მანანა თანდაშვილი
ზაქარია ფურცხვანიძე

კორპუსლინგვისტიკა

(გლოსარიუმი და რეგისტრი)

Manual of Corpus Linguistics

დიგიტალური ჰუმანიტარია – ქართველოლოგია
და XXI საუკუნის გამოწვევები

Frankfurt University Press

2014

წინამდებარე ნაშრომი ფრანკფურტის ლინგვისტური წრის ინიციატივით შეიქმნა. აღნიშნული წრე ნებაყოფლობითი სასწავლო-სამეცნიერო რგოლია და დაარსდა ფაკულტატიური სემინარების ბაზაზე, რომელსაც ფრანკფურტის გოეთეს სახ. უნივერსიტეტის ემპირიული ენათმეცნიერების ინსტიტუტში ატარებდა პროფ. მანანა თანდაშვილი კავკასიური ენების კვლევაში სტუდენტთა და დოქტორანტთა აქტიურად ჩაბმის მიზნით. სემინარებზე განხილული მრავალფეროვანი თემატიკიდან გამომდინარე, 2010 წლიდან სემინარი გადაიქცა მუდმივმოქმედ აკადემიურ ორგანოდ და ეწოდა **ფრანკფურტის ლინგვისტური წრე**. ლინგვისტური წრის სხდომები რეგულარულად იმართება ზამთრისა და ზაფხულის სემესტრების დროს, კვირაში ერთხელ (პარასკევობით) და საშუალებას აძლევს სტუდენტებსა და დოქტორანტებს სადისკუსიოდ წარმოადგინონ რეფერატები, მოხსენებები, საკურსო, სამაგისტრო და სადისერტაციო ნაშრომები.

2011 წლიდან ფრანკფურტის ლინგვისტური წრე საქართველოშიც იწყებს აქტიურ მოღვაწეობას და წელიწადში ერთხელ აწყობს საორიენტაციო შეხვედრებს გერმანიაში კავკასიოლოგიის შესწავლის მსურველთათვის. ფრანკფურტის ლინგვისტური წრის ერთ-ერთ მთავარ აქტივობას წარმოადგენს საზაფხულო სკოლების ორგანიზება საქართველოში. განსაკუთრებით ნაყოფიერი იყო ამ თვალსაზრისით ბათუმის შოთა რუსთაველის სახელობის უნივერსიტეტში 2012 და 2013 წლებში ჩატარებული ბათუმის საზაფხულო სკოლები, რომელიც მიეძღვნა კორპუსლინგვისტიკასა და ქართველოლოგიის სპეციალურ პრობლემს.

2011 წელს ფრანკფურტის ლინგვისტურმა წრემ გამოსცა სამეცნიერო კრებული Folia Caucasica, რომელიც მიეძღვნა პროფ. იოსტ გიპერტის 55 წლის იუბილეს. წინამდებარე ნაშრომი ფრანკფურტის ლინგვისტური წრის მეორე ბეჭდვითი გამოცემაა და შეიქმნა ბაკურიანის ზამთრის სკოლისათვის საერთაშორისო პროექტის "დიგიტალური ჰუმანიტარია – ქართველოლოგია და XXI საუკუნის გამოწვევები" ფარგლებში ივ. ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტთან თანამშრომლობით.

ავტორთა წინასიტყვაობა

სიტყვა **ეპოქა** დროის კონკრეტულ პერიოდს აღნიშნავს, რომელიც ადამიანის ცნობიერებისა და ყოფიერების სფეროში ახალი მიღწევებით, კაცობრიობის განვითარების სიახლეებით არის ნიშანდებული. ჩვენც კონკრეტული ეპოქის ადამიანები ვართ და ამ ეპოქას **დიგიტალური ეპოქა** ჰქვია – ჩვენი ყოველდღიურობა გაჯერებულია ციფრული სამყაროს თანამედროვე ატრიბუტიკით – ბინარული ციფრების – ნულისა (0) და ერთის (1) კომბინაციის უსასრულო ჯაჭვით.

დიგიტალურმა ეპოქამ მეცნიერების ყველა დარგი გააციფრა. მეტიც, მთელი რიგი დარგთაშორისი მიმართულებები წარმოშვა. კორპუსლინგვისტიკაც ამგვარი ინტერდისციპლინარული დარგია – ლინგვისტიკისა და ინფორმატიკის ზღვარზე აღმოცენებული.

საქართველოში კორპუსლინგვისტიკა ახლა იდგამს ფეხს. დარგის დამკვიდრება და განვითარება კი არა მარტო ახალგაზრდა სპეციალისტების გაზრდას გულისხმობს, არამედ დარგობრივი მეტაენის შექმნასაც ქართულად. გვინდა შეგახსენოთ, რომ გასული საუკუნის დასაწყისში, ეროვნული უნივერსიტეტის შექმნის პირველსავე წლებში, ქართველი მეცნიერები დაუშრომლად იღვწოდნენ მეცნიერების სხვადასხვა დარგებისათვის

მეტაენის ქართულად შესაქმნელად. წინამდებარე გლოსარიუმი ფრანკფურტის ლინგვისტური წრის მიერ გამოცემული ნაშრომია. ბათუმის საზაფხულო სკოლის ჩატარებამ კორპუსლინგვისტიკაში (2012-2013 წ.) გამოცდილება შეგვძინა ამ წიგნის ავტორებს და თვალნათლივ დაგვანახა, რომ საქართველოში კორპუსლინგვისტიკის განვითარების უპირველეს ამოცანას კორპუსლინგვისტიკის ტერმინების გაქართულება წარმოადგენს.

წინამდებარე ნაშრომი, რომელიც საგანგებოდ დაიწერა ბაკურიანის ზამთრის სკოლისათვის, კორპუსლინგვისტიკის ცნებების გაქართულებისა და ქართველი მკითხველისათვის მათი განმარტების პირველ ცდას წარმოადგენს.

ავტორები დიდ მადლობას უხდიან იოსტ გიპერტს, ნინო შარაშენიძეს, ლელა სამუშიას და ნათია დუნდუას მხარდაჭერის, თანამშრომლობისა და საინტერესო შენიშვნებისათვის.

ფრანკფურტი, 2014

განმარტებები მომხმარებელთათვის

წინამდებარე გლოსარიუმი კორპუსლინგვისტიკაში ორი ნაწილისაგან შედგება: გლოსარიუმისა და კორპუსის შერჩევითი რეგისტრისაგან.

გლოსარიუმში წარმოდგენილია კორპუსლინგვისტიკის საბაზისო ცნებები განმარტებებითურთ, რომლებიც დალაგებულია ანბანთრიგზე. გლოსარიუმში შეტანილ ყველა ცნებას (რამდენიმე გამონაკლისის გარდა) ახლავს შესაბამისი შესატყვისი ინგლისურად. მაგ.: **ლინგვისტური ანოტაცია linguistic annotation**.

ცნების განმარტებას საჭიროების შემთხვევაში ახლავს საილუსტრაციო მასალაც. მაგ.:

ანოტაციის დონეები annotation levels – ლინგვისტური მეტაინფორმაციის დამოუკიდებელ ბწკარებზე (ანოტაციის სხვადასხვა დონეებზე) ასახვის ფორმა. ანოტაციის დონეები იმავდროულად წარმოადგენენ გრამატიკულად რელევანტურ დესკრიფციულ პერსპექტივებს მონაცემთა ბაზასთან მიმართებაში. ანოტაციის სრულყოფილი ფორმაა მრავალდონიანი ანოტაცია, რომელიც ლინგვისტური ინფორმაციის სხვადასხვა ენობრივ დონეზე გადანაწილების ეფექტურ ფორმას წარმოადგენს.

მაგალითი:

L1	თხა-მ	ვენახ-ი	შეჭამ-ა
L2	*txa-m*	*venax-i*	*šeč̣am-a*
L3	goat	vineyard	eat
L4	N	N	V-tr.
L5	ERG.Sg.	NOM.Sg.	AOR.S3
L7	A	DO	PRD
L10	“The goat ate up the vineyard.”		

განმარტებებში გამოყენებული ცნებების შემთხვევაში მას ერთვის ე.წ. ბმული შესაბამის ცნებაზე. ბმული ჩასმულია ფრჩხილებში და მარკირებულია ბმულის აღმნიშვნელი საგანგებო ნიშნით – ↗. მაგ.:

> **ლინგვისტური ანოტაცია linguistic annotation –** ანოტირების სახე, ენობრივი მონაცემების ლინგვისტური ანალიზის შედეგად მოპოვებული ინფორმაცია, რომელიც ანოტაციის (↗ანოტაცია) ერთ ან რამდენიმე დონეზე (↗ანოტაციის დონეები) გადანაწილდება.

ნაშრომის მეორე ნაწილი წარმოადგენს სხვადასხვა ენებისათვის შექმნილი კორპუსების შერჩევით რეგისტრს. რელევანტური მონაცემები კორპუსის შესახებ – ე.წ. მეტამონაცემები – რამდენიმე მახასიათებლით არის წარმოდგენილი და ინფორმაციას აძლევს მომხმარებელს, რომელი ენისათვის არის შექმნილი იგი, რა მოცულობისაა კორპუსი, ვინ არიან კორპუსის ავტორები (პროექტის ან პიროვნებათა სახელების მითითებით). ასევე მოცემულია ვებგვერდის მისამართი და კორპუსის მოკლე აღწერა.

მაგ.:

Language:	Armenian
Indication:	Leiden Armenian Lexical Textbase
Kind:	Text and Corpora Meta Sites
Size:	80.000 Armenian lexemes and ten texts
Link:	http://www.sd-editions.com/LALT/home.htm
Description:	The complete Nor Bargirk, main sections of Adjarian's Root Dictionary, Bedrossian's Armenian-English Dictionary and other material are integrated in LALT...

შემოკლებები

ბერძ. – ბერძნული
ე.წ. – ეგრეთ წოდებული
იხ. – იხილე
ლათ. – ლათინური
მაგ. – მაგალითად
მრ. – მრავლობითი
შდრ. – შეადარე
vs. – versus

A	agens	აგენსი
ADJE	adjective	ზედსართავი სახელი
AOR	aorist	აორისტი
ANPH	anaphora	ანაფორა
ATT	attributive	ატრიბუტი
ADV	adverb	ზმნიზედა
DAT	dative	მიცემითი
DO	direct object	პირდაპირი ობიექტი
DPRON	dem. pronoun	ჩვენებითი ნაცვალსახელი
ERG	ergative	მოთხრობითი
L	level	დონე
intr	intransitive	გარდაუვალი
N	noun	არსებითი სახელი
NOM	nominative	სახელობითი
NP	nom. phrase	სახელური ფრაზა
PARTC	particle	პარტიკლი
PART	participle	მიმღეობა
Pl	plural	მრავლობითი რიცხვი
POS	part of speech	მეტყველების ნაწილი
PP	postposition	თანდებული
PRD	predicate	პრედიკატი
PRON	pronoun	ნაცვალსახელი
REF	referent	რეფერენტი

S	subject	სუბიექტი
Sg	singular	მხოლობითი
tr	transitive	გარდამავალი
V	verb	ზმნა

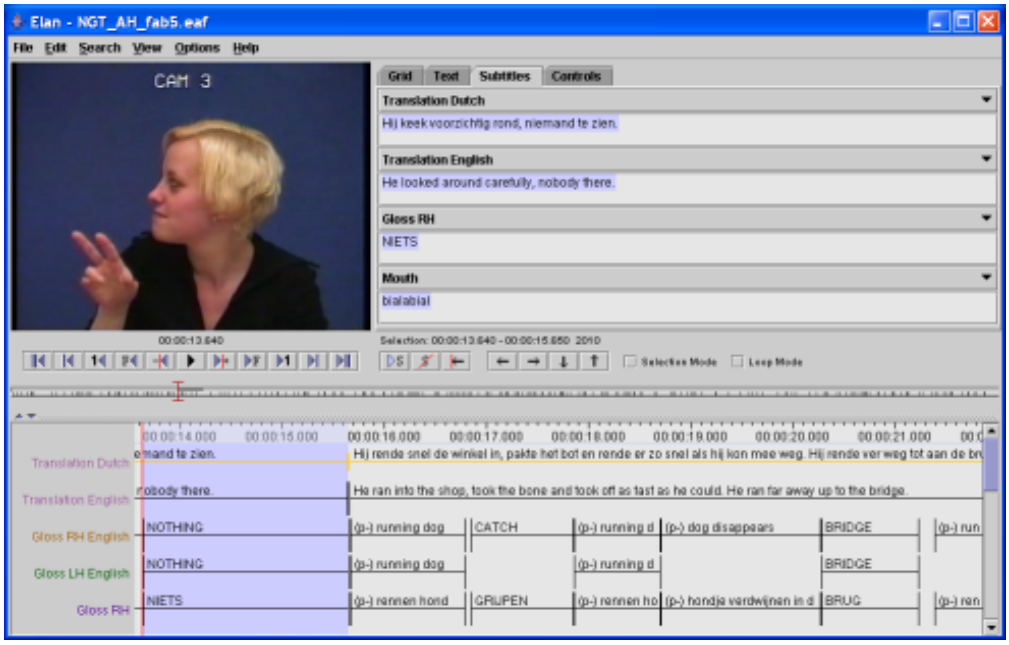
ანაფორა anaphora (ბერძ. ἀναφορά უკუმიმართება) – უკუმსწრები ელემენტი, რომელიც მიემართება წინადადებაში ან დისკურსში (↗დისკურსი) უკვე ნახსენებ სიტყვას ან სიტყვათა ჯგუფს. რიტორიკაში ანაფორა "გამეორებას" გულისხმობს და სტილისტურ საშუალებას წარმოადგენს. მაგ. "ქარი ჰქრის, ქარი ჰქრის ქრის, ქარი ჰქრის..." (გ. ტაბიძე). ენათმეცნიერებაში ანაფორა პრაგმატული დონის ელემენტია, ლექსიკური შინაარსისაგან არის დაცლილი და გამოიყენება როგორც სიტყვის ან სიტყვათა ჯგუფის შემცვლელი ფუნქციური ელემენტი. ქართულში ანაფორის ფუნქციით ძირითადად **დეიქტური ელემენტები** (↗დეიქსისი, ↗დეიქტიკონი) არის გამოყენებული – ნაცვალსახელი, ზმნიზედა, წინდებული. მაგ.: *„კაცი <u>სახლში</u> დაბრუნდა, მაგრამ **იქ** არავინ დახვდა“*; *„<u>მაგიდა</u> კარებთან იყო მიდგმული, **ზედ** არაფერი იდო.“* ისეთ შემთხვევაში, როდესაც დეიქტური ელემენტი მომდევნო სიტყვას ან სიტყვათა ჯგუფს მიემართება წინადადებაში ან დისკურსში, წინმსწრები ელემენტი წარმოადგენს **კატაფორას** (↗კატაფორა). მაგ.: *„**ის**, ვინც <u>ვერ შეძლებს რეფერატის მომზადებას</u>, სერტიფიკატს ვერ მიიღებს“*. კატაფორა ხშირად მთავარ წინადადებაში გვხვდება და განსაზღვრებითი დამოკიდებული წინადადება მას, როგორც მისამართ სიტყვას, მიემართება.

ანაფორული ანოტაცია anaphoric annotation – ანოტაციის ისეთი ფორმა, სადაც ანაფორული რეფერირების (↗რეფერენცია) გამომხატველი ელემენტები ცალსახადაა მარკირებული.

მაგ.

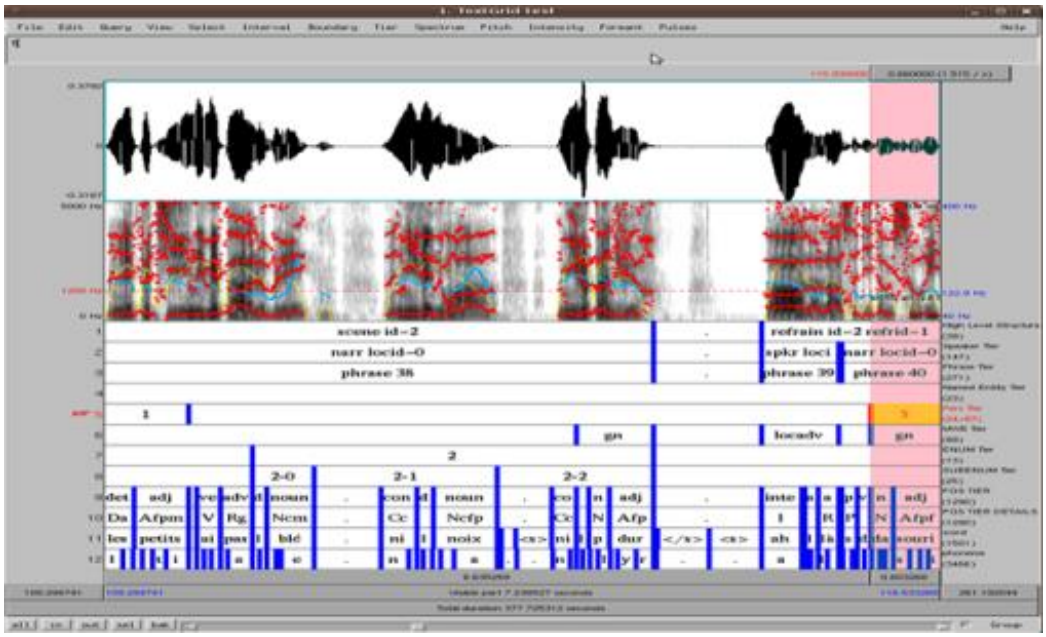
L1	კაცი	სახლში	დაბრუნდა,	მაგრამ	იქ	არავინ	დახვდა.
L2	*ķaci*	*saxlši*	*dabrunda,*	*magram*	*ik*	*aravin*	*daxvda.*
L3	man	hause	return	but	there	nobody	to be
L4	N	N	V-intr.	PARTC	ADV	PRON	V-intr.
L5	NOM.	DAT.PP	AOR.S3	-	–	NOM.	AOR.S3
L8	-	*REF*	-	–	*ANPH*	–	-

ამ წინადადებაში ზმნიზედა ***იქ***, რომელიც დეიქტური ელემენტის (↗დეიქსისი, ↗დეიქტიკონი) ფუნქციით არის გამოყენებული, მიემართება მთავარ წინადადებაში რეალიზებულ ადგილის გარემოებას – მიცემითში მდგარ თანდებულიან არსებით სახელს ***სახლში*** – და ანაფორას წარმოადგენს.



ანოტაცია annotation – კორპუსის შემადგენელი კომპონენტი, რომელიც კორპუსული კვლევის განხორციელებისათვის აუცილებელ სამი სახის ინფორმაციას შეიცავს: მეტამონაცემი (*metadata*), ტექსტის მარკირების ფორმა (*textual markup*) და ლინგვისტური ანოტაცია (*linguistic annotation*).

ლინგვისტური ანოტაცია linguistic annotation – ანოტირების სახე, ენობრივი მონაცემების ლინგვისტური ანალიზის შედეგად მოპოვებული ინფორმაცია, რომელიც ანოტაციის (↗ანოტაცია) ერთ ან რამდენიმე დონეზე (↗ანოტაციის დონეები) გადანაწილდება. ლინგვისტური ანოტაციის დროს განარჩევენ დესკრიფციულ და ანალიტიკურ ანოტაციას. დესკრიფციული ანოტაცია ძირითადად ფორმალური ნიშნების მიხედვით ახდენს ენობრივი მონაცემების (სიტყვაფორმების) დახასიათებას, ანალიტიკური ანოტაციის დროს ყურადღება ექცევა არაფორმალურ მახასიათებლებსაც. ლინგვისტური ანოტაცია ელექტრონული მონაცემის ლინგვისტური მეტაინფორმაციით გამდიდრების ერთ-ერთი ფორმაა და წარმოადგენს პირველად ენობრივ მონაცემებში ბუნებრივად კოდირებული ლინგვისტური ინფორმაციის ვიზუალიზაციის (↗ვიზუალიზაცია) ყველაზე გავრცელებულ ფორმას.



კორპუსლინგვისტიკის განვითარების ერთ-ერთ უმნიშვნელოვანეს ეტაპად მიჩნეულია ლინგვისტური ანოტაციის პროცესის პროგრამული ავტომატიზაცია. ამ მიზნის მისაღწევად კორპუსის პროგრამულ ინფრა-სტრუქტურაში ჩაშენებულია სპეციალური ინსტრუმენ-ტები, რომელთა საშუალებითაც ხორციელდება მიმარ-თება სხვადასხვა სახის პირველად მონაცემთა ბაზებთან (როგორც წესი, ასეთია, მაგ., ლექსიკონი და მორფემათა ბანკი), რაც ლექსემათა ავტომატური ანოტირების საშუალებას იძლევა. ამგვარი მექანიზმით აღჭურვილ კორპუსს ანოტირებული კორპუსი ეწოდება. ანოტირებული კორპუსი ყველაზე მოქნილ რესურსს წარმოადგენს ლინგვისტური კვლევისთვის ენის სხვადასხვა დონეზე. დღესდღეობით ანოტაციის ყველაზე გავრცელებული ფორმა არის POS-ანოტაცია (Part Of Speech Annotation). სამეცნიერო ნაშრომებში გამოყენებული ტექსტის ლინგვისტური ანოტაცია, კორპუსისაგან განსხვავებით, მანუალურად იქმნება და შრომატევადია. ილუსტრირებისათვის იხ. მანუალური ლინგვისტური ანოტაციის (↗მანუალური ანოტაცია) მაგალითი:

თხა-მ	ვენახ-ი	შეჭამ-ა
txa-m	*venax-i*	*šeč̣am-a*
goat:ERG.Sg.	vineyard:NOM.Sg.	eat:AOR.S3

“The goat ate up the vineyard“.

ანოტაციის დონეები annotation levels – ლინგვისტური მეტაინფორმაციის დამოუკიდებელ ბწკარებზე (ანოტაციის სხვადასხვა დონეებზე) ასახვის ფორმა. ანოტაციის დონეები იმავდროულად წარმოადგენენ გრამატიკულად რელევანტურ დესკრიფციულ პერსპექტივებს მონაცემთა ბაზასთან

მიმართებაში. ანოტაციის სრულყოფილი ფორმაა მრავალდონიანი ანოტაცია, რომელიც ლინგვისტური ინფორმაციის სხვადასხვა ენობრივ დონეზე გადანაწილების ეფექტურ ფორმას წარმოადგენს.

მაგალითი 1:

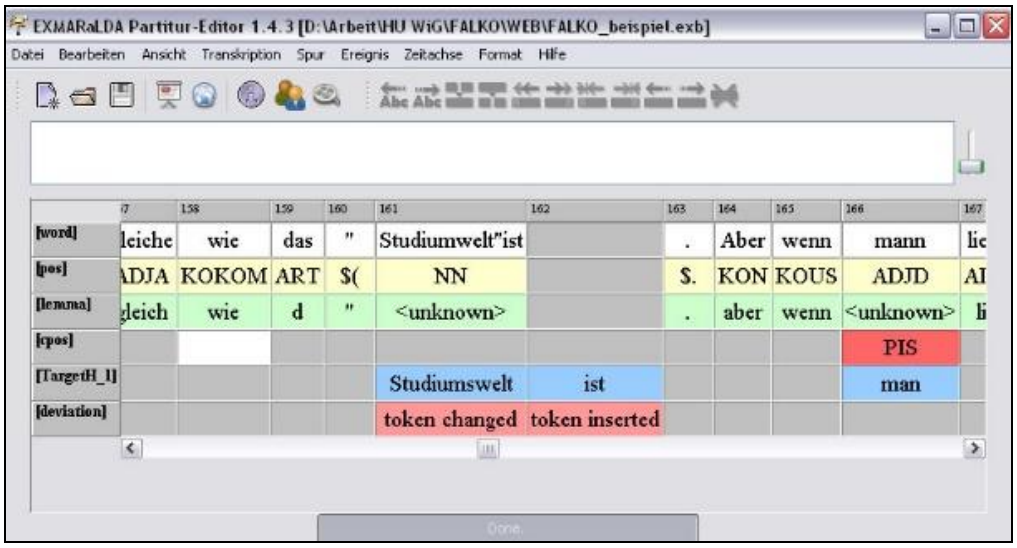
L1	თხა-მ	ვენახ-ი	შეჭამ-ა
L2	*txa-m*	*venax-i*	*šeč̣am-a*
L3	goat	vineyard	eat
L4	N	N	V-tr.
L5	ERG.Sg.	NOM.Sg.	AOR.S3
L7	A	DO	PRD
L10	“The goat ate up the vineyard.”		

მაგალითი 2:

L1	ეს	ვირუს-ის	მ-ა-ტარ-ებ-ელ-ი	კაც-ი
L2	*es*	*virus-is*	*m-a-ṭar-eb-el-i*	*ḳac-i*
L3	that	virus	carrying	man
L4	DPRON	N	PART.	N
L5	NOM.	GEN.	NOM.	NOM.
L6	ATT.	ATT.	ATT.	HEAD
L8	D1	-	-	

განმარტება: L1 – ენობრივი მონაცემი ორიგინალ ენაში, L2 – ტრანსკრიფციის დონე, L3 – ლექსიკური დონე, L4 – POS-დონე (მეტყველების ნაწილები), L5 – მორფოლოგიური დონე, L6 – მორფოსინტაქსური დონე, L7-სინტაქსური დონე, L8 – პრაგმატული დონე. საჭიროების შემთხვევაში შეიძლება გაიზარდოს დონეების რაოდენობა, მაგ., სემანტიკური დონე, პროსოდიული დონე და ა.შ. დღეისათვის არსებული კორპუსები ერთმანეთისაგან, როგორც წესი, განსხვავდებიან როგორც ანოტაციის დონეების რაოდენობის, ისე დონეების იერარქიული სტრუქტურის თვალსაზრისით.

მაგ., იხ. ანოტაციის დონეების მაგალითი EXMARaLDA-დან:



ანოტაციური მდგრადობა consistency of annotation – კორპუსის ანოტაციის მეთოდი, რომლის დროსაც ანოტაციის პრინციპები არ იცვლება და მუდმივად მოქცეულია მდგრად თეორიულ და კონცეპტუალურ ჩარჩოში. მაგალითად, არ ხდება დადგენილი შემოკლებებისა და ტრანსკრიფციის წესების გადახედვა. ანოტაციის მდგრადობის მისაღწევად მიზანშეწონილია ე.წ. ანოტაციის კონვენციის – წესებისა და პრინციპების შემუშავება. შდრ. **ლაიფციგის გლოსირების წესები Leipzig Glossing Rules**.

