ESTABLISHMENT OF A NATIONAL CORPUS THE UZBEK LANGUAGE IS A REQUIREMENT OF A NEW ERA

Guli Toirova Ibragimovna
Doctor of Philosophy, Associate Professor
Bukhara State Univer

2021

Keywords:

corpus, coding, transformation and graphic analysis, technological process, automatic marking, metadata, text information of a large array, morphological marking, spelling module, morphological module, linguistic module, phrase modules, word algorithm, formula algorithm, tabular algorithm, graphic algorithm.

In world linguistics, by the second decade of the twenty-first century, the creation of language corpora on the Internet is the main means of maintaining a particular language, expanding its field of research and demonstrating language skills. In particular, computer technology, which is the great invention of the twentieth century, opens the door to a wide range of possibilities for linguistics as well as other fields and poses important challenges for the computer language, the emergence of computational linguistics is critical to the success of natural languages.
In global language studies, the study of linguistic language modeling, the development of algorithms for lemming words and tags, as well as the electronic use of oral and written monuments, samples of spiritual heritage created in a specific language, in order to increase the use of national and cultural heritage. Particular attention is paid to the processing of information using computer technology, the development of the necessary methodological and software for the introduction of information resources, the development of the language corpus on the Internet and, based on this, the scientific and theoretical aspects of the national language.

On the technological process of building a body that provides the stages of the technological process VV Rykov, Yu.N. Marchuk, I. Melchuk, Sh. Khamroeva

1. Лексик таҳлил – матнни гаплар ва сўзларга ажратиш.
2. Сўзларнинг морфологик таҳлили – сўзлар контекстини (маъно устуворлигини) инобатга олган ҳолда нутқ қисмини, келишигини, турини (родини) ва бошқа грамматик белгиларини аниқлаш.
3. Лемматизация – сўзни дастлабки шаклига (леммега) келтириш.
4. Синтактик таҳлил (dependency parsing) – гаплардаги сўзлар боғланишини аниқлаш, эга ҳамда кесимни излаш ва ҳ.к.з.
5. Соддалаштирилган синтактик таҳлил (chunking) – гапларнинг эга, тўлдирувчи ва ҳол бўйича гуруҳларга ажратилиши.

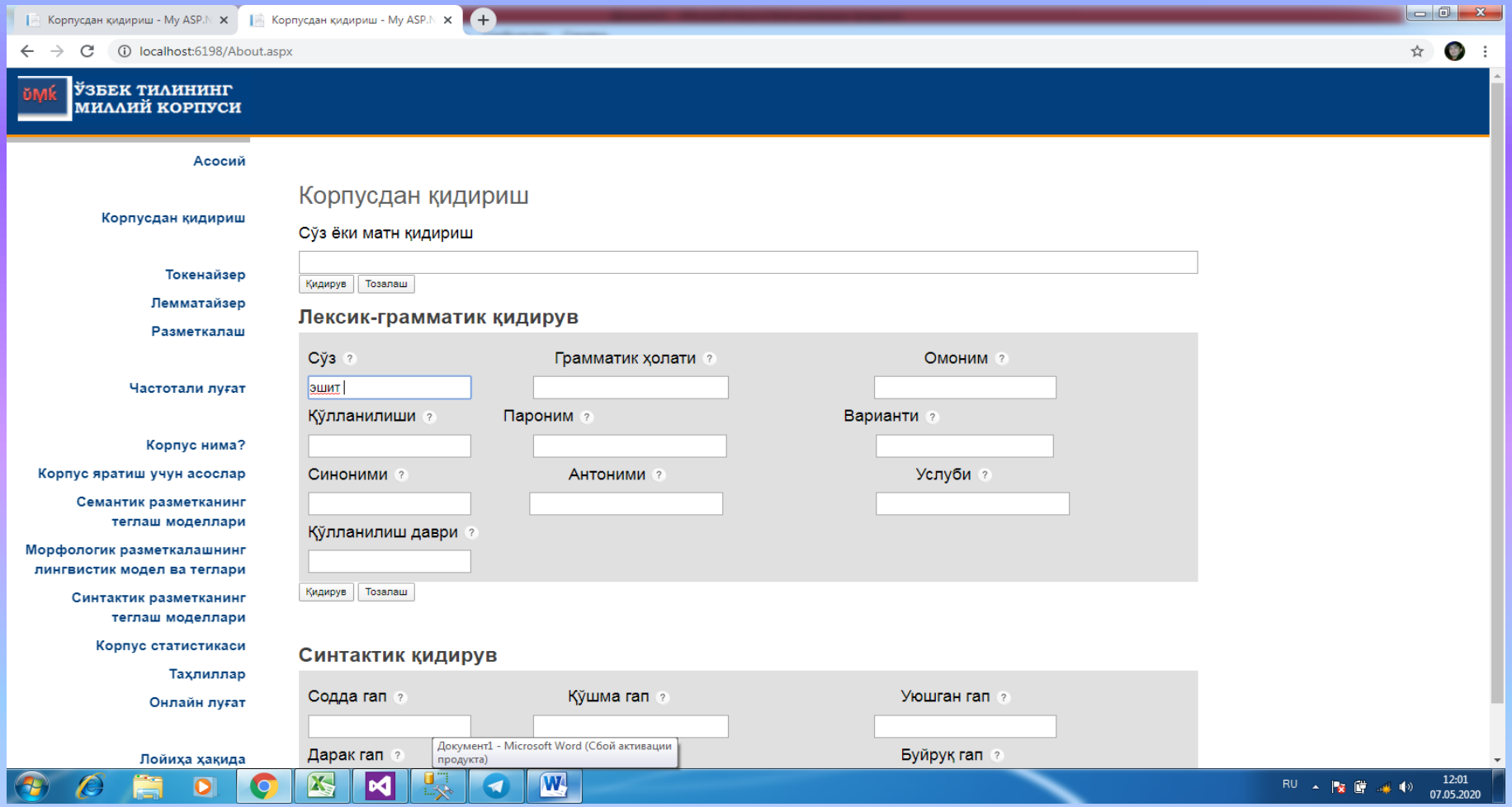
Компьютер технологиясида ишлатиладиган интерфейс қуйидаги турларда бўлади:

