FAKULTETI I FILOLOGJISË – PRISHTINË
FACULTY OF PHILOLOGY – PRISTINA

FAKULTETI HISTORI-FILOLOGJI – TIRANË
FACULTY OF HISTORY-PHILOLOGY – TIRANA

SEMINARI NDËRKOMBËTAR PËR GJUHËN, LETËRSINË DHE KULTURËN SHQIPTARE

INTERNATIONAL SEMINAR FOR ALBANIAN LANGUAGE, LITERATURE AND CULTURE

REVISTË / JOURNAL

36

ISSN 2521-3687

PRISHTINË, 36/2017

GJUHËSI / LINGUISTICS

Besim KABASHI

AlCo – NJË KORPUS TEKSTESH I GJUHËS SHQIPE ME NJËQIND MILIONË FJALË

Pse një korpus gjuhe?

Pothuajse është e pamundur të bëhen studime serioze në fushën e gjuhëve natyrore pa u bazuar në të dhëna empirike. Si mjet ndihmës që ofron këto të dhëna duke i përmbajtur ato në vete të paprekura, pra të pandryshuara, janë korpuset e materialeve gjuhësore. Ndonëse në raste të caktuara një korpus i tillë mund edhe të mos e përmbajë çdo fjalë(formë) që ekziston apo që është përfshirë në një fjalor të madh evidence, drejtshkrimor apo shpjegues, ai ofron të dhëna të llojeve të ndryshme për fenomene gjegjëse gjuhësore për secilën njësi, si dhe veçoritë ku ato janë pjesëmarrëse si p. sh. konteksti i tyre i përdorimit. Me rritjen e numrit të teksteve të korpusit, d. m. th. fjalëve, rritet edhe gjasa që fjalët e papërfshira më parë në të të gjenden në korpus. Ndër të tjera, korpuset mundësojnë edhe studime cilësore si dhe përshpejtojnë njëkohësisht punën e përdoruesve të tyre, për më tepër i largojnë ata nga të dhëna joempirike apo të konstruktuara e përshtatura sipas rastit e aty për aty. Që të shfrytëzohen këto të mira që ofron një korpus, më parë duhet krijuar një resurs i tillë, një punë jo e thjeshtë.

Fillimi i ndërtimit të korpusit AlCo – pak histori

Puna e cila ka rezultuar me ndërtimin e korpusit AlCo (Albanian [text and speech] Corpus), e cila vazhdon edhe sot e këtyre ditëve, ka filluar me një punim të vogël në universitet në vitin 1997. Qëllimi atë kohë ka qenë krahasimi i disa llojeve të fjalive të shqipes me ato përkatëse të gjermanishtes. Për të mbuluar sa më mirë në pikëpamje cilësore këtë temë duhej mbledhur material gjuhësor, d. m. th. tekste në sasi të mjaftueshme e në lloje të shumta për aq sa ishte e mundur teknikisht atë kohë. Pas kësaj, me rritjen e vazhdueshme të numrit të teksteve të gjuhës shqipe në formë elektronike, ka vazhduar edhe puna për të grumbulluar sa më shumë tekste që të jetë e mundur, gjithnjë me një dëshirë të pandërprerë për të ndërtuar një korpus të vogël tekstesh, i cili do të kishte cilësinë e duhur që të mund të merrej për bazë për studime

