistik korpusdan foydalangan holda frazeologik birliklar ma'lumotlar bazasini yaratish tildan olingan real misollar asosida frazeologik birliklarni to'plash, tahlil qilish va ma'lumotlar bazasini tuzishning samarali usuli hisoblanadi. Bu jarayon quyidagi algoritm asosida amalga oshadi:

1. Lingvistik korpusni tanlash. Bu jarayonda ma'lumotlar bazasi uchun manba bo'ladigan korpus tanlanadi. Turli xil matnli materiallarni (masalan, yangilik, badiiy adabiyot, ilmiy maqola) o'z ichiga oladigan lingvistik korpus tanlash maqsadga muvofiq. Frazeologik birlik (idiomani oson qidirish va chiqarib olish uchun maxsus korpus dasturiy ta'minotidan (masalan, Sketch Engine, AntConc) foydalanish mumkin.

2. Frazeologik birliklarni ajratib olish. Frazeologik birliklarni qidirish (turg'un iboralarni topish) uchun tanlangan korpusda qidirish va filtrlash funksiyalaridan foydalanish talab etiladi (masalan, "fe'l + ot" qolipidagi birikmalarni qidirish orqali).

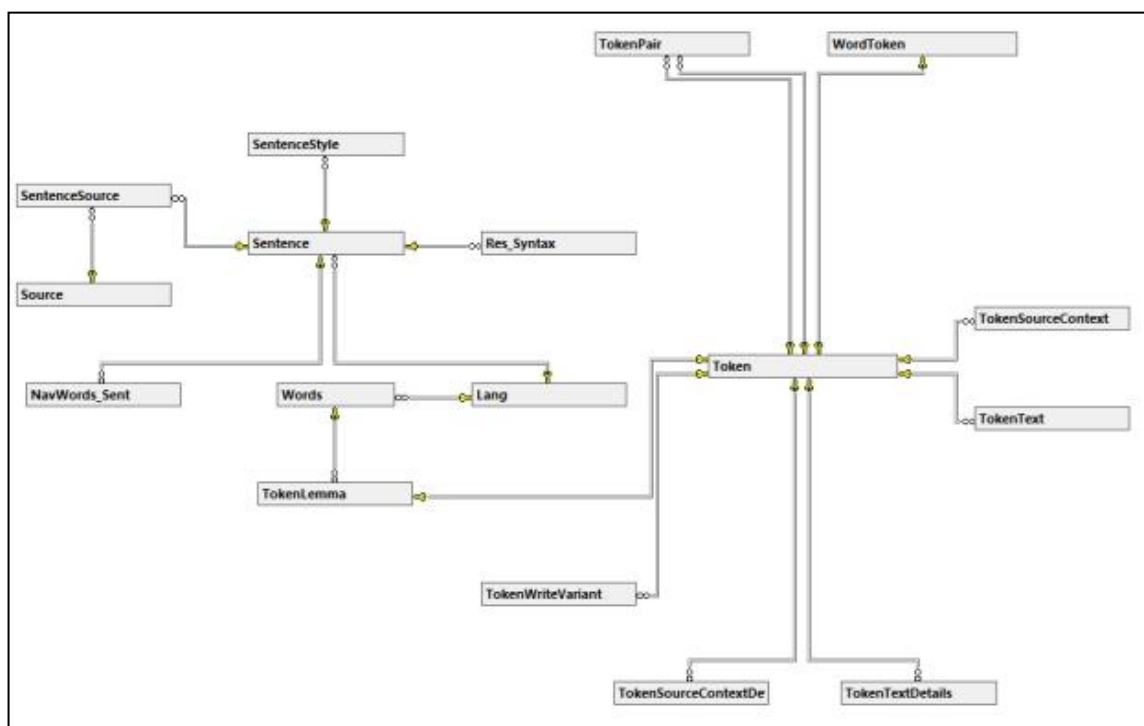
3. Qidiruvni avtomatlashtirish.

- skriptlar va algoritmlar: agar ob'ekt katta hajmdagi matn bo'lsa, oddiy iboralar yoki so'zni qayta ishlashning boshqa usullaridan foydalangan holda iboralarni avtomatik ravishda ajratib olish uchun Python yoki boshqa dasturlash tilidagi skriptlar orqali qidiruvni amalga oshirish mumkin;
- mashinali o'qitish: katta hajmdagi matnlardan idiomalarni tasniflash va ajratib olish uchun mashinali o'rganish usullaridan foydalanish yaxshi natija beradi.