Визуал. Мониторда намойиш этиладиган визуал тасвирлар ёрдамида маълумотларни узатадиган стандарт компьютер интерфейси.

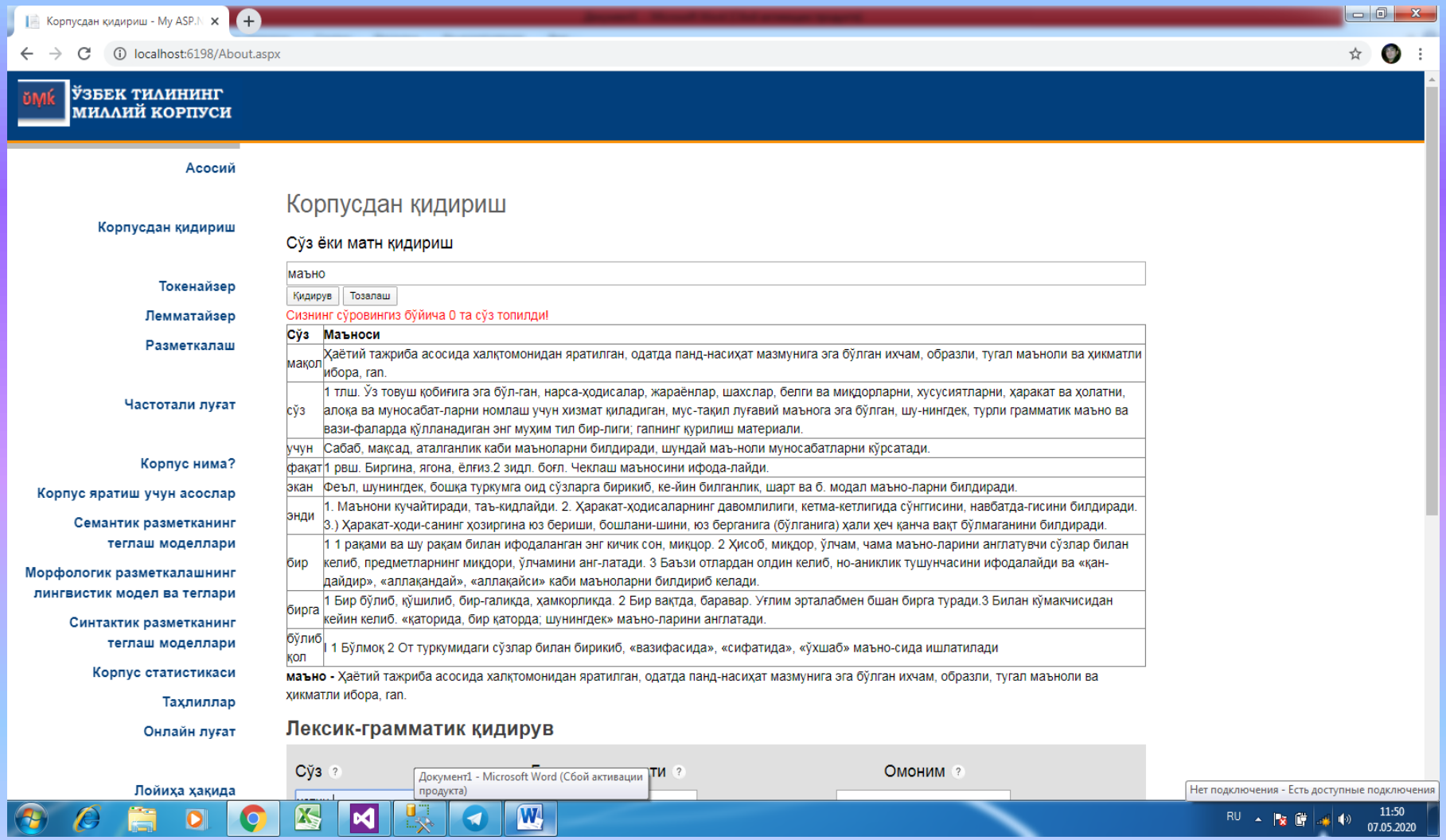
Имо-ишора. Қоида тариқасида, у телефонлар ёки планшетлар учун интерфейс бўлиб хизмат қилади. Кўпгина ҳолларда, бу тизимни бошқарадиган одамнинг бармоқларининг ҳаракатларига жавоб берадиган ва ҳар бир аниқ ҳаракатга маълум даражада жавоб берадиган сенсорли панел. Оддий визуал интерфейснинг соддалаштирилган версияси деб аташ мумкин.

Овоз. Ушбу турдаги интерфейс нисбатан яқинда пайдо бўлди. Овозли буйруқлар ёрдамида тизимни бошқариш имкониятини беради. Тизим, ўз навбатида, фойдаланувчи билан мулоқот орқали ҳам жавоб беради. Энг қизиғи шундаки, замонавий технологиялар бизга нафақат телефонлар ёки компьютерларнинг овозини, балки маиший техника ва ҳатто бортли компьютерларнинг овозини бошқаришга имкон беради.