ანოტირებული კორპუსი – იხ. ლინგვისტური ანოტაცია

ბალანსირება balancing – კორპუსის ერთ-ერთი თვისობრივი მახასიათებელი, რომელიც გულისხმობს კორპუსში ჩართული რესურსების ჟანრული მრავალფეროვნების გაწონასწორებას. ბალანსირება შესაძლებელია როგორც სინქრონული, ისე დიაქრო-

ნული თვალსაზრისით. დიაქრონული კვლევებისათვის მხოლოდ დიაქრონულად ბალანსირებული კორპუსის გამოყენებაა მიზანშეწონილი. სინქრონულად ბალანსირებული კორპუსი უნდა მოიცავდეს ენის რეალიზების მრავალფეროვან რეესტრს, უპირველეს ყოვლისა, დიალექტებსა და სოციოლექტებს. ენის ემპირიული ფორმის ყველა შესაძლო ვარიანტის ასახვა კორპუსში უზრუნველყოფს ბალანსირების ხარისხს. ემპირიულ ფორმებში იგულისხმება არა მარტო წერილობითი, არამედ ზეპირმეტყველების რეგისტრიც. ამ რეგისტრის შიგნით მრავალფეროვნება მიიღწევა, მაგ., ასაკობრივი ჯგუფების მრავალფეროვნებით. ბალანსირებულ კორპუსში ცალკე აისახება, მაგალითად, ბავშვის მეტყველება, ალკოჰოლის ზემოქმედების ქვეშ მყოფი ადამიანის მეტყველება, კომერციული დიალოგი, სპეციფიური სოციოლექტური დიალოგი და ასე შემდეგ.

The Brown Corpus sampling frame

Text categories		Broad genre	No. texts	% of corpus
A	Press: reportage	Press	44	8.8
B	Press: editorial	Press	27	5.4
C	Press: reviews	Press	17	3.4
D	Religion	General prose	17	3.4
E	Skills, trades and hobbies	General prose	36	7.2
F	Popular lore	General prose	48	9.6
G	Belles lettres, biography, essays	General prose	75	15
H	Miscellaneous (government & other official documents)	General prose	30	6
J	Learned and scientific writings	Learned	80	16
K	General fiction	Fiction	29	5.8
L	Mystery and detective fiction	Fiction	24	4.8
M	Science fiction	Fiction	6	1.2
N	Adventure and western fiction	Fiction	29	5.8
P	Romance and love story	Fiction	29	5.8
R	Humour	Fiction	9	1.8

ბრაუნის კორპუსის ბალანსირების ცხრილი

ბაიტი byte – ციფრული ინფორმაციის საზომი ერთეული; გამომთვლელი მანქანისათვის აღქმადი უმცირესი ციფრული ოდენობა, რომელიც უდრის 8 ბიტს (ამის გამო **octet-**საც უწოდებენ) და შეესაბამება ერთ კონკრეტულ ნიშანს, მაგალითად, ერთ ასოს ან სიმბოლოს.

Quantities of bytes						
Common prefix				Binary prefix		
Name	Symbol	Decimal SI	Binary JEDEC	Name	Symbol	Binary IEC
kilobyte	KB/kB	10^3	2^{10}	kibibyte	KiB	2^{10}
megabyte	MB	10^6	2^{20}	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	2^{30}	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	2^{40}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	2^{50}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	2^{60}	exbibyte	EiB	2^{60}
zettabyte	ZB	10^{21}	2^{70}	zebibyte	ZiB	2^{70}
yottabyte	YB	10^{24}	2^{80}	yobibyte	YiB	2^{80}

ბიტი bit/binary digit – ციფრული ინფორმაციის საზომი ერთეული; ბინარული ციფრი, რომელიც, როგორც წესი, „0“ და „1“-საგან შედგება. ცნების ავტორად ითვლება მათემატიკოსი ჯონ თიუქი (John W. Tukey). 1 ბიტი გულისხმობს ორ დასაშვებ მდგომარეობას (0 ან 1), 2 ბიტი 4-ს, 3 ბიტი 8 -ს ანუ **n** ბიტი უდრის 2^n-ს. 8 ბიტი უდრის 1 ბაიტს (8 ბიტი $=2^8 = 256$).

ბმული link – ელექტრონულ ტექსტში მოცემული რომელიმე ელემენტიდან შიდა ან გარე რესურსზე გადასვლის ტექნიკური შესაძლებლობა. ამის

შესაბამისად განასხვავებენ შიდა (ტექსტის შიგნით) და გარე (სხვა ტექსტზე გადამყვან) ბმულებს. იხ. **ჰიპერ-ბმული.**

გიგაბაიტი gigabyte – ციფრული ინფორმაციის საზომი ერთეული. 1 გიგაბაიტი უდრის 10^9 ბაიტს = 1.000.000.000 ბიტს.

გრამატიკალიზაცია grammaticalisation – დროში განფენილი ენობრივი პროცესის შედეგი, რომლის დროსაც ლექსიკური ელემენტი კარგავს სემანტიკურ მნიშვნელობას და იძენს გრამატიკულ ფუნქციას. ამ პროცესს, როგორც წესი, თან სდევს სიტყვის ფონეტიკური ეროზია (გაცვეთა). გრამატიკალიზაციის ნიმუშს ქართულში წარმოადგენს, მაგ., ციტირების (სხვათა სიტყვის) ნაწილაკი ***-მეთქი*** (< მე ვთქვი) წინადადებაში „ხომ გითხარით, ჯერ არსად ვმუშაობ-*მეთქი*“; მოდალური სიტყვა ***უნდა*** წინადადებაში „აუცილებლად *უნდა* წასულიყო, აქ ვეღარ დარჩებოდა“, შდრ. „ვერ გამიგია, მაგას რაღა *უნდა*“; არსებითი სახელი ***თავი*** რეფლექსური ნაცვალსახელის ფუნქციით წინადადებაში „*საკუთარ იდეალებს თავი უყოყმანოდ შესწირა*“; რიცხვითი სახელი ***ერთი*** მოდალური სიტყვის ფუნქციით წინადადებაში „*ერთი,* კარგად მოვილხინოთ, ბიჭებო!“

გრამატიკული თავი head – ფრაზის მორფოსინტაქსური და ფუნქციურ-სემანტიკური საყრდენი. ფრაზის იერარქიული სტრუქტურის ის წევრი, რომელიც, როგორც წესი, იერარქიის უმაღლეს საფეხურს იკავებს, ხოლო ფრაზის დანარჩენი წევრები მას ექვემდებარებიან და ახდენენ მის მოდიფიკაციას –

განავრცობენ რაიმე ნიშნის მიხედვით. სინტაქსური ანალიზის დროს ასეთი ელემენტი გამოდის გრამატიკული თავის როლში და განსაზღვრავს მთელი ფრაზის გრამატიკულ ნიშნებს. მაგალითად, ნომინალურ ფრაზაში (NP) „ჩემს პატარა კატას“ „კატა“ არის გრამატიკული თავი და განაპირობებს დაქვემდებარებული წევრების *„ჩემი“* (PRON) და *„პატარა“* (ADJ) გრამატიკულ გაფორმებას ბრუნვასა და რიცხვში.

დაგეგმილი ენა planned speech – ენის რეალიზების ის ფორმატი, რომელიც წერილობით დოკუმენტირებულ ტექსტებში ვლინდება და ასახავს ენის გარკვეულ რეესტრს – ნორმირებულ ენას. მისგან განსხვავებით, ზეპირმეტყველება ხასიათდება სპონტანურობით და ენობრივი ფორმების მოხმარების თავისუფლებით, რაც საშუალებას გვაძლევს, დავაკვირდეთ ენაში მიმდინარე ცვლილებებს ბუნებრივ გარემოში. ზეპირმეტყველების გზით რეალიზებულ ენას **დაუგეგმავი ენა unplanned speech** ეწოდება.

დეიქსისი deixis – (ბერძ. δεῖξις ჩვენება, მითითება) საკომუნიკაციო აქტის დეტერმინაციის ენობრივი საშუალება – ოპერაცია, რომელიც მიზნად ისახავს მოქმედების ადგილის, დროის, სივრცის და ა.შ. განსაზღვრას საკომუნიკაციო აქტთან მიმართებაში. დეიქსისის გამოსახატავად ენაში სხვადასხვა სახის დეიქტიკონები (↗დეიქტური ელემენტი) გამოიყენება: პირის, სივრცული, დროული, ტექსტური, სოციალური და სხვ. შესაბამისად, განასხვავებენ დეიქსისის რამდენიმე სახეს: პირის დეიქსისი (person deixis), დროის დეიქსისი (time deixis), სივრცული დეიქსისი (place deixis), დისკურსული დეიქსისი (discourse deixis), სოციალური

დეიქსისი (social deixis). დეიქსისს განიხილავენ როგორც კონტექსტური ან სიტუაციური სფეციფიკაციის აღწერის საშუალებას, რომელიც ემსახურება სივრცის, დროის და პირის დეტერმინაციას მთქმელის პოზიციის გათვალისწინებით. ხშირად მას **მე-აქ-ახლა** კორელაციასაც უწოდებენ. მაგ., წინადადებაში "**ახლა**, როცა **ამ** სტრიქონს ვწერ, შუადამე იწვის, დნება" (გ. ტაბიძე) ორი დეიქტიკონია გამოყენებული – **ახლა** (დროის დეიქტიკონი) და **აქ** (სივრცის დეიქტიკონი). დეიქსისი, კონკრეტული ლინგვისტური თეორიიდან გამომდინარე, განიხილება როგორც პრაგმატული ან სემანტიკური ფენომენი.

დეიქტიკონი deictic – იგივეა, რაც **დეიქტური ელემენტი**.

დეიქტური ელემენტი – დეიქტური ელემენტი ენის პრაგმატული დონის ფუნქციური ელემენტია, რომელიც ენაში გამოიყენება რეფერენციულობის კონკრეტული ფორმის, დეიქსისის გამოსახატავად. დეიქტური ელემენტი მიუთითებს სიტყვაზე ან სიტყვათა ჯგუფზე ერთი წინადადების ან დისკურსის ფარგლებში. ქართულში დეიქტური ელემენტების რამდენიმე ჯგუფი გამოიყოფა: პირის (პირის ნაცვალსახელები: მე, შენ, იგი), ობიექტზე ორიენტირებული (ჩვენებითი ნაჩვალსახელები: ეს, ეგ, ის), ადგილის (ზმნიზედები: აქ, მანდ იქ,), დროის (ზმნიზედები: ახლა, მაშინ). დეიქტური ელემენტები ქართულში სამსაფეხუროვან სისტემას გვიჩვენებენ: I საფეხურის დეიქსისი (მე, ეს, აქ), II საფეხურის დეიქსისი (შენ, ეგ, მანდ) და III საფეხურის დეიქსისი (იგი, ის, იქ).

დიგიტალური ჰუმანიტარია digital humanities – სამეცნიერო დარგი, რომელიც კვლევის დროს სისტემატურად იყენებს დიგიტალურ (ციფრულ) რესურსებს და კომპიუტერულ ტექნოლოგიებს ჰუმანიტარულ დარგებსა და კულტუროლოგიაში. დიგიტალური ჰუმანიტარია ინტერდისციპლინარულ დარგს წარმოადგენს და აღმოცენდა სამეცნიერო დისციპლინების – ჰუმანიტარული მეცნიერებისა და ინფორმატიკის მიჯნაზე. ამ დარგში მუშაობა მკვლევარისაგან მოითხოვს როგორც ჰუმანიტარულ მეცნიერებებში თეორიულ ცოდნას, ისე ინფორმატიკის საბაზისო და სტანდარტული კვლევითი მეთოდებისა და საშუალებების ფლობას. კორპუსლინგვისტიკის გარდა დიგიტალური ჰუმანიტარია მოიცავს კვლევის ისეთ მიმართულებებს, როგორებიცაა ე.წ. კრიტიკული დიგიტალური რედაქციების შექმნა, ტექსტის კვანტიტატიური ანალიზის საკითხები, მონაცემთა კომპლექსური სტრუქტურების ვიზუალიზაცია, თეორია დიგიტალური მედიების შესახებ და ა.შ. დიგიტალური ჰუმანიტარიის სინონიმად ინგლისურენოვან სამყაროში გამოიყენება ტერმინი e-Humanities. ორივე ინგლისური ტერმინი – digital humanities და e-Humanities – ახლად დანერგილი ტერმინებია და ჩაენაცვლა მანამდე არსებულ და ამჟამად მოძველებულ ტერმინებს computing in the humanities და humanities computing. დიგიტალური ჰუმანიტარიის სინონიმად ქართულში გამოიყენება ტერმინი **ციფრული ჰუმანიტარია**. *რესურსები:* იხ. პროექტები DIO (Deutsche Inschriften online, http://www.inschriften.net/) და Digital Humanities Hessen – Integrierte Aufbereitung und Auswertung textbasierter Corpora (http://www.dhhe.de/).

დისამბიგვირება disambiguation – სემანტიკური და ფორმალური ომონიმიის მოხსნა მოცემული კონტექსტის ფარგლებში.

დისკურსი discourse (ლათ. *discursus* გარემომცველი) – სამეცნიერო ტერმინი, რომელიც გამოიყენება ფილოსოფიაში, ენათმეცნიერებასა და ლიტერატურათმცოდნეობაში, რის გამოც მისი მნიშვნელობა დარგების მიხედვით განსხვავებულია. ფილოსოფოსი იურგენ ჰაბერმასი (Jürgen Habermas) დისკურსს კომუნიკაციური ქცევის თეორიის ფარგლებში განიხილავს და ერთმანეთისაგან განასხვავებს ნეიტრალურ სამეტყველო ქცევას (როგორც ურთიერთგაგებინებადობაზე ორიენტირებულ საკომუნიკაციო აქტს) და სტრატეგიულ სამეტყველო ქცევას (როგორც საკუთარი ინტერესებით განპირობებულ მეტყველების აქტს). თავის „ჭეშმარიტების თეორიაში“ დისკურსს იგი განმარტავს, როგორც კომუნიკაციის ისეთ ფორმას, რომელიც არგუმენტებით არის გამყარებული და რაციონალურობით გაჯერებული. პრიორიტეტულობისა და ძალაუფლების ნიშნით მარკირებული საკომუნიკაციო აქტი, მისი აზრით, ხელს უშლის კომუნიკაციის ძირითად ფუნქციას – ჩავწვდეთ ჭეშმარიტებას. დისკურსი არის მეტყველების ზეპირი ან წერილობითი ნაკადი, რომლის დროსაც ჭეშმარიტების დადგენის კონსტრუქციული წესები გამოიყენება და იდეალურ საკომუნიკაციო გარემოს ქმნის. ენათმეცნიერებაში დისკურსი განიხილება როგორც ენის ბუნებრივ ინტერაქციაში გამოყენებული ფორმა, რომელიც პრაგმატულად კოჰერენტულია და ხასიათდება სიგნიფიკანტურობით, რაც გულისხმობს საკომუნიკაციო ინტერაქციაში გამოყენებული ენობრივი

საშუალებების აუცილებლად მინიმალურსა და საკმარისად მაქსიმალურ ოდენობას.

დისტანციური სწავლება distance education – სასწავლო პროცესის ფორმატი, რომლის დროსაც სწავლების პროცესის ორივე მხარე – ლექტორი და სტუდენტი – სხვადასხვა სივრცულ გარემოში იმყოფება და ერთდროულად ან არაერთდროულად იყენებს დისტანციური სწავლების რესურსს. დისტანციური სწავლება არ გამორიცხავს ინტერაქციას ვიდეო-კონფერენციების სახით. იგი თავის თავში პრინციპულად მოიაზრებს სწავლების ელექტრონული ფორმატის გამოყენებას. სწავლების ამ ფორმატის უპირატესობა იმაში მდგომარეობს, რომ მას შეუძლია მრავალფეროვანი სასწავლო გარემოს შექმნა, იგი ხასიათდება მულტიმედიალური საშუალებების გამოყენების შეუზღუდაობით და ხელს უწყობს დამოუკიდებელი სწავლის უნარ-ჩვევების გამომუშავებას მომხმარებელში. დისტანციური სწავლების **distance education** სინონიმად ინგლისურში გამოიყენება **distance learning, dlearning, D-Learning.**

ელექტრონული გადამისამართება hyperlink – იხ. ჰიპერბმული.

ელექტრონული რესურსი – რესურსის ელექტრონული ფორმა, ელექტრონულ მატარებლებზე გადატანილი მონაცემი, რომელიც მისი თანამედროვე კომპიუტერული ტექნოლოგიებით დამუშავების საშუალებას იძლევა.

ელექტრონული სწავლება eLearning – სწავლების თანამედროვე მეთოდი, მიხაელ კერესის (Michael Kerres)

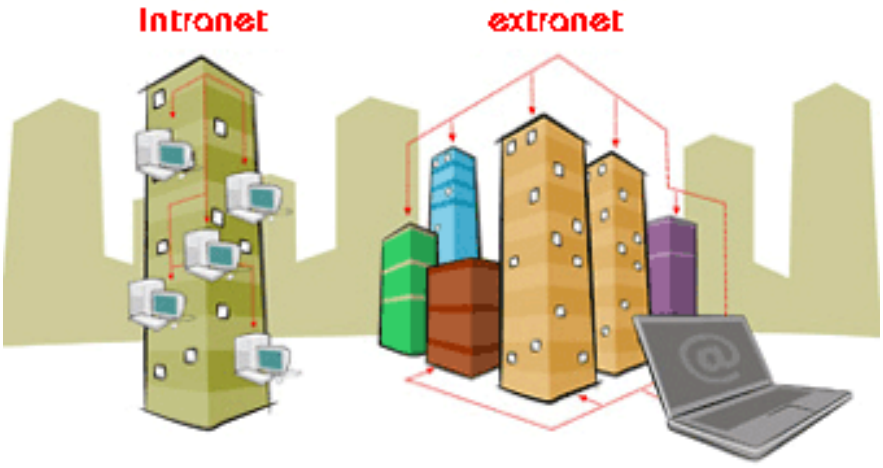
მიხედვით ელექტრონული სწავლება გულისხმობს სწავლების პროცესში ელექტრონული მედიებისა და ინფორმაციული და კომუნიკაციური ტექნოლოგიების გამოყენებას განათლებაში. ტერმინის **eLearning** სინონიმებად ინგლისურში ხშირად გამოიყენება ტერმინები multimedia learning, technology-enhanced learning (TEL), computer-based instruction (CBI), computer-based training (CBT), computer-assisted instruction ან computer-aided instruction (CAI), internet-based training (IBT), web-based training (WBT), online education, virtual education, virtual learning environments (VLE). ელექტრონული სწავლების ძირითადი მახასიათებლებია **ინტერაქციულობა** (გრისომის ექვსსაფეხურიანი იერარქიული სისტემა), **მულტიკოდირება** (განსხვავებულ კოდებში მოცემული რესურსების გამოყენების შესაძლებლობა), **მულტი-მედიალურობა** (სხვადასხვა მედიალური საშუალებების, ვიდეო, აუდიო, გრაფიკული, მულტიმედიალური საშუალებების გამოყენების შესაძლებლობა) და **მულტიმოდალურობა** (შემეცნების პროცესში აღქმის განსხვავებული საშუალებების გამოყენება, როგორიცაა აკუსტიკური, ვიზუალური და სხვ.). ელექტრონულ სწავლებაში უმნიშვნელოვანეს რგოლს ქმნის **სწავ-ლების მართვის სისტემა** (learning content management system). ელექტრონული სწავლების გამოყენება შესაძლებელია როგორც ტრადიციულ სასწავლო გარემოში (კლასი, აუდიტორია), ისე მის გარეშე. თანამედროვე ელექტრონულ სწავლებაში გამოიყენება როგორც ინდივიდუალური, ისე ჯგუფური სწავლების ფორმები: **ვირტუალური სწავლება virtual learning** (ინტერნეტის მეშვეობით მიმდინარე პროცესი, რომელიც არ საჭიროებს მონაწილეთა მხრიდან ერთდროულ ჩართულობას), **ინტეგრირებული სწავლება blended learning** (რომელიც აერთიანებს როგორც ვირტუალური

სწავლების მეთოდებს, ისე სიგნიფიკანტების სასწავლო სივრცეში ერთდროულად მუშაობის ელემენტებს), **მასალის გაზიარება content sharing** (გამოიყენება ძირითადად ფორუმებში, სადაც მომხმარებლები აქვეყნებენ და განავრცობენ სასწავლო მასალებს, როგორც თეორიულს, ისე პრაქტიკულს), **ჯგუფური სწავლება learning community** (ერთი და იმავე დარგის მომხმარებელთა მიერ იქმნება საკუთარი პლატფორმა, სადაც გროვდება დარგისათვის რელევანტური სასწავლო მასალა), **კომპიუტერულ ტექნოლოგიებზე დამყარებული კოოპერაციული სწავლება computer-supported cooperative learning** (რომლის დროსაც ძირითადად გამოიყენება ინფორმაციული და კომუნიკაციური სისტემები), **ვირტუალური საკლასო ოთახი virtual classroom** (სასწავლო პროცესში ინტერნეტი გამოიყენება როგორც საკომუნიკაციო მედიუმი ტერიტორიულად დაშორებულ მონაწილეებს შორის), ე.წ. **თეთრი დაფა whiteboard** (რომლის დროსაც სასწავლო პროცესის მონაწილეებს საშუალება აქვთ საერთო ქსელის მეშვეობით მოამზადონ სასწავლო თემის სამუშაო ვერსიები, გააკეთონ ჩანაწერები და ა.შ. ასეთ დროს გამოიყენება როგორც ტექსტური ფორმატის, ისე გრაფიკების, ნახაზებისა და სხვა სახის რესურსების შექმნის საშუალებები), **სამგანზომილებიანი სწავლება three-dimensional learning** (რომლის დროსაც გამოიყენება სამგანზომილებიანი ინფრასტრუქტურა სწავლების ეფექტურობის მისაღწევად). სასწავლო პროცესში ჩართულობის თვალსაზრისით ელექტრონული სწავლების ორი ცნება არის რელევანტური: სინქრონული და ასინქრონული სწავლება. სინქრონული სწავლების დროს ადგილი აქვს სწავლების ცოცხალ პროცესს

უშუალო ინტერაქციის სახით ორივე მხრიდან, რაც გამორიცხულია ასინქრონული სწავლების დროს.

ენის სინთეზი – ინფორმატიკაში ენის სინთეზი გულისხმობს ხელოვნურ ინტელექტს **AI** (**artificial intelligence**). ენის სინთეზის დროს ადგილი აქვს ენობრივი ფორმების (ბგერების, ფრაზების, წინადადებების) ხელოვნურ წარმოქმნას. ენის სინთეზის ტექნოლოგია ეყრდნობა ენის კომპიუტერულ მოდელირებას. მეტყველების რეპროდუქციისათვის აუცილებელია წინასწარ ჩაწერილი ბუნებრივი მეტყველების რეპრეზენტაციული ბაზა.

ეტერნეტი ethernet – კაბელური ტექნოლოგია, ლოკალური ქსელის ტიპი. განკუთვნილია ინფორმაციის გასაცვლელად ქსელში ჩართულ კომპიუტერებს შორის. კომპიუტერები ერთმანეთთან კავშირს ამყარებენ კოაქსიალური (ან ოპტიკური) კაბელების მეშვეობით გაგზავნილი რადიოსიხშირული სიგნალებით. ეტერნეტი განეკუთვნება **LAN**-ტექნოლოგიას (**local area network**), რომელიც პირველად შეიქმნა ფირმის Xerox Palo Alto Research Center მიერ.



ეფიციენტურობა efficiency – რესურსების ოპტიმალური და ხარისხობრივი ექსპლუატაცია.

ექსტრანეტი extranet – ინტრანეტის გაფართოებული სახე; ლოკალური ან გლობალური ქსელი, რომელიც იყენებს TCP/IP, HTML, SMTP და სხვა ღია ინტერნეტზე დაფუძნებულ სტანდარტებს ინფორმაციის გადასაცემად. ექსტრანეტი მისაწვდომია მხოლოდ წინასწარ განსაზღვრული ჯგუფებისათვის ორგანიზაციების შიგნით ან მის ფარგლებს გარეთ. იხ. **ინტრანეტი**.

ვერიფიკაცია, ვერიფიცირება verification – (ლათ. *verus* ჭეშმარიტი და *facere* კეთება) – ვითარების ან მდგომარეობის ჭეშმარიტების პოზიტიური დასაბუთება. მეცნიერული მეთოდი, რომელიც მიზნად ისახავს ცალსახა ექსპლიკატის მეშვეობით მეცნიერული თეზის ან ვარაუდის შემოწმებას. ინფორმატიკაში უმეტესად გამოიყენება ბინარული ცნება **ვერიფიკაცია** და **ვალიდურობა verification and validation (V&V)**, რაც არსებითად პროგრამული ხარისხის კონტროლს გულისხმობს. ფილოსოფიაში ვერიფიკაცია მეთოდოლოგიურ საშუალებას წარმოადგენს და ლოგიკური

პოზიტივიზმის ერთ-ერთი საკვანძო ცნებაა. ეს უკანასკნელი აღიარებს მხოლოდ იმ გამონათქვამებს, რომელიც გადამოწმებადია, ანუ ვერიფიკაციას ექვემდებარება. ენის ფილოსოფიაში წინადადების აზრობრივი მნიშვნელობა მხოლოდ მაშინ არის მისაღები, თუ იგი ვერიფიცირებადია. კორპუსლინგვისტიკაში ვერიფიკაცია არის სამეცნიერო დებულების ან წესის კორპუსში გადამოწმების მეთოდი, ინფორმაციული ტექნოლოგიების გამოყენებით წარმოებული რევიზია. გადამოწმების შედეგად დებულების არდადასტურების შემთხვევაში საქმე გვაქვს დებულების მეთოდური გზით უარყოფასთან – **ფალსიფიკაციასთან falsification** (↗ფალსიფიკაცია). განმავრცობელი ცნებებია: ლოგიკური ემპირიზმი, პოზიტივიზმი.

ვიზუალიზაცია visualisation – აბსტრაქტული მიმართებების ოპტიკური საშუალებებით ასახვა. კორპუსლინგვისტიკაში იგი ხშირად გამოიყენება დებულებების ან შედეგების თვალსაჩინოდ წარმოდგენის მიზნით.

თეზაურუსი thesaurus (ბერძ. θησαυρός განძი, განძთსაცავი) – ენის წერილობით ძეგლებში დადასტურებული სიტყვაფორმების ამსახველი ბანკი. თეზაურუსი ენის ან ენების ერთგვარ ელექტრონულ საცავს წარმოადგენს. ფართო მნიშვნელობით ცნება **თეზაურუსი** გამოიყენება ***ცოდნის საცავის*** აღსანიშნად, როგორიცაა, მაგ., ლექსიკონი ან ენციკლოპედია. თეზაუსური პირველად ჩნდება 1552 წელს, როდესაც ჰენრიკუს სტეფანუსმა (Henricus Stephanus) ბერძნული ენის თეზაურუსის "Thesaurus Graecae Linguae" ხუთტომეული გამოსცა. თეზაურუსი განსაკუთრებულ

მნიშვნელობას იძენს ინდექსაციის პროცესში, როდესაც აუცილებელია მითითება სიტყვაფორმის წყაროზე. საინფორმაციო-საძიებო კონტექსტში თეზაურუსი პირველად ჰანს პეტერ ლუნმა (Hans Peter Luhn) გამოიყენა 1957 წელს. ინდოევროპული ენებისათვის დღეისათვის არსებული რესურსებიდან ყველაზე დიდი და რეპრეზენტაციულია პროფ. იოსტ გიპერტის მიერ შექმნილი თეზაურუსი ↗**TITUS**-ი (იხ. ინდოევროპული ტექსტებისა და ენობრივი მასალების თეზაურუსი).