gjuhësore. Caku i parë ka qenë një milion fjalë teksti. Aq fjalë teksti përmbajnë disa nga korpuset e para elektronike, khs. p. sh. për anglishten "Korpusin [e Universitetit] Brown" (Brown Corpus, 1963)[1] e për gjermanishten korpusin "LiMaS" (LiMaS-Korpus, 1971)[2]. Me këtë sasi teksti të korpusit AlCo u arrit edhe një njësi matëse e shqipes me anglishten a gjermanishten, në përgjithësi me gjuhët që kanë një korpus të tillë tekstesh. Por ky cak i arritur në vitin 2002 u bë kalimtar për t'u harruar shumë shpejt meqë mbledhja e teksteve vazhdonte. Një nxitje për të vazhduar punën ishte gjenerata e dytë e korpuseve gjuhësore, po ashtu e arritur së pari në hapësirën anglisht-folëse, me njëqind milionë fjalë teksti, khs. p. sh. BNC (British National Corpus).[3] Dëshira për këtë cak të ri, njëqindfishi i cakut të parë, përbën padyshim një dimension krejtësisht tjetër për korpusin. Kjo sasi tekstesh është e lidhur me vëllimin e punës, sidomos për t'i përzgjedhur dhe bashkërenditur ato. Së pari, pas shumë vitesh, në mes të vitit 2017 kjo sasi tekstesh u mblodh. Në këtë moment korpusi ka arritur *versionin beta*, d. m. th. puna nuk ka përfunduar, por është material bruto, i cili do edhe shumë punë. Sidomos përzgjedhje, klasifikimi dhe baraspeshimi i materialit gjuhësor marrin shumë kohë e mund, por do edhe shumë njohuri të fushës së gjuhësisë së korpuseve. Materiali gjuhësor mund të përdoret, por ende mungojnë shumë çelësa të përzgjedhjes. Për ta përdorur korpusin në këto raste kërkohen njohuri shtesë në teknologji kërkimi e gjuhësore.

Materiali gjuhësor: tekstet

Në fillim mbledhja e teksteve ka qenë pak a shumë e kufizuar. Tekstet e para kanë qenë tekste të përzgjedhura, kryesisht lajmesh, të ofruara për publikun nga Qendra për Informim e Kosovës (njohur si QIK). Atyre teksteve iu janë shtuar kohë pas kohe tekste të skanuara e të cilat janë korrigjuar me dorë dhe krahasuar me origjinalin e shtypur meqë njohja e tekstit shqip përmes programeve OCR (Optical Caracter Recognition), pra programeve që njohin tekstin, ishte jocilësore, d. m. th. pas skanimit teksti elektronik përmbante shumë gabime dhe duhej korrigjuar me dorë. Në të njëjtën kohë paralelisht, me rritjen e sasisë se teksteve elektronike të pranishme në internet, është rritur edhe mundësia e një përzgjedhjeje më cilësore e më të madhe të teksteve të reja për korpusin. Për shembull, tani tekstet e lajmeve merren kryesisht nga Agjencia Telegrafike e Shqipërisë (ATSH, *https://www.ata.gov.al)*. Një pjesë e madhe e

[1] Shih. [FRANCIS/KUČERA 1964].
[2] Shih. [GLAS 1975].
[3] Shih. [BNC-XML 2007] dhe *http://www.natcorp.ox.ac.uk/* dhe *http://ota.ox.ac.uk/desc/2554/* (XML edition).

teksteve është mbledhur, përmes lidhjeve private, nga dashamirë të gjuhës shqipe, studiues albanologë, shkrimtarë, autorë tekstesh të fushave të ndryshme – pra, nga shumë burime.

Llojet e teksteve të përzgjedhura për korpusin janë të ndryshme në shumë pikëpamje. Kjo çështje është e papërmbyllur, pasi ende vazhdimisht tekste shtohen, hiqen përkatësisht zëvendësohen, varësisht nga mosbarazpesha e teksteve ndër vete ose, për më tepër, mungesa e ndonjë lloji tekstesh. Një tekst i cilësisë së mirë, apo i një cilësie të veçantë zëvendëson një tekst të korpusit që nuk i ka këto veçori. Këta parametra përcaktohen nga vlerësimi i përgjithshëm e tërësor për korpusin. Qëllimi këtu është që korpusi të jetë sa më cilësor që të jetë e mundur. Llojet (domenet) e teksteve janë nga fusha të ndryshme si *ekonomia*, *mjekësia*, *politika*, *temat shoqërore*, *arti e kultura* etj. të ndara në nënfusha të veçanta, momentalisht, gjithsej 55. Ky numër, me gjetjen e teksteve të reja, mund të shkojë deri në rreth 80. Pos kësaj ndarjeje, materiali gjuhësor është i përzgjedhur edhe në bazë të shumë kushteve të tjera, p. sh. tekstet (e një lloji/kategorie) nga autorë të ndryshëm kanë përparësi ndaj atyre të vetëm një autori. Pastaj, kushtet tjera janë të tilla si *gjinia e autorit*, *mosha*, *prejardhja*, por dhe veti të tjera rreth tekstit për aq sa është e mundur. Pra korpusi AlCo në këtë moment nuk është i përfunduar sa i përket përzgjedhjes dhe bashkërenditjes së teksteve si dhe llojeve të tyre, përkatësisht numrit të teksteve brenda një lloji.

