La elaborazion di un corpus etichetât de lenghe furlane scrite: esperiments e prospetivis

Sandri Carrozzo, Franz Feregot – Serling soc. coop. – CLAAP
publicât in Ce fastu?, 2012-2, pag. 241-262

1. Definizion di corpus etichetât

Te linguistiche computazionâl, vâl a dî te linguistiche che si poie sul ûs di imprescj informatics, si dopre la espression *corpus etichetât*, o plui dispès ancje dome *corpus*, par dî une racuelte di tescj li che cierts elements interessants a vegnin leâts a etichetis, o notis, che a permetin di identificâ e dopo di cjatâ in maniere sistematiche informazions linguistichis (o ben metalinguistichis)[1].

I corpus a puedin jessi formâts di tescj di une lenghe sole o di plui lenghis; a puedin jessi di lenghe scrite o di lenghe fevelade; daûr dal interès dai linguiscj che a produsin il corpus e des finalitâts che al à, la etichetadure e pues centrâsi su informazions di un cualsisei nivel di studi: fonetiche, fonologjie, morfosintassi, semantiche e v.i.

La utilitât dai corpus (o corpora) e à rivoluzionât la linguistiche contemporanie, in particolâr tai ultins trente agns, lant a pâr cu la cressite imburide des gnovis tecnologjiis: di fat chestis racueltis di tescj a permetin une ricercje precise e une vore svelte di fenomens linguistics testemoneâts intune espression spontanie e a puedin jessi doprâts ancje par aplicazions pratichis, par esempli par fâ o miorâ dizionaris, coretôrs ortografics, tradutôrs automatics, sintetizadôrs vocalics e programs di ricognossiment di tescj scrits, di vôs e une vore di altris.

In dutis lis lenghis dal mont la formazion di un o di plui corpus etichetâts e fâs passâ un idiome a une fase di tratament tecnologjic che e vierç gnovis prospetivis e che, soredut tal câs di lenghis menaçadis, e da plui risorsis par ribaltâ i fenomens di sostituzion linguistiche che a podaressin puartâlis a sparî. In cheste maniere la elaborazion di corpus e cjape une impuartance strategjiche te planificazion linguistiche[2].

2. Prins esperiments

Tal 2002 al jere publicât un prin studi di frecuencis lessicâls de lenghe furlane: tal presentâlu i autôrs, Alessandra Burelli e Marino Miculan, de Universitât dal Friûl, a sclarivin che al derivave de elaborazion pe base di dâts dal coretôr ortografic[3] de Societât Sientifiche e Tecnologjiche Furlane. Chest strument al otignive la sô base di dâts daûr dal sisteme statistic, vâl a dî dant dongje une cuantitât di tescj e selezionant pal coretôr chês plui frecuentis. La racuelte di tescj e je

[1] Par une definizion plui complete dal concet di corpus e par une esplicazion plui detaiade des carateristichis e de utilitât dai corpus cfr. Petkevič 2002.

[2] Te politiche linguistiche si fevele di "corpus" o di "planificazion dal corpus" di une lenghe intun sens diferent rispiet ae linguistiche computazionâl, che nol va sconfondût e che al è complementâr ae "planificazion dal status": pe planificazion linguistiche il corpus al è il complès di pussibilitâts espressivis di une lenghe, duncje in chest cjamp a jentrin i intervents tant che fissazion di une grafie, di une variante di riferiment, il slargjament dal lessic, il slargjament e il consolidament di regjistris e v.i.. Il status invezit al è in sostance la considerazion che la int e à de lenghe. La elaborazion di un "corpus etichetât" e jentre tal cjamp de planificazion da corpus, stant che e permet ricercjis e aplicazions che a rinfuarcin lis pussibilitâts espressivis, ma ancje in chel de planificazion dal status, pa vie che leant la lenghe aes gnovis tecnologjiis i da une imagjin plui moderne.

[3] Intal stes timp, anzit za publicât tal 2001, al jere stât elaborât il Coretôr Ortografic Furlan de Cooperative di Informazion Furlane: ducj i doi i prodots a àn cjapât la sigle di COF, stant che in chest scrit si fasarà riferiment a ducj i doi al è ben tignî presint che a son doi programs dâts dongje cun sistemis diviers.

stade doprade dai autôrs ancje par fâ cheste rassegne, che cun onestât a definissin dome come une prime prove, su lis peraulis plui dopradis te lenghe furlane scrite. Tal articul (Burelli, Miculan 2002) publicât tal "Gjornâl Furlan des Siencis" la liste di tescj furlans che a àn doprât e ven definide ancje "corpus", ma chest tiermin al è di intindi in sens larc, stant che sui tescj nol jere stât fat nissun tratament. In ogni câs chest studi al à di une bande il merit di fâ intraviodi une aplicazion pussibile cul tratament informatic de lenghe naturâl, di chê altre al rive al limit, che si evidenzie bessôl, dal fat di doprâ racueltis di tescj no etichetâts.

Un esperiment di etichetadure al vignive fat, impen, in prevision di une pussibile elaborazion di un Vocabolari dal Furlan Antîc, par iniziative dal professôr Giovanni Frau, de Universitât dal Friûl: une prime prove si faseve tal 2001, etichetant il test dai Rodui dai Cjaliârs di Udin (agns 1401-1407) cul program GATTO, elaborât di Domenico Iorio-Filli e doprât pal Corpus OVI (Opera del Vocabolario Italiano) dal talian antîc. Tal 2005 si zontavin altris 20 tescj, pal plui regjistris, letaris, esercizis di version, ma ancje cualchi test leterari (*Biello Dumnlo, Piruç myo doç inculurit...*). Si etichetavin cussì tescj par un totâl di 66.516 ocorencis, che si podevin assegnâ a 1.834 lemis, articolâts in 3.509 formis. Il numar bas di lemis e formis di chest corpus al derivave soredut de nature dai tescj plui luncs, che a jerin di caratar contabil; pal stes motîf, tra chei altris lemis, si regjistravin 409 antroponims e 105 toponims.

Daûr des carateristichis dal program GATTO, coerent cul obietîf di doprâ il corpus par un vocabolari dal furlan antîc, la etichetadure e zontave a ogni forme dome la informazion dal leme di partignince e la sô categorie gramaticâl, intindint cun chest la indicazion di ce part dal discors che si tratàs; no jere dade impen nissune definizion morfosintatiche, che no je previodude di GATTO. Classificant cheste etichetadure daûr des racomandazions dal progjet EAGLES[4] si pues dî che e jere di nivel L0, vâl a dî chel plui elementâr e obligatori.

Dopo chescj esperiments, il progjet di un corpus pal furlan antîc, cul non di *Dizionario Storico Friulano*, al è passât sot dal coordenament di Federico Vicario, cressint une vore in cuantitât: a risultin 89 documents, di divierse lungjece, par un totâl sul ordin des 600.000 peraulis, li che a son atestadis 14.044 formis, di chestis 4.918 antroponims e 1.424 toponims, e i risultâts si puedin viodi vuê publicâts te pagjine web www.dizionariofriulano.it

Chest sît al permet ricercjis une vore impuartantis pal studi dal furlan antîc, cuntune consultazion par lessic, par document e par bibliografie. In particolâr al permet ancje di viodi e di scjariâ ducj i documents in formât PDF, pandint materiâl che se di no al sarès une vore plui dificil di consultâ.

Un limit di chest strument al è tal sisteme di etichetadure, che nol dopre plui il program GATTO e che nol lee plui in maniere automatiche lis formis ai lemis, duncje al ridûs une vore la utilitât propit tal cjamp che al varès di jessi la sô aplicazion primarie, la lessicografie e la redazion di dizionaris: di fat chei che tal interface de pagjine web a vegnin definîts tant che "lemma" in realtât a son formis e cussì peraulis tant che *spendei* o *spendè*, che a son formis dal leme *spindi*, a vegnin consideradis ognidune tant che leme autonom. Par corezi cheste carateristiche, che di fat e inderede ricercjis sistematichis, e ven furnide in ogni câs une liste di leams a altris formis dal stes leme. Te compilazion de liste però si evidenzie il fat che no je automatiche, ma manuâl e cussì e je cualchi incompletece: par esempli tra dôs variantis grafichis tant che *espendey* e *spendegi* la prime e ven leade a altris 18 formis, la seconde dome a 13. L'ûs di risorsis informatichis tant che

[4] Sigle che e vâl par "Expert Advisory Group on Language Engineering Standards", si trate di un progjet de Comission Europeane par rivâ a nivei di riferiment comuns par risorsis di linguistiche computazionâl di largje scjale, di mût che si puedin condividi e riciclâ dâts e programs.