იდიომი idiom – მყარი შესიტყვება ან ხატოვანი გამოთქმა, რომლის მნიშვნელობა არ უდრის შემადგენელი სიტყვების მნიშვნელობათა ჯამს. იდიომებს მიეკუთვნება, მაგალითად, ანდაზები, კულტურულად დეტერმინირებული მეტაფორები.

ინდექსირება indexing – პროცესი, რომლის დროსაც პროგრამა უზრუნველყოფს კორპუსული ტექსტის კონკრეტული ინდექსით აღჭურვას.

ინტელექტუალური საკუთრება intellectual property – გონებრივი შემოქმედების ნაყოფი, რომელიც კანონის შესაბამისად დაცულია, როგორც მისი შემოქმედის საკუთრება.

ინტერნეტი Internet – ღია, გლობალური ქსელი, რომელიც სხვადასხვა ტიპის ქსელებს შორის მონაცემთა გაცვლის საშუალებას იძლევა. ხშირად გამოიყენება როგორც WWW-ს (World Wide Web) სინონომი, რაც არასწორია. WWW წარმოადგენს საინტერნეტო ქსელის მეშვეობით ფუნქციონირებად სისტემას, რომელიც ვებგვერდებისაგან – ელექტრო-

ნული ჰიპერტექსტური დოკუმენტებისაგან შედგება. ინტერნეტი შეიქმნა სამეცნიერო პროექტის Arpanet-ის (Advanced Research Project Agency) ბაზაზე და მიზნად ისახავდა ძვირადღირებული კომპიუტერული აპარატურის ეფიციენტურობის (↗ეფიციენტურობა) გაზრდას მონაცემთა გაცვლის თვალსაზრისით. ქსელის გამოყენების ზრდას განსაკუთრებით შეუწყო ხელი მომხმარებლებისთვის ერთ-ერთი მნიშვნელოვანი აპლიკაციის – ელექტრონული ფოსტის (e-Mail) შექმნამ. 1990 წლიდან ამერიკის ეროვნული სამეცნიერო ფონდის გადაწყვეტილებით ქსელი ღია რესურსად გამოცხადდა.

ინტრანეტი intranet – მონაცემთა გაცვლის ჩაკეტილი ქსელი. **LAN**-ისაგან (local area network) და **GAN**-ისაგან (global area network) განსხვავებით ინტრანეტი ჩაკეტილია არა გეოგრაფიული თვალსაზრისით, არამედ დასაშვებ მომხმარებელთა თვალსაზრისით.

ინტროსპექცია introspection – ენობრივი მონაცემის ანალიზის მეთოდი, რომელიც ეყრდნობა ინტუიციას. ფართოდ გამოიყენება სემანტიკური ანალიზის დროს, ასევე ფსიქოლოგიაში და ფსიქოლინგვისტიკაში.

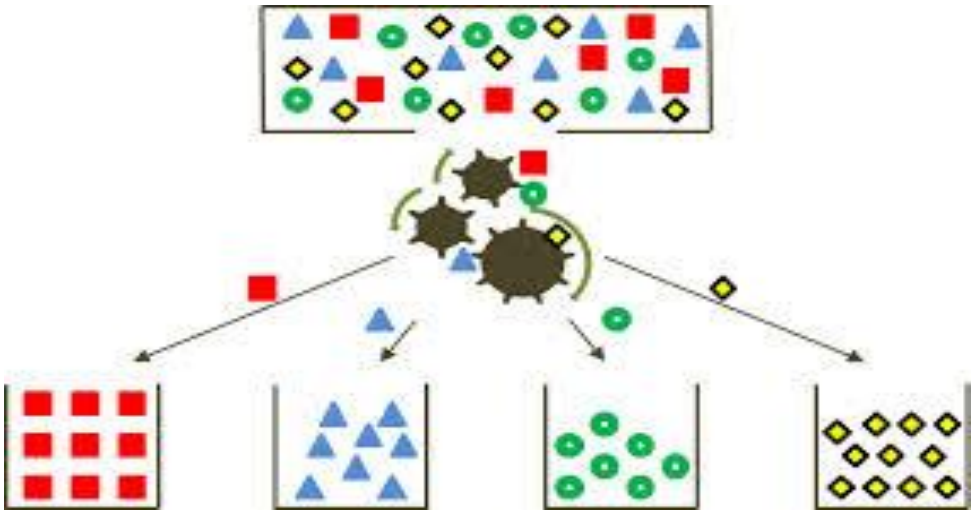
ისტორიული კორპუსი historical corpus – კორპუსის ტიპი, რომელიც მოიცავს მოცემული კონკრეტული ენის ან ენათა ისტორიული არსებობის პერიოდში შექმნილ დოკუმენტთა ფართო ჟანრობრივ რეპერტუარს (↗ჟანრი) და დიაქრონული კვლევის საშუალებას იძლევა. ისტორიული კორპუსი შეიძლება რომელიმე ქრონოლოგიურ პერიოდზე იყოს შეზღუდული. მაგ., ინგლისური ენის ისტორიული კორპუსი ARCHER (A Representative Corpus of Historical English Registers) მხოლოდ

1650-1999 პერიოდის მასალებს მოიცავს და 1,8 მილიონი ტოკენისაგან (↗ტოკენი) შედგება. ისტორიული კორპუსი არ გამორიცხავს ჟანრულ მრავალფეროვნებას. იგივე ARCHER მულტიჟანრობრივ კორპუსს წარმოადგენს. ისტორიული კორპუსები სხვადასხვა მოცულობის შეიძლება იყოს. მაგ., ინგლისური ენის მეორე ისტორიული კორპუსი Penn Parsed Corpora of Historical English უფრო ვრცელ პერიოდს მოიცავს (1150-1914) ARCHER-თან შედარებით და 3.9 მილიონი ტოკენისაგან შედგება. ქართული ენის ეროვნული კორპუსი დღევანდელი მონაცემების მიხედვით ისტორიულ კორპუსს წარმოადგენს. იხ. **ქართული ენის ეროვნული კორპუსი**.

კილობაიტი kilobyte – ინფორმაციის საზომი ერთეული. 10^3 ბიტი = 1.000 ბაიტი = 1 კილობაიტი.

კლასტერული ანალიზი cluster analysis / clustering – სტატისტიკური ანალიზის მეთოდი, რომელიც დიდ მონაცემთა ბაზებში ახდენს მსგავს ობიექტთა ჯგუფების – ე. წ. კლასტერების (↗კლასტერები) დახარისხებას. ამგვარად შერჩეული ერთგვაროვანი კლასტერები ერთიანდებიან იერარქიულ (დაქვემდებარებულ) ან აგლომერაციულ (თანაბარდონიან) სისტემაში. კლასტერული ანალიზის მიზანია ახალი ჯგუფების გამოვლენა და არა ჯგუფების კლასიფიკაცია.

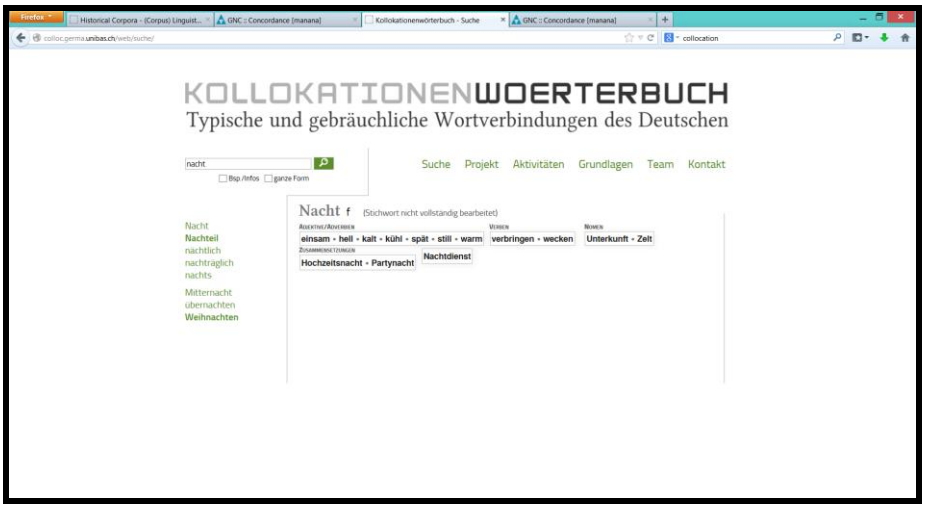
კლასტერები clusters – ***n*-gram-**ის ალტერნატიული ტერმინი (↗ *n*-gram).



კლასტერული ანალიზის ვიზუალიზაცია

კოლოკაცია collocation – უშუალო კონტექსტუალური თანაარსებობა, გვერდიგვერდ პოვნიერება. კოლოკაცია, როგორც ცნება, ჯონ რუპერტ ფერსმა (John Rupert Firth) გამოიყენა პირველად ენათმეცნიერებაში. კორპუსლინგვისტიკაში კოლოკაცია აღნიშნავს ყველაზე ხშირად გამოყენებულ, სტატისტიკურად დაანგარიშებად ლექსიკურ ელემენტებს კორპუსში. კოლოკაციის მაგალითი GEKKO-დან **KWIC-**ისათვის (↗ KWIC) „ღამე“ (კრიტერიუმებით „ინფორმაციულობა“ + “სიხშირე“) არის:

1. *მთელი ღამე*
2. *ვარსკვლავიანი ღამე*
3. *ბნელი ღამე*
4. *დღე და ღამე*
5. *მთვარიანი ღამე*
6. *უკუნეთი ღამე*
7. *წყვდიადი ღამე*



გერმანული ენის კოლოკაციების ინტერნეტლექსიკონი
სიტყვაფორმის „ღამე“ კოლოკაციები

კოლოკაცია განპირობებულია ისეთი ფაქტორებით, როგორიცაა სემანტიკური თავსებადობა, ლოგიკური განპირობებულობა, ფრაზეოლოგიზმები და სტერეოტიპები. შდრ., მაგ., *მიზნის მიღწევა*, მაგრამ *შედეგის მიღება*. კოლოკაციის ფენომენს ვალტერ პორციგი (Walter Porzig) ისეთი ცნებების საშუალებით აღწერდა, როგორებიცაა *არსებითი შინაარსობრივი კავშირი* და *მნიშვნელობის (საერთო) სინტაქსური არე*, ევგენი კოსარიუსთან (Eugenio Coseriu) კი მას *ლექსიკური სოლიდარობა* ეწოდება.

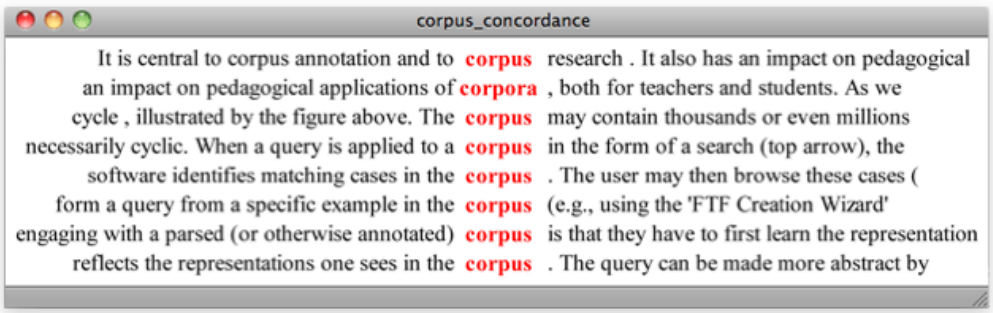
კომპაილერი compiler – დამჯამებელი. იხ. **კომპილირება**.

კომპილირება compilation – დაჯამება; პროგრამული კოდის „თარგმნა“ მანქანის (პროგრამულ) ენაზე. პროგრამა ფუნქციონირებას იწყებს მხოლოდ მისი კომპილირების შემდეგ. ამგვარი „თარგმნის“ დროს ხდება პროგრამული კოდის შემოწმება ფორმალური ცდომილებებისა (სინტაქსი) და თავსებადობის

(სემანტიკური ექვივალენტურობის) თვალსაზრისით. პირველი კომპაილერი დააპროგრამა ამერიკელმა მათემატიკოსმა გრეის ჰოპერმა (Grace Hopper) 1952 წელს.

კონკორდანსერი concordancer – კომპიუტერული პროგრამა (ალგორითმი), რომელიც კონკორდანსის (↗ კონკორდანსი) გენერირებას ახდენს.

კონკორდანსი concordance (ლათ. *concordare* გაერთიანება, ჰარმონიზაცია) – თავდაპირველად გამოიყენებოდა ბიბლიათმცოდნეობაში, შემდგომ ლიტერატურათმცოდნეობაში. დღეისათვის წარმოადგენს კორპუსლინგვისტიკის ერთ-ერთ ძირეულ ცნებას. კონკორდანსი არის ანბანურ რიგზე დალაგებული სიტყვების ან ფრაზების სია. მის სინონიმებად გამოიყენება აგრეთვე **რეგისტრი** ან **ინდექსი**. კორპუსლინგვისტიკაში გამოყენებული ცნება *კონკორდანსი* იმით განსხვავდება სიტყვების უბრალო ჩამონათვალისაგან, რომ იგი ცალკე ამოღებული სიტყვების ჩამონათვალი კი არ არის, არამედ გვიჩვენებს მის ადგილს იმ კონტექსტში, რომელშიც იგი გვხვდება. კონკორდანსის საყრდენს წარმოადგენს სიტყვაფორმა, რომელსაც კორპუსლინგვისტიკაში ↗ **KWIC** (key word in context) ეწოდება.



კონკორდანსის მაგალითი British National Corpus-იდან

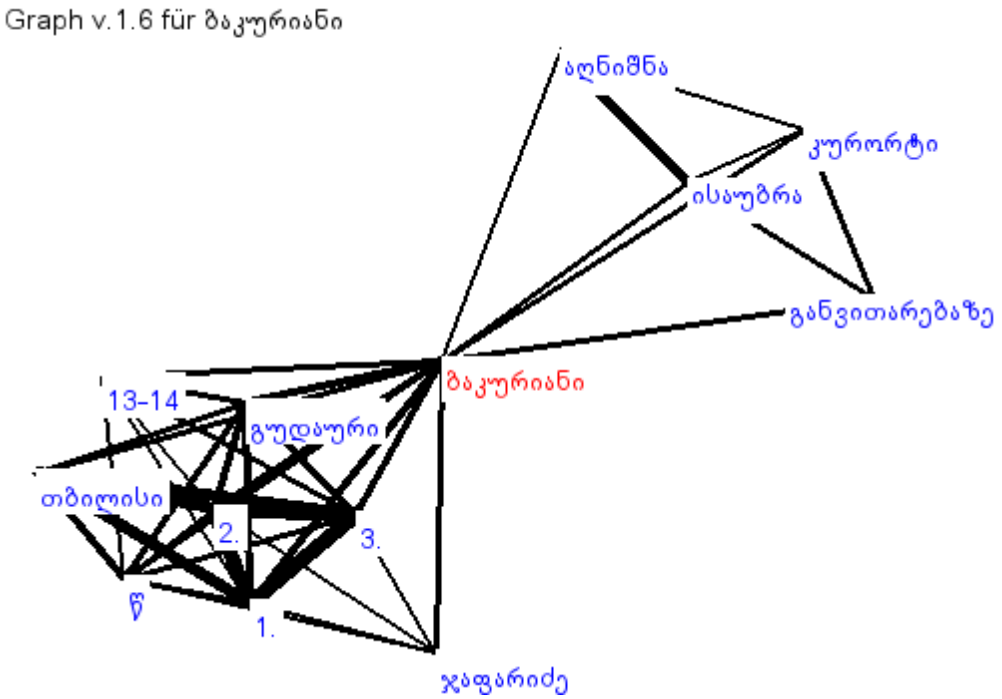
კოდირება (character) encoding – ტექსტის ასახვა კომპიუტერული მახსოვრობისათვის გასაგებ ნიშნებში – ბიტებში. კოდირების პირველი, ე. წ. 8-ბიტიანი კოდირების საერთაშორისო სტანდარტი (ISO-8859) სულ მალე 16-ბიტიანმა კოდირების სტანდარტმა შეცვალა, რომელიც სტანდარტების საერთაშორისო კონსორციუმმა უნიკოდმა (↗ უნიკოდის კონსორციუმი) 1991 წელს გამოაქვეყნა. კოდირების ახალმა სტანდარტმა შესაძლებელი გახადა ეროვნული ანბანების საყოველთაო სტანდარტიზაცია, მათ შორის, ჩინური იეროგლიფებისაც. დღეისათვის ქართული ანბანის სამივე სახე არის უნიკოდში ასახული: ასომთავრული, ნუსხა-ხუცური და მხედრული. განსაკუთრებული წვლილი ქართული ანბანის ნაირსახეობების სტანდარტიზაციაში მიუძღვით იოსტ გიპერტსა და ირაკლი ღარიბაშვილს.

კონოტაცია connotation – ლექსემის დენოტაციის თანამდევი პერიფერიული (კონტექსტუალური) სემანტიკა. მაგალითად: ლექსიკური ერთეულის *როჯა* სემანტიკური მნიშვნელობა იგივეა, რაც *სახე,* მაგრამ მისგან განსხვავებით იგი უარყოფითი კონოტაციით არის დატვირთული – პეიორატივია.

კოოკურენტი co-occurring term – ლექსემის ლოგიკურ-კოჰერენტული რელაცია. იხ. კოოკურენცია.

კოოკურენცია co-occurrence – კორპუსლინგვისტიკაში ცნება კოოკურენცია აღნიშნავს ორი ტოკენის (↗ტოკენი) თანაობას კონტექსტში. კოოკურენცია ეყრდნობა Zipf-ის კანონს. თუ f ∗ r ≈ k (f ~ ერთი ტოკენის სიხშირე კორპუსში, r ~ რანგი ნუმერირებულ სიაში),

მაშინ f ≈ k / r, ანუ კორპუსში სიტყვის სიხშირე უკუპროპორციულია მის რანგთან სიხშირის სიაში.
სიტყვის „ბაკურიანი“ კოოკურენტები და მათი ვიზუალიზირების (↗ვიზუალიზაცია) შედეგი: გუდაური (146.033), *3.* (73.0919), *2.* (71.7552), *1.* (71.4885), *წ.* < წელი (64.4273), *13-14* (42.625), *კურორტი* (41.3009), *განვითარებაზე* (33.7972), *ჯაფარიძე* (33.1845), **,** (25.4459), **(** (24.1829), *ისაუბრა* (22.0791), *თბილისი* (20.0721).



სიტყვა „ბაკურიანის“ კოოკურენტები და მათი ვიზუალიზაცია

კორპორა corpora – ცნების ↗ *კორპუსი* მრავლობითი რიცხვის ფორმა.

კორპუსზე ორიენტირებული კვლევა – სამეცნიერო კვლევის თანამედროვე მეთოდი. კორპუსზე ორიენტირებული კვლევა გულისხმობს ემპირიული ენათმეცნიერების მეთოდების გამოყენებით რელევანტური ფორმე-

ბისა და გრამატიკული მოვლენების კორპუსის ბაზაზე ვერიფიკაციას.

კორპუსზე დამყარებული ლინგვისტური კვლევა vs. კორპუსით განპირობებული ლინგვისტური კვლევა corpus-based studies vs. corpus-driven studies

– კორპუსზე დამყარებული ლინგვისტური კვლევა გულისხმობს მეთოდს, რომლის საშუალებითაც შესაძლებელი ხდება ლინგვისტური ჰიპოთეზის ვალიდურობის შემოწმება კორპუსის ბაზაზე. ხაზგასასმელია ის ფაქტი, რომ ჰიპოთეზის ჩამოყალიბება ამ შემთხვევაში არ უკავშირდება კორპუსულ კვლევას. კორპუსის ბაზაზე ხდება მისი სისწორისა და დეტალურობის ხარისხის ანუ ვალიდურობის დადგენა. ამის საპირისპიროდ **კორპუსით განპირობებული ლინგვისტური კვლევა** გულისხმობს მეთოდს, რომლის დროსაც ლინგვისტური ჰიპოთეზის ჩამოყალიბება იმთავითვე უკავშირდება კორპუსს, ანუ მეცნიერული პრობლემის დასმის წყაროს წარმოადგენს კორპუსში მონაცემების ანალიზი. ზოგიერთი ავტორი (მაგ., Tognini-Bonelli 2001: 84–5) ამ მეთოდზე დაყრდნობით თვლის, რომ ენობრივი კორპუსი თავად „ქმნის“ ენის თეორიას (განმვრცობი ცნებები: პოლისისტემატიზმი, John R. Firth).

კორპუსი corpus (ლათ. *corpus*, მრ. *corpora* სხეული) – შემოკლებული აღნიშვნა ცნებისათვის **ტექსტკორპუსი** ან **ტექსტური კორპუსი**. ლინგვისტიკაში იგი აღნიშნავს რომელიმე ენის წერილობითი ძეგლების ან წერილობით დოკუმენტირებული ზეპირმეტყველების მასალების ნაკრებს. კორპუსების შექმნამ განსაკუთრებული მნიშვნელობა შეიძინა ისეთი ჰუმანიტარული დარგებისათვის, როგორიცაა ენათმეცნიერება, ლიტერატუ-

რათმცოდნეობა, ისტორიოგრაფია, თუმცა აქტიურად გამოიყენება სამართალმცოდნეობაშიც. კორპუსი იძლევა ცალკეული საკითხებისა და პრობლემების სისტემური კვლევის საშუალებას როგორც სინქრონულ, ისე დიაქრონულ ასპექტში. დიაქრონული კვლევის განსახორციელებლად შექმნილ მიზნობრივ კორპუსს ↗**ისტორიული კორპუსი** ეწოდება. კორპუსი გარკვეული სამეცნიერო კრიტერიუმების შედეგად შექმნილი ტექსტების სტრუქტურირებული კრებულია, რომელიც აერთიანებს გარკვეული **ჟანრისა** (↗ჟანრი) და **ტიპის** ტექსტებს და აღჭურვილია მართვის სპეციალური სისტემით – **კორპუსის მენეჯერით**. კორპუსის აგების მეთოდოლოგიის ჩამოყალიბებამ და მონაცემთა მანქანური დამუშავების შესაძლებლობების სწრაფმა განვითარებამ ხელი შეუწყო ახალი დარგთაშორისი სამეცნიერო დისციპლინის **კორპუსლინგვისტიკის** ჩამოყალიბებას. ენობრივი ანუ პირველადი მონაცემების გარდა კორპუსის აუცილებელ ატრიბუტს წარმოადგენს **მეტამონაცემები** (↗მეტამონაცემი) – მონაცემები ენობრივი მონაცემების შესახებ.

კორპუს-ძიება – მიზანმიმართული ძიება კორპუსში, რომლის ეფექტურობა მიემართება ე.წ. კორპუსის ინჟინერიის პრერეკვიზიტებს ანუ წინდაწინ დადგენილ პირობებს. კორპუსის პრერეკვიზიტებიდან გამომდინარე შესაძლებელია შემდეგი სახის კორპუს-ძიება:

a. ცალკეული სიტყვაფორმის (მაგ.: "კაცისაგან")
b. ცალკეული ფუძის (მაგ.: "კაც")
c. მორფოლოგიური ელემენტის (მაგ.: "ი" ~ Nom)
d. POS (Tag) დაფა (მაგ.: "N" ~ noun)
e. სინტაქსური სტრუქტურის (კვანძის) (მაგ.: "NP" ~ noun phrase)

f. სემანტიკური კონცეფციის (მაგ.: "მარილზე წასვლა" ⇔ "სიკვდილი")

კორპუსის პრესტრუქტურირებიდან გამომდინარე კორპუს-ძიება შესაძლებელია მხოლოდ ანოტაციაში მითითებული ფენომენების მეშვეობით. ამიტომ შეუძლებელია ისეთი გამოტოვებული ელემენტების საძიებო ცნებად ფორმულირება, როგორიცაა ელიფსი, Ø-ით მარკირებული ელემენტები და ა.შ. მაგ.: *ელენე ყავას სვამს, გიორგი ჩაის* ან *ვიღაც დავინახე, მაგრამ არ ვიცი ვინ.*

კორპუს-საძიებო სისტემა – კორპუს-ძიებაზე მომართული ექსპლორაციული ალგორითმები (ხელსაწყოები ~ tool). მაგ.: TIGERSearch – (წყარო: König, Esther; Lezius, Wolfgang 2003: The TIGER language – A Description Language for Syntax Graphs, Formal Definition. Technical report IMS, Universität Stuttgart, Germany).

კორპუსის სახეები – კორპუსის სტრუქტურისა და კორპუსის გამოყენების მიზნებიდან გამომდინარე განასხვავებენ კორპუსის სხვადასხვა სახეს: ელექტრონული კორპუსი vs. ქაღალდის კორპუსი, ფრაგმენტული კორპუსი vs. რეფერენციული კორპუსი, სტატიკური კორპუსი vs. მონიტორული (↗მონიტორული კორპუსი), ანოტირებული კორპუსი, ერთენოვანი (↗მონოლინგვური კორპუსი) vs. მრავალენოვანი (↗მულტილინგვური კორპუსი) vs. პარალელური კორპუსი (↗პარალელური კორპუსი), სპეციალური ანუ თემატური კორპუსი.

კორპუსის მახასიათებლები – კორპუსის ფუნქციონირებისათვის აუცილებელი რელევანტური ნიშნები. როგორც წესი, კორპუსის დახასიათების დროს

მიუთითებენ რამდენიმე მნიშვნელოვან მახასიათებელზე, როგორიცაა:

- სერვერული განვრცობა (ღია რეჟიმის ქსელური კვანძი);
- ზომა&მოცულობა (ტოკენების რაოდენობა);
- ბალანსირება (ჟანრული მრავალფეროვნება);
- საძიებო სისტემა (ნიმუშის ამოცნობის პროგრამული ინსტრუმენტი ე. წ. pattern matching);
- მეტაინფორმაციული ბანკი;
- სტატისტიკური ანალიზის ალგორითმი;
- მომხმარებლის დინამიური ბიბლიოთეკა.

კორპუსის მოცულობა თვისობრივად და რაოდენობრივად – კორპუსული კვლევის ერთ-ერთ მეთოდურ პრობლემას წარმოადგენს მომხმარებლის მიერ დასმული ლინგვისტური საკითხის შესაბამისი ოპტიმალური მასალის კორპუსში მოძიება. თუ ზეპირმეტყველებას ვიკვლევთ, საძიებო სისტემა შესაბამისად ენის ამ რეგისტრით უნდა შემოიფარგლოს, თუ ენის ისტორიული ეტაპის რომელიმე თვალსაჩინო გრამატიკულ ფენომენს ვიკვლევთ, მაშინ კორპუსი საშუალებას უნდა იძლეოდეს, ძებნის შედეგად ზედაპირზე ამოიტანოს ამ ფენომენთან ასოცირებული ყველა კონტექსტი. სპორადულ, არასისტემურ მაგალითებზე დაყრდნობით შეუძლებელია სოლიდური მეცნიერული დასკვნის გამოტანა. კორპუსის აგება, მონაცემთა შეგროვება და მათი მეტაინფორმაციით აღჭურვა თანამედროვე კორპუსლინგვისტიკის ერთ-ერთ საკვანძო საკითხს წარმოადგენს. არსებობს ამ საკითხის გადაჭრის ორგვარი მიდგომა: **მონიტორული კორპუსი** (Sinclair 1991: 24–6), რომელიც მუდმივად განივრცობა და

ბალანსირებული (↗ბალანსირება) ანუ რეპრეზენტაციული **კორპუსი** (Biber 1993, Leech 2007), სადაც ენის სისტემა აღიბეჭდება სახასიათო ექსპლიკატების მეშვეობით.