Numri aktual i teksteve është 48.181, i fjalëve 100.120.118, me rreth njëqind milionë. Kështu i bie që mesatarisht një tekst të këtë përafërsisht 2.078 fjalë. Por ky numër është relativ në këtë moment. Nga këto tekste, ka të tilla që kanë vetëm dy fjali, p. sh. një tekst për motin/kohën ka 17 fjalë teksti, e të tillë që janë të gjatë, p. sh. një tekst letrar ka 10154 fjalë teksti.

Pjesa e gjuhës së folur është shumë e vogël dhe kjo duhet ndryshuar për të arritur të paktën te 10 % e numrit të përgjithshëm të fjalëve, pra 10 milionë fjalë. Deri më tani ka vetëm 21 tekste të shkurtra të përfituara/përkthyera nga gjuha e folur.

Anotimi i teksteve të korpusit AlCo

Ndërtimi i një korpusi arrihet kur të jenë mbledhur tekstet dhe të jenë bashkërenditur e klasifikuar ato. Materiali gjuhësor, pra *të dhënat burimore*, angl. *raw data*, mundësojnë që të dhënat të përdoren për qëllime dhe detyra të ndryshme studimi. Kjo gjendje e korpusit, parë nga këndvështrimi i përdorimit, nuk është optimale. P. sh. nëse dikush do të kërkojë fjalën *dhe* në korpus do të merr si rezultat kërkimi fjalën *dhe*, ndër të tjera, si lidhëz (*dhe*), por edhe si emër (*dhé*). Dhe nëse dikujt i intereson fjala *dhe* vetëm si emër, atij i duhet që të gjithë emrat t'i ndajë/dallojë nga

lidhëzat e nga rastet tjera të mundshme. Kjo gjë mund të marrë shumë kohë e mund, aq më tepër, kur duhet të nxirren edhe statistika.

Që të tejkalohet kjo, është e nevojshme që materiali gjuhësor i korpusit të përpunohet. Nëse secilës fjalë i shtohet informacioni i pjesës së ligjëratës, angl. *Parts-of-speech*, shkurt *POS*, në rastin e përmendur, fjalës *dhe*, kërkimi do të ishte (më) i saktë, po ashtu edhe rezultatet e tij. Do të kërkohej fjala *dhe* si emër, dhe të gjitha ato raste me cilësinë e lidhëzës do të anashkaloheshin. Edhe më mirë do të ishte sikur secila fjalë të kishte edhe të dhëna tjera, p. sh. ato morfologjike, si numri, rasa, shquarsia te emrat ose veta, koha, etj. te foljet. Kështu mund të kërkohej një një rast të caktuar p. sh. një emër vetëm në rasën emërore.

Për të bërë këtë, d. m. th. *analizën morfologjike*, tanimë ekziston një gramatikë/program, khs. [KABASHI 2015]. Analiza morfologjike mundëson qasje precize dhe të detajuar në veçoritë e një fjale/fjalëforme. Një shembull është analiza morfologjike e fjalëformës *ishte* e cila është ishte jam+V+3P+Sg+Ind+Impf+Act+NonAdm. Secila pjesë e informacionit (e ndarë me shenjën +) mund të përdoret si qelës kërkimi, pra nëse do të donim format e vetës së tretë do t'i gjenim duke përdorur si qelës kërkimi 3P. Ato mund të specifikohen me të dhëna të tjera të kombinuara, si p. sh. 3P Sg ose 3P Sg Impf. Ky kombinim informacioni mundëson qasje të gjithanshme duke dhënë një liri kërkimi në materialin gjuhësor. Për më tepër, ky informacion mund të kombinohet me informacionin tjetër për të përfituar një sinergji nga i gjithë informacioni i përdorshëm për qëllime emërtimi/shenjimi/-etiketimi/tagimi/anotimi gjuhësor.