Bular frazeologik birliklar ma'lumotlar bazasini tuzishning umumiy algoritm va tavsifi bo'lib, mazkur bo'limda o'zbek tili korpusi asosida frazeologik birliklarning ma'lumotlar bazasini ishlab chiqishning aniq algoritmiga to'xtalamiz.

Oldingi bo'limlarda aytilganidek, tadqiqot ob'ekti o'zbek tili korpusi bo'lib, frazeologik ma'lumotlar bazasi mazkur korpus asosida shakllantirildi. Buning uchun quyidagi ketma-ketlikda so'rov berish natijasida har bir frazeologik birlikning ro'yxati grammatik paradigmasi topildi, keyinchalik shular asosida ma'lumotlar bazasi shakllantirildi.

O'zbek tili korpusi ma'lumotlar bazasidan frazeologik birliklarni aniqlashda quyidagi ER sxemadan foydalanamiz:



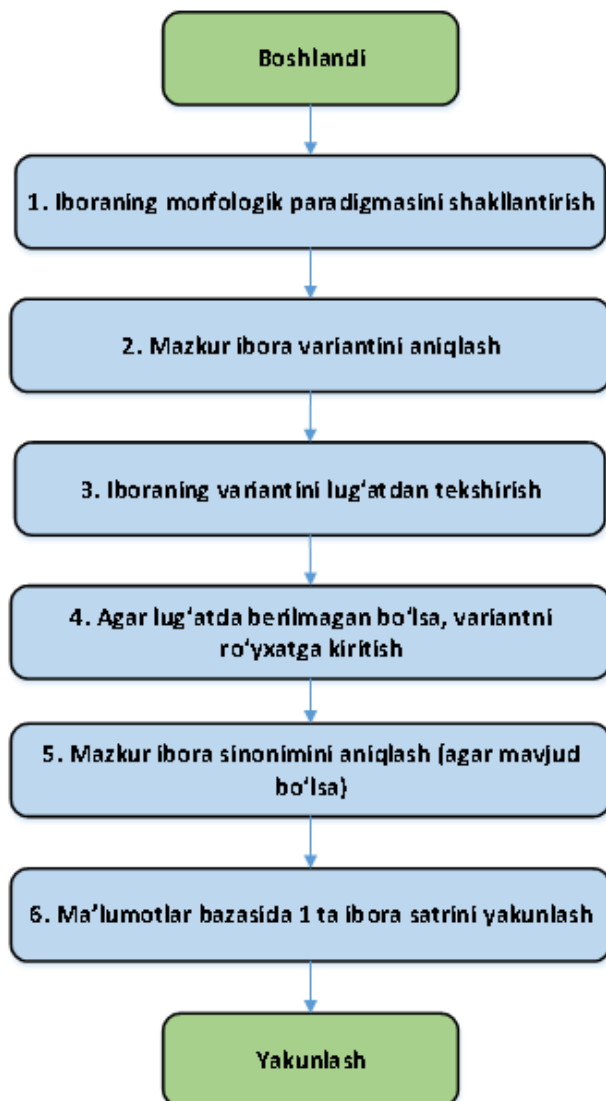
2.-rasm. Frazeologik birliklarni aniqlashda quyidagi ER sxemasi

Ushbu maqolada turli chastotali frazeologik birliklar ma'lumotlar bazasidan qanday o'rin egallashi, ma'lumotlar bazasining tuzilishi va tarkibi, o'zbek tili frazeologik birliklarining statistikasi masalasini tahlil qilamiz.

```
SELECT dbo.Token.Id, dbo.Token.Value, dbo.Token.CorrectVersion,
dbo.WordToken.LemmaValue, dbo.WordToken.LemmaSuffixAffsPart,
dbo.WordToken.StemValue, dbo.WordToken.StemPrefixAffsPart,
dbo.WordToken.StemSuffixAffsPart, dbo.WordToken.RootValue
FROM dbo.Token INNER JOIN dbo.WordToken ON dbo.Token.Id =
dbo.WordToken.TokenId
```

WHERE dbo.WordToken like % “ajali yet(moq)”

Ma'lumotlar bazasini shakllantirish jarayonini ushbu ketma-ketlikda quyidagi biznes-jarayonda ifodalaymiz (Qarang: 3-rasm).



3-rasm. Ma'lumotlar bazasini shakllantirish jarayoni

Demak, ma'lumotlar bazasida mazkur frazeologik birlikning birinchi (original) shakli, morfologik paradigmasi, varianti va sinonimini kiritish mumkin.