კორპუსის ტიპები – კორპუსის ტიპი განისაზღვრება იმის მიხედვით, თუ ელექტრონული რესურსების რა სახის მასალებს აერთიანებს იგი.

კორპუსის შექმნა corpus construction – კორპუსის შექმნის პროცესი: ტექსტების შეგროვება, მონაცემების კოდირება და შენახვა, კორპუს-მენეჯერით აღჭურვა, ანოტირება, ადმინისტრირება და ა. შ.

კორპუსლინგვისტიკა corpus linguistics – დარგთაშორისი სამეცნიერო დისციპლინა, რომელიც იკვლევს კორპუსების აგების, მართვისა და გამოყენების მეთოდებს. კორპუსლინგვისტიკა ანგლო-ამერიკულ გარემოში ჩამოყალიბდა გასული საუკუნის 60-იან წლებში და უკავშირდება პირველი ტექსტური კორპუსის – **Brown Corpus-**ის – შექმნას. მიუხედავად იმისა, რომ კორპუსების შექმნა გაცილებით ადრე დაიწყო, კორპუსლინგვისტიკამ, როგორც მეცნიერების დარგმა, სამეცნიერო წრეებში მხოლოდ გასული საუკუნის 90-იანი წლებიდან მოიპოვა აღიარება. კორპუსლინგვისტიკის მიზანია ენობრივი ფაქტებისა და მოვლენების სისტემური კვლევა. თეორიული ენათმეცნიერების პარადიგმაში კორპუსლინგვისტიკა უპირისპირდება გენერატიულ ლინგვისტიკას – ამ უკანასკნელისაგან განსხვავებით იგი ძირითადად ინდუქციურ-ემპირიულ მეთოდებს იყენებს (კონკრეტულიდან ზოგადისაკენ), მაშინ როდესაც გენერატიული

ლინგვისტიკა ძირითადად დედუქტიურ დებულებებს ეყრდნობა. კორპუსლინგვისტიკა ენას განიხილავს არა მხოლოდ როგორც აბსტრაქტულად არსებულ სისტემას („langue“ ფერდინანდ დე სოსიურის მიხედვით, ან „competence“ ნოამ ჩომსკის მიხედვით), არამედ როგორც ენის გამოვლენის ნებისმიერ ფორმათა ერთობლიობასაც („parole“ ფერდინანდ დე სოსიურის მიხედვით, ან „performance“ ნოამ ჩომსკის მიხედვით). კორპუსლინგვისტიკის მიზანია ენის რეალიზების ნებისმიერი ემპირიული ფაქტის კვლევა და ენათმეცნიერებაში გავრცელებული დიქოტომიის „Langue“ - „Parole“-ის მოხსნა. ამდენად, მისი კვლევის ობიექტს წარმოადგენს არა მხოლოდ **დაგეგმილი** (↗დაგეგმილი ენა), არამედ **დაუგეგმავი ენის** (↗დაგეგმილი ენა) ფაქტებიც. კორპუსლინგვისტიკის სამეცნიერო დისციპლინად აღიარებას ხშირად ხელს უშლის გავრცელებული აზრი იმის შესახებ, რომ კორპუსლინგვისტიკას კვლევის საკუთარი მეთოდები არ გააჩნია. ამიტომ კორპუსლინგვისტიკის ერთ-ერთ მნიშვნელოვან ამოცანას კვლევის საკუთარი მეთოდური სისტემის ჩამოყალიბება წარმოადგენს.

ლაიფციგის გლოსირების წესები Leipzig Glossing Rules – იხ. დანართი 1.

ლემა lemma – სიტყვაფორმის საწყისი, ძირითადი ფორმა. სახელებისათვის ქართულში ლემა არის სახელობითი ბრუნვის ფორმა, ზმნებისთვის – საწყისი.

ლემატიზაცია lemmatisation – კორპუსის ანოტაციის ფორმა, რომლის დროსაც ტოკენის ფორმა დაიყვანება ლემაზე (↗ლემა).

მაგ., წინადადებაში „ქათმებმა საკენკი აკენკეს“ სამი ლემაა:

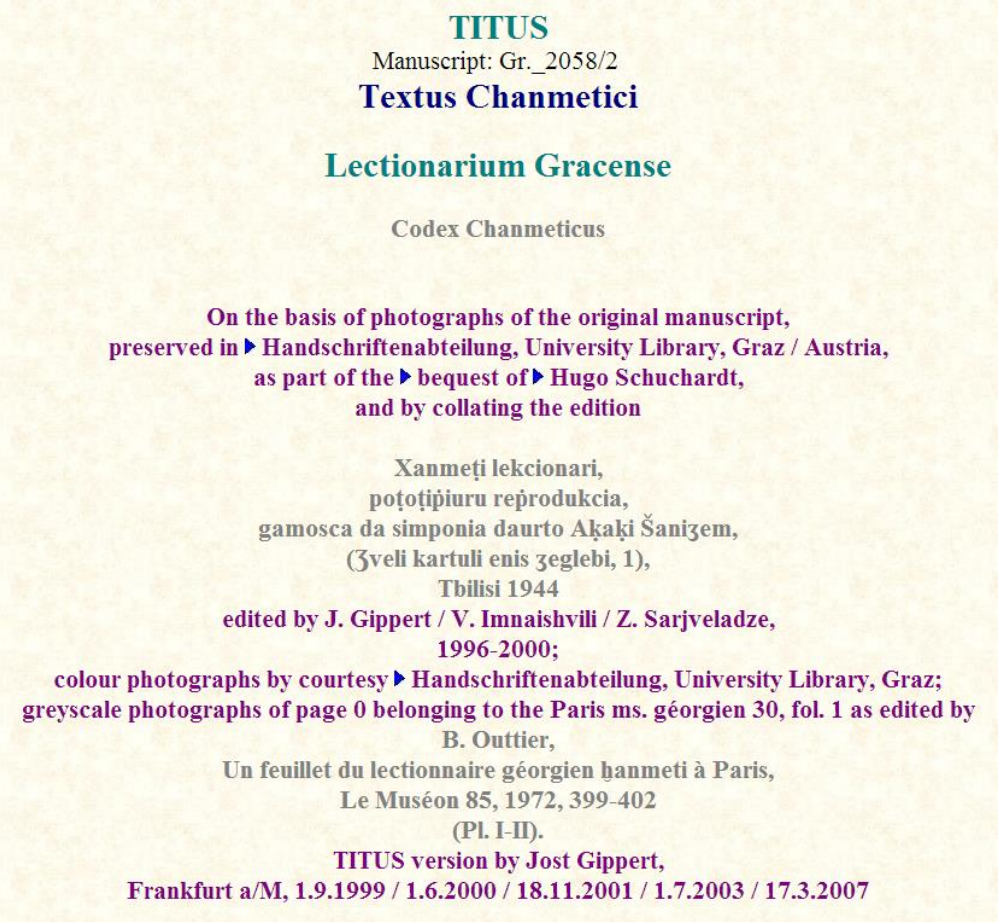
1. ქათმებმა (ERG.Pl.) > ქათამი
2. საკენკი (NOM.Sg.) > საკენკი
3. აკენკეს (AOR.S3.Pl.) > აკენკვა.

ლექსიკური ერთეული lexical item – ლემის ზოგადი სახელი კორპუსლინგვისტიკაში, ლექსიკური მნიშვნელობის მქონე ერთეული ენაში. უტოლდება სალექსიკონო ერთეულს ლექსიკოლოგიაში. იხ. **ლემა**.

მანუალური ანოტაცია – ხელით შესრულებული, არაავტომატიზებული ანოტაცია (იხ. ანოტაცია). ძირითადად გამოიყენება სამეცნიერო სტატიებში მოყვანილი მაგალითების ანოტირებისას (↗ანოტაცია).

მეგაბაიტი megabyte/MB – ინფორმაციის საზომი ერთეული. 1.000.000 ბაიტი ანუ 10^6 (ან 1000^2) გვაძლევს 1 მეგაბაიტს. იხ. **ბიტი**.

მეტამონაცემი metadata – მონაცემი ენობრივი მონაცემის შესახებ. მეტამონაცემი წარმოადგენს მეორად, განმავრცობელ ინფორმაციას პირველადი ენობრივი მონაცემის – ენობრივი ფაქტების შესახებ. ამგვარი ინფორმაციის მაგალითებია, მაგ., მონაცემი პირველადი მონაცემის (მაგ., ხელნაწერის) გადამწერის შესახებ, ავტორის ვინაობის შესახებ, შექმნის ადგილის შესახებ, პირველად მონაცემთან დაკავშირებული რელევანტური თარიღების შესახებ და ასე შემდეგ. მეტამონაცემი შესაძლოა თან ერთვოდეს პირველად მონაცემს ან სპეციალური ფორმატის დამოუკიდებელ ბაზაში (header) განთავსდეს.


TITUS
Manuscript: Gr._2058/2
Textus Chanmetici

Lectionarium Gracense

Codex Chanmeticus

On the basis of photographs of the original manuscript,
preserved in ▸ Handschriftenabteilung, University Library, Graz / Austria,
as part of the ▸ bequest of ▸ Hugo Schuchardt,
and by collating the edition

Xanmeṭi lekcionari,
poṭoṭipiuru reṗrodukcia,
gamosca da simponia daurto Aķaķi Šaniʒem,
(ʒveli kartuli enis ʒeglebi, 1),
Tbilisi 1944
edited by J. Gippert / V. Imnaishvili / Z. Sarjveladze,
1996-2000;
colour photographs by courtesy ▸ Handschriftenabteilung, University Library, Graz;
greyscale photographs of page 0 belonging to the Paris ms. géorgien 30, fol. 1 as edited by B. Outtier,
Un feuillet du lectionnaire géorgien ḫanmeti à Paris,
Le Muséon 85, 1972, 399-402
(Pl. I-II).
TITUS version by Jost Gippert,
Frankfurt a/M, 1.9.1999 / 1.6.2000 / 18.11.2001 / 1.7.2003 / 17.3.2007

მეტამონაცემის მაგალითი TITUS-იდან

დღესდღეობით მეტამონაცემების ასახვა ხდება სპეციალურ ფორმატში. ყველაზე გავრცელებული ფორმატია TEI (Text Encoding Initiative). მეტამონაცემების ეს ფორმატი 1995 წლიდან ფართოდ გამოიყენება ბიბლიოთეკებში, არქივებსა და მუზეუმებში. ქართული ენის ეროვნული კორპუსი იყენებს TEI ფორმატს.

მეხსიერების ჩხირი USB (universal serial bus) stick – ელექტრონული რესურსების გადატანის საშუალება, რომელსაც შედარებით დაბალი მეხსიერების მოცულობა გააჩნია.

მეტყველების ნაწილების ანოტაცია POS (part-of-speech) tagging – ანოტაციის ერთ-ერთი ყველაზე გავრცელებული ფორმა, რომლის დროსაც ყველა

სიტყვაფორმას მიემართება მეტყველების შესაბამისი ნაწილის აღმნიშვნელი თეგი (↗მეტყველების ნაწილების თეგი). მაგ., არსებითი სახელი აღინიშნება თეგით N, ზმნა – V, ნაცვალსახელი – PRON, ზმნიზედა – ADV და ა.შ. იხ. **ლინგვისტური ანოტაცია**.

მეტყველების ნაწილების თეგი POS (part-of-speech) tags – კოდი, რომელიც გამოხატულია კონვენციური შემოკლების სახით და აღნიშნავს მეტყველების ნაწილს. მაგ., ქართული ენის ეროვნულ კორპუსში ნაცვალსახელებისათვის ძველ ქართულში გამოიყენება შემდეგი აღნიშვნის კოდები:

მარკერი	განმარტება	მაგალითი
Pron	ნაცვალსახელი	*მე, თჳსი*
Pers	პირის ნაცვალსახელი	*მე, ესე (იტყჳს)*
Dem (Demonstrative)	ჩვენებითი ნაცვალსახელი	*ესე (კაცი იტყჳს)*
Poss (Possessive)	კუთვნილებითი ნაცვალსახელი	*ჩემი, თჳსი*
Interr (Interrogative)	კითხვითი ნაცვალსახელი ვინ → ვინ+Pron+Interr+Hum+Nom ან ვინ+Pron+Interr+Hum+Erg	*ვინ?*
Poss Interr	კითხვით–კუთვნილებითი ნაცვალსახელი ვისი → ვის–ი+ Pron+Poss+Interr+Hum+Nom	*ვისი?*
Rel (Relative)	მიმართებითი ნაცვალსახელი რომელიცა → რომელი+Pron+Rel+ Nom+Sg+Encl:ცა	*რომელიცა// რომელი (ძვ.)*
Det (Determinative)	განსაზღვრებითი ნაცვალსახელი (აქვე შემოდის ნაწევარი)	*თჳთ, ყოველი, სხუად, თავადი, კაცად–კაცადი (კაცი ესე, ეგე, იგი)*
Indef (Indefinite)	განუსაზღვრელობითი ნაცვალსახელი	*რომელიმე// რომელი (ძვ.)*

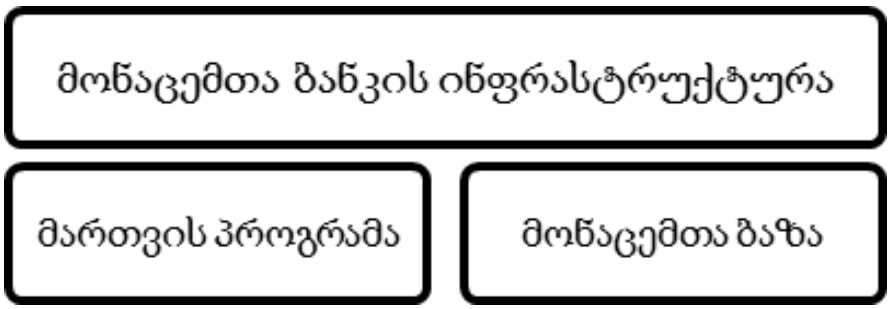
მარკერი	განმარტება	მაგალითი
Neg (Negative)	უარყოფითი ნაცვალსახელი ნუვინ → ნუვინ+Pron+Neg+Hum+Nom+Imp (Imperative – ბრძანებითი)	*ნუვინ*
Recip (Reciprocal)	ურთიერთობითი ნაცვალსახელი	*ურთიერთას*
Refl (Reflexive)	უკუქცევითი ნაცვალსახელი თავი → თავი+Pron+Refl+ Nom +Sg	*თავი*

მომხმარებლის ნიღაბი user interface – მანქანისა და ადამიანის ინტერაქციის პროგრამული ჭრილი (განმავრცობელი ცნება usability engineering).

მონაცემი data – დისკრეტული რიცხვების ბინარული რიგი, რომელიც შეიცავს ინფორმაციას. მაგ.: 12356-ის ბინარული კოდი არის [11110001001000000]; სიტყვაფორმის „ბაკურიანი“, როგორც ციფრული მონაცემის ბინარული კოდია:

[11100001 10000011 10010001 11100001 10000011 10010000 11100001 10000011 10011001 11100001 10000011 10100011 11100001 10000011 10100000 11100001 10000011 10011000 11100001 10000011 10010000 11100001 10000011 10011100 11100001 10000011 10011000].

მონაცემთა ბანკი database, data bank – იგივე *მონაცემთა ბანკ-სისტემა,* ელექტრონული მონაცემების მართვის ინფრასტრუქტურა. მისი მიზანია მონაცემთა სიგნიფიკანტური ბაზის ეფიციენტური (↗ეფიციენტურობა), ხანგრძლივი და კოჰერენტული (არაურთიერთწინააღმდეგობრივი) შენახვა. მონაცემთა ბანკი შედგება ორი სტრუქტურული ნაწილისაგან: **მართვის პროგრამა** და თვით **მონაცემი** (რომელსაც **მონაცემთა ბაზასაც** ეძახიან). მართვის პროგრამა უზრუნველყოფს მონაცემების სტრუქტურირებულ შენახვას და აკონტროლებს საძიებო ბრძანებების წაკითხვას და შესრულებას. მონაცემთა პირველი ბანკები გასული საუკუნის 60-იანი წლებიდან ყალიბდება. მონაცემთა ბანკ-სისტემების მაგალითებია Oracle, IBM-DB2, mySQL.



მონოლინგვური კორპუსი monolingual corpus – ერთი მოცემული ენის რესურსების ბაზაზე შექმნილი კორპუსი. უკავშირდება „ენის“, როგორც ასეთის, დეფინიციის პრობლემას.

მონიტორული კორპუსი monitor corpus – მონიტორული კორპუსის მონაცემთა ბაზა იმთავითვე გათვლილია პერმანენტული განვრცობისათვის. კორპუსში დამატებული მასალის ჟანრობრივი ბალანსი არ არის კონსტანტური და შესაძლოა დროდადრო შეიცვალოს. მონიტორული კორპუსის საუკეთესო

მაგალითს წარმოადგენს BoE (The Bank of English). მისი შექმნა დაიწყო გასული საუკუნის 80-იან წლებში და მუდმივად განივრცობა დღემდე. იგი შეიცავს დაახლოებით ნახევარ მილიარდ ტოკენს (Hunston 2002: 15). მეორე, ასევე ცნობილი მონიტორული კორპუსია COCA (The Corpus of Contemporary American English, Davies 2009) და გამოირჩევა ზუსტი კორპუსდიზაინით ტექსტურ ჟანრებთან მიმართებით.

მულტილინგვური კორპუსი multilingual corpus – ერთზე მეტი ენის რესურსების ბაზაზე შექმნილი კორპუსი. ამგვარი კორპუსების შექმნა მოტივირებულია თარგმანის ავტომატიზირებისა და შედარებითი კვლევის პრობლემებით. უკავშირდება „ენის“, როგორც ასეთის, დეფინიციის პრობლემას.

მყარი დისკი hard disk – ელექტრონული დამახსოვრების ტექნოლოგია, რომელიც ფერომაგნეტური რემანენტულობის პრინციპს ეყრდნობა. ინფორმაციის ჩაწერის დროს ადგილი აქვს მაგნეტური დისკის ზედაპირის მაგნეტური ძაბვით გაჯერებას.

ნიმუში sample – ცალკეული ტექსტი ან ტექსტის ექსპლიკატი, რომელიც კორპუსისათვის შეირჩა კონკრეტულ კრიტერიუმებზე დაყრდნობით. ამ ცნებას გააჩნია სტატისტიკური გაგებაც. Sampling-ით ხშირად აღინიშნება მონაცემების შეგროვების წინა ფაზა, როცა დოკუმენტირებული მასალა მოწმდება რეპრეზენტაციულობისა და რაოდენობრივი სიგნიფიკანტურობის კრიტერიუმებით.

ოპერატიული მეხსიერება RAM – random access memory – კომპუტერში განთავსებული ინფორმაციის დროებითი საწყობი.

ოპორტუნისტული კორპუსი opportunistic corpus (ხშირად იხმარება **კანიბალური კორპუსი cannibalistic corpus**) – ოპორტუნისტული კორპუსი ითვლება რეფერენციალური კორპუსის ალტერნატივად და გამოდის იქიდან, რომ კორპუსის ბალანსირება და რეპრეზენტაციულობა კორპუსის ინჟინერიის არარეალური მიზნებია. ამიტომ იმთავითვე მიჩნეულია, რომ ყოველი კორპუსი დაუბალანსებელია. ეს იდეა საზრდოობს დევიზით: *რაც უფრო დიდია კორპუსი, მით უკეთესი.* ამიტომ სახეზე გვაქვს მონაცემების მასიური, არასისტემური დაგროვება. ოპორტუნისტული კორპუსის დადებით მხარეს წარმოადგენს მისი განსაკუთრებული მოქნილობა ზეპირი მეტყველების კორპუსის შექმნის დროს.

ორთოგრაფიული ტრანსკრიფცია orthographic transcription – აუდიო მასალის ტრანსკრიფციის ფორმა, რომლის დროსაც მხედველობაში არ მიიღება მთქმელის ინდივიდუალური მეტყველების თავისებურებები და ტრანსკრიბირების (↗ტრანსკრიბირება) დროს გათვალისწინებულია სტანდარტული ენის ნორმები.

პარალელური კორპუსი parallel corpus – კორპუსი, რომელიც აერთიანებს ერთი და იმავე ტექსტის ვარიაციებს ერთი ან რამდენიმე ენის ფარგლებში. პარალელური კორპუსი შეიძლება შეიქმნას ერთი რომელიმე კონკრეტული ძეგლის რედაქციების

სინქრონიზაციის ან ტექსტის/ტექსტების სხვადასხვა ენაზე არსებული კორპუსების სინქრონიზაციის შედეგად.

პირველი-, მეორე-, მესამე-, მეოთხე თაობის კონკორდანსი – პირველი თაობის კონკორდანსი უკავშირდება Roberto Busa-ს სახელს, რომელმაც 1952 წელს პირველი ავტომატური გენერატორი დააპროგრამა. აღნიშნულ გენერატორს შეეძლო სწორხაზოვანი კონკორდანსის გამოტანა ლოკალური ბაზის საფუძველზე.

მეორე თაობის კონკორდანსი უკავშირდება 80-იან წლებში პერსონალური კომპიუტერის გავრცელებას და პირველად ართმევს თავს მონაცემთა რაოდენობრივად გაზრდილ ბაზებს.

მესამე თაობის კონკორდანსი გამოირჩევა პროგრამული ინსტრუმენტების მრავალფეროვნებით (მაგ., WordSmith: Scott 1996; MonoConc: Barlow 2000; AntConc: Anthony 2005; Xaira) და პროცესორის სისწრაფით.

მეოთხე თაობის კონკორდანსი მიემართება ინტერნეტის ქსელს, მონაცემთა ექსპონენციალურ რაოდენობას და გამოირჩევა განსაკუთრებული სისწრაფით. ახალი პროგრამული ინსტრუმენტები კორპუსის ლოკალური გამართვისა და აღჭურვის საშუალებას იძლევა (corpus.byu.edu: Davies 2005; Wmatrix: Rayson 2008; SketchEngine: Kilgarriff et al. 2004; BNCweb: Hoffmann et al. 2008; CQPweb).

პოსტი post – ელექტრონული შეტყობინების (ტექსტური, ვიზუალური, აუდიო, ვიდეო) განთავსება საჯარო ფორუმზე.

პოსტმოდიფიკაცია postmodification – სახელური ფრაზის დეტერმინაცია მსაზღვრელ-საზღვრულის ადგილმდებარეობის მიხედვით. მსაზღვრელი საზღვრულის (↗გრამატიკული თავი) დეტერმინაციას, იგივე მოდიფიკაციას, ახდენს მისგან მარჯვენა პოზიციაში. მაგ., წინადადებაში *„ღმერთო, ღმერთო, ეს **ხმა ტკბილი** გამაგონე ჩემს მამულში!“* (ი. ჭავჭავაძე) *ტკბილი* მიემართება გრამატიკულ თავს *(ხმა)* და ახდენს მის პოსტმოდიფიკაციას.

პროსოდიული ანოტაცია prosodic annotation – კორპუსის ანოტაციის (↗ანოტაცია) ფორმა, რომლის დროსაც ხდება ზეპირმეტყველების მასალაში პროსოდიული ფენომენების მარკირება, მაგ.: პაუზა, ინტონაცია, მახვილი, ტონი და ასე შემდეგ.

პრემოდიფიკაცია premodification – სახელური ფრაზის დეტერმინაცია მსაზღვრელ-საზღვრულის ადგილმდებარეობის მიხედვით. მსაზღვრელი საზღვრულის (↗გრამატიკული თავი) დეტერმინაციას, იგივე მოდიფიკაციას ახდენს მისგან მარცხენა პოზიციაში. მაგ. წინადადებაში *"**ჭრელი პეპელა** დაათრო და გააბრუა იამა"* (ა. წერეთელი) *ჭრელი* არის სახელური ფრაზის გრამატიკული თავის *(პეპელა)* ატრიბუტი და ახდენს მის პრემოდიფიკაციას.

პროსოდია prosody – ზეპირმეტყველების რელევანტური ფონეტიკური სეგმენტების ჯამი (პაუზა, ინტონაცია, მახვილი, ტონი).

ჟანრი genre – კომპოზიციის კატეგორიზებისთვის კრიტერიუმების თავისუფალი ერთობლიობა, როგორც

წესი, ვრცელდება ხელოვნებასა და კულტურაზე, განსაკუთრებით ლიტერატურაზე. ჟანრები კონვენციური ხასიათისაა და დარგების მიხედვით სპეციფიცირდება. კორპუსლინგვისტიკაში ჟანრი გულისხმობს ტექსტის სახეს – ლიტერატურული კორპუსი, იურიდიული დოკუმენტაციის კორპუსი, სამეცნიერო კორპუსი, ჩატის კორპუსი, მინიგზავნილების (**SMS**) კორპუსი, იდიომატურ გამოთქმათა კორპუსი, ანეკდოტების კორპუსი და ა.შ.

ჟანრული მრავალფეროვნება – იხ. **ბალანსირება.**

რეგულარული გამოსახულება regular expression (RegEx) – სუროგატული სიმბოლოების ჯაჭვი, რომელიც დალაგებულია ფორმალური სინტაქსური წესებით და იძლევა გარკვეული მიმართებების პოვნის საშუალებას მონაცემთა ბანკში. მაგალითად, რეგულარული გამოსახულება [„კაც.*“ „features=Aor“] GEKKO-ში მოიძიებს სიტყვის „კაცი“ ყველა შესაძლებელ ფლექსიურ ფორმას (*კაცმა, კაცს, კაცის* და ა.შ.), რომელიც წინ უსწრებს აორისტში მდგომ ზმნას.

რეპრეზენტაციულობა representativeness – კორპუსის ერთ-ერთი თვისობრივი მახასიათებელი, რომელიც გულისხმობს ტექსტური მასალის სელექციას წინდაწინ დადგენილი კრიტერიუმის (sampling frame) მიხედვით. მაგალითად, გარკვეული დროის განმავლობაში დაბეჭდილი ყველა გაზეთის მაგივრად კორპუსის ბაზაში განთავსდება ამ გაზეთების მხოლოდ ერთი პატარა ნაწილი, რომელსაც ენიჭება კონკრეტული კორპუსული ინდექსი ამ პერიოდის ბეჭდვითი ენისათვის.

რეფერენცია reference – ცალმხრივი მიმართება ორ ელემენტს შორის ერთი წინადადების ან დისკურსის ფარგლებში (იხ. ანაფორა, კატაფორა, დეიქსისი).