GATTO o altris programs al permetarès di sistematizâ chest corpus e di fâ cressi une vore la sô utilitât[5].

3. Un etichetadôr semiautomatic: dal COF a Jude e di Jude al etichetadôr

Se pal studi de lenghe antighe une etichetadure come chê previodude di GATTO e pues bastâ, e ancje chê no sistematiche dal *Dizionario Storico Friulano* e à in ogni câs une grande utilitât, pal studi de lenghe moderne e contemporanie, pes aplicazions tal studi statistic, sintatic, leterari, te corezion ortografiche, te traduzion automatiche, te lessicografie, e v.i. e covente une etichetadure morfosintatiche plui complete[6].

Tai ultins agns, profitant de competence che si è disvilupade in altris progjets e di materiâl linguistic computazionâl prontât par altris imprescj, chest lavôr nol è plui impussibil o une vore dificil, ma al pues deventâ sistematic e avonde svelt.

Une sielte fondamentâl e je stade fate dal 2001, cuant che par fâ il Coretôr Ortografic Furlan di Informazion Furlane, invezit di doprâ il sisteme statistic, si à vût miôr di doprâ chel morfologjic: invezit di partî di tescj scrits par vê lis formis plui dopradis (come tal coretôr de Societât Sientifiche e Tecnologjiche Furlane), si è partîts di une liste di lemis, che a son stâts leâts cun regulis ognidun al so paradigme, ven a stâi a une liste di terminazions par produsi dutis lis formis.

Cussì par esempli il leme *cjase* contrassegnât cu la etichete "sf" al rispuindeve aes istruzions di gjavâ la desinence *–e* par vê il teme e di zontâ *-e, -is, -ute, -utis,-uce,-ucis, -ine, -inis, -one, -onis, -ate, -atis, -ace, -acis*... formant *cjase, cjasis, cjasute, cjasutis* e v.i.

Cun chest materiâl al sarès za stât pussibil fâ un program che, suntun test in grafie uficiâl e in lenghe standard, al fos plui svelt di GATTO e che invezit di domandâ volte par volte di segnâ il leme di apartignince al varès podût dâ ae persone une sielte tra i lemis pussibii: in dut câs ta chei agns la produzion di un program dal gjenar no jere une prioritât e i fonts a disposizion a permetevin a pene la sperimentazion inviade dal professôr Frau, che par altri, centrantsi su tescj antîcs, e veve reson di doprâ GATTO, pensât propit par tratâ tescj cuntune alte variabilitât ortografiche e dialetâl.

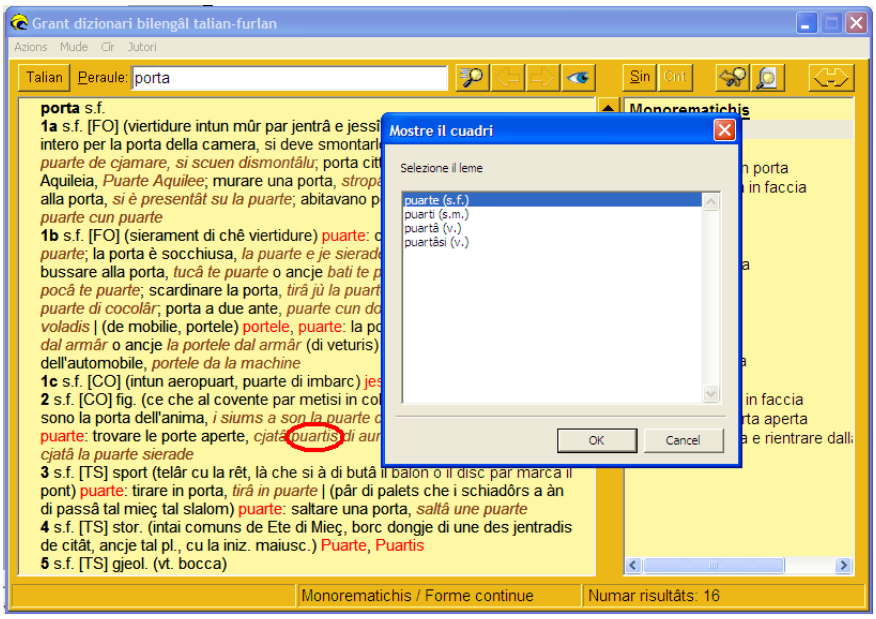
Une aplicazion dal gjenar e veve invezit di nassi te suaze dai lavôrs pal *Grant Dizionari Bilengâl Talian-Furlan*, par une iniziative no contemplade tai plans (e nancje metude a belanç). Pôc prime che si sierassin i lavôrs pe edizion dal dizionari di base (la version dal 2005 cun plui o mancul 6.500 lemis), si à pensât che al sarès stât pussibil zontâ i cuadris flessionâi pe lenghe furlane: cheste funzion e varès permetût al utent dal dizionari di viodi la flession di ogni leme e e jere, ancje se in forme divierse, pal talian tal program informatic dal *Grande Dizionario dell'Uso* (GDU) di Tullio De Mauro.

Prin di dut, cul materiâl dal coretôr ortografic di Informazion Furlane, si à vût di realizâ un ricognossidôr di formis che lis leàs al leme di là che a podevin derivâ; po dopo cu lis desinencis des formis principâls si à formât tabelis ordenadis che a vegnin visualizadis dal utent. Cussì clicant suntune peraule furlane cualsisei dal GDBTF il program al propon il leme o i lemis

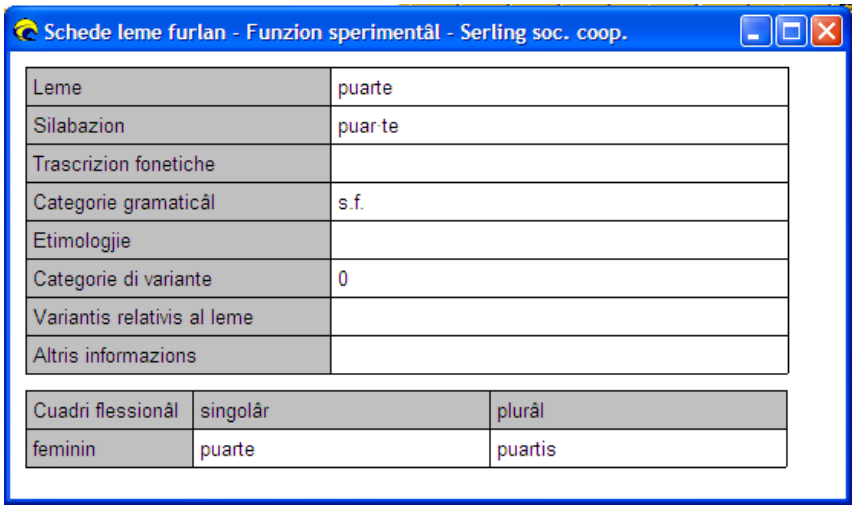
[5] In ogni câs al è di calcolâ che un corpus di documents antîcs, tant che chel dal Dizionario Storico Friulano al fronte problemis complès: di fat plui che di documents par furlan antîc, si scuen fevelà di documents antîcs scrits in Friûl, dulà che il furlan si messede cun altris codiçs linguistics, tant che il latin e il toscovenit. Une etichetadure in chestis situazions e devente pardabon complesse, stant che e scuen segnalâ se une forme e parten a une lenghe o a une altre, ma une vore dispès fâ cheste distinzion al è dificil o impussibil.

[6] In chest câs, tal progjet EAGLES, si fevele di nivel di etichetadure L1.

pussibii di riferiment. Par esempli clicant su *puartis* al proponarà di visualizâ il paradigme di *puarte*, *puarti*, *puartâ* o *puartâsi*.