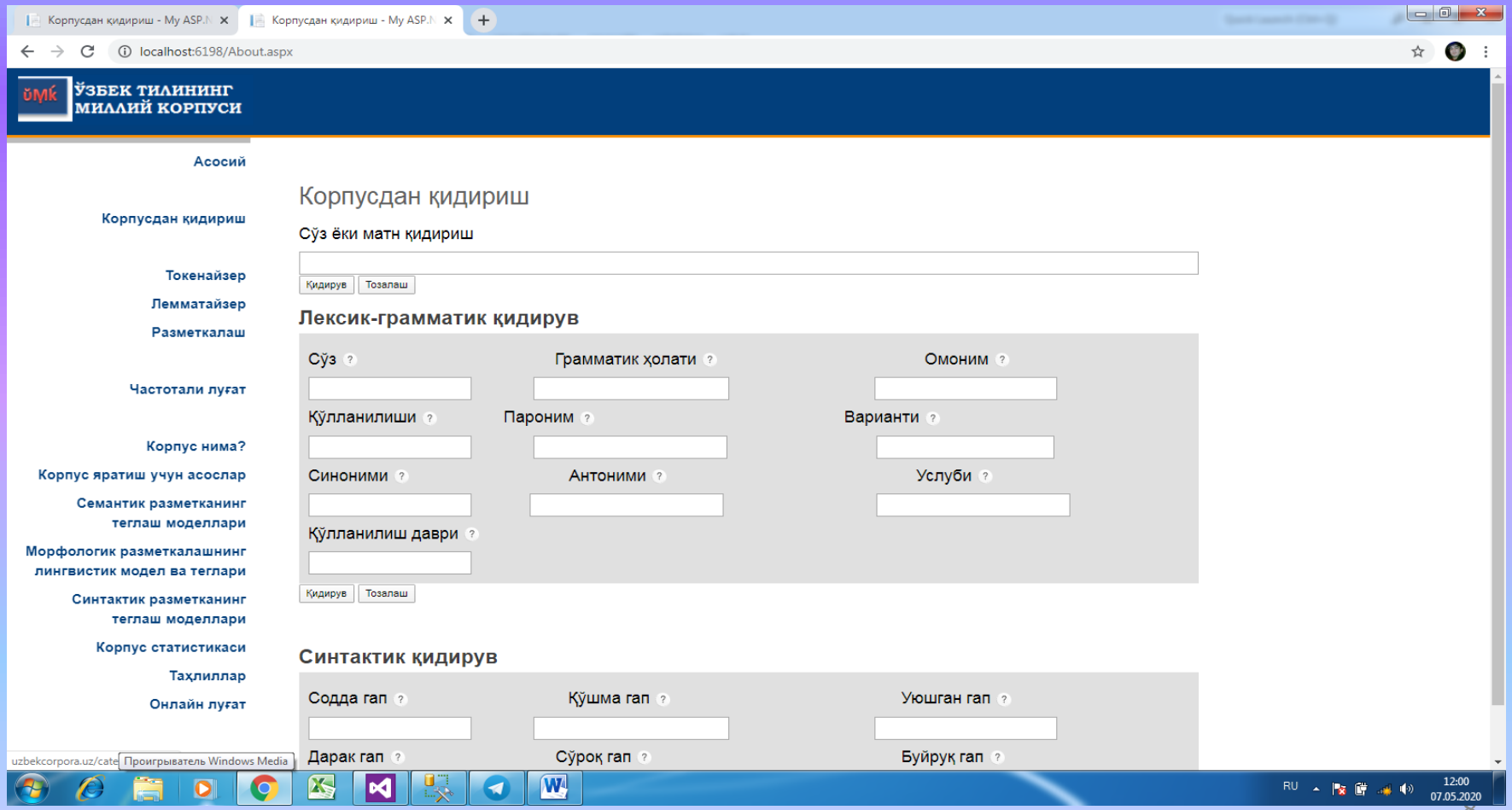
This is the search window of the case, i.e. the interface. "Body search" - this is where you can search for a word or phrase.