სასწავლო ელექტრონული რესურსი – ელექტრონულ ფორმატში წარმოდგენილი სასწავლო რესურსი, ელექტრონული და დისტანციური სწავლების აუცილებელი კომპონენტი. ელექტრონული რესურსების გამოყენება ხელს უწყობს პროდუქტიული სასწავლო გარემოს შექმნას.

სასწავლო კორპუსი learner corpus – წერილობითი და ზეპირი ტექსტების კომპიუტერული კოლექციები, რომლებიც უცხო ენის შემსწავლელთა მეტყველებას ასახავს. სასწავლო კორპუსების ერთ-ერთი მთავარი მახასიათებელია შეცდომების გამოვლენაზე ორიენტირება: მათი ტიპიზირება, ანალიზი და აღბეჭდვა. ასეთებია, მაგ., LINDSEI (Louvain International Database of Spoken English Interlanguage), ICLE (International Corpus of Learner English) ინგლისურისათვის, FALKO – Ein fehlerannotiertes Lerner-korpus des Deutschen გერმანულისათვის.

სატიტულო გვერდი / თავფურცელი home page – ვებგვერდის ქუდი.

სტატისტიკა statistics – კვანტიტატიური ინფორმაციის (მონაცემების) ასახვის მეთოდი. იგი წარმოადგენს თეორიასა და ემპირიას შორის კოჰერენტული რელაციის ასახვის ერთ-ერთ ფორმას. კორპუსლინგვისტიკა აქტიურად იყენებს სტატისტიკის მეთოდებს და ოპერირებს იმგვარი თემებით,

როგორებიცაა სკალირება, კომბინატორიკა, მონაცემთა კვანტიტატიური განაწილების ფორმები, კონტინგენტურობა და კოეფიციენტი.

სტატისტიკური სიგნიფიკანტურობა statistical significance – სტატისტიკური სიგნიფიკანტურობა გულისხმობს კვანტიტატიურ შედეგს, რომლის ცდომილება, როგორც წესი, 5%-ზე დაბალია. მის გასაზომად გამოიყენება ტესტების სხვადასხვა ფორმა.

სემანტიკური თეგირება semantic tagging – ანოტაციის მეთოდი, რომლის დროსაც კორპუსის ენობრივი მონაცემის კოდირება ხდება მისი სემანტიკური როლიდან გამომდინარე (აგენსი, პაციენსი, რეციპიენსი, ექსპირიენსი და ასე შემდეგ).

სემანტიკური ორმნიშვნელობიანობის მოხსნა / დისამბიგვირება disambiguation – მორფოლოგიური და სემანტიკური ომონიმიის მოხსნა კონტექსტის ფარგლებში.

სემანტიკური პრეფერენცია semantic preference – გარკვეული სიტყვის კონტექსტუალური ასოციაცია გარკვეულ სემანტიკურ როლთან. მაგალითად, სიტყვა „ცემა“ ასოცირდება აგენსის და პაციენსის სემანტიკურ როლებთან, მაშინ, როდესაც „წყურვილი“ – ექსპირიენსთან.

სემანტიკური პროსოდია semantic prosody – ზოგიერთი იდიომატური ერთეული (ანდაზა, დისკურსული ნაწილაკი...), რომელიც პროსოდიის მუდმივად მდგრად ტიპს ავლენს და უკავშირდება მის მიერ

გამოხატულ შინაარსსა და გამონათქვამის პრაგმატულ ფუნქციას.

სიხშირული (კვანტიტატიური) ანალიზი quantitative analysis – კორპუსის მონაცემებზე დამყარებული სტატისტიკური კვლევის ფორმა, რომელიც ასახავს ტოკენის (↗ტოკენი) სიხშირულ მაჩვენებელს.

სწავლების მართვის სისტემა learning content management system – ელექტრონულ სწავლებაში გამოყენებული ტექნიკური საშუალება და ფუნქციონირებადი რგოლი, რომელიც უზრუნველყოფს სასწავლო მასალის (**content**) შექმნას, მოძიებას, ტრანსპორტირებას და გამოყენებას. სწავლების მართვის სისტემა საშუალებას აძლევს მის ავტორებს გასცენ სასწავლო მასალებით სარგებლობის უფლება და დიფერენცირება მოახდინონ მომხმარებლებს შორის მასალაზე წვდომის, მისი გადამუშავების ან მართვის თვალსაზრისით.

ტერაბაიტი terabyte – ინფორმაციის საზომი ერთეული. 1 ტერაბაიტი უდრის 10^{12} ბაიტს = 1.000.000.000.000 ბიტს.

ტექსტი text – კრებითი ცნება წერილობითი დოკუმენტების, არტეფაქტებისა და დოკუმენტირებული ზეპირი მეტყველების აღსანიშნავად.

ტექსტური მარკერი textual markup – *(ტექსტური მანიშნებელი)* – გამოიყენება კორპუსში ტექსტის ფორმატის კოდირებისათვის. ტექსტური მარკერის კოდი იძლევა ორიგინალური მონაცემის მახასიათებ-

ლების ადეკვატური კოდირების საშუალებას. მაგალითად, ტოპონიმის, ანთროპონიმის ასევე ტექსტის სტრუქტურის (თავები, პარაგრაფები) დეტალურ მონიშვნას. ყველაზე გავრცელებული ტექსტუალური მარკირების ენაა XML (Extensible Markup Language *მინიშნების განვრცობადი ენა*).

ტიპი type – კორპუსლინგვისტიკის ერთეული, გამოიყენება სტატისტიკური მიზნით. ტრადიციულ ლინგვისტიკაში მას შეესაბამება ცნება **სიტყვაფორმა**. კორპუსში ერთი ტიპი, როგორც წესი, უდრის n-ტოკენს. მაგალითისათვის – „ქარი ჰქრის, ქარი ჰქრის, ქარი ჰქრის ...“ (გ. ტაბიძე) შედგება 6 ტოკენისაგან, მაგრამ 2 ტიპისაგან – *ქარი* და ჰ*ქრის*. იხ. **ტოკენი**.

ტოკენი token – კორპუსის უმცირესი შემადგენელი ელემენტი, ტრადიციულ ლინგვისტიკაში იგი შეესაბამება ცნებას **სიტყვაფორმა**. იხ. **ტიპი**.

ტიპ-ტოკენის შესაბამისობა type–token ratio (TTR) – კორპუსის გაზომვადი თვისება, რომელიც გამოიხატება მასში არსებული მონაცემების ტიპებისა და ტოკენების რაოდენობების ურთიერთმიმართების ინდექსში. რაც უფრო ახლოსაა ინდექსი 1-თან, მით უფრო მრავალფეროვანია კორპუსის მონაცემთა ბაზა ფორმობრივი თვალსაზრისით.

ტრანსკრიბირება transcription – ტექსტის წინასწარ შეთანხმებული წესების მიხედვით გადმოწერა. წინადადების ტრანსკრიბირების მაგალითია:

თხა-მ	ვენახ-ი	შეჭამ-ა
txa-m	*venax-i*	*šeč̣am-a*

უნიკოდი Unicode – მანქანურად წაკითხვადი ტექსტის კოდირებისა და დეკოდირების სტანდარტული სისტემა, რომელიც საერთაშორისო სტანდარტის ISO-10646-ის ნაწილს წარმოადგენს და კოდირების 16-ბიტიან სისტემას იყენებს.

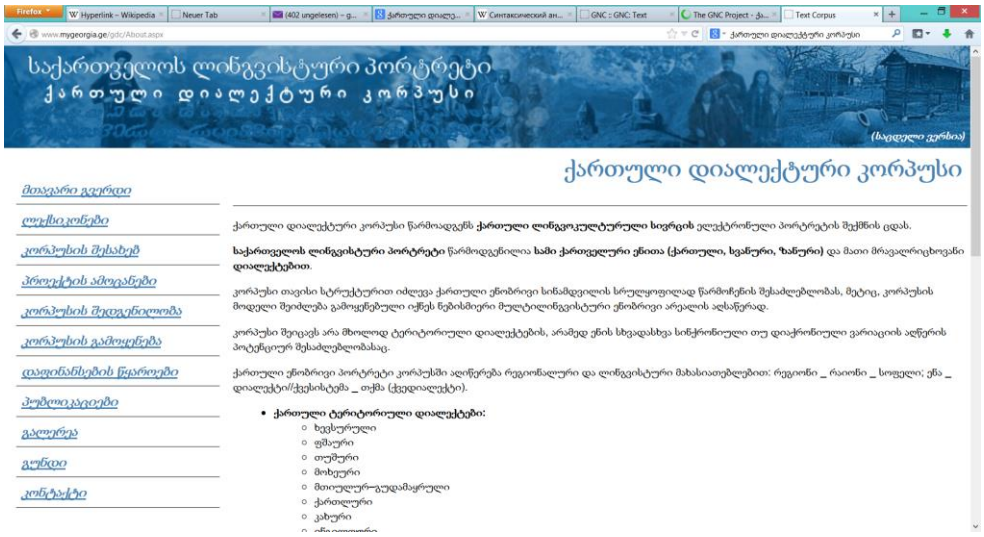
უნიკოდის კონსორციუმი Unicode consortium – უნიკოდის (↗ უნიკოდი) სტანდარტის საერთაშორისო ორგანიზაცია.

ფალსიფიკაცია, ფალსიფიცირება falsification – (ლათ. *falsus* მცდარი და *facere* კეთება) ექსპერიმენტული შედეგები ან დასკვნები, რომლებიც ვარაუდის მცდარობას ასაბუთებენ. იხ. **ვერიფიკაცია** (განმავრცობელი ცნება: ანტითეზა).

ფონეტიკური ტრანსკრიფცია phonetic transcription – ზეპირი მეტყველების ანოტაციის სახეობა, რომლის დროსაც აუდიო მასალის ჩაწერის ფორმა ეყრდნობა საერთაშორისო ფონეტიკურ ანბანს IPA (International Phonetic Alphabet).

ქართული ენის დიალექტური კორპუსი GDC (Georgian Dialect Corpus) – თემატური კორპუსი, რომელიც შეიქმნა არნ. ჩიქობავას სახ. ენათმეცნიერების ინსტიტუტში მარინე ბერიძის მიერ. ქართული ენის დიალექტური კორპუსის რესურსები იქმნებოდა სამეცნიერო პროექტის "საქართველოს ლინგვისტური პორტრეტი" ფარგლებში და ფინანსდებოდა „ღია საზოგადოება საქართველოსა“ და შოთა რუსთაველის ეროვნული სამეცნიერო ფონდის მიერ. დიალექტური

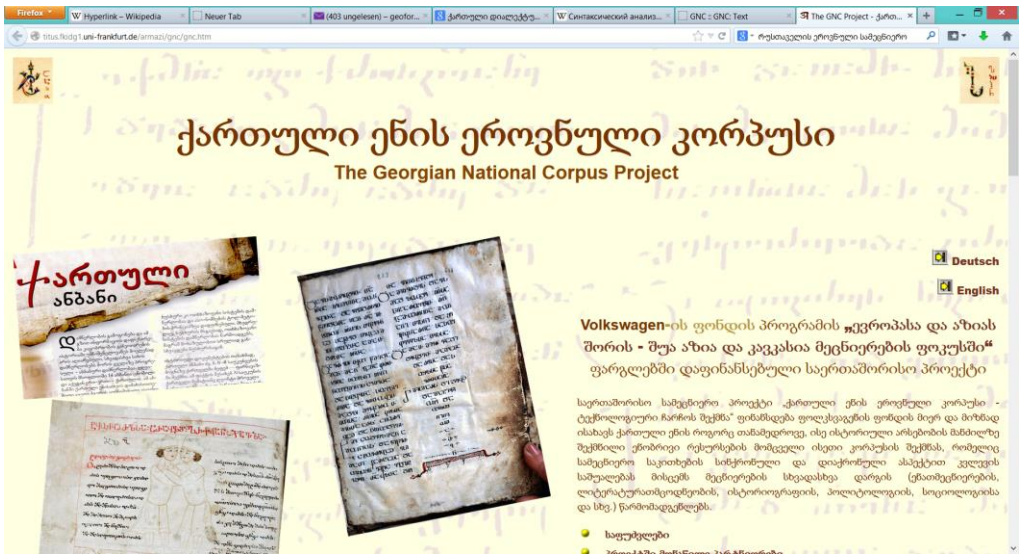
კორპუსის რესურსები შესულია ქართული ენის ეროვნული კორპუსის ძირითად რესურსებში. ქართული ენის დიალექტური კორპუსი წარმოადგენს ქართული ლინგვოკულტურული სივრცის ელექტრონული პორტრეტის შექმნის ცდას. საქართველოს ლინგვისტური პორტრეტი წარმოდგენილია სამი ქართველური ენითა (ქართული, სვანური, ზანური) და მათი მრავალრიცხოვანი დიალექტებით. კორპუსი თავისი სტრუქტურით იძლევა ქართული ენობრივი სინამდვილის სრულყოფილად წარმოჩენის შესაძლებლობას, კორპუსის მოდელი შეიძლება გამოყენებულ იქნას ნებისმიერი მულტილინგვური ენობრივი არეალის აღსაწერად. კორპუსი შეიცავს არა მხოლოდ ტერიტორიული დიალექტების, არამედ ენის სხვადასხვა სინქრონიული თუ დიაქრონიული ვარიაციის აღწერის პოტენციურ შესაძლებლობასაც. ქართული ენობრივი პორტრეტი კორპუსში აღიწერება რეგიონალური და ლინგვისტური მახასიათებლებით: რეგიონი – რაიონი – სოფელი; ენა – დიალექტი//ქვესისტემა – თქმა (ქვედიალექტი). ქართული ენის დიალექტური კორპუსი მოიცავს როგორც საქართველოს ტერიტორიაზე გავრცელებული, ისე მის ფარგლებს გარეთ ირანში, თურქეთსა და საინგილოში არსებული ენობრივი ვარიანტების სამეტყველო ნიმუშებს. ქართული დიალექტური კორპუსი დღეისათვის 1 მილიონზე მეტი ტოკენისგან შედგება.



ქართული ენის დიალექტური კორპუსი

ქართული ენის ეროვნული კორპუსი GNC (Georgian National Corpus) – ქართული ენისათვის შექმნილი კორპუსი, რომელიც აერთიანებს ქართული ენის ეროვნულ საგანძურს – ქართული ენის ისტორიული არსებობის პერიოდში არსებულ წერილობით და ზეპირმეტყველების დოკუმენტაციის შედეგად შექმნილ ელექტრონულ რესურსებს. კორპუსში გაერთიანებული ელექტრონული რესურსების დაგროვება გასული საუკუნის 80-იანი წლებიდან დაიწყო როგორც საქართველოში, ისე მის ფარგლებს გარეთ – ფრანკფურტის გოეთეს სახ. უნივერსიტეტში ცნობილი გერმანელი ქართველოლოგის იოსტ გიპერტის მიერ. ქართული ენის ეროვნული კორპუსი იქმნება ფოლკსვაგენის ფონდის ფინანსური ხელშეწყობით და ეყრდნობა ოთხ ელექტრონულ რესურსს: TITUS, ARMAZI, GEKKO და GDC. ქართული ენის ეროვნული კორპუსის შექმნაში დღეისათვის მონაწილეობენ საქართველოსა და უცხოეთის კვლევითი ინსტიტუტები და სამეცნიერო ჯგუფები: ფრანკფურტის გოეთეს სახ. უნივერსიტეტი

(ფრანკფურტი), არნოლდ ჩიქობავას სახელობის ენათმეცნიერების ინსტიტუტი (თბილისი), ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი, ილიას სახელმწიფო უნივერსიტეტი (თბილისი), საქართველოს ტექნიკური უნივერსიტეტი (თბილისი), შოთა რუსთაველის სახელობის უნივერსიტეტი (ბათუმი), აკაკი წერეთლის სახელმწიფო უნივერსიტეტი (ქუთაისი), საქართველოს პარლამენტის ეროვნული ბიბლიოთეკა (თბილისი), საქართველოს ეროვნული სამეცნიერო ბიბლიოთეკა (თბილისი), საქართველოს ეროვნული მუზეუმი (თბილისი), CLARINO (↗GEKKO) – ბერგენი. ქართული ენის ეროვნული კორპუსი დღეისათვის 130 მილიონი ტოკენისაგან (↗ტოკენი) შედგება და წარმოადგენს დიაქრონული კორპუსის ტიპს. 2013 წლის 28 დეკემბრიდან საქართველოს პარლამენტის თავმჯდომარის დავით უსუფაშვილის ინიციატივით ქართული ენის ეროვნულმა კორპუსმა საქართველოს პარლამენტის ეროვნულ ბიბლიოთეკაში დაიდო ბინა.



ქართული ენის ეროვნული კორპუსი

შედარებითობა – მულტილინგვური კორპუსების გაზომვადი ღირებულება. კორპუსული შედარებითობის ფენომენი უკავშირდება ვიკიპედიის წარმოშობას, რომელიც წარმოადგენს ყველაზე მრავალენოვან ფართოდ გავრცელებულ "კორპუსს". სპეციალური ალგორითმის საშუალებით გამოთვლილი ორი სხვადასხვა ენის კორპუსული შედარებითობა საშუალებას იძლევა დაიხვეწოს ავტომატური თარგმნა. (განმავრცობელი ცნება: შედარებითობის ხარისხი, შედარებითობის კოეფიციენტი).

ციფრული ჰუმანიტარია – იხ. **დიგიტალური ჰუმანიტარია**

ხარისხობრივი (კვალიტატიური) ანალიზი qualitative analysis – კორპუსის მონაცემებზე დამყარებული ლინგვისტური კვლევის ფორმა, რომელიც ასახავს განსხვავებული ფორმების მოხმარების სპეციფიურ შემთხვევებს.

ხაზთაშორისი ანოტაცია interlinear glossing – ენობრივი მონაცემის ანალიზის ასახვის მეთოდი, სიტყვასიტყვითი ანალიზი, მისი სინონიმია **გლოსირება.** ხაზთაშორისი ანოტაციის წაკითხვის გაადვილებას ემსახურება **ტრანსკრიფცია**.

კაც-ი	სახლ-ში	მიდი-ს.
ķaci	*saxlši*	*midis*
man:NOM	Home:DAT.PP	go:PRES. 3S.Sg.

“The man is going home.”

ჰიპერბმული hyperlink – (იგივე **ელექტრონული გადამისამართება**) ელექტრონულ რესურსში

საგანგებოდ მონიშნული ობიექტი (სიტყვა, ფოტო, გრაფიკული გამოსახულება, აკრონიმი), რომელიც საშუალებას აძლევს მომხმარებელს გადავიდეს სხვა ინფორმაციაზე ერთი ჰიპერტექსტის ფარგლებში, რომელიც გაცილებით ვრცელ ინფორმაციას შეიცავს მოცემული ობიექტის შესახებ. იგი პირველად შემოიტანა ტედ ნელსონმა 1960 წელს. HTML-ის სინტაქსური წესების მიხედვით ჰიპერბმული ასე ჩაიწერება:

ქართული ენის ეროვნული კორპუსი

ამგვარ ჰიპერბმულს მომხმარებელი გადაჰყავს ქართული ენის ეროვნული კორპუსის ვებგვერდზე, რომლის მისამართია:

http://titus.fkidg1.uni-frankfurt.de/armazi/gnc/gnc.htm

ჰიპერტექსტი hypertext – (იგივე ზეტექსტი). წარმოადგენს ქსელური სტრუქტურის მქონე ტექსტს შემდგარს ობიექტებისაგან, რომლებსაც ერთმანეთთან ბმულობა გააჩნიათ. მისი უპირატესობა მდგომარეობს ინფორმაციის მრავალგანზომილებიან ურთიერთწვდომაში, რაც კომპლექსური შინაარსის ასახვისა და მოხმარების გაადვილებას იწვევს.

ჰიპერტექსტის მარკირების ენა HTML (Hyper Text Markup Language) – ციფრული ინფორმაციის სტრუქტურის მონიშვნის პროგრამული საშუალება, რომელიც განკუთვნილია ვებგვერდების შესაქმნელად და ინფორმაციის გასავრცელებლად ინტერნეტის საშუალებით. HTML დაწერილია HTML ელემენტების ფორმით, რომელიც შედგება თეგებისგან. თეგი მოქცეულია კუთხიან ფრჩხილებში (როგორც <b>). HTML-თეგები ძირითადად წყვილის სახით არის წარმოდგენილი,

როგორც <h1> და </h1>, თუმცა ზოგიერთი თეგი ცარიელ კენტ ელემენტს წარმოადგენს და არა ჰყავს მეწყვილე, მაგალითად, <img> ან
. წყვილის პირველი თეგი არის საწყისი თეგი – ხსნის ბრძანებას, ხოლო მეორე – დამასრულებელი, ანუ ზღვარს უდებს ბრძანების მოქმედების არეს. წყვილი თეგების შუაში თავსდება ტექსტი, სურათი ან სხვა რესურსი, რამდენიმე თეგის ერთდროულად გამოყენების შემთხვევაში კი სხვა თეგები. მაგ. "ბაკურიანის ზამთრის სკოლა – ***დიგიტალური ჰუმანიტარია***" ასე ჩაიწერება: [ბაკურიანის ზამთრის სკოლა ****დიგიტალური ჰუმანიტარია ** **].

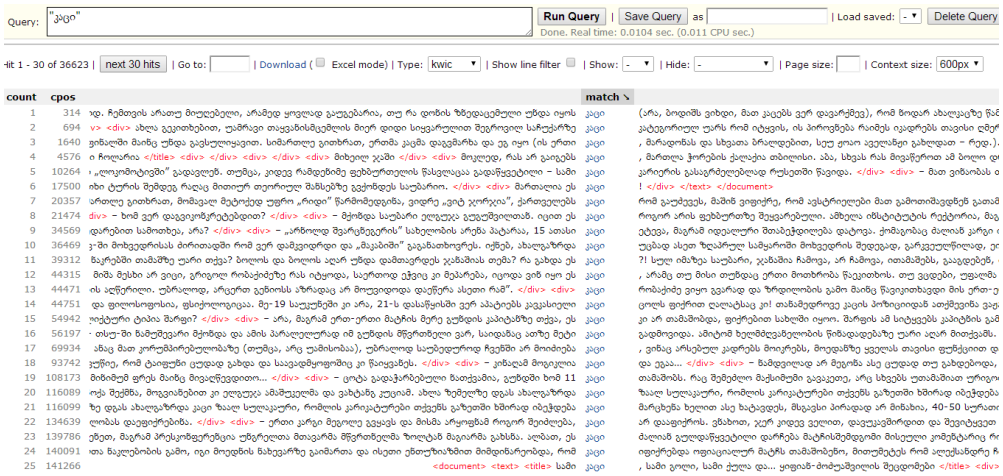
ARMAZI (Caucasian Languages and Cultures: Electronic Documentation) – აკრონიმი საერთაშორისო სამეცნიერო პროექტისა "კავკასიური ენები და კულტურები: ელექტრონული დოკუმენტაცია". პროექტს ხელმძღვანელობდნენ ფრანკფურტის გოეთეს სახ. უნივერსიტეტის პროფესორები იოსტ გიპერტი და მანანა თანდაშვილი. პროექტს აფინანსებდა **ფოლკსვაგენის ფონდი** (**Volkswagen Stiftung**). პროექტი მიზნად ისახავდა კავკასიური ენების ელექტრონული პორტალის შექმნას და საქართველოს ჰუმანიტარული დარგის წამყვანი სამეცნიერო კვლევითი ინსტიტუტებისა და კვლევით-საგანმანათლებლო დაწესებულებების საერთაშორისო საინფორმაციო ქსელში (**WWW**) ჩართვას. **ARMAZI** აერთიანებდა 11 სამეცნიერო და კვლევით დაწესებულებას. მათ შორის იყო: საქართველოს მეცნიერებათა აკადემია, თბილისის ივ. ჯავახიშვილის სახ. სახელმწიფო უნივერსიტეტი (ძველი ქართული ენის კათედრა, სამეცნიერო ლაბორატორია "ორიონი", სტრუქტურული და გამოყენებითი ლინგვისტიკის კათედრა), არნ. ჩიქო-

ბავას სახ. ენათმეცნიერების ინსტიტუტი, კ. კეკელიძის სახ. ხელნაწერთა ინსტიტუტი, შოთა რუსთაველის სახელობის ქართული ლიტერატურის ინსტიტუტის ფოლკლორის არქივი, გელათის მეცნიერებათა აკადემია, გ. წერეთლის აღმოსავლეთმცოდნეობის ინსტიტუტი, შოთა რუსთაველის სახელმწიფო კომისია, ა. წერეთლის სახ. ქუთაისის სახელმწიფო უნივერსიტეტი. **არმაზის** რესურსების შექმნამ საშუალება მისცა პროექტში მონაწილე სამეცნიერო კვლევით ინსტიტუტებსა და კვლევით-საგანმანათლებლო დაწესებულებებს, მათ ხელთ არსებული ცალკეული მასალები გადაექციათ ელექტრონულ რესურსებად და დაუფლებოდნენ თანამედროვე ტექნოლოგიური საშუალებებით კვლევის მეთოდებს. პროექტმა უდავოდ შეუწყო ხელი კავკასიის ენობრივი და კულტურული მემკვიდრეობის მრავალფეროვნებისა და ქართული სამეცნიერო სკოლის წარმოჩენას საერთაშორისო ქსელში და ხელმისაწვდომი გახადა **არმაზის** მასალები როგორც დასახელებული სამეცნიერო კვლევითი ცენტრების, ისე საერთაშორისო სამეცნიერო საზოგადოებისათვის.

ARMAZI-ს ვებგვერდი

GEKKO – ქართული ენის ელექტრონული კორპუსი, რომლის პროგრამულ ჩარჩოს წარმოადგენს კორპუსის მართვის პლატფორმა Corpuscle. პლატფორმა განკარგავს კორპუსის ძებნისა და ანალიზის ყველა ძირითად ხელსაწყოს (კორკონდანსი, კოლოკაცია, სიტყვაფორმების სია, დისტრიბუცია) და მიემართება სტრუქტურირებულ მონაცემებს (XML). **GEKKO-**ს ავტორია პაულ მოირერი, ნორვეგიაში მოღვაწე მეცნიერი. იხ. http://iness.uib.no/gekko.

KWIC-ინდექსი – (ასევე პერმუტაციული ინდექსი) საძიებო ცნებისა და კონტექსტის დალაგების (სორტირების) გავრცელებული ფორმა, რომლის დროსაც მოძიებული კონტექსტები ციკლურად (პერმუტაციულად) აღბეჭდავს ყველა ელემენტს **KWIC**-ის (↗KWIC) სახით. ციკლს საფუძვლად უდევს ანბანური რიგი, რომელსაც **KWIC**-სიასაც უწოდებენ.



KWIC- ინდექსის მაგალითი GEKKO-დან

KWIC (key word in context) – საძიებო სიტყვაფორმა კონტექსტში. არსებითად განსხვავდება KWOC -საგან (key word out of context საძიებო სიტყვა კონტექსტის გარეშე). აღნიშნული ცნება პირველად გამოიყენა გერმანელმა მათემატიკოსმა ჰანს პეტერ ლუნმა (Hans Peter Luhn),

რომლის სახელი ასევე უკავშირდება ალგორითმს "modulus 10".