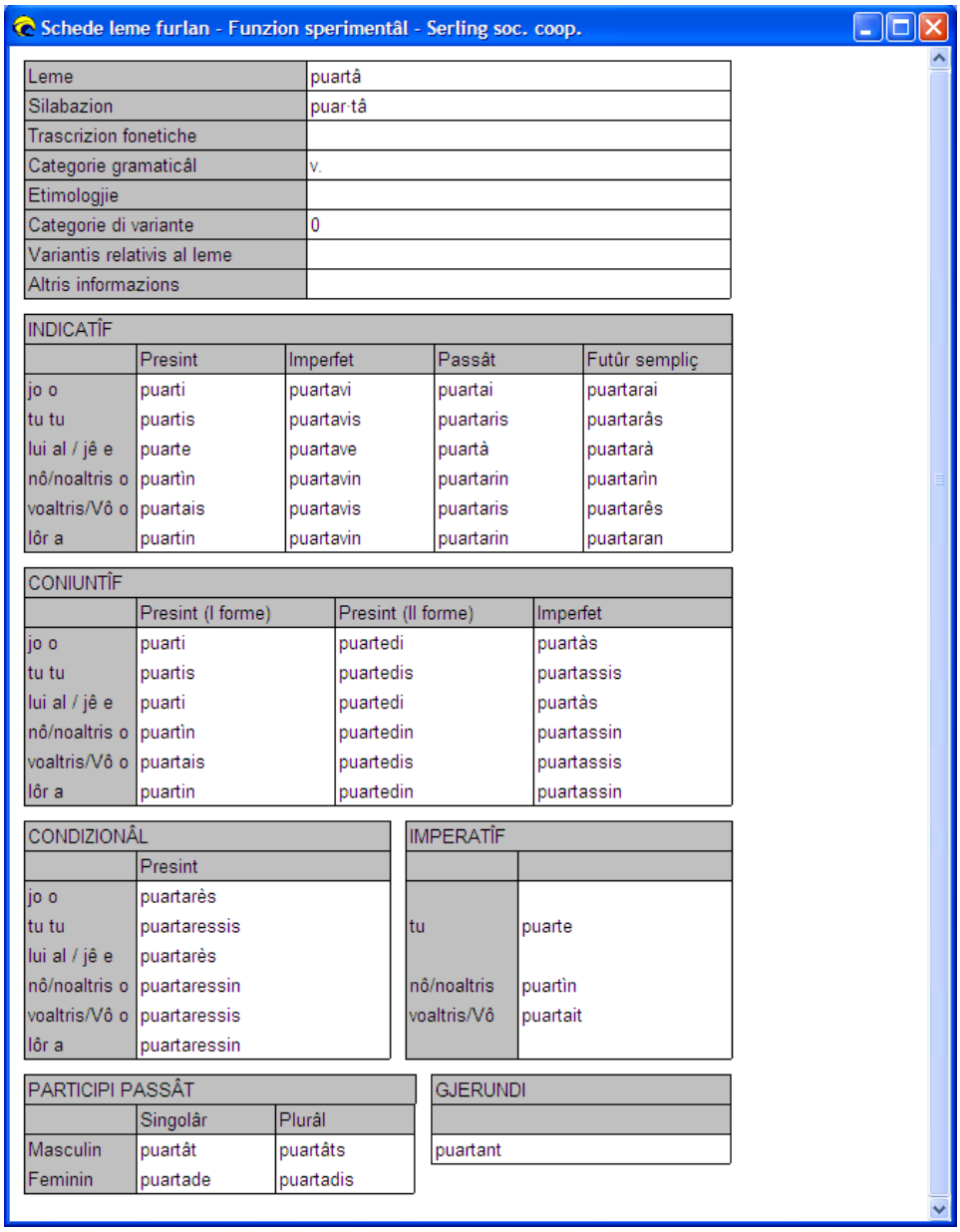
Une volte sielt il leme che al interesse si pues viodi il paradigme principâl, duncje tal câs di *puarte*:



Leme	puarte
Silabazion	puar·te
Trascrizion fonetiche	
Categorie gramaticâl	s.f.
Etimologjie	
Categorie di variante	0
Variantis relativis al leme	
Altris informazions	

Cuadri flessionâl	singolâr	plurâl
feminin	puarte	puartis

O ben tal câs di *puartâ*:



Schede leme furlan - Funzion sperimentâl - Serling soc. coop.

Leme	puartâ
Silabazion	puar·tâ
Trascrizion fonetiche	
Categorie gramaticâl	v.
Etimologjie	
Categorie di variante	0
Variantis relativis al leme	
Altris informazions	

INDICATÎF

	Presint	Imperfet	Passât	Futûr sempliç
jo o	puarti	puartavi	puartai	puartarai
tu tu	puartis	puartavis	puartaris	puartarâs
lui al / jê e	puarte	puartave	puartà	puartarà
nô/noaltris o	puartìn	puartavin	puartarin	puartarìn
voaltris/Vô o	puartais	puartavis	puartaris	puartarês
lôr a	puartin	puartavin	puartarin	puartaran

CONIUNTÎF

	Presint (I forme)	Presint (II forme)	Imperfet
jo o	puarti	puartedi	puartàs
tu tu	puartis	puartedis	puartassis
lui al / jê e	puarti	puartedi	puartàs
nô/noaltris o	puartìn	puartedin	puartassin
voaltris/Vô o	puartais	puartedis	puartassis
lôr a	puartin	puartedin	puartassin

CONDIZIONÂL

	Presint
jo o	puartarès
tu tu	puartaressis
lui al / jê e	puartarès
nô/noaltris o	puartaressin
voaltris/Vô o	puartaressis
lôr a	puartaressin

IMPERATÎF

tu	puarte
nô/noaltris	puartìn
voaltris/Vô	puartait

PARTICIPI PASSÂT

	Singolâr	Plurâl
Masculin	puartât	puartâts
Feminin	puartade	puartadis

GJERUNDI

puartant

I derivâts e lis formis verbâls cun pronons enclitics par resons di spazi no jentravin tes tabelis, ma il fat di frontâ chest lavôr in ogni câs al à ispirât un mudament di plante fûr de base di dâts, elaborade aromai de cooperative Serling (che in dì di vuê e à i dirits su di chê): dutis lis terminazion, no dome chês des formis di visualizâ tal GDBTF, a son stadis ordenadis in tabelis sistematichis a trê nivei (flession des formis principâls, flession derivade, flession derivade de derivade), daûr di un sisteme di etichetadure che al permet une mapadure rigorose e che pal plui al è compatibil cui sistemis di etichetadure standard a nivel internazionâl, tant che cun adataments une vore piçui al podarès jessi doprât rispuindint ad implen aes racomandazions di EAGLES.

Une des primis aplicazions che e à cjatât cheste base di dâts e je stade la realizazion di un tradutôr automatic dal talian al furlan. Une base di dâts paralele e je stade creade ancje par ce

che al tocje la lenghe taliane e metint in relazion lis dôs basis intun sisteme di cubiis di leme talian-leme furlan al è stât elaborât il tradutôr automatic a trasferiment superficiâl Jude3[7].

In Jude3, come in cualsisei tradutôr automatic a trasferiment, un dai modui plui impuartants pal bon funzionament al è chel di disambiguazion: vâl a dî che une part dal program e cîr di distingui in maniere automatiche a ce leme assegnâ une peraule de lenghe di origjin (in chest câs il talian). Di fat *porta* al pues vignî di *porta*, *portare*, *portarsi* e *porgere* e al pues vê diviers valôrs morfosintatics. Chest modul al funzione cuntun algoritmi che al cjale prime o dopo de peraule ambigue e daûr dal contest al prove a induvinâ la pussibilitât juste: se lis regulis a disposizion dal program a son buinis il ricognossiment al è coret intune percentuâl che e pues rivâ tor dal 95%.

Cemût che si diseve, la esperience fate in chescj progjets e part dal materiâl linguistic computazionâl a son stadis impleadis te prospetive di formâ un corpus etichetât del lenghe furlane.

Par fâ un tant si à cjapât la base di dâts furlane dal tradutôr Jude3 (liste di lemis furlans etichetâts e cuadris flessionâi complets) e si à realizât un program di ricognossiment semiautomatic (che in pratiche al funzionave, ancje se cuntun materiâl plui grês, za tal GDBTF) e di etichetadure: cun chest program, che al ricognòs lis formis proponint, in câs di ambiguitât, lis soluzions pussibilis a nivel lessicâl e morfosintatic, al è stât etichetât un corpus di prove di un 100.000 peraulis.

4. La etichetadure di un corpus di prove

Tal moment di inviâ la etichetadure di un corpus la prime robe di fâ e je di definî la sô finalitât e di sielzi i tescj te maniere che le puedi sodisfâ te maniere miôr.

Tal câs dal furlan si à pensât che lis primis bisugnis a jerin chês di fâ un strument che al judàs in particolâr oparis di lessicografie, terminologjie, ricercjis su la sintassi; cun di plui si podeve previodi un interès ancje par ricercjis comparadis di plui varietâts furlanis (di timps diferents e di lûcs diferents) e la analisi leterarie.

Za tal 2009 la Serling e veve puartât al president de Agjenzie Regjonâl pe Lenghe Furlane, Lorenzo Zanon, une letare di propueste par lâ indevant cuntun progjet dal gjenar, ma cheste struture e jere impegnade in altris iniziativis e soredut intune riforme interne, duncje no veve podût dâ une rispueste a chest stimul. Di consecuence il disvilup dal progjet al è lât indevant par iniziative privade.