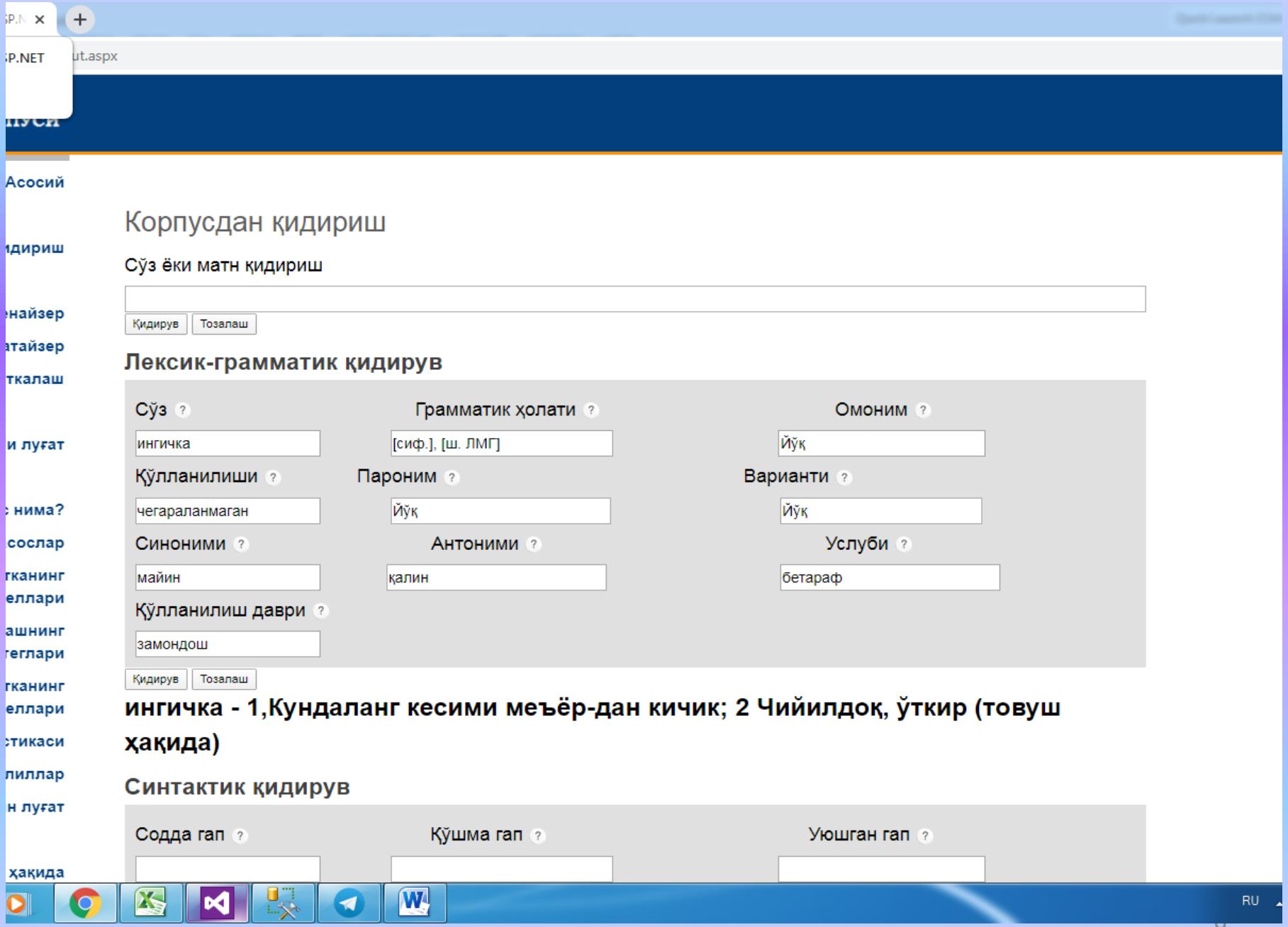


In this case, the result will look like this :