KWIC-ი წარმოადგენს წინაკომპიუტერული ეპოქის მეთოდს, რომელიც ბიბლიოთეკებსა და არქივებში გამოიყენებოდა მონაცემთა სათაურების ინდექსირებისათვის. იგი უადვილებდა მომხმარებელს მოსაძებნი წყაროს მიგნებას თემატური ველის მითითების საფუძველზე. კომპიუტერულ ეპოქაში **KWIC**-ი წარმოადგენს მონაცემთა კომპრესიის (კომპრიმირების) ერთერთ ფორმას, რომლის დროსაც საძიებო ცნება კონტექსტთან შედარებით გამოყოფილია ცენტრალური პოზიციითა და განსხვავებული გრაფიკული ნიშნით (სიმუქე, ზომა და ა.შ.).

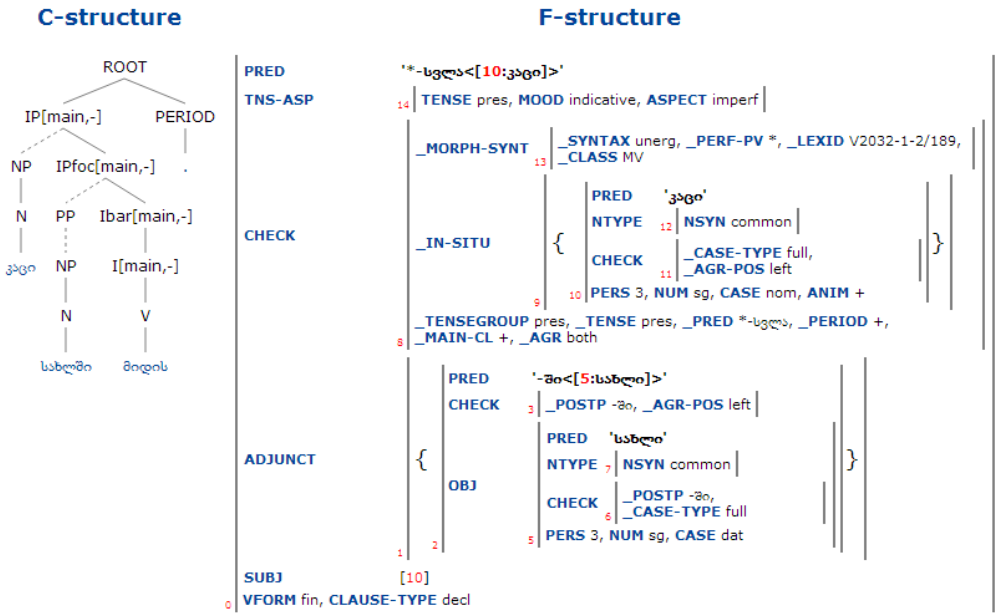
n*-gram** – ტექსტის დანაწევრება ფრაგმენტებად, რომელიც მეორე ეტაპზე **N**-ელემენტებად (n*-gram**) ჯგუფდება. ამგვარ დაჯგუფებას შესაძლოა საფუძვლად ედოს ასოების, მორფემების, ლექსემების მსგავსებები. ***n*-gram**-ის სახეებია მონოგრამები, ბიგრამები და ტრიგრამები შესამაბისად მაკონსტრუირებელი ნიშნის რაოდენობის მიხედვით. ასევე საუბარია მულტიგრამებზეც. ***n*-gram**-ის ფორმალური განსაზღვრება: თუ **Σ** სასრული ანბანია და **n** წარმოადგენს მთლიან დადებით რიცხვს, მაშინ ***n*-gram**-ი არის **Σ** ანბანისაგან შემდგარი ერთი **w** სიტყვა **n** სიგრძით –$w=(w_1,\ldots,w_n) \in \Sigma^n$. იხ. მაგალითი ქართულისათვის:

Monogramm/Unigramm	1	ა
Bigramm	2	აი
Trigramm	3	სიო
Tetragramm	4	თათი

Pentagramm	5	ვაშლი
Hexagramm	6	თამასა
Heptagramm	7	თბილისი
Oktogramm	8	წიგნების
…	…	…
Multigramm	N	კორპუსლინგვისტიკა

online vs. offline – ქსელური მომსახურების რეჟიმი vs. პერსონალური მომსახურების რეჟიმი.

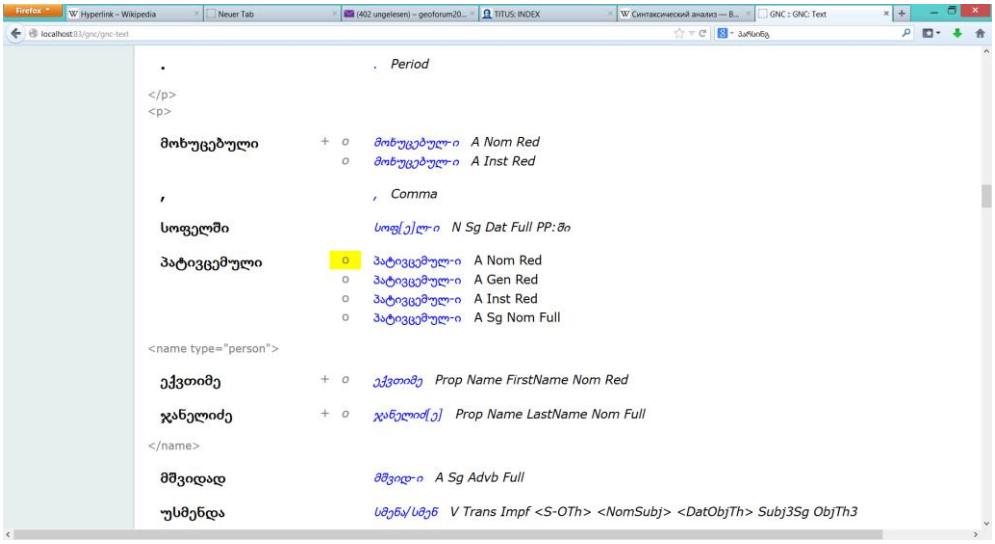
Parsing – კორპუსლინგვისტიკის ცნება ავტომატური სინტაქსური დანაწევრების პროცესის აღსანიშნად. პროცესი გულისხმობს ლექსემათა (კორპუსში ტოკენთა) ხაზობრივი თანმიმდევრობის ავტომატურ შეპირისპირებას ბუნებრივი ან ფორმალური ენის ფორმალურ გრამატიკასთან (ფორმალური წესების ბანკთან) სემანტიკური ორმნიშვნელიანობის (↗დისამბიგვირება) გათვალისწინებით. ლექსიკური ანალიზატორის პარალელურად, როგორც წესი, გამოიყენება მორფოლოგიური და სინტაქსური ანალიზატორიც. **Parsing**-ის შედეგად ხდება წინადადების სინტაქსური სტრუქტურის ვიზუალური მოდელის გენერირება ე. წ. ლინგვისტური (სინტაქსური) ხის ფორმატში.



Parsing-ის მაგალითი GEKKO-დან

Parsed corpus – სინტაქსური წესების მიხედვით დანაწევრებული და სემანტიკურად დისამბიგვირებული მონაცემთა ბაზა. ამგვარ კორპუსებში, როგორც წესი, დაძლეულია მორფოლოგიური ომონიმიის პრობლემა. მაგ., ტოკენისათვის (↗ტოკენი) „პატივცემული“ GEKKO-ს მორფოლოგიური ანალიზატორი გვთავაზობს შემდეგ ვარიანტებს:

პატივცემული – სახ. ბრუნვა (< პატივცემული ექვთიმე ჯანელიძე)
პატივცემული – ნათ. ბრუნვა (<პატივცემული ექვთიმე ჯანელიძის)
პატივცემული – მოქ. ბრუნვა (<პატივცემული ექვთიმე ჯანელიძით)



დისამბიგვირების პროცესი GEKKO-ში

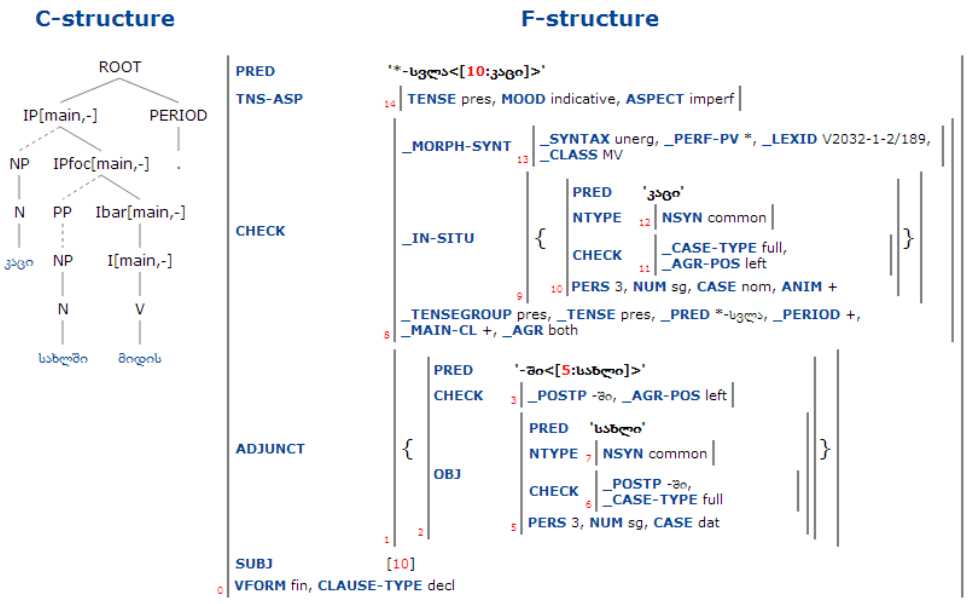
TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien) – აკრონიმი საერთაშორისო სამეცნიერო პროექტისა "ინდოგერმანული ტექსტებისა და ენობრივი მასალების თეზაურუსი". პროექტის ინიციატორი და ხელმძღვანელი იყო პროფ. იოსტ გიპერტი, რომელმაც პრაღის უნივერსიტეტთან თანამშრომლობით პირველმა შექმნა კვლევის თანამედროვე ტექნოლოგიური საშუალებებით აღჭურვილი უპრეცენდენტო კორპუსი ინდოევროპული ენებისათვის. **TITUS-**ი აერთიანებს ინდოევროპული ენების ელექტრონულ რესურსებს, როგორიცაა ინდური ენები (ვედური, სანსკრიტი, ფალი, ჰინდი, მალდივური), ირანული ენები (ავესტა, ძველი, საშუალი და ახალი სპარსული, პართიული, სოგდური, ხოტან-საკური, ბაქტრული, ოსური, ზაზაკი), ანატოლიური (ხეთური, ლუვიური, ფალაური), თოხარული A და B (აღმოსავლური და დასავლური თოხარული), სომხური, ბალტიური ენები (ძველი პრუსიული, ლატვიური, ლიტვური), სლავური ენები (ძველი საეკლესიო სლავური, ჩეხური, პოლონური, ხორვატული, სორბული, ბულგარული),

გერმანიკული ენები (გოთური, ძველი ფრიზული, ძველი საქსონური, ქვემო გერმანული, ძველი, საშუალი და ახალი გერმანული, ძველი ინგლისური, ძველი და საშუალი ჰოლანდიური), ბერძნული (მიკენური და კლასიკური ბერძნული), იტალიკური ენები (ლათინური, ფრანგული, იტალიური, პორტუგალიური), კელტური ენები (ძველი ირლანდიური, საშუალუელსური, საშუალბრეტონული, ლეპონტური), სემიტური ენები (ებრაული, არაბული, სირიული, ეთიოპური), ფინო-უნგრული ენები (ფინური, უნგრული), თურქული ენები (ძველი თურქული, ყარაჩაული, ბალყარული, აზერბა-იჯანული), დრავიდული ენები (თამილი), სინოტიბეტური ენები (ჩინური). **TITUS-**ი, გარდა ინდოევროპული ენებისა, კავკასიური ენების რესურ-სებსაც მოიცავს. მათ შორისაა ქართულის, მეგრულის, ლაზურის, სვანურის, წოვა-თუშურის, უდიურის რესურსები.

TITUS-ის ვებგვერდი

Treebank – სინტაქსური ხეების ბანკი (განშტოებული დიაგრამების ფორმით). წარმოადგენს სინტაქსურად და სემანტიკურად ანოტირებულ ტექტს, რომელიც დანაწევრებულია მოცემული ენის სინტაქსის

ფორმალური წესების (↗Parsing) მიხედვით. ცნება **parsed corpus** ხშირად გამოიყენება ცნების **Treebank** სინონიმად.



ქართული მაგალითი The INESS treebanking environment-იდან

დანაწევრებული კორპუსის (↗ Parsed corpus) კონსტრუქცია მონაცემთა დიდი ბაზების გასაანალიზებლად პირველად ადრეულ 90-იან წლებში გამოიყენეს და წარმოადგენდა კორპუსლინგვისტიკის საეტაპო მიღწევას.

t-test – სტატისტიკური ტესტი, რომლიც გამოიყენება კოლოკაციების დასაანგარიშებლად.

World Wide Web როგორც „კორპუსი“ – WWW-ის კორპუსად გაგების იდეა (Kilgarriff & Grefenstette 2003) ძალიან წააგავს მონიტორული კორპუსის იდეას. აქაც საქმე ეხება მონაცემების მასიური შეგროვებისა და ბაზის მუდმივი ექსპანსიის კორპუსულ ჩარჩოში მოქცევას და მის გამოყენებას ლინგვისტური კვლევისათვის. ამის საუკეთესო მაგალითია ანტონიმების Web-

ზე დაყრდნობით კვლევა სტივენ ჯონსის მიერ (Jones et al. 2007). ყველაზე ცნობილი საძიებო სისტემის google-ის გარდა სპეციალურად Web-ში საძიებლად შეიქმნა დამატებითი ნიუანსირებული საძიებო სისტემები, როგორიცაა, მაგალითად, WebCorp (Renouf 2003). Web-ს, როგორც კორპუსს, გააჩნია რამდენიმე, სხვა კორპუსებისაგან განმასხვავებელი თავისებურება. იგი შედგება ბევრი რედაქტირებულ-კომენტირებული და უფრო მეტი ზერელე ტექსტისაგან. მეტამონაცემების (↗მეტამონაცემი) არქონის გამო ეს ტექსტები არ იძლევიან ავტომატური დაჯგუფების საშუალებას ჟანრების (↗ჟანრი) მიხედვით, რაც ქაოტურ, არარეპრეზენტაბელურ შედეგებს იძლევა. მთავარ პრობლემად კი რჩება ის, რომ Web-ში ძალიან ბევრი შეცდომა ფიქსირდება. ლექსიკონის ამ სტატიის წერის დროს სიტყვისთვის *receive* მივიღეთ 517.000.000 შედეგი მაშინ, როდესაც ამავე სიტყვის მცდარი დაწერილობისათვის *recieve* – 8.940.000, რაც შეცდომისათვის ძალიან მაღალი ციფრია. თუმცაღა თავისთავად ეს ნაკლიც კი საინტერესოა მართლწერის ტენდენციების კვლევის თვალსაზრისით.

24 საათიანი სასწავლო სივრცე – დიდაქტიკური სცენარი თანამედროვე სწავლების მეთოდიკაში, ერთგვარი ვირტუალური სასწავლო სივრცე, რომელიც ღიაა მომხმარებლისათვის 24 საათის განმავლობაში. სწავლების ეს კონცეფცია არ ზღუდავს სასწავლო პროცესს გარკვეულ, წინასწარ დადგენილ და შეთანხმებულ დროსა და სივრცეში. ლექტორის მხრიდან სასწავლო მასალის ელექტრონულ ფორმატში მომზადება და მისი წვდომადობა ინტერნეტის საშუალებით სტუდენტს საშუალებას აძლევს

სასწავლო დროის მენეჯმენტი საკუთარი სურვილისამებრ დაგეგმოს.

ნაწილი II

კორპუსების

შერჩევითი რეგისტრი

Language: Apache
Indication: The University of Virginia Electronic Text Center
Kind: Electronic Texts
Size: Unknown
Link: http://www.scholarslab.org/
Description: An on-line archive of tens of thousands of SGML and XML-encoded electronic texts and images with a library service that offers hardware and software suitable for the creation and analysis of text.

Language: Armenian
Indication: Leiden Armenian Lexical Textbase
Kind: Text and Corpora Meta Sites
Size: 80.000 Armenian lexemes and ten texts
Link: http://www.sd-editions.com/LALT/home.htm
Description: The complete Nor Bargirk, main sections of Adjarian's Root Dictionary, Bedrossian's Armenian-English Dictionary and other material are integrated in LALT. There is a Greek-Armenian lexicon (20000 entries), and aligned Armenian-Greek texts. LALT will be updated at regular intervals. Also, LALT easily is able to integrate additional material and welcomes contributions of other scholars. I have been asked about fonts: LALT is written in xml and uses unicode. Any unicode font will be able to read it, provided this font contains the glyphs (screen images) for Armenian and Greek. One such font is Titus Cyberbit, which is used within LALT itself.

Language: Armenian
Indication: Eastern Armenian National Corpus
Kind: Corpora
Size: 90 million tokens
Link: http://www.eanc.net/
Description: Eastern Armenian National Corpus (EANC) is a comprehensive linguistic database of annotated texts in Standard Eastern Armenian (SEA), the language spoken in the Republic of Armenia. EANC

is: – a comprehensive corpus with about 90 million tokens – a powerful search engine for making complex lexical morphological queries – a learner's corpus including English translations for frequent tokens – a diachronic corpus covering SEA texts from the mid-19th century to the present – a mixed corpus consisting of both written discourse and oral discourse – an open-ended corpus with new texts being added continuously – an annotated corpus with morphological and metatext tagging – an open access corpus – an electronic library with full access to over 100 Armenian classic titles Another important feature is the Glossed output: typologists and language learners can now work with a text format similar to interlinear morphological glosses. In this format, wordforms are supplied with lemmas, lexical and grammatical categories, and translations, vertically aligned below each wordform. Also possible is switching to Latin transliteration from the Armenian alphabet.

Language: Catalan-Valencian-Balear
Indication: Corpus de Català Contemporani de la Universitat de Barcelona (CCCUB)
Kind: Corpora
Size: unknown
Link: http://www.ub.edu/cccub/
Description: Spoken language corpora developed for the study of geographical, functional and socio-cultural variation in Catalan. The texts are in .pdf. The sound files are not yet available through the web, but they have been published in CD-ROM and can be purchased. The CCCUB is also available through RECERCAT (Dipòsit de la Recerca de Catalunya): http://www.recercat.net/handle/2072/8925).

Language: Catalan-Valencian-Balear
Indication: IULA's UPF Textual, plurilingual, specialized Corpus

Kind: Corpora
Size: unknown
Link: http://www.iula.upf.edu/corpus/corpusuk.htm
Description: The main goal of the Corpus project is the construction and exploitation of a textual, plurilingual and specialized corpus. The languages involved are the following: Catalan, Spanish, English, German and French. The areas of interest include: economics, law, computer science, medicine and enviromental science. This corpus is the main support for teaching and research at our institut. Some of the research activities envisaged against this corpus include the following ones: terminology detection, parallel texts alignment, partial parsing, (semi)automatic extraction of several levels of linguistic information for building computational systems (for example, subcategorization patterns), language variation studies.

Language: Chinese Mandarin
Indication: Linguistic Data Consortium
Kind: Text and Corpora Meta Sites
Size: unknown
Link: http://www.ldc.upenn.edu/
Description: Creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes.

Language: Chinese Mandarin
Indication: Chinese Text Project
Kind: Electronic Texts
Size: unknown
Link: http://ctext.org/
Description: The Chinese Text Project is a web-based e-text system designed to present ancient Chinese texts, particularly those relating to Chinese philosophy, in a well-structured and properly cross-referenced manner, making the most of the electronic medium

to aid in their study and understanding.

Language: Chinese Mandarin
Indication: The University of Virginia Electronic Text Center
Kind: Electronic Texts
Size: unknown
Link: http://www.scholarslab.org/
Description: An on-line archive of tens of thousands of SGML and XML-encoded electronic texts and images with a library service that offers hardware and software suitable for the creation and analysis of text.

Language: Chinese Mandarin
Indication: Chinese Gigaword Second Edition
Kind: Corpora
Size: unknown
Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T14
Description: Chinese Gigaword Release Second Edition is a comprehensive archive of newswire text data in Chinese that has been acquired over several years by the LDC. This release includes all of the contents in the first release of the Chinese Gigaword corpus (LDC2003T09), material from one new source, as well as new materials from the other two sources. Thus, the corpus contains three distinct international sources of Chinese newswire – Central News Agency, Taiwan, Xinhua News Agency, and Zaobao. Some minor updates to the documents from the first release have been made.

Language: Croatian
Indication: Croatian National Corpus
Kind: Electronic Texts
Size: unknown
Link: http://www.hnk.ffzg.hr/default.htm
Description: The starting point for each linguistic research is the

corpus. As Croatian does not have a systematically compiled corpus the objective of this project is the compilation and analysis of the representative Croatian texts -- both older and contemporary -- in the form of the corpus the usage of which is applicable for all kinds of Croatistic, lexicographic and lexicological research.

Language: Croatian
Indication: Croatian Language Corpus
Kind: Corpora
Size: the corpus indexes more than 100 k tokens
Link: http://riznica.ihjj.hr/
Description: The Croatian Language Corpus is the result of various projects at the Institute of Croatian Language and Linguistics and the Linguistics Department of the University of Zadar. There is an online interface based on Philologic at the given URL. The current status is that the corpus indexes more than 100 k tokens, and the base is growing continuously. It is annotated in TEI XML P5, its annotation is being enriched with morphological segmentation, lemmatization, phonemic transcription, morphosyntactic annotation and syntactic parses. The online interfaces are subject to change and extension for the improvement of access to various corpus properties.

Language: Szech
Indication: Czech Academic Corpus v. 1.0
Kind: Corpora
Size: 600,000 words
Link: http://ufal.mff.cuni.cz/rest/CAC/cac_10.html
Description: The Czech Academic Corpus version 1.0 is a corpus with a manual morphological annotation of morphology of the Czech language consisting of approximately 600,000 words in continuous texts.

Language: Danish
Indication: Korpus 2000
Kind: Electronic Texts
Size: unknown
Link: http://www.scandinavie-vertalingen.nl/
Description: Translation agency for translations from and into the Scandinavian languagues (Swedish, Finnish, Norwegian, Danish).

Language: Danish
Indication: 16. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages
Kind: Corpora
Size: 49 million words per language
Link: http://langtech.jrc.it/JRC-Acquis.html
Description: The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 22 Languages – New: Version 3.0 almost tripled in size: The JRC-Acquis Version 3.0 is a unique and freely available parallel corpus containing European Union (EU) documents of mostly legal nature. It is available in the 23 official EU languages, with the exception of Irish. The corpus consists of about 23,000 documents per language, with an average size of 49 million words per language, totalling to over one Billion words. Pair-wise paragraph alignment information produced by two different aligners (Vanilla and HunAlign) is currently available for a subset of 8000 documents in 210 language pair combinations. Pair-wise alignment for all texts in all 231 language pairs will be available soon. Most texts have been manually classified according to the EUROVOC subject domains so that the collection can also be used to train and test multi-label classification algorithms and keyword-assignment software. The corpus is encoded in XML, according to the Text Encoding Initiative Guidelines. Due to the large number of parallel texts in many languages, the JRC-Acquis is particularly suitable to carry out all

types of cross-language research, as well as to test and benchmark text analysis software across different languages (for instance for alignment, sentence splitting and term extraction).

Language: Dutch
Indication: Het Corpus Gesproken Nederlands
Kind: Text and Corpora Meta Sites
Size: 900 hours of spoken Dutch
Link: http://tst-centrale.org/
Description: The Corpus Gesproken Nederlands, (Spoken Dutch Corpus), or CGN is a collection of approximately 900 hours of spoken Dutch from Flemish and Dutch speakers. All recordings have been aligned with an orthographic transcription and each word has been given a POS tag and a lemma. Part of the data has been enriched with syntactic, prosodic and/or phonetic information.

Language: Dutch
Indication: IFA Dialog Video corpus
Kind: Text and Corpora Meta Sites
Size: 5 hours of speech
Link: http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/
Description: The IFA Dialog Video corpus is a collection of annotated video recordings of friendly Face-to-Face dialogs licensed under the GNU General Public License (GPLv2). It is modeled on the Face-to-Face dialogs Spoken Dutch Corpus (CGN). The procedures and design of the corpus were adapted to make this corpus useful for other researchers of Dutch speech. For this corpus 20 dialog conversations of 15 minutes were recorded and annotated, in total 5 hours of speech. To stay close to the very useful Face-to-Face dialogs in the CGN, pairs of well acquainted participants, either good friends, relatives, or long-time colleagues were selected. The participants were allowed to talk

about any topic they wanted.

Language: Ega
Indication: Ega XML Lexicon
Kind: Corpora
Size: unknown
Link: http://emeld.org/school/case/ega/lexicon.html
Description: Digitized, online lexicon of Ega, a language of the Ivory Coast as provided by the late Prof. Eddy Aimé Gbery.

Language: English
Indication: CONCISUS Corpus of Event Summaries
Kind: Text and Corpora Meta Sites
Size: unknown
Link: http://www.taln.upf.edu/pages/concisus/index.html
Description: The CONCISUS Corpus is an annotated dataset of comparable Spanish and English event summaries in four application domains. The CONCISUS Corpus covers for the time being the following domains: aviation accidents, train accidents, earthquakes, and terrorist attacks. The dataset contains: comparable summaries, comparable automatic translations, and comparable full documents.

Language: English
Indication: HATII and DCC Release KRYS I Corpus to Aid Research
Kind: Text and Corpora Meta Sites
Size: 6434 documents
Link: http://www.krys-corpus.eu/
Description: The Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow and the Digital Curation Centre (DCC) are delighted to announce the release of the KRYS I Corpus for genre classification research.

http://www.krys-corpus.eu The corpus, consisting of 6434 documents labelled with document genres, is expected to become a major research resource among text processing and data and information management researchers. In particular, we encourage the use of the corpus for the research of: – Automated Text Classification (TC) – Digital curation and metadata extraction – Natural Language Processing (NLP) – Computational Linguistics (CL) Despite the potential of document genre classification as a supporting step in language processing, document management, and information retrieval (e.g. the linguistic style and the vocabulary of a document varies distinctively across document genres), to date, there has been a severe lack of genre-labelled document corpora with which researchers can experiment. It is, therefore, with great pleasure that the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow and the Digital Curation Centre (DCC) makes the KRYS I Corpus available to researchers around the globe. The Corpus originated as part of the ongoing Semantic Metadata Extraction research at the Digital Curation Centre (http://www.dcc.ac.uk) and the HATII at the University of Glasgow (http://www.hatii.arts.gla.ac.uk). The metadata extraction research evolved into a study of automated genre classification, reflecting the observation that the genre of a document (e.g. whether a document is a scientific article or a letter) is characterised by the form and structure of a document, the understanding of which would facilitate further extraction of metadata from within the document. Further details about the development of the KRYS I corpus are available via the website (http://www.krys-corpus.eu). Specifically, researchers will find a detailed account of the document collection process, the reclassification of the documents in the corpus, and

the initial findings with regard to human classification of the documents. We encourage researchers to make full use of this corpus for their own research activity and recommend that you consider contributing towards the ongoing development of the corpus by adding your own documents to the database. Instructions as to how to contribute to the corpus are provided at http://www.krys-corpus.eu. Comments and/or feedback on the KRYS I Corpus are invited. Contacts details can be found on the website. Please feel free to distribute this announcement to any interested colleagues.