Par meti in vore lis primis fasis dal progjet si à destinât di rivâ a etichetâ un corpus di prove di 100.000 peraulis: cheste cuantitât e je il limit che si met pal solit tra un corpus piçul e un medi, ma al pues jessi suficient pal "alenament" dal etichetadôr, vâl a dî che se si etichetin in maniere corete tescj par cheste cuantitât di peraulis, analizant lis ocorencis e i lôr contescj te frase, dopo l'algoritmi al è bon di disambiguâ in maniere automatiche cuntune precision che in gjenerâl e je plui alte dal 90%: cheste percentuâl e pues cambiâ daûr dal test che al ven frontât, ma ancje daûr dal algoritmi[8] che al ven doprât, che tal nestri câs al è chel di Brill, e dal sisteme di etichetadure che si sielç: cemût che si à dite in chest câs la complessitât e je avonde alte, in pratiche dal nivel

[7] Pes fasis di elaborazion dal tradutôr Jude e pes sôs carateristichis plui in detai cfr. Carrozzo, Feregot, Mistrut 2010.

L1 dal progjet EAGLES, e cussì e je avonde alte ancje la dificoltât di disambiguâ in maniere automatiche.

Par jessi sigûrs di vê un bon risultât, si veve di dâ dongje tescj che a vessin chê cualitât che si clame "rapresentativitât" ven a stâi jessi esemplis di plui stîi, di plui regjistris e di plui argoments. Par cheste prime fase a son stâts cjapâts dome tescj scrits in lenghe furlane standard, de seconde metât dal secul XX al dì di vuê. La sielte linguistiche e grafiche e je determinade, in plui che de volontât di rispietâ lis normis uficiâls, de impussibilitât tecniche di fâ un etichetadôr che al funzioni daurman cun dutis lis varietâts furlanis: invezit fasint un etichetadôr pal furlan di riferiment e in grafie uficiâl, dopo si podarès doprâ chestis regulis tant che scjalin intermedi, par fâlu funzionâ ancje cu lis varietâts.

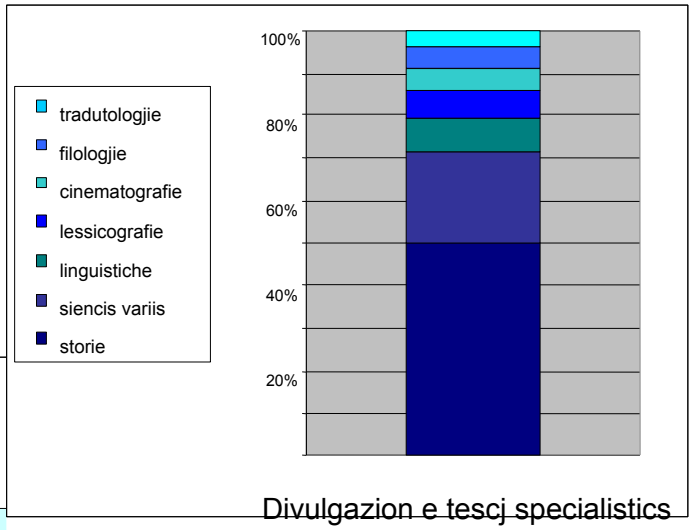
Tal câs di tescj che a fossin un esempli interessant ma che no fossin in grafie uficiâl si à fat une conversion ortografiche. In ogni câs ducj i tescj a son stâts uniformâts a nivel ortografic.

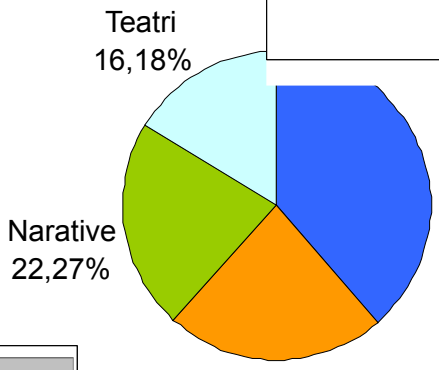
Pe sielte dai gjenars e dai autôrs, cirint juste apont une buine rapresentativitât de lenghe scrite, si è rivâts ae composizion mostrade di chescj schemis.

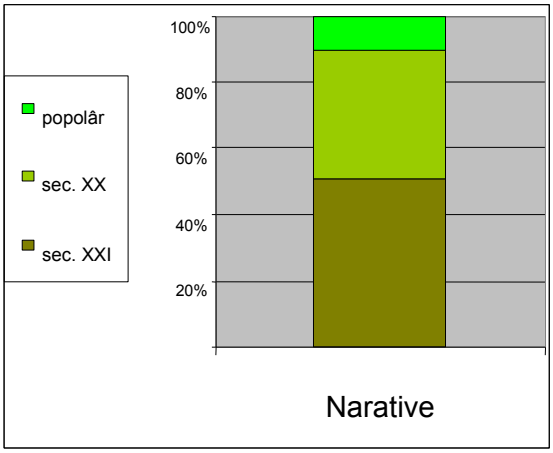
[8] I disambiguadôrs dai tradutôrs automatics, che a sopuartin in gjenar un margjin di erôr plui alt, par vie che cualchi volte l'erôr te traduzion no si viôt o par vie che al pues jessi coret in altris fasis dal program, a puedin funzionâ ancje cuntun corpus di alenament une vore plui piçul, tor des 30.000 peraulis; tal progjet CORIS il corpus di alenament al è stât di 84.000 peraulis e al è stât provât cun plui algoritmis, a regulis e stocastics, rilevant margjins di erôr te individuazion des parts dal discors di tescj talians tra il 4,81% e il 12,5%, (6,95% pal algoritmi di Brill, cfr. Tamburini 2000, 2009); altris corpus di alenament a puedin jessi une vore plui grancj, ancje di 1.000.000 di peraulis.

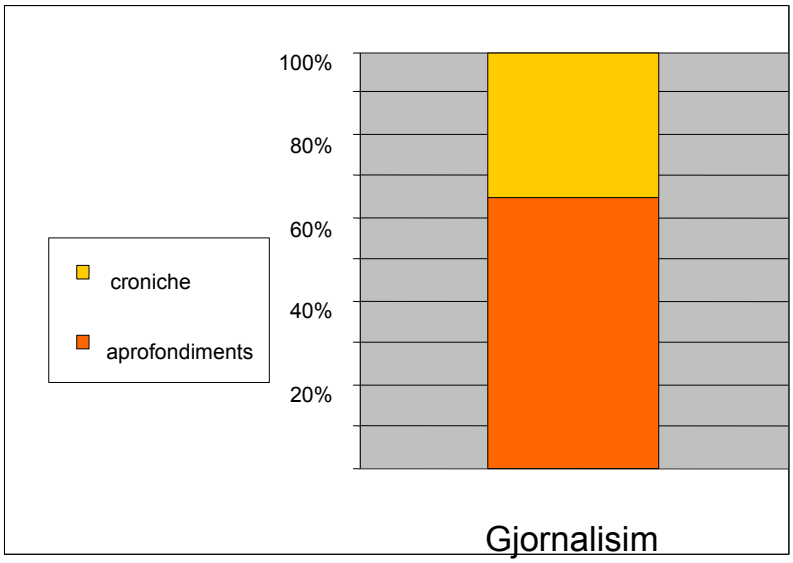
Composizion dal corpus di prove
de lenghe furlane scrite
cun sudivision par gjenars
e sotgjenars