Pos informacionit morfologjik, tashmë shumë i përhapur dhe po ashtu i dëshiruar është edhe informacioni morfo-sintaksor – për shkak se një fjalë apo formë e saj varësisht nga pozita në fjali mund të këtë kuptime/interpretime të ndryshme. Për shembull, fjalëforma *mund*, e cila mund të jetë folje modale ose emër i pashquar. Një emër/shenjë/etiketë/tag/notë morfo-sintaksor(e) do të mundësonte dallimin mes përdorimeve të ndryshme. Për gjuhën shqipe, kjo është e rëndësishme edhe për faktin se në shumë raste fjalët e reja formohen nga bashkimi i dy a më shumë fjalëve themelore, si p. sh. *i mirë* (← *i* + *mirë*) vs. *mirë*.

Për këtë qëllim është zhvilluar edhe një set, d. m. th. bashkësi e tagëve, angl. *tagset*, morfo-sintaksorë, khs. këtu [KABASHI/PROISL 2016]. Emrat/shenjat/-etiketat/tagët/notat janë një lloj alfabeti i fushës përkatëse, i cili duhet të mbulojë çdo veçori e funksion të saj, d. m. th. secili emër mbulon veçoritë dhe funksionin e të emërtuarit të tij. Rreth 32 000 fjalë teksti janë emërtuar/shenjuar/etiketuar/taguar/-anotuar me dorë për të bërë një korpus i cili shërben si shembull, apo ndryshe i quajtur *standardi i artë*, angl. *gold standard*. Me këtë janë testuar në mënyra të ndryshme

dhe përmirësuar emrat/shenjat/etiketat/tagët/notat deri sa rezultatet e testimit kanë arritur nivelet e pranueshme a të dëshirueshme. Me këtë korpus të emërtuar/-shenjuar/etiketuar/taguar/anotuar me dorë është trajnuar një program emërtues/-shenjues/etiketues/tagues/anotues, angl. *tagger*, i cili pastaj është përdorur për të emërtuar/shenjuar/etiketuar/taguar/anotuar në mënyrë automatike pjesën tjetër të korpusit, khs. për imtësi [KABASHI/PROISL 2016].

Përdorimi i korpusit

Për të përdorur një korpus ka mundësi dhe mënyra të ndryshme. Në disa raste duhen njohuri paraprake në lëmin e informatikës, gjë që jo secili që merret me studime gjuhësore i ka. Për të mbushur këtë zbrazëti janë zhvilluar programe të cilat mundësojnë kërkimin dhe gjurmimin e të dhënave të një korpusi. Një nga programet e tilla është CQPweb, khs. këtu [HARDIE 2012]. Për korpusin AlCo është vendosur të përdoret programi i lartcekur, i cili ofron mundësi të mjaftueshme për të nxjerrë të dhënat e dëshirueshme, shih. pamjet 1 dhe 2.[4] CQPweb i përngjan shumë programit BNCweb, khs. këtu [HOFFMANN ET AL. 2008], dhe si i tillë është i mirënjohur në shumë qarqe gjuhësore.