Language: English
Indication: Linguistic Data Consortium
Kind: Text and Corpora Meta Sites
Size: unknown
Link: http://www.ldc.upenn.edu/
Description: Creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes.

Language: English
Indication: NPS Chat Corpus
Kind: Text and Corpora Meta Sites
Size: 10,567 posts
Link: http://faculty.nps.edu/cmartell/NPSChat.htm
Description: The NPS Chat Corpus, Release 1.0 consists of 10,567 posts gathered from various online chat services in accordance with their terms of service. The posts have been: 1) Hand privacy masked; 2) Part-of-speech tagged; and 3) Dialogue-act tagged.

Language: English
Indication: Linguistic Data Consortium
Kind: Text and Corpora Meta Sites
Size: unknown

Link: http://www.phon.ox.ac.uk/files/apps/IViE/
Description: An intonationally transcribed corpus covering seven dialects of English from the British Isles. Subjects were secondary school students. The corpus covers short read sentence, a read story, a retold story, map tasks and free conversation.

Language: English
Indication: Alex: A Catalogue of Electronic Texts on the Internet
Kind: Electronic Texts
Size: unknown
Link: http://infomotions.com/alex/
Description: A collection of public domain documents from American and English literature as well as Western philosophy.

Language: English
Indication: Penn-Helsinki Parsed Corpus of Early Modern English
Kind: Electronic Texts
Size: 600,000 words
Link: http://www.ling.upenn.edu/hist-corpora/
Description: The Penn-Helsinki Parsed Corpus of Early Modern English is a 1.8 million word parsed corpus of text samples of Early Modern English. It includes the text samples of the Helsinki Corpus of Historical English, which consists of 600,000 words of genre balanced text and two extension samples of the same size, balanced for genre in the same way. It is a sister corpus of the Penn-Helsinki Parsed Corpus of Middle English and the two corpora are distributed together.

Language: English
Indication: Electronic Texts
Kind: Text and Corpora Meta Sites
Size: over 400 British Library texts

Link: http://www.bl.uk/learning/

Description: The collection of classified, annotated and (partially) downloadable texts from the British Library's collection – good for both teaching and research. Here's the introduction: Texts in Context is a rich and unusual collection of over 400 British Library texts. You can find menus for medieval banquets and handwritten recipes scribbled inside book covers. You can browse the first English dictionary ever written and explore the secret language of the Georgian underworld. You can study the East India Company's shopping lists and practise sentences from colonial phrasebooks. You can learn smugglers' songs, listen to rare dialect recordings, and examine the logbooks of 17th century trading ships.

Language: English
Indication: British National Corpus
Kind: Corpora
Size: 100 million word
Link: http://www.natcorp.ox.ac.uk/

Description: A 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written.
Buckeye Corpus: This corpus contains high-quality recordings of conversational American English speech from 40 speakers in Columbus, OH, USA. The speech has been orthographically transcribed and phonetically labeled. Currently the audio files and transcriptions for 20 talkers are available.

Language: English
Indication: Centre for English Corpus Linguistics
Kind: Corpora
Size: 3.7 million words
Link: http://www.uclouvain.be/en-cecl.html

Description: ICLEv2 contains 3.7 million words of writing from higher intermediate to advanced learners of English representing 16 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, Tswana). It differs from the first version published in 2002 not only by its increased size and range of learner populations, but also by its interface, which contains two new functionalities: built-in concordancer allowing users to search for word forms, lemmas and/or parts-of-speech tags and breakdown of the query results according to the learner profile information. The accompanying ICLEv2 Handbook contains a detailed description of the corpus, a user's manual and an overview of the ELT situation in the countries of origin of the learners. There are three types of licence (for non-profit research purposes only): single user, multiple-user (2-10) and multiple-user (11-25). The corpus can be ordered online at http://www.i6doc.com

Language: English
Indication: Centre for English Corpus Linguistics
Kind: Corpora
Size: unknown
Link: http://www.linguateca.pt/COMPARA/
Description: COMPARA is bi-directional parallel corpus based on an open-ended collection of Portuguese-English and English-Portuguese source-texts and translations. Access is free and requires no registration.

Language: English
Indication: Corpora at ICAME
Kind: Corpora
Size: unknown
Link: http://icame.uib.no

Description: International Computer Archive of Modern and Medieval English.

Language: English
Indication: CSLU: Spelled and Spoken Words
Kind: Corpora
Size: 3647 callers
Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S15
Description: The CSLU: Spelled and Spoken Words corpus consists of spelled and spoken words. 3647 callers were prompted to say and spell their first and last names, to say what city they grew up in and what city they were calling from, and to answer two yes/no questions. In order to collect sufficient instances of each letter, 1371 callers also recited the English alphabet with pauses between the letters. Each call was transcribed by two people, and all differences were resolved. In addition, a subset of 2648 calls has been phonetically labeled.

Language: English
Indication: EF Cambridge Open Language Database
Kind: Corpora
Size: 412,000 scripts
Link: http://corpus.mml.cam.ac.uk/
Description: EFCamDat contains writings submitted to Englishtown, EF's online school, accessed daily by thousands of learners worldwide. The database currently contains 412,000 scripts from 76,000 learners summing up 32 million words. (More information: http://linguistlist.org/issues/24/24-2935.html#1)

Language: English
Indication: English Accents and Dialects

Kind: Corpora
Size: unknown
Link: http://www.bl.uk/cbhasmoved.html
Description: Extracts from the Survey of English Dialects and the Millennium Memory Bank document how we spoke and lived in the 20th century.

Language: English
Indication: International Corpus of English (British Component)
Kind: Corpora
Size: unknown
Link: http://www.ucl.ac.uk/english-usage/projects/ice-gb/index.htm
Description: The British Component of the International Corpus of English (ICE-GB) contains one million words of spoken and written British English. The material is fully tagged and parsed and the associated syntactic treebank is searchable with dedicated exploration software. The spoken material can be listened to.

Language: English
Indication: IULA's UPF Textual, plurilingual, specialized Corpus
Kind: Corpora
Size: unknown
Link: http://www.iula.upf.edu/corpus/corpusuk.htm
Description: The main goal of the Corpus project is the construction and exploitation of a textual, plurilingual and specialized corpus. The languages involved are the following: Catalan, Spanish, English, German and French. The areas of interest include: economics, law, computer science, medicine and enviromental science. This corpus is the main support for teaching and research at our institut. Some of the research activities envisaged against this corpus include the following ones: terminology detection, parallel texts alignment, partial parsing, (semi)automatic extraction of several levels of linguistic information for building

computational systems (for example, subcategorization patterns), language variation studies.

Language: English
Indication: MDE RT04 Training Data Speech
Kind: Corpora
Size: unknown
Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S16
Description: DE RT-04 Training Data Speech was created to provide training data for the RT-04 Fall Metadata Extraction (MDE) Evaluation, part of the DARPA EARS (Efficient, Affordable, Reusable Speech-to-Text) Program. The goal of MDE is to enable technology that can take raw Speech-to-Text output and refine it into forms that are of more use to humans and to downstream automatic processes. In simple terms, this means the creation of automatic transcripts that are maximally readable. This readability might be achieved in a number of ways: flagging non-content words like filled pauses and discourse markers for optional removal; marking sections of disfluent speech; and creating boundaries between natural breakpoints in the flow of speech so that each sentence or other meaningful unit of speech might be presented on a separate line within the resulting transcript. Natural capitalization, punctuation and standardized spelling, plus sensible conventions for representing speaker turns and identity are further elements in the readable transcript. LDC has defined a SimpleMDE annotation task specification and has annotated English telephone and broadcast news data to provide training data for MDE.

Language: English
Indication: N4 NATO Native and Non-Native Speech
Kind: Corpora
Size: unknown
Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S13
Description: The N4 NATO Native and Non-Native Speech corpus was developed by the NATO research group on Speech and Language Technology in order to provide a military oriented database for multilingual and non-native speech processing studies. The NATO Speech and Language Technology group decided to create a corpus geared towards the study of non-native accents. The group chose naval communications as the common task because it naturally includes a great deal of non-native speech and because there were training facilities where data could be collected in several countries. Speech data was recorded in the Naval transmission training centers of four countries (Germany, The Netherlands, United Kingdom, and Canada). The material consists of native and non-native speakers using NATO English procedure between ships and reading from a text.

Language: English
Indication: Penn Parsed Corpora of Historical English
Kind: Corpora
Size: 3.3 million words
Link: http://www.ling.upenn.edu/hist-corpora/
Description: The Penn Parsed Corpora of Historical English are a collection of three annotated corpora of historical British English: the Penn-Helsinki Parsed Corpus of Middle English (1.2 million words), the Penn-Helsinki Parsed Corpus of Early Modern English (1.7 million words) and the Penn Parsed Corpus of Modern British English (currently 1 million words). The corpora are genre-balanced and consist of POS-tagged and syntactically annotated text

samples, including all of the samples in the Middle and Early Modern English sections of Helsinki Corpus of Historical English (1.1 million words).

Language: English
Indication: Penn-Helsinki Parsed Corpus of Early Modern English
Kind: Corpora
Size: 600,000 words
Link: http://www.ling.upenn.edu/hist-corpora/
Description: The Penn-Helsinki Parsed Corpus of Early Modern English is a 1.8 million word parsed corpus of text samples of Early Modern English. It includes the text samples of the Helsinki Corpus of Historical English, which consists of 600,000 words of genre balanced text and two extension samples of the same size, balanced for genre in the same way. It is a sister corpus of the Penn-Helsinki Parsed Corpus of Middle English and the two corpora are distributed together.

Language: English
Indication: SCoSE – Saarbrücken Corpus of Spoken English
Kind: Corpora
Size: unknown
Link: http://www.uni-saarland.de/fak4/norrick/scose.html
Description: The SCoSE consists of five parts: Part 1: Stories Part 2: Indianapolis Interviews Part 3: Jokes Part 4: Complete Conversations Part 5: Drawing Experiment You can download each of the five parts as a .pdf file.

Language: English
Indication: Scottish Corpus of Texts and Speech (SCOTS)
Kind: Corpora
Size: unknown
Link: http://www.scottishcorpus.ac.uk/

Description: SCOTS is an AHRC-funded project, creating a corpus of texts in the languages of Scotland, in the first instance Scots and Scottish English, of all available genres. Spoken texts (orthographic transcription plus accompanying audio/video files) make up 20% of the complete corpus. The corpus is fully searchable online, and the website also contains a description and instructions.

Language: English
Indication: SMULTRON – The Stockholm Multilingual Treebank
Kind: Corpora
Size: 1000 sentences
Link: http://www.ling.su.se/DaLi/research/smultron/index.htm
Description: SMULTRON is a parallel treebank developed by the Computational Linguistics Group at the Department of Linguistics, at Stockholm University. The parallel treebank contains around 1000 sentences each in English, German and Swedish. The sentences have been PoS-tagged and annotated with phrase structure trees. The trees have been aligned on sentence, phrase and word level. Additionally, the German and Swedish monolingual treebanks contain lemma information.

Language: English
Indication: Speech Controlled Computing
Kind: Corpora
Size: unknown
Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S30
Description: The Speech Controlled Computing corpus was designed to support the development of small footprint, embedded ASR applications in the domain of voice control for the home. It consists of the recordings of 125 speakers of American English from four regions, three age groups and

two gender groups, pronouncing isolated words. The recordings were conducted in a sound-attenuated room, and a high-quality microphone was used. Each speaker read a randomized word list consisting of 2100 words (100 distinct words appearing 21 times each). NOTE: Nonmembers may obtain a commercial rights license to Speech Controlled Computing for US$7000 by signing the LDC User License Agreement for Speech Controlled Computing. For-Profit Membership to the LDC is not required.

Language: English
Indication: The Bergen Corpus of London Teenage Language (COLT)
Kind: Corpora
Size: unknown
Link: http://torvald.aksis.uib.no/colt/
Description: The Bergen Corpus of London Teenage Language (COLT) is the first large English Corpus focusing on the speech of teenagers. It was collected in 1993 and consists of the spoken language of 13 to 17-year-old teenagers from different boroughs of London. The complete corpus, half a million words, has been orthographically transcribed and word-class tagged, and is a constituent of the British National Corpus.

Language: English
Indication: Timebank 1.2
Kind: Corpora
Size: 183 news articles
Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08
Description: The TimeBank 1.2 corpus contains 183 news articles that have been annotated with temporal information, adding events, times and temporal links between events and times. The annotation follows the TimeML 1.2.1 specification. The

most recent information on TimeML is always available at www.timeml.org. TimeML aims to capture and represent temporal information. This is accomplished using four primary tag types: TIMEX3 for temporal expressions, EVENT for temporal events, SIGNAL for temporal signals, and LINK for representing relationships. Timebank 1.2 is distributed via web download. Nonmembers may license this data at no cost – please note that a signed copy of our generic nonmember user agreement is required.

Language: English
Indication: VOICE: Vienna-Oxford International Corpus of English
Kind: Corpora
Size: unknown
Link: https://www.univie.ac.at/voice/page/corpus_availability
Description: The Vienna-Oxford International Corpus of English (VOICE) 1.0 Online is available as a free-of-charge resource for non-commercial research purposes. VOICE comprises naturally occurring, non-scripted face-to-face interactions in English as a lingua franca (ELF). The recordings made for VOICE are keyboarded by trained transcribers and stored as a computerized corpus. The speakers recorded in VOICE are experienced ELF speakers from a wide range of first language backgrounds. The ELF interactions recorded cover a range of different speech events in terms of domain (professional, educational, leisure), function (exchanging information, enacting social relationships), and participant roles and relationships (acquainted vs. unacquainted, symmetrical vs. asymmetrical).

Language: English
Indication: Word Frequency Lists and Dictionary for American

English
Kind: Corpora
Size: 400 million word
Link: http://www.wordfrequency.info/
Description: This site contains what we believe are the most accurate and hopefully the most useful word frequency lists of (American) English. Our data is based on the only large, genre-balanced, up-to-date corpus of American English -- the 400 million word Corpus of Contemporary American English. You can be sure that the words in these lists and in this dictionary – sorted from most to least frequent – are really the most common ones that you will encounter in the real world. The frequency data comes in a number of different formats: * An eBook containing up to the 20,000 most frequent words, along with the 20-30 most frequent collocates (nearby words) and the synonyms for each word. * A printed book (from Routledge) with the top 5,000 words (including collocates) and thematic lists. * A free word list -- top 5,000 words, but no collocates or synonyms. * Simple word lists of the top 10,000 or 20,000 words, but without collocates or synonyms. * Lists with the top 200-300 collocates for each of the 20,000 words, for up to 5,000,000 node word / collocate pairs.

Language: Finnish
Indication: Scandinavië Vertalingen
Kind: Corpora
Size: unknown
Link: http://www.scandinavie-vertalingen.nl/
Description: -

Language: Finnish
Indication: 50. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages
Kind: Corpora

Size: unknown
Link: http://ipsc.jrc.ec.europa.eu/index.php?id=198
Description: -

Language: French
Indication: Corpus de français parlé au Québec (CFPQ)
Kind: Corpora
Size: unknown
Link: http://recherche.flsh.usherbrooke.ca/cfpq/
Description: Le Corpus de français parlé au Québec (CFPQ) vise à refléter le français québécois spontané en usage dans les années 2000. Il est susceptible d'aider tout chercheur qui s'intéresse à la variation en français, notamment sous des angles lexicologique, sémantique ou pragmatique. Sa taille actuelle est de sept sous-corpus, ce qui correspond à plus de dix heures de conversations informelles à quatre ou cinq locuteurs. Les informations requises pour une exploitation optimale des données sont disponibles sur le site.

Language: French
Indication: Corpus of Remarks on the French language (17th Century)
Kind: Corpora
Size: unknown
Link: http://www.classiques-garnier.com/numerique-en/index.php?option=com_content&view=article&id=139%3Acorpus-of-remarks-on-the-french-language-17th-century&catid=33%3Acatalogue-bases-dicenc&Itemid=30
Description: The authors of Remarks treat all aspects of usage – pronunciation, spelling, morphology, syntax, vocabulary and style – but drop the traditional format of grammars. This corpus is an indispensable instrument, not only for specialists of 17th century language and literature, but also

for all those interested in the history of the French language, of its codification and standardization. This data-base contains the classic texts (the remarks of Vaugelas, Ménage and Bouhours); collections which adopt an alphabetical presentation (Alemand, Andry de Boisregard) ; texts which criticise Vaugelas and call for greater freedom of usage (Dupleix, La Mothe Le Vayer) ; the volumes which emanate from circles close to the Academy (the Academy's comments on Vaugelas, and its decisions collected by Tallemant), as well as some less prestigious texts (Buffet's observations addressed to a female audience, the compilation by Macé which completes his general and critical grammar). For easy use and exploitation, the Corpus of remarks on the French language is accompanied by a number of research instruments: full-text search, a thesaurus of authors (5 categories) and of titles (3 categories), thesaurus of examples and quotations. The user can constitute his/her own corpus, extract and export results. This set of instruments will promote new research in the fields of the history of the French language and linguistic conceptions.

Language: French
Indication: Corpus TCOF
Kind: Corpora
Size: unknown
Link: http://www.cnrtl.fr/corpus/tcof/
Description: Le projet « Traitement de Corpus Oraux en Français » (TCOF) de l'ATILF (UMR 7118, Université de Lorraine & CNRS) met à disposition de la communauté des corpus oraux alignés texte-son (Transcriber). Le corpus TCOF comporte deux grandes catégories : des enregistrements de corpus d'interactions adultes / enfants (126 enregistrements actuellement) et des enregistrements d'interactions entre adultes (102

enregistrements actuellement). Ce corpus est enrichi régulièrement. L'accès aux données, via le site du CNRTL, est facilité par une interface de recherche qui permet aux utilisateurs de choisir les corpus en fonction de leurs objets d'étude (adultes, enfants, homme, femme, situations professionnelles, genre de discours, etc.).

Language: French
Indication: IULA's UPF Textual, plurilingual, specialized Corpus
Kind: Corpora
Size: unknown
Link: http://www.iula.upf.edu/corpus/corpusuk.htm
Description: The main goal of the Corpus project is the construction and exploitation of a textual, plurilingual and specialized corpus. The languages involved are the following: Catalan, Spanish, English, German and French. The areas of interest include: economics, law, computer science, medicine and enviromental science. This corpus is the main support for teaching and research at our institut. Some of the research activities envisaged against this corpus include the following ones: terminology detection, parallel texts alignment, partial parsing, (semi)automatic extraction of several levels of linguistic information for building computational systems (for example, subcategorization patterns), language variation studies.

Language: Georgian
Indication: Georgian National Corpus
Kind: Corpora
Size: ca. 130 mio. tokens
Link: http://titus.fkidg1.uni-frankfurt.de/armazi/gnc/gnc.htm
Description: International partnership project, supported by the Volkswagen Foundation within the program Between Europe and the Orient - A Focus on

Research and Higher Education in/on Central Asia and the Caucasus. The project, which has been funded by the Volkswagen Foundation since 2012, aims to develop a comprehensive corpus which makes the Georgian language in all its diachronic and synchronic diversity accessible to scientific investigations from various perspectives (linguistics, literary studies, history, political and social sciences etc.).

Language: Georgian
Indication: The Corpus of Georgian Dialects
Kind: Corpora
Size: ca. 1 mio. tokens
Link: http://www.mygeorgia.ge/gdc/
Description: The Georgian Language world is represented by three Kartvelian languages and more than 25 dialects. The project "The Linguistic Portrait of Georgia" is aimed to associate the problem of documentation and researching of the Georgian dialects to the achievements of the corpus linguistics. The corpus is now under development, in which quite vast textual collection is integrated. It involves all the dialectal texts published during the last 100 years, archive material obtained in all the dialectological field expeditions which took place in the second half of the last century, and additionally dialectal texts recorded by the project group in Georgia and its neighboring countries (Azerbaijan, Iran).

Language: German
Indication: Freiburger Anthologie
Kind: Corpora
Size: 1200 poetry
Link: http://freiburger-anthologie.de/
Description: Die 1200 bekanntesten deutschen Gedichte in einer durchsuchbaren Datenbank.

Language: German
Indication: COSMAS Corpus Archive
Kind: Corpora
Size: 1181 Mio words
Link: http://corpora.ids-mannheim.de/ccdb/
Description: The largest German corpus archive, free-of-charge online search in 1181 Mio words of running text (1846 Mio words for invited guests).

Language: German
Indication: dlexDB
Kind: Corpora
Size: 100 million words
Link: http://dlexdb.de/
Description: dlexDB is a new lexical statistical database for German. It is based on the DWDS-Kerncorpus, a balanced collection – over time and text genre – of 100 million words of texts of the 20th century. dlexDB provides frequencies for types, lemmas, syllables, characters, orthographic neighbors and more. These measures are of considerable interest for research in psycholinguistics and psychology (e.g., studies on visual word recognition) as well as for general linguistics and lexicography. During the course of the project, more levels of linguistic representation will be added.

Language: German
Indication: DWDS Corpora and Dictionaries
Kind: Corpora
Size: 100 million words
Link: http://dlexdb.de/
Description: A lexical information system of German, based on very large corpora and dictionaries. It contains the DWDS-Kerncorpus, a balanced collection – over time and text genre – of 100 million words of texts of the 20th century, various newspaper and special

corpora (~650 million words of text publicly available).

Language: German
Indication: German Political Speeches Corpus and Visualization
Kind: Corpora
Size: unknown
Link: http://perso.ens-lyon.fr/adrien.barbaresi/corpora/index.html
Description: This corpus consists of speeches by the German Presidents, Chancellors and a few ministers, all gathered from official sources. It can be freely republished. The two main corpora are released in XML format with metadata. POS-tags will be added to it. There is also a basic visualization tool enabling users to get a first glimpse of the resource.

Language: German
Indication: 60. German Speech Errors
Kind: Corpora
Size: 474 German speech errors
Link: http://staff-www.uni-marburg.de/~wiese/German-errors.html
Description: Collection of 474 German speech errors by Richard Wiese.

Language: German
Indication: IULA's UPF Textual, plurilingual, specialized Corpus
Kind: Corpora
Size: unknown
Link: http://www.iula.upf.edu/corpus/corpusuk.htm
Description: The main goal of the Corpus project is the construction and exploitation of a textual, plurilingual and specialized corpus. The languages involved are the following: Catalan, Spanish, English, German and French. The areas of interest include: economics, law, computer science,

medicine and enviromental science. This corpus is the main support for teaching and research at our institut. Some of the research activities envisaged against this corpus include the following ones: terminology detection, parallel texts alignment, partial parsing, (semi)automatic extraction of several levels of linguistic information for building computational systems (for example, subcategorization patterns), language variation studies.

Language: German
Indication: SMULTRON – The Stockholm Multilingual Treebank
Kind: Corpora
Size: 1000 sentences
Link: http://www.ling.su.se/DaLi/research/smultron/index.htm
Description: SMULTRON is a parallel treebank developed by the Computational Linguistics Group at the Department of Linguistics, at Stockholm University. The parallel treebank contains around 1000 sentences each in English, German and Swedish. The sentences have been PoS-tagged and annotated with phrase structure trees. The trees have been aligned on sentence, phrase and word level. Additionally, the German and Swedish monolingual treebanks contain lemma information.

Language: Greek, Ancient
Indication: Leiden Armenian Lexical Textbase
Kind: Text and Corpora Meta Sites
Size: unknown
Link: http://www.sd-editions.com/LALT/home.html
Description: -

Language: Greek, Modern
Indication: Hellenic National Corpus
Kind: Text and Corpora Meta Sites
Size: 32,000,000 words
Link: http://hnc.ilsp.gr/en/
Description: HNC is a corpus of written Modern Greek texts, available over the Internet, for research use only. It is based on the General Language corpus developed by the Institute of Language and Speech Processing and is fully available on the Internet since 2000. It currently contains about 32,000,000 words of written texts from several media (books, periodicals, newspapers etc.), which belong to different genres (articles, essays, literary works, reports, biographies etc.) and various topics (economy, medicine, leisure, art, human sciences etc.). The HNC users can make the following queries concerning the lexicon, morphology, syntax and usage of Modern Greek: – specific words (e.g. child), – lemmas (e.g. child as a lemma produces every inflected type of the word), – parts of speech and – up to three combinations of all the above, in which users can specify the distance among lexical items (e.g. word + word, lemma + word, lemma + word + word, lemma + part of speech). Users can define their own sub-corpus within the HNC. This sub-corpus may cover one or more media, genres and/or topics and may also be saved for further reference by the users. Query results are presented as whole sentences, within which the query objects are highlighted. Alternatively, concordances of query results are presented, where the query object is centred on the page. Finally, HNC users can make queries concerning word, lemma and/or parts of speech frequencies within the HNC texts. Statistical information about the 100 and 1,000 most frequent words and lemmata in these texts is also available.

Language: Greek, Modern
Indication: Corpus of Greek Texts
Kind: Corpora
Size: 30 million words
Link: http://sek.edu.gr/
Description: The Corpus of Greek Texts (CGT) is now available via an alternative webpage interface at the University of Athens. Access is free of charge, provided that users are registered with a valid e-mail address. The Corpus of Greek Texts (CGT) is the first electronic corpus of Greek texts designed for linguistic research in a wide range of Modern Greek genres. CGT includes 30 million words from spoken and written texts produced between 1990 and 2010. It has been created by co-operation between the Universities of Athens and Cyprus and was funded by the Research Committee of the University of Cyprus (For more info see: www.ucy.ac.cy/sek) and the programme Pythagoras (co-funded by the EU and Greek sources) (For more info see: http://greekcorpora.isll.uoa.gr/gr/Default.aspx).
The alternative webpage interface was funded by the research programme Kapodistrias of the National and Kapodistrian University of Athens (Programme No: 70/4/760, Dionysis Goutsos). The Corpus of Greek Texts (CGT) has as an exclusive aim the scientific linguistic research into Greek through language data. The use of the webpage is strictly restricted for academic purposes and for non-profit exploitation and has as a sole precondition that researchers will inform CGT developers of any output in the form of papers, dissertations, presentations or publications arising from its analysis. For acknowledgments, please quote Γούτσος, Δ. (2003). Σώμα Ελληνικών Κειμένων: Σχεδιασμός και υλοποίηση. Πρακτικά του 6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας, Πανεπιστήμιο Κρήτης, 18-21 Σεπτεμβρίου 2003.

Electronic publication: http://www.philology.uoc.gr/conferences/6thICGL/gr.htm.