100% = 100.970 peraulis

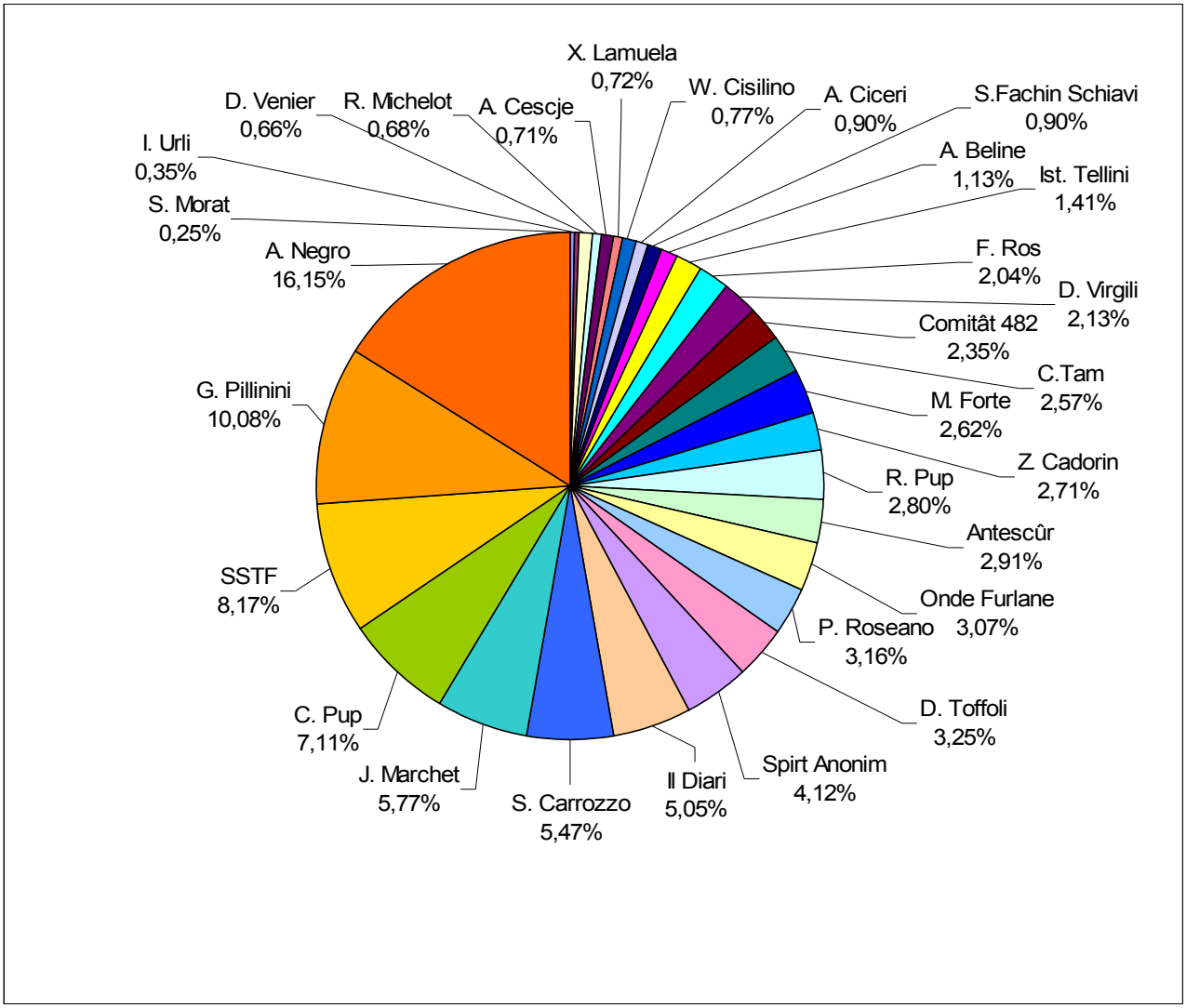








Composizion dal Corpus di prove
de lenghe furlane scrite
cun sudivision par autôrs



Par mantignî la rapresentativitât tun corpus no tant grant si à vût di cjapâ tescj avonde piçui: in dut 144, duncje cuntune lungjece medie di pôc plui di 700 peraulis. La lungjece massime e je une vore plui alte, 16.332 peraulis, tal câs de opare teatrâl *Int di masnade*, e e sbelance l'ecuilibri par autôrs, dant un grant rilêf a Alviero Negro, ma chest difiet al ven compensât dal fat di vê ancje un test teatrâl, che al documente formis di dialic diret che se di no a son raris in altris gjenars di produzion scrite.

I autôrs a son uns trente, ma si à di notâ che in cierts câs, tant che i tescj di sintesi dai articui de Societât Sientifiche e Tecnologjiche Furlane, i articul di Il Diari e lis gnovis di Radio Onde Furlane, nol è pussibil cjatâ un autôr, alore i tescj a son stâts assegnâts ae testade responsabile.

Al è di tignî cont che chestis sudivisions a àn un ecuilibri avonde diferent di chel racomandât tai progjets di nivel internazionâl, tant che LE-PAROLE, che a dan une vore di plui di spazi ae lenghe di comunicazion e mancul a chê leterarie[9]. Cheste carateristiche e diven di une riflession, che in efiets no si fonde su dâts sientifics, stant che no son a disposizion, su la cundizion sociolinguistiche dal furlan: la cuantitât dai scrits gjornalistics par furlan no je disvilupade tant che in altris lenghis e salacor no à nancje il stes pês tal complès de produzion scrite e de fruizion dai letôrs, duncje dant plui pês a chest setôr, dulà che i autôrs in fin dai fats a son avonde pôcs, si varès limitât la rapresentativitât dal corpus invezit di fâle cressi. Cun di plui la grande part des notiziis di croniche no vegnin di une formulazion spontanie, ma a son traduzion di tescj di

[9] I margjins in chest câs si riferissin ae origjin dai tescj: libris 16-22%; informatîfs 58-72%; rivistis 4-10%; variis 8-12%. Cfr. R. Marinelli e altris, 2003.

partence par talian. In ogni câs, se si volarà, si podarà corezi l'ecuilibri par ordin che la cuantitât di tescj etichetâts e cressarà.

In plui che par formâ un imprest di alenament pal algoritmi di disambiguazion, cheste prime prove di etichetadure e à zovât par individuâ i ponts critics te assegnazion de categorie gramaticâl, par zontâ câs no previodûts o par ordenâ cun altris criteris cualchi detai dai cuadris flessionâi e ancje par zontâ lemis ae base di dâts di COF e Jude.

In particolâr a meretin une riflession i ponts critics te assegnazion de categorie gramaticâl: di fat il tratament automatic dal lengaç al domande regulis claris e rigorosis, ma la lenghe invezit e je alc di vîf e variabil, che in fin dai fats e scjampe de volontât dai gramatics e de necessitât dai struments informatics di inscjaipulâle in categoriis fissis.

Cun di fat no je stade fate nissune descrizion linguistiche complete, corete e rigorose a nivel teoric, che si rivi a doprâ ancje a nivel pratic cuntun program automatic.

Un esempli clâr, che si cjate une vore dispès tai corpus[10], e je la dificoltât di distingui tra participi passât e adietîf e no dome in maniere automatiche, ma ancje a nivel personâl: se intune frase tant che "o ai viert il barcon" si pues dî dal sigûr che *viert* al è participi passât e invezit inte espression "tal mâr viert" si pues dî che *viert* al è adietîf, intune vore di altris espressions la assegnazion no je sigure.

Ancje altris câs a fasin capî che te etichetadure dai corpus no si pues pretindi une precision assolude e che no si pues fevelâ de linguistiche tant che di une sience esate, ma tant che di une dissipline, li che l'impuartant al è declarâ i criteris che si doprin e cirî di jessi coerents, permetint al letôr o al utent di orientâsi, ma savint che si movisi intun cjamp li che a son margjins di insigurece[11].

Duncje la prime etichetadure e je stade ancje une esplorazion dai câs dubis che si puedin cjatâ te lenghe furlane: su la fonde di cheste si podarà elaborâ criteris plui definîts che, ancje se cualchi volte a scugnaran jessi arbitraris, a sedin plui rigorôs e coerents che si pues e tal stes timp pragmatics.

5. Il program di consultazion in linie

Par dimostrâ il funzionament dal corpus etichetât al è stât realizât un program di gjestion de base di dâts che al permet consultazions in linie. Stant che il progjet nol è ancjemò finanziât e che la realizazion di un program di gjestion e consultazion complet e compuartarès un invistiment privât masse alt, il program di cumò al è di considerâ dome un piçul prototip.

Pal moment al permet di fâ ricercjis jemplant une gridele ordenade daûr dai cjamps di forme, leme, categorie gramaticâl, definizion flessionâl. In particolâr chest ultin cjamp, al è une vore util, ma al è ancjemò une vore dificil di doprâ par un utent cualsisei, stant che lis opzions pussibilis a son lis siglis dal codiç doprât dal program e no je stade metude a disposizion dal utent, pal moment, une leiende o une descrizion: se lis siglis de categorie gramaticâl a son mancul di vincj e di significât intuitîf, lis pussibilitâts desinenziâls, che a previodin ducj i câs di derivazion e di enclisi, a son passe mil e dusinte. I cjamps di ricercje, inte version di cumò, a puedin jessi jemplâts par une secuence di trê elements.