Kërkimi i thjeshtë: kërkimi i një fjale, vargu fjalësh apo një fjalie

Në rastet më të shpeshta gjatë punës (së pari) kërkohet një fjalë për të parë p. sh. se a është në korpus, si është përdorur ajo ose për të parë dendurinë e përdorimit të saj. Ky funksion, siç edhe pritet nga një program i tillë, është themelor dhe pothuajse i pakalueshëm. Kërkimi mund të bëhet duke dhënë fjalë(formë)n e plotë ose duke përdorur (edhe) shprehjet e rregullta, angl. *regular expressions*, p. sh. mund të kërkohet pikërisht në këtë formë, siç është shkruar, fjala *lexon,* ose në formën *lexo** (si shprehje e rregullt) për të gjetur të gjitha trajtat e mundshme që fillojnë me *lexo*, pra *lexo*, *lexoj*, *lexon*, *lexoni*, *lexojnë*, e. k. m. r. Rezultat i një kërkimi është paraqitja e radhitur e gjetjeve në korpus. Fjalët janë të paraqitura bashkë me kontekstin e tyre, khs. pamjen 2. Pos kërkimit të një fjale/fjalëforme të vetme, mund të kërkohen edhe fjalë/fjalëforma njëra pas tjetrës, pra në varg, deri te një fjali e tërë ose edhe më tej.

[4] Në këtë artikull nuk ka hapësirë për t'i përshkruar të gjitha funksionet e programeve të sapo-përmenduara – as që është qëllim i këtij artikulli. Për këtë arsye këshillojmë lexuesit që në rast nevoje të përdorin literaturën e dhënë më sipër.

Kërkimi i zgjeruar: kërkimi i bazuar në veçori të fjalëave

Në shumë raste një kërkim si ai i përshkruar më lart është i papërshtatshëm apo i pamjaftueshëm. Po të kujtojmë këtu shumëkuptimësinë (ambiguitetin) e shumë fjalëve sidomos fjalëformave, si p. sh. *hap* (folje) – *hap* (emër), atëherë në rast të një kërkimi të foljes *hap/-a/-ur,* për studime gjuhësore, nuk ka nevojë që të shqyrtohet emri *hap/-i/-a/-at*. Për të mundësuar këtë duhet që të dhënat burimore (vetë fjalët) të jenë të anotuara, d. m. th. të analizuara.

Me këtë shprehje [pos="PCl"] [pos="V" & word=".+oj"] kërkohet përkatësisht gjendet një trajtë e shkurtër e përemrit vetor (PCl) dhe një folje (V) që mbaron me *oj* si p. sh. *i/e* dhe *kujtoj/lejoj*, pra një vargu me dy fjalë si *i kujtoj*.[5]

Në rastet ku p. sh. fjala (pjesa e ligjëratës) *e* duhet kërkuar duke dalluar kontekstin e saj të përdorimit, janë të nevojshme shprehjet si [pos= "Art" & word = "e"], [pos="ConjC" & word="e"] ose [pos="PCl" & word="e"]. Shprehja e parë gjen fjalën *e*, kur ajo ka funksionin e nyjës (Art). E dyta gjen atë, kur ajo është lidhëz bashkërenditëse (ConjC), ndërsa e treta, kur kemi të bëjmë me trajtë të shkurtër të përemrit vetor (PCl).

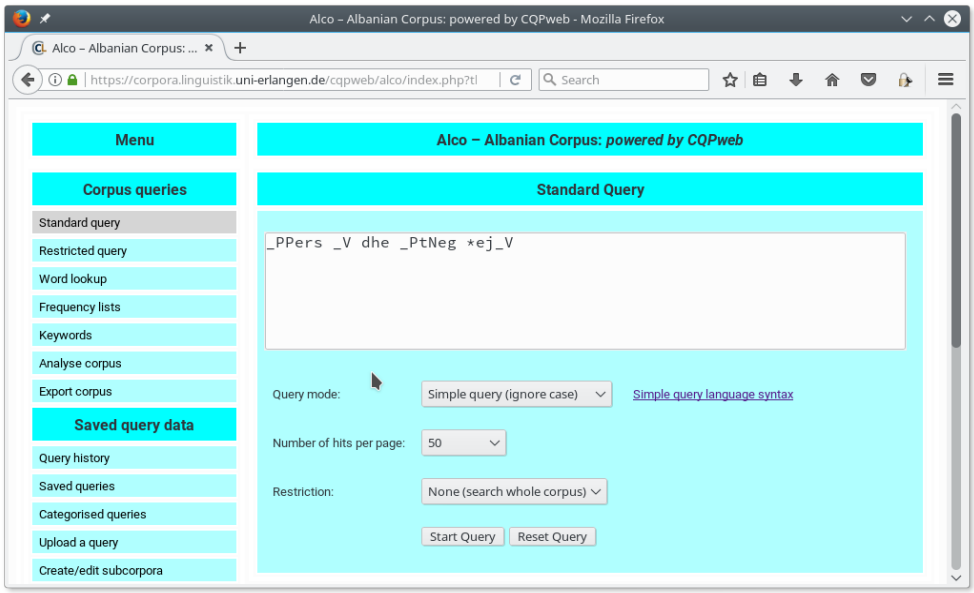
Kërkimi i një vargu fjalëformash përkatësisht një konstruksioni sintaksor bazuar në tagset