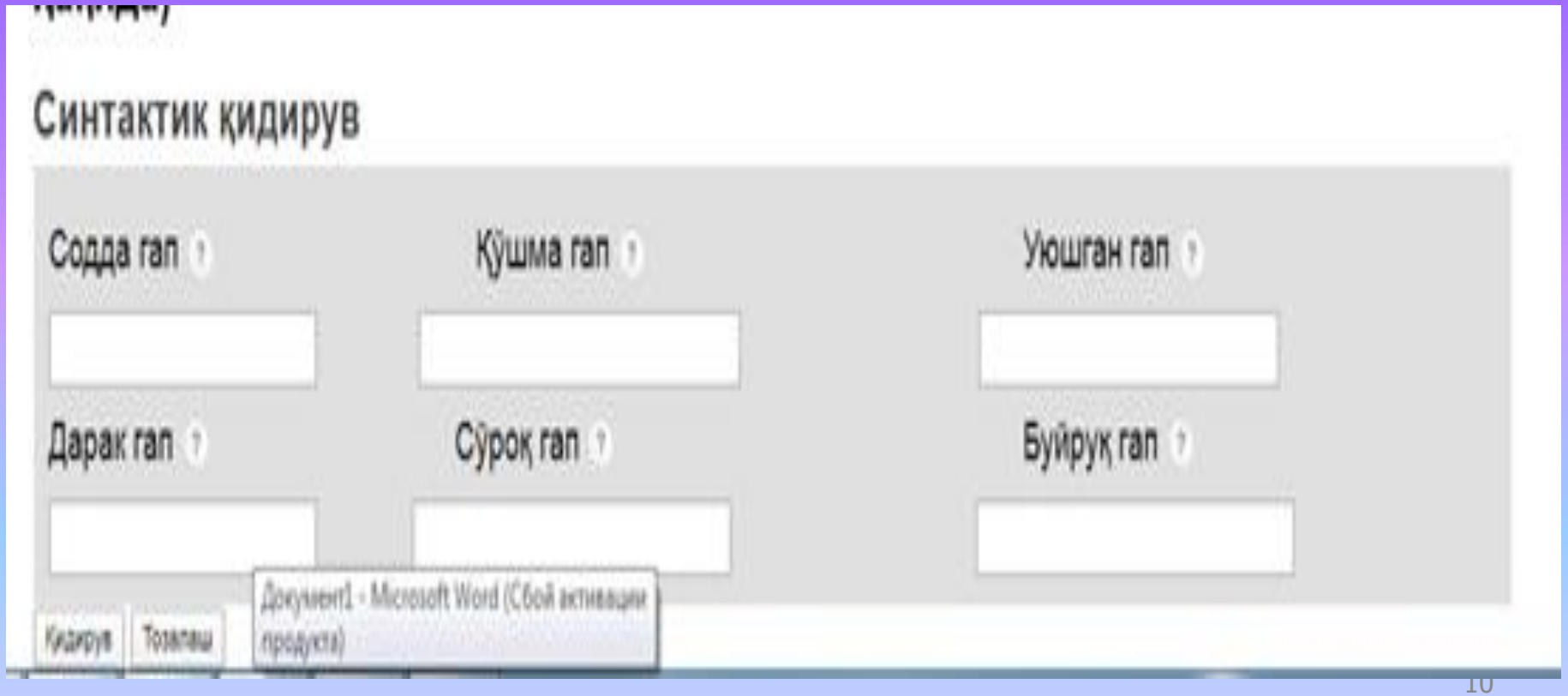


Корпуснинг “Лексик-грамматик қидирув” тугмасидан сўзнинг қуйидаги хусусиятларини билиб олиш мумкин: Сўзнинг маъноси, грамматик ҳолати, синонимлиги, омонимлиги, паронимлиги, антонимлиги, сўзнинг варианти, сўзнинг қўлланилиш даври, сўзнинг қўлланилиш услуби кабилар.




Асосий
идириш
енайзер
атайзер
ткалаш
и луғат
нима?
сослар
ткaнинг
еллари
ашнинг
теглари
ткaнинг
еллари
стикаси
лиллар
н луғат
ҳақида
Корпусдан қидириш
Сўз ёки матн қидириш
Қидирув
Тозалаш
Лексик-грамматик қидирув
Сўз
ингичка
Грамматик ҳолати
[сиф.], [ш. ЛМГ]
Омоним
Йўқ
Қўлланилиши
чегараланмаган
Пароним
Йўқ
Варианти
Йўқ
Синоними
майин
Антоними
қалин
Услуби
бетараф
Қўлланилиш даври
замондош
Қидирув
Тозалаш
ингичка - 1,Кундаланг кесими меъёр-дан кичик; 2 Чийилдоқ, ўткир (товуш ҳақида)
Синтактик қидирув
Содда гап
Қўшма гап
Уюшган гап
RU

Syntactic search button



According to the structure of the sentence, simple sentences are divided into the following groups:

Бир таркибли содда ёйиқ гаплар	Икки таркибли йиғиқ содда гаплар	Икки таркибли ёйиқ содда гаплар
–Бўлсанг-чи! Ҳали замон кун ёйилиб кетади. Далага чиқиш керак.	*Инсон – Ҳазрат! Инсон – Кабир! Бу гапни аммо уққаним йўқ. На падарим ва на онамдан.*	Жиянингизни олиб кетганим келдим (шахси аниқ).
Кеча дунёнинг қиргинли туни экан.	*Уммагали индамади. Ўрнидан туриб у ёқдан–бу ёққа юра б.*	–Гулзорни бузиб ташладилар (шахси ноаниқ).
Адабини бериш керак эди. Ё кўпчилик орасида хўрлаб сўкамизми?	*Ҳаво иссиқ. Кун бўйи югуриб-елиб чарчаган одамлар ва отлар ҳализамон чанқаб сув қидиришга муқаррар.*	Бугунги ишни эртага қўйма (Мақол) (Шахси умумлашган).
Соатли графикдан тўғри фойдаланилсин.	*Айби катта. Уни шундай қолдириб бўлмайди.*	*Кула-кула ўлибмиз. Буни Кифояхоннинг ўзи боплаб айтади.*
Уйланишни кўп кечиктирмаслик керак, иним. Ёшлик ҳам ғанимат (Сўзлашувдан).	*Хонзода бегим йигирма беш ёшда. Айни узатиладиган пайтлари.*	*Начора, ихтиёринг. Нима қилмоқчи бўлаётганингни ўзинг ҳам тушунмасанг керак.*
Эй, қўҳна Шарқимизнинг ҳеч сўнмас машъали.	*Китоб икки юзу ўттиз бетлик экан. Жуда қизиқарли.*	*Ҳамма газета, радио ва телевидение сизнинг ихтиёрингизда!*
Ўқтин-ўқтин шаҳарга тушиб Думасига етиб бориш керак.	*Баҳор сеникидир.*	*Ёзинг! Гапиринг! (ҳозирги замон, буйруқ-истак майли)*
Ҳамманғни қатор қўйиб битта-битта пешонангдан отиш керак.	*Гап шу. Ашрафжон шундай деди-да, ўрнидан туриб кетди, сўнг анчагача тўғон атрофида айланиб юрди.*	*... Ҳа, ҳа, хунуксан! (ҳозирги-келаси замон, аниқлик майли)*
– Тўғри, йўқ. Демак, ойимларнинг устларидан ҳеч қаерга шикоят қилиб бўлмайди.	*Бугун юрсак, эртага яхши йўлларга чиқамиз. Йўл анча.*	*–Қисқаси, бахтиёрсизлар. (ҳозирги-келаси замон, аниқлик майли)*
	Низомжон индамади. Юрагида тўлиб-тошган гапларни	

O'ZBEK TILI MILLIY KORPUSI

Alisher Navoiy
1441-1501

ROMANLARI

HIKOYALARI

PYESA

PORTRET

FELYETON

•Alisher Navoiy hayoti va ijodi
• O`zbek mumtoz adabiyoti

FOTO ALBOM
✦ Buyuk siymolar

TAQDIMOT
✦ Taqdimot

Korpus nima...

Korpus — til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo'ysundirilgan matnlar majmui, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og'zaki, kompьyuterlashtirilgan qidiruv tizimiga dasturiy ta'minot asosida joylashtirilgan on-line yoki off-line tizimda ishlaydigan matnlar jamlanmasi[1]. U elektron shaklda mavjud bo`lgan matnlarni, hujjatlarni ma'lumotlarni qayta ishlash, ularni avtomatik tahlil, ya'ni morfologik sintaktik va semantik tahlil qilish, morfologik analiz va sintez qilish bilan birga, yarim avtomatik rejimda ma'noni buzmasdan ishonchli nutq materialini moslashtirish darajasini tekshirish ushbu sohadagi til bilan qiziqishni oshirishga xizmat qiladi. Korpus – tilni tadqiq etish, til o`rganish, lug`at