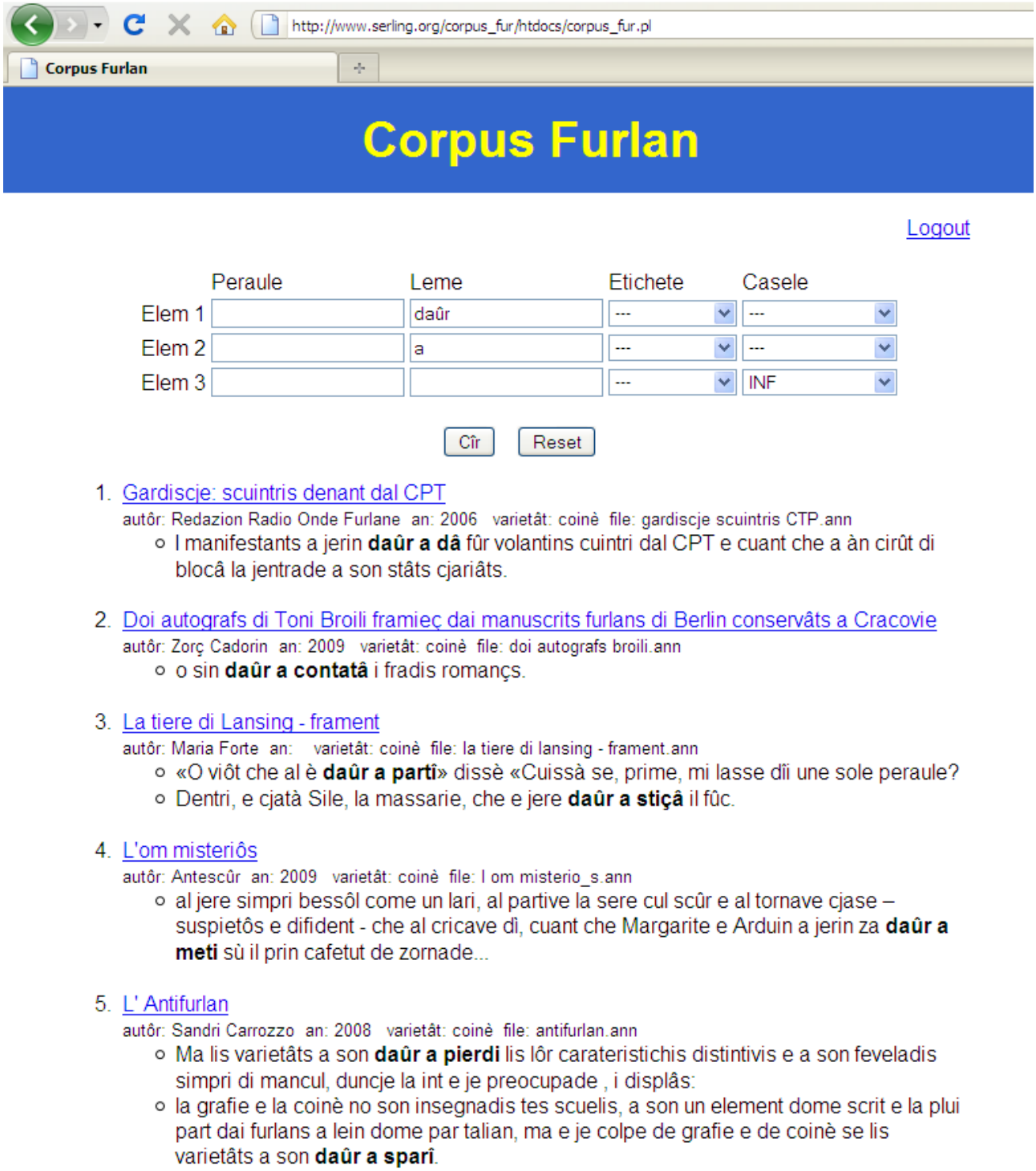
Daûr dai elements volûts il program al cîr tal corpus e al mostre al utent la frase dulà che si cjatin, segnalant autôr, titul dal test, an di edizion, variante furlane (pal moment a son ducj de coinè). Al è ancje pussibil scjarià dut il test in formât txt, ma cheste funzion, une volte che il program al sarà viert al public e varà di jessi limitade par rispietâ lis normativis sui dirits di autôr. Il program in ogni câs nol è ancjemò di acès public, ma si pues jentrâ dome passant un control di non di utent e peraule segrete.

Si zonte ca sot la imagjin di un esempli di ricercje di une costruzion sintatiche interessante pe lenghe furlane, ven a stâi "daûr a" + infinît.

---

[10] Si pues lei par esempli i criteris di etichetadure dal CNR tal corpus LIP, dulà che si dîs "La distinzione tra uso aggettivale e uso participiale dei participi passati è stata frequentemente problematica." e dopo di vê presentât une casistiche si conclût "Questi test non sono stati sufficienti pertanto a discriminare tutti i casi, ma sono stati usati in caso di dubbio."

[11] Dentri dal stes sisteme proponût di EAGLES, che al vûl jessi un riferiment standard, si cjatin incoerencis ancje profondis te aplicazion a diviersis lenghis, in câs che però a varessin di jessi compagns: cussì i numerâi ordenâi a vegnin considerâts pronons (P) o determinants (D) te aplicazion dal sisteme al talian, ma adietîfs normâi (A) te aplicazion al catalan, intant che tal ûs lis dôs lenghis si compuartin te stesse maniere.



6. Prins risultâts

Dopo di vê etichetât il corpus di prove a son stadis gjeneradis lis regulis che a fasin funzionâ il disambiguadôr automatic. Cun chestis gnovis regulis a son stâts etichetâts trê tescj par une sume di 2.022 peraulis. Chescj tescjs a son stâts sielzûts in maniere di partignî ai gjenars za cjapâts dentri te formazion dal corpus di prove, ma di autôrs che no jerin presints[12]. Fasint une revision di cheste etichetadure automatiche al è risultât che la completece te etichetadure e je stade dal 99,61%: vâl a dî che lis formis furlanis no cognossudis a son stadis dome 8, in 4 câs neologjisims leteraris, in 3 formis che a jerin za regjistradis dome tant che adietîf, ma che tai tescj si cjatavin tant che sostantîfs, intun sôl câs un leme tecnic (*metropolitic*).

[12] Une conte curte di Gianfranco Pellegrini, un editoriâl dal diretôr de Patrie dal Friûl Andrea Valcic, un scrit storiografic di Luzian Verone.

La precision te assegnazion de etichete e je dal 91,15%. Cheste però e cjape dentri ancje erôrs che si riferissin a distinzions plui altis rispiet al nivel L1 di EAGLES, tant che la diferenziazion tra verps transitîfs, intransitîfs e pronominâi. Tignint cont dome di ce che si pues riferî al nivel L1, la precision dal etichetadôr automatic e va sù, tal piçul campion cjapât in esam, al 92,39%.

In plui des regulis gjeneradis in cheste maniere, cemût che al sucêt tal tradutôr Jude3, il program al previôt che si puedin zontâ regulis di disambiguazion subietivis, no individuadis tal corpus di alenament, ma compiladis dal linguist: in cheste maniere la precision dal disambiguadôr e podarà cressi ancjemò di plui, no dome cu la cressite cuantitative dal corpus di riferiment, ma ancje cul intervent diret dai responsabii dal program, che a podaran sveltî la cressite des prestazions dal program, soredut tai erôrs su lis formis plui difondudis.

Par instaurâ un aboç di dialic cul studi su lis frecuencis za publicât te riviste “Siencis par furlan” (Burelli, Miculan 2002) si à fat une ricercje parele sul corpus atuâl. In plui di rilevâ la frecuence des formis, però, si à rilevât ancje la frecuence dai lemis. Il sisteme gnûf al fâs par altri che i risultâts si rivin a comparâ dome in part cu la ricercje dal 2002: di fat cumò si pues, e si scuen, se si vûl vê risultâts significatîfs, distingui tra formis diferentis ancje se a son omografis, o meti adun peraulis che, a ben che scritis in manieris diferentis, a son in sostance une forme sole.

Stant che in ogni câs un confront detaiât nol jentre tai fins di chest scrit si ripuarte ca sot dome lis listis des primis 100 formis e dai prins 100 lemis.