Shprehja mos _PCl _V _N mundëson gjetjen e një vargu apo një konstruksioni sintaksor, të përbërë nga fjala *mos*, një trajtë e shkurtër e përemrit vetor (PCl), një folje (V) dhe një emër (N). Gjatësia e vargut dhe forma/përbërja e tij është pothuajse e pakufizuar. Kjo do të thotë se mund të kërkohen fjali të tëra, të çfarëdo lloji. Kjo mënyrë e përgjithësuar kërkimi (bazuar në pjesë të ligjëratës e më gjerë/imët, në 77 emra/shenja/etiketa/tagë/nota) është e mundur pasi korpusi AlCo është i emërtuar/-shenjuar/etiketuar/taguar/anotuar.

Kërkimi është mjaft imtësor dhe si i tillë mundëson gjetjen e njësive të ndryshme morfologjike e sintaksore, këto të fundit duke bashkërenditur njësitë elementare. Për shembull, mund të kërkohet për një konstruksion përkatësisht kompleks foljor si _PtNeg _PCl _VAux _Vpart, i cili nxjerr nga korpusi rezultate si *nuk i keni parë* apo *nuk e kishte parë*. Shprehja _PtProh _PCl _V _N mundëson gjetjen e

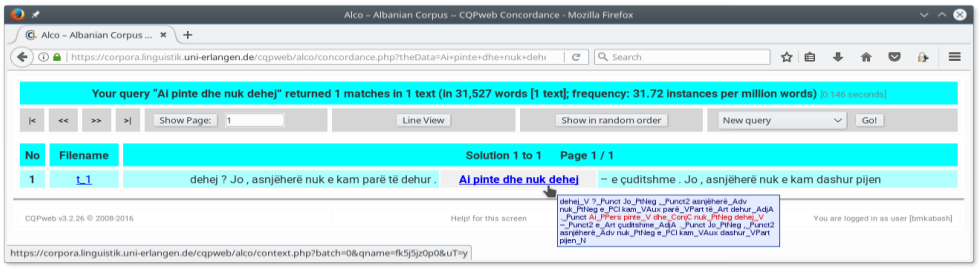
[5] Për të mos e ngarkuar artikullin me pamje (figura) të shkëputura nga AlCo/CQPweb, në vazhdim ato vetëm do të përshkruhen. Lexuesi mund t'i paramendojë/pasqyrojë ato si në figurat 1 dhe 2.

shembujve *mos e lëshoni vetën* ose *mos i harroni premtimet*, ndërsa fundi i _N nxjerr rezultate si *fundi i dimrit* ose *fundi i fundit.*



Pamja 1: Kërkimi i vargut *_PPers _V dhe _PtNeg *ej_V* në korpusin AlCo me CQPweb.

Në pamjen 1 janë ndërthurur mënyra të ndryshme kërkimi, emrat/shenjat/-etiketat/tagët/notat (khs. *_PPers*, *_V*, *_PtNeg* dhe *_V*), fjalët e thejshta (khs. *dhe*), si dhe shprehjet e rregullta (khs. **ej*). Rezultati i këtij kërkimi është paraqitur në pamjen 2, khs. fjalinë (e nënvizuar) dhe analizën (e theksuar me ngjyrë të kuqe) dhënë pranë.



Pamja 2: Rezulatati i kërkimit nga pamja 1 në korpusin AlCo me CQPweb.