tuzishda eng zamonaviy, keng imkoniyatli dasturlashtirilgan tizimdir. Korpus - millionlab so`z kontekstlariga havola qiluvchi, maxsus qidiruv tizimi asosida ishlaydigan elektron matnlar yig`indisi.U internet tizimidagi elektron kutubxona, lug`at va lingvistik grammatikadir. Korpus - tabiiy (real) tilning elektron shaklda bo`lib, uni to`ldirish, tuzatish, tahrir qilish imkoniyatiga ega[2]. Korpus bu so`zlar, so`z birikmalari, grammatik shakllari ma'nosini ma'lum bir izlash tizimi orqali topishning elektron shakldagi matnlar to`plamidir. Korpuslarning har xil turlari mavjud. Masalan, bir muallif korpusi[3], bir kitob korpusi (jumladan dastlabki korpuslar "Bibliya" uchun qilingandir). Ma'lum bir til Milliy korpusi shu tilning xayotining barcha qirralarini, janrlarini, usullarini, xududiy va ijtimoiy variantlarini o`zida ifoda etadi. Masalan, o`zbek tili Milliy korpusini Internetda amalga qo`yilsa, 140 mln so`z ishlatmalaridan iborat shu tildagi barcha turdagi matnlarni o`zida mujassamlashtira oladi. Kelajakda o`zbek tili Milliy korpusi o`zida 300 mln so`z ishlatmalarini jamlashi lozim bo`ladi. Boshqa milliy tillar korpuslari kabi o`zbek tili Milliy korpusi xam ikkita muxim xususiyatga egadir. Birinchidan u ancha salobatli va barcha yo`nalishlar uchun (turli badiiy janrlar: publitsistik,o`quv,ilmiy, ish yuritish, nutk - so`zlashuv, shevalar va boshqalar uchun) bir xilda muvofiqlashgandir. Bu matnlar o`z davrida tildagi ta'sir kuchiga qarab proportsional ravishda korpusda jamlangan bo`ladi. Ikkinchidan korpus matnlarga xos bulgan alohida jihatlarni qo`shimcha ma'lumot sifatida o`zida aks ettiradi (masalan ma'lum bir jixatdan belgilanishlar va annotatsiyalar). So`zlar va boshka birikmalar belgilanishi - korpusning bosh xarakteristikasidir (tasnifidir); U korpusni hozirda Internetni to`ldirib yuborgan boshqa matnlar kollektsiyalari va kutubxona matnlaridan farqlaydi. Matnlar belgilanishi qanchali boy va turli tuman bo`lsa korpusning ilmiy ahamiyati shuncha yuqori bo`ladi. O`zbek tili Milliy korpusini belgilanishlarning uch turi e'tiborga olinish kerak bo`ladi: metamatnli (u matnni muallif qarashi, janri va boshka xususiyatlari bo`yicha to`liq ifodalaydi), morfologik va semantik belgilanishlar (morfologik va semantik belgilanishlar matnni emas balki alohida bir so`zni tasniflab beradi). ADABIYoTLAR: 1. Grudeva V.A. Korpusnaya lingvistika. Uchebnoe posobie. 2-e izd., stereotip.– M.: Flinta, 2012. – 165 s. 2. Zaxarov V.P. Korpusnaya lingvistika: Uchebno-metod. posobie. – SPb., 2005. – 48 s. 3. Zaxarov V.P., Bogdanova S.Yu. Korpusnaya lingvistika.– Irkutsk: IGLU, 2011. – 161 c. 4. Toirova G.I. Importance of Interface in Creating Corpus. International Journal of Recent