Ordin	**Forme**	**Leme**	**Etichete**	**Ocorencis**	**Tescj**
1	di	di	PREP	4543	143
2	e	e	CONIUN	3115	143
3	al	lui	PRONP	2334	134
4	che	che	CONIUN	1934	132
5	il, l'	il	ART	1754	134
6	la	la	ART	1750	138
7	a	a	PREP	1644	129
8	dal	di	PREP	1271	128
9	e	jê	PRONP	1168	134
10	de, da la	di	PREP	1134	127
11	a	lôr	PRONP	1082	117
12	par	par	PREP	1078	117
13	in	in	PREP	1035	125
14	i	il	ART	921	125
15	un	un	ART	789	118
16	une	un	ART	691	116
17	dai	di	PREP	668	109
18	che	che	PRON	634	109
19	lis	la	ART	618	120
20	si	si	PRONP	604	107
21	no	no	AV	587	93
22	ma	ma	CONIUN	579	94
23	ancje	ancje	AV	537	98
24	plui	plui	AV	436	97
25	è	jessi	VINTR	425	76
26	a	lôr	PRONP	392	93
27	tal	ta	PREP	391	86
28	cun	cun	PREP	374	82
29	des, da lis	di	PREP	373	86
30	che	che	PRON	372	87
31	che	che	PRON	353	84
32	o	jo	PRONP	339	38
33	à	vê	VAUX	320	78
34	se	se	CONIUN	315	63
35	je, è	jessi	VINTR	311	76
36	nol	lui	PRONP	272	61

37	jere	jessi	VINTR	270	45
38	o	o	CONIUN	260	66
39	son	jessi	VINTR	255	66
40	al	a	PREP	250	74
41	o	nô	PRONP	240	50
42	si	lui	PRONP	235	62
43	te, ta la	ta	PREP	231	67
44	che	che	PRON	217	78
45	ae, a la	a	PREP	213	65
46	dome	dome	AV	205	66
47	chest	chest	ADI	205	60
48	à	vê	VTR	197	60
49	lu	lui	PRONP	197	48
50	fâ	fâ	VTR	184	62
51	pe, par la	par	PREP	180	56
52	chel	chel	PRON	179	58
53	sô	so	ADI	174	55
54	speziâr	speziâr	SMF	174	1
55	chei	chel	PRON	173	48
56	lenghe	lenghe	SF	173	31
57	ai	a	PREP	172	46
58	chest	chest	ADI	170	63
59	come	come	CONIUN	170	49
60	simpri	simpri	AV	169	60
61	ce	ce	PRON	166	54
62	i	lui	PRONP	163	41
63	fûr	fûr	AV	162	75
64	cul	cun	PREP	158	56
65	so	so	ADI	156	48
66	jo	jo	PRONP	155	24
67	pal	par	PREP	154	60
68	tant	tant	AV	154	57
69	sul	su	PREP	153	56
70	agns	an	SM	151	59
71	cuant	cuant	AV	150	52
72	Friûl	Friûl	SM	149	41
73	è	jessi	VAUX	147	62
74	chel	chel	ADI	145	59
75	àn	vê	VAUX	143	54
76	chê	chel	PRON	142	58
77	di un	di	PREP	138	59
78	di une	di	PREP	138	54
79	cence	cence	PREP	137	48
80	sù	sù	AV	136	65
81	cu la	cun	PREP	133	56
82	po	po	AV	133	46
83	come	come	AV	132	45
84	furlan	furlan	SM	132	20
85	no	jê	PRONP	130	47
86	veve	vê	VAUX	129	29
87	Sgobar	Sgobar	SMF	129	1
88	àn	jessi	VAUX	125	52
89	mi	jo	PRONP	124	25
90	ben	ben	AV	122	44
91	lui	lui	PRONP	121	24

92	vie	vie	AV	117	47
93	tu	tu	PRONP	117	22
94	le	jê	PRONP	116	43
95	veve	vê	VTR	116	43
96	Udin	Udin	SM	116	34
97	dut	dut	ADI	114	41
98	là	là	AV	113	46
99	cumò	cumò	AV	112	50
100	su la	su	PREP	112	43

Ordin	**Leme**	**Etichete**	**Ocorencis**	**Tescj**
1	di	PREP	8287	144
2	lui	PRONP	3424	138
3	e	CONIUN	3115	143
4	il	ART	2675	140
5	a	PREP	2390	136
6	la	ART	2368	142
7	lôr	PRONP	2043	136
8	che	CONIUN	1934	132
9	jessi	VINTR	1910	132
10	jê	PRONP	1636	135
11	che	PRON	1576	136
12	par	PREP	1566	132
13	un	ART	1457	133
14	in	PREP	1303	131
15	vê	VAUX	960	115
16	ta	PREP	908	111
17	jessi	VAUX	907	117
18	cun	PREP	902	111
19	vê	VTR	865	109
20	jo	PRONP	847	44
21	si	PRONP	604	107
22	no	AV	587	93
23	ma	CONIUN	579	94
24	fâ	VTR	563	95
25	ancje	AV	537	98
26	chel	PRON	537	97
27	chest	ADI	480	95
28	nô	PRONP	467	58
29	plui	AV	436	97
30	so	ADI	436	78
31	su	PREP	429	99
32	dî	VTR	401	68
33	chel	ADI	398	73
34	dut	ADI	351	74
35	lâ	VINTR	346	70
36	podê	VINTR	325	82
37	se	CONIUN	315	63
38	altri	ADI	291	77
39	tu	PRONP	265	29
40	o	CONIUN	260	66
41	lenghe	SF	238	39
42	vignî	VINTR	232	71
43	an	SM	214	66
44	dome	AV	205	66

45	savê	VTR	205	52
46	dut	PRON	184	50
47	viodi	VTR	179	61
48	furlan	ADI	175	38
49	speziâr	SMF	174	1
50	come	CONIUN	170	49
51	dâ	VTR	169	61
52	simpri	AV	169	60
53	ce	PRON	166	54
54	fûr	AV	162	75
55	stâ	VINTR	161	55
56	meti	VTR	159	67
57	tant	AV	154	57
58	volê	VTR	154	44
59	cuant	AV	150	52
60	Friûl	SM	149	41
61	gno	ADI	149	30
62	chest	PRON	145	46
63	doi	ADI	143	57
64	rivâ	VINTR	142	57
65	cence	PREP	137	48
66	sù	AV	136	65
67	po	AV	133	46
68	cjase	SF	132	39
69	come	AV	132	45
70	furlan	SM	132	20
71	volte	SF	132	58
72	cjapâ	VTR	129	58
73	Sgobar	SMF	129	1
74	cualchi	ADI	128	48
75	voaltris	PRONP	128	19
76	ben	AV	125	46
77	câs	SM	121	35
78	prin_a	ADI	120	46
79	femine	SF	119	19
80	siôr	SMF	117	10
81	vie	AV	117	47
82	Udin	SM	116	34
83	là	AV	113	46
84	nestri	ADI	113	33
85	plevan	SM	113	7
86	un	PRON	113	39
87	cumò	AV	112	50
88	ancjemò	AV	109	47
89	cussì	AV	109	50
90	ogni	ADI	108	46
91	leç	SF	106	18
92	man	SF	106	39
93	vignî	VAUX	105	38
94	vore	SF	105	49
95	za	AV	105	37
96	bon	ADI	103	37
97	Lucrezie	SF	103	1
98	cui	PRON	101	32
99	li	AV	101	49

100	lôr	ADI	101	38

Limitantsi ae analisi di chestis formis e di chescj lemis, si pues dî che di une bande si cjate une conferme sostanziâl di ce che al jere za stât publicat tal 2002, tal sens che si viôt che preposizions, coniunzions, articui, pronons personâi, averbis, verps ausiliârs, verps iregolârs (*fâ*, *dî*, *lâ*, *podê*, *vignî*, *savê*, *dâ*, *stâ*, *volê*...) a son simpri tra i prins puescj. Une diference e je che l'articul *il*, al contrari de ricercje de SSTF, al risulte cumò pôc plui frecuent dal articul *la*: la reson e je che *il* (1280 ocorencis) e *l'* (474 ocorencis) a son stâts zontâts e calcolâts une sole forme, invezit tal 2002 a jerin calcolâts dôs formis diferentis; cun di plui, stant che i tescj no jerin stâts corets a nivel ortografic, sot de forme *la* a jerin sedi l'articul definît feminin, sedi il pronon personâl di tierce persone singolâr feminine tant che complement diret, che daûr des regulis de lenghe standard al sarès *le*.

Tra lis peraulis no gramaticâls si confermin presintis tra lis primis 100, sedi te ricercje dal 2002, sedi in cheste, *lenghe* (n. 56 tes formis, n. 41 tai lemis) e *furlan* (n. 85 tes formis e n. 48 tai lemis): segnâl che tai scrits in lenghe furlane cjapâts in esam la lenghe e la identitât a son argoments une vore dibatûts.