Language: Hebrew
Indication: Hebrew Corpus of Arutz7 Newswires
Kind: Corpora
Size: unknown
Link: http://www.mila.cs.technion.ac.il/english/resources/corpora/a7corpus/index.html
Description: The Corpus containing news and articles from Arutz 7 since 2001, which updates daily. Text is available in HTML, plain ascii text, tokenized text in XML format. It is possible to obtain an XML version of the text morphologically annotated (with all possible analyses) and morphologically disambiguated (with the correct morphological analysis in context). Every day, the front page of Arutz 7 is being scanned for updated news and articles and new material is being downloaded. The relevant text is being extracted from the downloaded pages, and then analyzed for document structure (paragraph, sentence and token segmentation). The texts are then being represented in XML. The resources are free but require a username and a password that can be obtained by sending an email to Shlomo Yona <shlomo@cs.haifa.ac.il>.

Language: Hungarian
Indication: The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages
Kind: Corpora
Size: unknown
Link: http://ipsc.jrc.ec.europa.eu/index.php?id=198
Description: -

Language: Italian

Indication: Italian Linguistics
Kind: Text and Corpora Meta Sites
Size: unknown

Link: http://www.sspina.it/
Description: The Information (in Italian) on Italian linguistics and corpora.

Language: Italian

Indication: CLIPS, Corpus of Spoken Italian
Kind: Corpora
Size: about 100 hours of speech

Link: http://www.clips.unina.it/it/
Description: CLIPS is a corpus of spoken Italian, freely available at www.clips.unina.it. The corpus (audio files, annotation and documentation) are fully downloadable from the website via ftp, free for research purposes. CLIPS consists of about 100 hours of speech, equally represented by female and male voices. A section of the corpus is transcribed orthographically, a smaller section has been phonetically labeled. Recordings were made in 15 Italian cities, selected on the basis of linguistic and socio-economic principles of representativeness: Bari, Bergamo, Bologna, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Venezia. For each of the 15 cities different text typologies have been included: a) radio and television broadcasts (news, interviews, talk shows); dialogue (240 dialogues collected using the map task procedure and the "spot the difference" game. In this set: 30 dialogues are phonetically labeled, 90 orthographically transcribed); c) read speech from non professional speakers (20 sentences each,

covering medium-high frequency Italian words); d) speech over the telephone (conversations between 300 speakers and a simulated hotel desk service operator), e) read speech from 20 professional speakers (160 sentences, covering all phonotactic sequences and medium-high frequency Italian words) recorded in an anechoic chamber. Documentation, corpus collection and annotation follow the EAGLES guidelines.

Language: Italian
Indication: CORIS/CODIS – Corpus of Contemporary Written Italian
Kind: Corpora
Size: 130 million words

Link: http://corpora.dslo.unibo.it/
Description: An updated version of CORIS/CODIS, the synchronic corpus of written Italian designed and developed at the University of Bologna, is now accessible online for research purposes. The new version contains around 130 million words and is updated to 2010. The corpus covers a wide range of text varieties, chosen by virtue of their representativeness of contemporary Italian, and ranges from the 1980s to 2010. The following features are now available: – annotation for part-of-speech and lemma; – user-friendly interface; – advanced IMS-CWB query language; The corpus is freely accessible online for research purposes only.

Language: Italian
Indication: Corpus e Lessico di Frequenza dell'Italiano Scritto
Kind: Corpora
Size: 3.150.075 lexical occurrences

Link: http://alphalinguistica.sns.it/BancheDati.htm

Description: CoLFIS (Corpus e Lessico di Frequenza dell'Italiano Scritto) [Corpus and Frequency Lexicon of Written Italian] produced by Pier Marco Bertinetto°, Cristina Burani*, Alessandro Laudanna^*, Lucia Marconi+, Daniela Ratti+, Claudia Rolando+, Anna Maria Thornton§ ° Scuola Normale Superiore, Pisa * Istituto di Scienze e Tecnologie della Cognizione, CNR, Roma ^ Università di Salerno + Istituto di Linguistica Computazionale,Unità Staccata di Genova, CNR, Genova § Università de L'Aquila The reference corpus consists of excerpts from newspapers, magazines and books. It includes 3.150.075 lexical occurrences. The corpus was designed as the best approximation to the Italians' average preferred readings, as mirrored by official statistics. The lexicon consists of two main components: the forms repertoire and the lemmas repertoire. In the latter, all identical forms belonging to different lemmas are disambiguated, while syntagmatic words (such as table's leg) are treated as single entries. The lexical lists (both forms and lemmas) are presently available for free download at http://alphalinguistica.sns.it/BancheDati.htm http://www.istc.cnr.it/material/database/colfis/ They are organized according to a number of possibilities: frequency rank, inverse alphabetical ordering, with or without capital / non-capital distinction, etc. The entire corpus is not yet available. We hope to put it on-line as soon as we obtain the necessary authorizations. The work has been produced with CNR (Consiglio Nazionale delle Ricerche) support. With the help of willing users, this product will hopefully be enriched with further facilities.

Language: Italian

Indication: Database of spoken Italian (BADIP)

Kind: Corpora

Size: 500,000 word

Link: http://badip.uni-graz.at/en/
Description: Contains an online edition of the 500,000 word LIP-Corpus. The edition is being enriched with POS-tags and lemmata, more data are being added continuously. Other corpora of spoken Italian will be included in the database as soon as possible. Access to BADIP is free. The database is part of the LanguageServer of the University of Graz (Austria).

Language: Italian
Indication: Italian Attribution Corpus
Kind: Corpora
Size: 37.000 tokens

Link: http://homepages.inf.ed.ac.uk/s1052974/resources.php
Description: This is a corpus annotated for attribution relations according to an annotation schema developed from the one adopted for the PDTB corpus. It comprises 50 articles drawn from the ISST corpus of Italian. The overall number of tokens is 37.000. Overall, 461 attribution relations are annotated, using MMAX2. The corpus is available for download and research use at: http://homepages.inf.ed.ac.uk/s1052974/resources.php

Language: Japanese
Indication: The University of Virginia Electronic Text Center
Kind: Electronic Texts
Size: unknown

Link: http://www.scholarslab.org/
Description: An on-line archive of tens of thousands of SGML and XML-encoded electronic texts and images with a library service that offers hardware and software suitable for the creation and analysis of text.

Language: Jola-Fonyi

Indication: Linguistic and Folklore materials from the Kujamaat Jóola

Kind: Text and Corpora Meta Sites

Size: unknown

Link: http://people.virginia.edu/~ds8s/Kujamaat-Joola/

Description: The site which will eventually grow to have an extensive collection of Kujamaat linguistic and folklore materials. It currently contains a dictionary (already listed), two folktales (text, translation and sound) and verses from extemporaneous funeral songs (text, sound, translation, commentary).

Language: Karagas

Indication: Tofa Videos and Texts

Kind: Text and Corpora Meta Sites

Size: unknown

Link: http://emeld.org/school/case/tofa/texts.html

Description: The Tofa stories available here were recorded by Dr. K. David Harrison in 2000 and 2001, for a project funded by a grant from Volkswagen-Stiftung.

Language: Kayardild

Indication: Searchable Kayardild Lexicon

Kind: Corpora

Size: unknown

Link: http://emeld.org/school/search/searchlang/searchadvanced.cfm?lang=2172

Description: Searchable lexicon of Kayardild, collected by Dr. Nicholas Evans and hosted by E-MELD.

Language: Korean

Indication: Korean Treebank Annotations Version 2.0

Kind: Electronic Texts
Size: 647 articles

Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T09

Description: The Korean Treebank Annotations Version 2.0 is an extension of the Korean English Treebank Annotations corpus, LDC2002T26 (2002). It is essentially an electronic corpus of Korean texts annotated with morphological and syntactic information. The original texts for the Korean Treebank 2.0 were selected from The Korean Newswire corpus published by LDC, catalog number LDC2000T45, which is a collection of Korean Press Agency news articles from June 2, 1994 to March 20, 2000. Korean Treebank 2.0 is based on the March 2000 portion of the corpus and includes 647 articles. The annotated corpus can find many uses, including training of morphological analyzers, part-of-speech taggers and syntactic parsers.

Language: Korean

Indication: Korean Propbank
Kind: Corpora
Size: Ca. 30.000 annotated predicate tokens

Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T03

Description: Korean Propbank is a semantic annotation of the Korean English Treebank Annotations and Korean Treebank version 2.0. Each verb and adjective occurring in the Treebank has been treated as a semantic predicate and the surrounding text has been annotated for arguments and adjuncts of the predicate. The verbs and adjectives have also been tagged with coarse grained senses. There are two basic components to Korean Propbank: * The Verb Lexicon. A frames file, consisting of one or more frame sets, has been created for each predicate

occurring in the Treebank. These files serve as a reference for the annotators and for users of the data. 2,749 such files have been created. * The Annotation. There are two annotation files. The virginia-verbs.pb file has 9,588 annotated predicate tokens. These predicate tokens include all those occurring in over 54 thousand words of the Korean English Treebank Annotations, totaling ~791 KB of uncompressed data. The newswire-verbs.pb file has 23,707 annotated predicate tokens. These predicate tokens include all those occurring in over 131 thousand words of the Korean Treebank version 2.0.

Language: Latin

Indication: CGL

Kind: Corpora

Size: 14,000 literary quotations

Link: http://htl2.linguist.jussieu.fr:8080/CGL/index.jsp

Description: The corpus of texts known as Grammatici Latini comprises the Latin grammar manuals written between the 2nd and 7th centuries AD and edited by Heinrich Keil in Leipzig, from 1855 to 1880. The corpus has several points of interest: By assembling the main sources, it allows for the reconstruction of the history of ideas in Western linguistics. From the Middle Ages on, these texts (Donatus' and Priscian's artes in particular) were the basis for the later linguistic tradition. The corpus comprises more than 14,000 literary quotations as grammatical examples, a large number of which being fragments (literary, philosophical) of works that are now lost, or passages which one can compare with the direct tradition of extant texts. Some tendencies in late Latin are given prominence, such as the proscription of expressive forms foreign to classical use.

Language: Macedonian
Indication: Digital Archive of the Macedonian Language
Kind: Text and Corpora Meta Sites
Size: unknown

Link: http://damj.manu.edu.mk/
Description: The Digital Archive of the Macedonian Language is a growing collection of digitized, searchable texts in Modern Macedonian from the nineteenth and twentieth centuries and is completely free for anyone who would like to use, search through, and/or download the materials. web address: http://damj.manu.edu.mk Macedonian Academy of Sciences and Arts Research Center for Areal Linguistics Project Coordinator: Prof. Marjan Markovik e-mail: marjan@manu.edu.mk.

Language: Mien, Biao-Jiao
Indication: Searchable Biao Min Lexicon
Kind: Corpora
Size: nearly 3,000 lexical items

Link: http://emeld.org/school/search/searchlang/searchadvanced.cfm?lang=713
Description: The Biao Min Lexicon, housed on the E-MELD site, consists of nearly 3,000 lexical items from Biao Min documentation collected by David Solnit.

Language: Mocovi
Indication: Searchable Mocoví Lexicon
Kind: Corpora
Size: unknown

Link: http://emeld.org/school/search/searchlang/searchadvanced.cfm?lang=3951

Description: Searchable Mocoví Lexicon, based on data provided and collected by Dr. Verónica Grondona.

Language: Mongolian, Halh
Indication: General Corpus of the Modern Mongolian language
Kind: Corpora
Size: 1 155 583 words

Link: http://web-corpora.net/MongolianCorpus/search/?interface_language=en
Description: General Corpus of the Modern Mongolian language (GCML) contains 966 texts, 1 155 583 words. The processor analyzes effectively 97 % of textual word forms which correspond to 76 % word forms from the inputs of the concordance to the GCML.

Language: Norvegian, Bokmal
Indication: Scandinavië Vertalingen
Kind: Corpora
Size: unknown

Link: http://www.scandinavie-vertalingen.nl/
Description: Translation agency for translations from and into the Scandinavian languagues (Swedish, Finnish, Norwegian, Danish).

Language: Ossetic
Indication: Ossetic National Corpus
Kind: Corpora
Size: unknown

Link: http://corpus.ossetic-studies.org/search/index.php?interface_language=en

Description: The Ossetic National Corpus, which contains about 5 million wordforms, is now freely available online. All the texts in the corpus have been automatically annotated and contain English translations of most lexemes. The percentage of annotated wordforms is more than 75%. The corpus supports automatic Latin transliteration of search results.

Language: Persian, Iranian

Indication: Persian Linguistic Database (PLDB)
Kind: Corpora
Size: unknown

Link: http://pldb.ihcs.ac.ir/
Description: This is the first on-line database for the contemporary (Modern) Persian designed and developed by Dr. S. M. Assi at the Institute for Humanities and Cultural Studies (IHCS), Iran. The database contains a huge selected corpora of all varieties of the Modern Persian language in the form of running texts. Some of the texts are annotated with grammatical, pronunciation and lemmatisation tags. A special and powerful software provides different types of search and statistical listing facilities through the whole database or any selective corpus made up of a group of texts. The database is constantly improved and expanded.

Language: Polish

Indication: IPI PAN Corpus of Polish
Kind: Corpora
Size: unknown

Link: http://korpus.pl/
Description: The 2nd edition of the IPI PAN Corpus of Polish, developed at the Institute of Computer Science of the Polish Academy of Sciences (PAS), is available

at the web pages of: – the Institute of Computer Science PAS: http://korpus.pl/en/ – the Institute of Polish Language PAS: http://corpus.ijp-pan.krakow.pl/en/ To the best of our knowledge, this is currently the largest searchable morphosyntactically annotated corpus of Polish available to the public. The whole corpus consists of over 250 million segments (about 200 million orthographic words) and it is not balanced, but a balanced sample of over 30 million segments is also available. These corpora can be directly searched at the above addresses (do read the query syntax cheatsheet at http://korpus.pl/en/cheatsheet/index.html) or downloaded in a binary form to be used with a standalone version of the corpus search engine Poliqarp (announced separately on the 'corpora' list and available from http://korpus.pl/en/).

Language: Portuguese

Indication: Lexicographical Corpus of Portuguese
Kind: Text and Corpora Meta Sites
Size: unknown

Link: http://clp.dlc.ua.pt/inicio.aspx
Description: The Lexicographical Corpus of Portuguese is a database of electronic texts in Portuguese. It contains the electronic transcription and edition of some of the most important dictionaries from the 17th and 18th centuries. The selected texts are generally considered the most important monuments of portuguese dictionary tradition for their dimension, reception and documental value. – Jerónimo Cardoso, Dictionarium iuventuti studiosae (1562, aliás 1551); Dictionarium ex lusitanico in latinum sermonem (1562); Dictionarium Latinolusitanicum (1569/70); Breve dictionarium vocum ecclesiasticarum (1569); De monetis (1569, aliás 1561) – Pedro de Poiares, Diccionario Lusitanico-Latino de Nomes Proprios

(1667) – Bento Pereira, Prosodia Tesouro (1697) – Rafael Bluteau, Vocabulario Portuguez e Latino (1712-1728) – António Franco (F. Pomey), Indiculo Universal (1716) With this project we made available, in a digital format, the complete text of those dictionaries and it's now possible to retrieve indexes off all the words, which are found in them.

Language: Portuguese
Indication: Arquivo Dialetal do CLUP
Kind: Corpora
Size: unknown

Link: http://cl.up.pt/arquivo/
Description: Arquivo Dialetal do CLUP (Dialectal Archive of the Center of Linguistics of the University of Porto) is a database of recordings of European Portuguese collected during the last two decades, spanning both Mainland Portugal and islands. Apart from the recordings themselves, this resource provides detailed maps, exhaustive narrow phonetic and orthographic transcriptions and information about dialectal phenomena. Arquivo Dialetal do CLUP is a growing project, and we welcome any comments, suggestions and contributions.

Language: Portuguese
Indication: COMPARA – Portuguese-English Parallel Corpus
Kind: Corpora
Size: unknown

Link: http://www.linguateca.pt/COMPARA/
Description: COMPARA is bi-directional parallel corpus based on an open-ended collection of Portuguese-English and English-Portuguese source-texts and translations. Access is free and requires no registration.

Language: Portuguese

Indication: CSLU Spoltech Brazilian Portuguese
Kind: Corpora
Size: 2540 utterances

Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S16

Description: The CSLU Spoltech Brazilian Portuguese corpus contains microphone speech from a variety of regions in Brazil with phonetic and orthographic transcriptions. The utterances consist of both read speech (for phonetic coverage) and responses to questions (for spontaneous speech). The corpus contains 477 speakers and 8080 separate utterances. A total of 2540 utterances have been transcribed at the word level (without time alignments), and 5479 utterances have been transcribed at the phoneme level (with time alignments).

Language: Portuguese

Indication: LumaLiDa – Resources for Child Language
Kind: Corpora
Size: unknown

Link: http://ww3.fl.ul.pt//LaboratorioFonetica/LumaLiDa.htm

Description: LumaLiDa is a family of database resources for the study of child language. It includes LumaLiDaOn (the Linguistic Diary of Luma, an European Portuguese Child), LumaLiDaOnLexicon (the lexicon used by the child in LumaLiDaOn, types and tokens), LumaLiDaAudy (transcribed audio files of child speech), and LumaLiDaAudyLexicon (the lexicon used by the child in LumaLiDaAudy, types and tokens).

Language: Patawatomi
Indication: Searchable Potawatomi Lexicon
Kind: Corpora
Size: unknown
Link: http://emeld.org/school/search/searchlang/searchadvanced.cfm?lang=5046

http://ipsc.jrc.ec.europa.eu/index.php?id=198
Description: Online, searchable Potawatomi lexicon, utilizing data provided to the E-MELD School of Best Practices by Dr. Laura Buszard-Welcher.

Language: Russian
Indication: Russian National Corpora
Kind: Corpora
Size: unknown
Link: http://www.ruscorpora.ru/
Description: Searchable Lexicon of the Saliba language, utilizing data provided to the E-MELD School of Best Practices by Nancy Morse.

Language: Scots
Indication: Scottish Corpus of Texts and Speech (SCOTS)
Kind: Corpora
Size: unknown
Link: http://www.scottishcorpus.ac.uk
Description: SCOTS is an AHRC-funded project, creating a corpus of texts in the languages of Scotland, in the first instance Scots and Scottish English, of all available genres. Spoken texts (orthographic transcription plus accompanying audio/video files) make up 20% of the complete corpus. The corpus is fully searchable online, and the website also contains a description and instructions.

Language: Sicilian

Indication: Corpus Artesia – Archivio Testuale del Siciliano Antico

Kind: Corpora

Size: unknown

Link: http://artesia.ovi.cnr.it/

Description: The Artesia Corpus is part of a larger research project, 'Artesia – Archivio testuale del Siciliano Antico' (Text Archive of Ancient Sicilian), a production of the Department of Modern Philology of the University of Catania, in close cooperation with the Centro di Studi Filologici e Linguistici Siciliani, Palermo (http://www.csfls.it). Among the other contributing national research projects, the Opera del Vocabolario Italiano (OVI) provided the software for creating and managing the full-text database. Our aim is to supply a well-structured research tool for the study of Medieval Sicilian (14th-16th centuries) from a Romance perspective and to account for its whole textual production. In particular, Artesia:

- makes accessible and searchable a philologically reliable and periodically up-to-date corpus of literary and non-literary Sicilian texts, from the earliest attestations (14th cent.) to the latest (mid-16th cent.);
- provides a brief, yet scholarly presentation for each author and text;
- documents the individual works by putting them into a historical and critical context, highlighting relationships with and comparing it to different Latin and Romance textual traditions (Catalan, Tuscan, etc.);
- brings a fundamental contribution to realize a Medieval Sicilian Dictionary;
- publishes philological studies and linguistic researches concerning Medieval Sicilian, both in electronic and paper format (see Quaderni di Artesia, Catania: Ed.it, http://www.editpress.it).

Language: Slovak
Indication: Slovak National Corpus
Kind: Corpora
Size: 30 million of words

Link: http://korpus.juls.savba.sk/
Description: Slovak National Corpus is built as a general monolingual corpus, which in the first phase (year 2003) started to compile written texts originated in years 1990 – 2003, containing about 30 million of words with a lemmatisation, morphological and source (bibliographical and style-genre) annotation. During the second phase (up to 2006) the representative span of written texts will be extended to other periods of the contemporary language (1955 – 2005) to the amount of 200 million words and its selected sample will be syntactically annotated. Simultaneously, specific sub-corpora of diachronic and dialectological texts will commence to be built, as well as a terminological and lexicographical database. Slovak National Corpus is provided primarily to lexicographers (dictionary creation), complements grammar and stylistic research (grammar and orthographical handbooks; varieties of the national language and their usage in communication). We suppose that it will also find its use at schools (preparing of orthography, grammar and style textbooks; teaching Slovak as a foreign language). Specific sub-corpora of historical and dialectological texts will help to preserve an important part of our cultural heritage in a long-term perspective.

Language: Spanish
Indication: Linguistic Data Consortium
Kind: Text and Corpora Meta Sites
Size: unknown

Link: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006S16

Description: Creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes

Language: Spanish

Indication: CONCISUS Corpus of Event Summaries

Kind: Text and Corpora Meta Sites

Size: unknown

Link: http://www.taln.upf.edu/pages/concisus/index.html

Description: The CONCISUS Corpus is an annotated dataset of comparable Spanish and English event summaries in four application domains. The CONCISUS Corpus covers for the time being the following domains: aviation accidents, train accidents, earthquakes, and terrorist attacks. The dataset contains: comparable summaries, comparable automatic translations, and comparable full documents.

Language: Spanish

Indication: El Grial Corpus of Spanish

Kind: Text and Corpora Meta Sites

Size: 100 million words

Link: http://www.elv.cl/prontus_linguistica/site/edic/base/port/grial.htm

Description: El Grial Corpus of Spanish (www.elgrial.cl) is a growing collection of eight corpora (almost 100 million words) with approximately 700 documents of contemporary Spanish, developed by the members of the Escuela Lingüística de Valparaíso (www.linguistica.cl) at the Pontificia Universidad Católica de Valparaíso, Chile. Also, there is a tagger and parser for Spanish Language available on the web site. These corpora have been collected under

specific methodological principles, identifying specialized/non-specialized, written/spoken registers and text types (academic, professional, technical, etc). Detailed description of each corpus is available in the web site. All documents have been tagged and parsed. Part of the data has been enriched with deep syntactic information. To the best of our knowledge, this is currently the largest searchable morphosyntactically annotated and register-diversified corpus of Spanish available to the public, with online tools that help analyze the collected data. Users can define their corpus of study and search the data using a wide variety of resources. Query results are presented in different formats, depending on the kind of research questions. El Grial users can make queries concerning word, lemma and/or parts of speech frequencies. One of the last tools developed is El Manchador de Textos, an online resource that "spots" and puts color to the words or sequences under study; statistical information about the co-occurrences is also available.

Language: Spanish
Indication: Sociolingüística Andaluza
Kind: Corpora
Size: unknown
Link: http://www.sociolinguisticaandaluza.us.es/
Description: Grupo de investigación en sociolingüística. Universidad de Sevilla.

Language: Spanish
Indication: Sociolingüística Andaluza
Kind: Corpora
Size: 500,000
Link: http://clic.ub.edu/corpus/ancora

Description: AnCora: Syntactically and Semantically Annotated Corpora (Spanish, Catalan) CLiC (Centre for Language and Computation) of the University of Barcelona, together with the Natural Language Processing group of the Polytechnic University of Catalonia, have created two new language technology resources: AnCora-Esp for Spanish and AnCora-Cat for Catalan, consisting of 500,000 words each. They are two treebanks enriched with different kinds of semantic information: 1) each function has its argument and thematic role; 2) each verb belongs to a semantic class according to its event structure and diathesis alternations; 3) each noun has its WordNet sense; and 4) each named entity (i.e. persons, organisations, locations, dates, etc.) is identified and categorized. The annotation process has also resulted in two verbal lexicons with approximately 2,000 entries for each language with information about verbal semantic classes and their syntactic subcategorization, their argument structure and the thematic roles for each sense. The AnCora corpora as well as the derived verbal lexicons (AnCora-Verb) are freely available (queries and downloads) from: http://clic.ub.edu/ancora/.

Language: Spanish
Indication: Sociolingüística Andaluza
Kind: Corpora
Size: unknown
Link: http://www.sociolinguisticaandaluza.us.es/
Description: Grupo de investigación en sociolingüística. Universidad de Sevilla.

Language: Spanish
Indication: Corpus del español
Kind: Corpora
Size: 100 million words

Link: http://www.corpusdelespanol.org/
Description: An online, searchable corpus of diachronic Spanish texts (100 million words, 13th century to present).

Language: Sumerian
Indication: The Sumerian Text Archive
Kind: Electronic Texts
Size: unknown
Link: http://www.bibliotheek.leidenuniv.nl/
Description: A growing collection of texts in the Sumerian language.

Language: Swedish
Indication: Scandinavië Vertalingen
Kind: Corpora
Size: unknown
Link: http://www.scandinavie-vertalingen.nl/
Description: Translation agency for translations from and into the Scandinavian languagues (Swedish, Finnish, Norwegian, Danish).

Language: Swedish
Indication: SMULTRON – The Stockholm Multilingual Treebank
Kind: Corpora
Size: 1000 sentences
Link: http://www.ling.su.se/DaLi/research/smultron/index.htm
Description: SMULTRON is a parallel treebank developed by the Computational Linguistics Group at the Department of Linguistics, at Stockholm University. The parallel treebank contains around 1000 sentences each in English, German and Swedish. The sentences have been PoS-tagged and annotated with phrase structure trees. The trees have been aligned on sentence, phrase and word level. Additionally, the German and Swedish

monolingual treebanks contain lemma information.

Language: Tamil
Indication: Digital Tamil Literature
Kind: Corpora
Size: unknown
Link: http://www.sanskrit-lexicon.uni-koeln.de/scans/MWScan/tamil/index.html
Description: Searchable Tamil Digital Text Archive.

Language: Tanaina
Indication: Online Dena'ina Qenaga Lexicon
Kind: Corpora
Size: unknown
Link: http://www.sanskrit-lexicon.uni-koeln.de/scans/MWScan/tamil/index.html
Description: Searchable online wordlist of Dena'ina Qenaga (Tanaina).

Language: Tu
Indication: Monguor – Online Texts
Kind: Corpora
Size: unknown
Link: http://emeld.org/school/case/monguor/texts.html
Description: Digitized online texts of Monguor, an endangered language spoken in the People's Republic of China, as provided to the E-MELD School of Best Practices by Dr. Wang Xianzhen.

Language: Turkish
Indication: TS Corpus
Kind: Corpora
Size: 491 million POSTagged tokens

Link: http://tscorpus.com/en

Description: TS Corpus is a Turkish Corpus project. TS Corpus is a general-purpose corpus containing 491 million POSTagged tokens (491,360,398 million). TS Corpus is a tagged corpus. TS Corpus aims to combine former Turkish computational linguistics studies and other corpus linguistics studies from around the world.

Literatur:

http://www.ugr.es/~jsantana/modelos/READINGS/R03_Types_of_Corpora_McEnery-Xiao-Tono-ChA7.pdf