SO`Z QIDIRISH

Hikoya nomi:
Kartina
Qidiruv

LEKSIK-GRAMMATIK QIDIRUV

- So'z
- Grammatik holati
- Omonim
- Qo`llanilishi
- Paronim
- Varianti
- Sinonimi
- Antonimi
- Uslubi
- Qo`llanilish davri

Qidiruv

SINTAKTIK QIDIRUV

- Gapning tuzilishiga ko`ra turlari
- Gapning maqsadiga ko`ra turlari

Qidiruv

MUALLIFLAR

Mualliflar

DASTUR HAQIDA

Ushbu korpus qidiruv natijasi shu hiqoya matni asosida amalga oshiriladi. Korpus

Conclusion

1. Corpus linguistics is the most advanced branch of linguistics, and corpus is a necessary tool for linguists; oral, written monuments are a source of information reflecting the national and cultural heritage. A corpus is a set of texts to be searched for, and a well-defined corpus serves as a stable linguistic base to ensure the effectiveness of linguistic research. As an artificial intelligence product, the linguistic corpus includes an electronic dictionary, translation portal, terminology database, virtual (electronic) library, electronic government, electronic publications, electronic textbooks and manuals. Linguistic electronic sources, which are a product of artificial intelligence, are considered raw materials for creating a certain linguistic corpus.

2. The creation of the National Corpus is carried out in two stages: determination of the list of sources and digitization of texts (transformation into computer form). Its technological process consists of: creating a dictionary of repetitions of lexemes and word forms based on selected texts; view the text for any unit of the received dictionary of repetitions; divide a graphic word into syllables and compose a dictionary of repetitions of syllables; sorting word resources; simultaneous processing of an unlimited number of files; create text corpora with external symbols; calculation of statistical data for the corpus of created texts and individual texts included in the corpus; work with source texts in txt, doc i rtf format, automatic encoding setting, etc.

3. The most effective standards for encoding corpus data have been selected. The presentation of its data is based on an SGML / XML text layout. When encoding, lexical information is adapted to HTML / XML rules. The texts selected for the National Corpus are taken from various sources and presented in different formats: plain text, HTML, RTF, PDF.

Thank you for your attention !!!