Rispiet al 2002 invezit al cole fûr de liste des primis 100 *an*: tal 2002 però al veve di jessi tra lis primis formis dome parcè che no si distingueve tra il sostantîf *an* e la tierce persone plurâl dal verp *vê*, che di fat, cu la scriture *àn*, e je in posizion n. 75 te tabele atuâl des formis (par altri il sostantîf *an*, come leme, pal plui in graciis dal plurâl *agns*, al è il leme n. 43).

A jessin des primis 100 formis, ma cuntune diference avonde limitade, ancje *peraule* (che però come leme e je tal n. 116, duncje no tant lontane de posizion documentade tal 2002), *mont* (che tal 2002 però e sumave lis formis di sostantîf masculin, sostantîf feminin e adietîf), *vite* (forme n. 192 e leme n. 166), *tiere* (forme n. 198 e leme n. 136), *om* (forme n. 241 e leme n. 128), *timp* (forme n. 153 e leme n. 105).

Invezit diferencis plui impuartantis a segnalin disecuilibris problematics, sedi te base di dâts dal 2002 sedi tal corpus che si presente cumò: tal 2002 tra lis primis 100 formis a risultavin *Signôr* (tal puest n. 24), *Diu* (n. 28) *Gjesù* (n. 70), *vanzeli* (n. 100). Cumò *Gjesù* nol è nancje documentât tai tescj cjapâts in esam e chês altris formis a son dopo de posizion n. 1.000. Une diference cussì grande e je la spie che i tescj di caratars religjôs (*Lezionari pes domeniis e pes fiestis* e *Sants, madonis e meracui*) a vevin pardabon masse pês percentuâl sul complès, fasint dam a une buine rapresentativitât.

Par cuintri tal corpus di prove di cumò, par vie de sielte za motivade di etichetâ *Int di masnade* di Alviero Negro, a son formis che a perturbin insot l'ecuilibri dal corpus, di fat i nons di persone *Sgobar* e *Lucrezie* e il sostantîf *speziâr*, ducj tai prins 100 lemis, in realtât a son documentâts dome ta chel toc teatrâl. Pe stesse reson ancje *plevan*, cundut che al è presint in altris 6 tescj, al à un pês sproporzionât pal fat di jessi un dai personaçs di *Int di masnade*.

Rispiet al 2002 si pues fâ cualchi altre considerazion statistiche: tal corpus, formât di 100.970 a son documentâts 8.053 lemis furlans, lis formis invezit a son 13.520, lis formis no furlanis (pal plui nons propris, o ancje citazions o elements varis in altris lenghis e v.i.) a son 1.317.

Une altre aplicazion interessante, cundut che ancjemò lontane di vê une significance sientifiche, e je stade chê dal confront di frecuencis cui lemis cul GDBTF: chest dizionari di fat al segnale, cjapantlis dal GDU di Tullio De Mauro, lis marcjis di ûs dai lemis talians.

Lis primis dôs fassis individuadis di Tullio De Mauro (FO, vâl a dî lemis fondamentâi, e ÛA, vâl a dî lemis di ûs alt) a son stadis otignudis cuntune osservazion di frecuence statistiche e a cuvierzin tor dal 96% dal ûs dal talian scrit e fevelât: daûr dai studis fats in plui lenghis, a son simpri plui o mancul 5.000 i lemis che a cuvierzin chê percentuâl tai scrits e tai discors.

Par fâ un esperiment si à fat une liste dai lemis furlans che a son presints tal GDBTF tant che ecuivalents dai lemis talians FO e ÛA: cheste liste e je risultade di 6.621 lemis furlans[13].

Incrosant cheste liste cui dâts dal corpus, si viôt che lemis furlans corispondents a talians FO e ÛA a cuvierzin tor dal 91% dal ûs. Chest valôr, plui bas rispiet al 96% dal fondamentâl e di alt ûs de lenghe taliane, al pues cjatâ une reson tal fat che i studis sul talian si son fondâts sedi sul scrit che sul fevelât. Invezit il corpus furlan di prove al è componût dome di tescj scrits, che in gjenerâl a àn une ricjece e variabilitât lessicâl plui alte che no la espression orâl, ridusint il pês dai lemis plui doprâts a favôr di chei altris.

[13] Il fat che chest numar al sedi plui alt di 5.000 al derive dal fat che par cuvierzi a plen i cjamps semantics di un leme talian, i lessicografs, soredut dulà che la ecuivalence no jere perfete, a àn vût miôr di proponi plui pussibilitâts cirint la completece, stant che, mancjant dâts sigûrs su la frecuence dal lessic furlan, no podevin abadâ a chest aspiet.

In ogni mût chescj prins risultâts a àn un valôr a pene dimostratîf, par vie de dimension dal corpus e dal fat che, cundut dal sfuarç in chest sens, la rapresentativitât no je garantide: risultâts plui significatîfs a podaran vignî cun cuintriprovis plui sistematichis e cu la cressite cuantitative e cualitative dal corpus etichetât.

7. Prospetivis

Cui struments che a son za pronts e in vore si pues previodi di rivâ a vê un corpus etichetât avonde grant (cualchi milion di peraulis) tal zîr di pôcs agns e cuntune precision e sistematicitât che no si podaressin garantî cence il tratament automatic dai tescj.

Di chê altre bande al covente fat un lavôr avonde profont par:

- perfezionâ l'etichetadôr semiautomatic, in particolâr te funzion di disambiguazion, ancje se i risultâts di cumò a son za avonde bogns;
- zontâ une funzion par tratâ ancje tescj scrits in varietâts e grafiis diviersis di chês di riferiment;
- completâ il program di gjestion e di ricercje: a coventin tescj descritîfs dal program, formulis di ricercje plui semplice par utents pôc esperts, ma ancje pussibilitâts di ricercje plui complesse e plui fine par specialiscj.

8. Conclusions

Dopo de realizazion dal *Grant Dizionari Bilengâl Talian-Furlan*, dal Coretôr Ortografic Furlan, dal tradutôr automatic Jude, la gnove sfide de linguistiche computazionâl aplicade ae lenghe furlane e je chê di un grant corpus etichetât. I struments par fâlu, ancje se si à di perfezionâju, a son za e, cemût che al jere sucedût altris voltis, a son nassûts par iniziative privade.

Di chê altre bande la impuartance di un corpus etichetât e la dibisugne che al puedi jessi doprât in maniere libare, in gjenerâl di ducj e in particolâr de comunitât sientifiche, e ancje la mancjance di un marcjât che al permeti di sostignî lis spesis di une imprese cussì grande, a fasin che chest progjet al vedi di deventâ public, stant che al è dal sigûr une des prioritâts de politiche linguistiche pe lenghe furlane.

Bibliografie

AA. VV., *Grant Dizionari Bilengâl Talian-Furlan*, ARLeF, Udin, 2011.

Brill E., *A Simple Rule-Based Part of Speech Tagger*, in *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, Trento, 1992

Burelli A., Miculan. M., *Frecuencis lessicâls dal furlan scrit*, in "Gjornâl Furlan des Siencis", 1, 2002.

Calzolari N., *Linguistica Computazionale e Risorse Linguistiche*, in Cicchese G., Pettorossi A., Crespi Reghizzi S., Senni V. (per cura di), *Scienze informatiche e biologiche. Espistemologia e ontologia*, Sefir Città Nuova, Roma, 2011.

Marinelli R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A., *The Italian PAROLE corpus: an overview*. In A. Zampolli, N. Calzolari, L. Cignoni, (eds.). In Linguistica Computazionale, Special Issue, XVI-XVII, 2003. Pisa-Roma, IEPI. Tomo I, pp. 401-421.

De Mauro T. (a cura di), Grande Dizionario dell'Uso In Linguistica Computazionale

Petkevič V., *Corpus lenghistics*, in "Gjornâl Furlan des Siencis", 1, 2002.

Picco L., *Ricercje su la condizion sociolenghistiche dal furlan-Ricerca sulla condizione sociolinguistica del friulano*, Forum, Udine, 2001.

Tamburini F., *Annotazione grammaticale e lemmatizzazione di corpora in italiano, Linguistica e informatica: multimedialita', corpora e percorsi di apprendimento*, Rossini Favretti R. (a cura di), Bulzoni, Roma, 57-73.

Tamburini F., *PoS-tagging Italian texts with CORISTagger. In Proc of EVALITA 2009. AI*IA Workshop on Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December 2009.

Riferiments web:

ctilc.iec.cat
gattoweb.ovi.cnr.it
www.claap.org
www.dizionariofriulano.it
www.ge.ilc.cnr.it/

www.ilc.cnr.it/EAGLES96/browse.html#wg2