Shënime për gjendjen e tashme në fushën e gjuhësisë së korpuseve për shqipen

Pos korpusit të paraqitur në këtë punim, për dijeninë tonë, ekzistojnë edhe dy korpuse tjera, ai nga Caka & Caka [2012], dhe korpusi i gjuhës shqipe i Universitetit Shtetëror të Shën-Petërburgut (në Federatën Ruse) me emrin *Korpusi Kombëtar i Shqipes*, angl. *Albanian National Corpus*, khs. këtu [ARKHANGELSKIJ ET AL. 2012]. I dyti përfshin rreth një milion (ndërtuar sipas modelit *Brown Corpus*), ndërsa i treti rreth 16 milionë fjalë teksti. I dyti është në procesin e emërtimit/shenjimit/etiketimit/tagimit/-anotimit, pra të përpunimit. I treti përmban shënime morfologjike, pra është i taguar/-anotuar në masë të madhe, përafërsisht për rreth 70-80% të fjalëve të tekstit, por jo në aspektin morfo-sintaksor. Do të thotë, shumëkuptimësia gramatikore dhe semantike në rrafshin morfo-sintaksor nuk është e zgjidhur. Si rezultat kërkimi paraqiten thjesht të gjitha rastet përkatësisht mundësitë. Pos të dhënave morfologjike, si veçori pozitive të korpusit në fjalë – duke pasur në mendje madhësinë e tij – duhet theksuar edhe mundësinë e përzgjedhjes së teksteve sipas autorit, apo edhe një fjale sipas pjesës së ligjëratës e veçorive morfologjike të saj.

AlCo në krahasim me dy korpuset e përmendura (1) ka numër shumë më të madh të fjalëve, dhe (2) është i emërtuar/shenjuar/etiketuar/taguar/anotuar në aspektin morfo-sintaksor. Një numër më i madh i fjalëve ofron gjasë më të madhe përkatësisht mundësi më të mëdha mbulimi të më shumë fenomeneve gjuhësore. Të dhënat statistikore bëhen më domëthënëse, më të qëndrueshme – përforcohen. Shënimet morfo-sintaksore, të paraqitura më lart, janë pothuajse të domosdoshme, të paanashlënshme, në ditën e sotshme gjatë një pune me korpus. Ato i japin korpusit një dimension shtesë, duke i rritur vlerën.

Disa mendime për të ardhmën

Paraqitja dhe përshkrimi i shkurtër i korpusit AlCo në këtë fazë ka për qëllim të tërheqë vëmendjen e studiuesve të shqipes për rëndësinë e krijimit të resurseve gjuhësore dhe të përdorimit të tyre gjatë punës hulumtuese-shkencore. Meqë korpusi, me gjithë punën e deritashme, numrin e madh të fjalëve, veglave kryesore ndihmëse si analiza automatike morfologjike, khs. [KABASHI 2015], tagseti khs. [KABASHI/PROISL 2016], modelet e trajnuara për tagues, si dhe disa vegla të tjera më pakdomëthënëse, është në një fazë që do edhe shumë punë.

Në të ardhmen e afërt duhet një kategorizim më i imët i teksteve, gjë që merr shumë kohë, gati sa vetë mbledhja e teksteve, shpesh sepse është e vështirë ose raste-raste edhe e pamundshme gjetja mjaftueshme përkatësisht e dëshirueshme e të

dhënave rreth teksteve apo autorëve të tyre. Pastaj zhvillimi i veglave për përpunimin e materialit gjuhësor dhe anotimit të shumëllojshëm të korpusit, punë, e cila duhet pothuajse patjetër të bëhet, kërkon shumë përkushtim, dije dhe forca pune.

Do të ishte mirë që e ardhmja ta gjente korpusin në një nga akademitë a institutet që si detyrë i kanë vënë vetës studimin e gjuhës shqipe. Një punë të tillë do të ishte dashur që ato ta kishin filluar para 40, gjithsesi më së voni para 15 vitesh. Mungesa e një korpusi gjuhësor për shqipen pas vitit 2010 nuk mund të arsyetohet – assesi.

Literatura

[ARKHANGELSKIJ ET AL. 2012] Timofej Arkhangelskij / Mikhail Daniel / Maria Morozova / Alexandar Rusakov: "Korpusi i gjuhës shqipe: Drejtimet kryesore të punës". 635–642. Në: [ISMAJLI 2012].

[BNC-XML 2007] British National Corpus. Version 3. BNC XML Edition. Distributed under license by Oxford University Computing Services on behalf of the BNC Consortium, 2007.

[CAKA/CAKA 2012] Nebi Caka / Ali Caka: "Korpusi i gjuhës shqipe: Drejtimet kryesore të punës". 643–656. Në: [ISMAJLI 2012].

[FRANCIS/KUCERA 1964] Nelson W. Francis / Henry Kučera: Manual of Information to Accompany "A Standard Corpus of Present-Day Edited American English". Brown University: Dept. of Linguistics. Providence 1964.

[GLAS 1975] Glas, Reinhold: „Das *LIMAS*-Korpus, ein Textkorpus für die deutsche Gegenwartssprache". Në: *Linguistische Berichte 40* (1975), 63–66.

[HARDIE 2012] Andrew Hardie (2012) " CQPweb – combining power, flexibility and usability in a corpus analysis tool". Në: *International Journal of Corpus Linguistics* 17 (3): 380–409.

[HOFFMANN ET AL. 2008] Sebastian Hoffmann / Stefan Evert / Nicholas Smith / David Lee / Ylva Berglund-Prytz: *Corpus Linguistics with* BNCweb – *a Practical Guide*. Frankfurt në Main etk.: Peter Lang, 2008.

[ISMAJLI 2012] Rexhep Ismajli (Red.): *Shqipja dhe gjuhët e Ballkanit*. Prishtinë, 10–11 nëntor 2011. Prishtinë/Tiranë: Akademia e Shkencave dhe e Arteve e Kosovës / Akademia e Shkencave e Shqipërisë, 2012.

[KABASHI 2015] Besim Kabashi: *Automatische Verarbeitung der Morphologie des Albanischen*. XVIII, 211. Erlangen, FAU University Press, 2015.

[KABASHI 2016] Besim Kabashi (2016): "Building an Albanian Text Corpus for Linguistic Research". Kumtesë në konferencën "Corpus-Based Approaches to the Balkan Languages and Dialects". 5–7 dhjetor 2016. Instituti i Shkencave Gjuhësore i Akademisë Ruse të Shkencave, Shën Petersburg. Khs. programin në faqen *https://iling.spb.ru/confs/balkan2016/schedule.html* dhe përmbledhjen e kumtesës në faqen *https://iling.spb.ru/confs/balkan2016/abstracts/7_kabashi.pdf*.

[KABASHI/PROISL 2016] Besim Kabashi / Thomas Proisl: "A Proposal for a Part-of-Speech Tagset for the Albanian Language". 4305–4310. Në: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorozh. Botuar nga Nicoletta Calzolari ettj. European Language Resources Association (ELRA) Paris.

Besim KABASHI

AlCo – A ONE HUNDERT MILLION WORD CORPUS OF ALBANIAN

Abstract

It is impossible to do serious studies in the field of natural languages without consulting empirical data. Natural language corpora offer this data in its original form. Apart from the fact that in some cases a corpus does not cover every possible word of a language, e. g. evidence, spelling or definition dictionary does, it offers the possibility to explore the data in every imaginable form, e. g. based on the actual context. A big corpus offers more data and the possibilities to cover more words and more phenomena than a small corpus does. Corpora make it possible to accurately study language in a quality not possible without empirical data. To use these benefits from corpora, it is necessary to create them first.

We present an Albanian Corpus (AlCo) that contains a hundert million word tokens (text words), the first Albanian corpus of this size. The corpus covers different domains of language and contains different text types – it is a reference corpus. At this moment the work is still in progress, some texts still need to be replaced or recategorised. The corpus is annotaed with a morpho-syntactic tagset of 77 tags, since 2015. We use CQPweb, a web based corpus analysis system, to explore the corpus data.