Il corpus etichetât de lenghe furlane: risultâts e prospetivis

Sandri Carrozzo, Franz Feregot,
Teresa Suñol Ribas

1. Presentazion gjenerâl

Chest articul al presente la costruzion di un corpus scrit in lenghe furlane. Un corpus al è une racuelte di tescj, o di files audio tal cas di corpus orâi, che a son analizâts in diviers nivei: morfologjic, semantic, sintatic o a nivel di discors. Si pues dâ dongje i corpus cun finalitâts diviersis: par esempli par resons etnografichis, linguistichis o sociolinguistichis.

Il corpus che si presente chi al à finalitâts linguistichis e al met adun tescj in lenghe furlane che a son stâts etichetâts a nivel morfosintatic.

Il fat di vê un volum di tescj no di pôc, soredut par jessi une lenghe minorizade, al vierç la pussibilitât di fâ un grant salt dentri di diviersis areis di studi e lavôr: la elaborazion di un corpus etichetât, e je un passaç di fonde par fâ ricercjis statistichis, linguistichis e leterariis e par fâ o perfezionâ altris prodots tant che dizionaris gjenerâi e setoriâi, tradutôrs automatics, coretôrs ortografics, programs di completament di digjitazion, ricercje des informazions, bibliotechis digjitâls e une vore di altris. Ducj i linguiscj computazionâi a son in cunvigne su la centralitât dai corpus etichetâts tal tratament des lenghis.[1]

[1] Cf. Petkevič 2002, Brill 1992, Burelli/Miculan 2002, Calzolari 2011, Marinelli et al. 2003.

Ladinia XLI (2017), 131–142.
ISSN 1124–1004; © Istitut Ladin Micurá de Rü, San Martin de Tor (BZ).

Tal corpus etichetât de lenghe furlane, cemût che al è stât fat tai ultins agns, a ogni peraule si lee une etichete, vâl a dî une informazion formade, tal nestri câs, di trê elements: ce leme che al è, ce part dal discors che e je, ce categorie gramaticâl che e à.

Par esempli la forme *è* e podarà vê la etichete *jessi/VAUX/IP3: cheste informazion e vûl dî che è al* ven dal leme *jessi*, che al è un verp ausiliâr (*VAUX*), che al è un indicatîf presint di tierce persone singolâr (*IP3*).

Par aplicâ lis etichetis al ven doprât un program specific, che al à dentri une grande liste di lemis e lis regulis par produsi dutis lis lôr formis; cun di plui al à sistemis di disambiguazion par cirî di dâ in automatic la etichete juste. Di fat a son formis li che la etichete e pues jessi dome une, par esempli *puarts* al pues vê dome *puart/MP/MS*. In dut câs une vore di peraulis (par furlan il 70% dai câs), ancje di chês che si doprin plui dispès, a podaressin vê plui di une etichete: par esempli *lis* al pues jessi *la/ART/FP*, duncje *la*, articul definît feminin plurâl, o ancje *lôr/PRONP/OA6f*, vâl a dî *lôr,* pronon personâl, di tierce persone plurâl feminine in funzion di obiet aton, e ancje *lis/ADI/MS* e *lis/ADI/MP*, vâl a dî masculin singolâr o plurâl dal adietîf *lis*, variante di *slis*. Tun câs come *lis* l'etichetadôr al met in vore regulis che a cjalin il contest: par esempli se dopo al è un verp, al è sigûr che *lis* al sarà pronon, se dopo al è un sostantîf o un adietîf feminin plurâl, al è sigûr che *lis* al sarà articul.

Lis regulis di disambiguazion a funzionin cun plui o mancul precision daûr di trop complès che al è il test, ma ancje daûr dal nivel di detai des etichetis che si decît di meti. Par esempli, il secont cjamp, chel de part dal discors, al podarès vê dome la informazion */V/* vâl a dî "verp", o ben al pues lâ plui insot, indicant se il verp al è ausiliâr */VAUX/*, modâl */VMOD/*, transitîf */VTR/*, intransitîf */VINTR/*... Plui che la definizion e va insot tal descrivi la lenghe, plui che al è dificil che il program automatic al eticheti just.

Cun di plui al è di savê che ciertis decisions a restin arbitrariis in ogni câs: par dî un esempli la distinzion jenfri adietîfs e participis in ciertis situazion e je une vore discutude e discutibile, e no dome pal furlan ma ancje in chês altris lenghis neolatinis.

L'etichetadôr cumò al funzione cuntune version che e ricognòs 75.000 lemis che si fletin in plui di 8.700.000 formis pussibilis, rispuindint a plui di 300 paradigmis flessionâi diviers (cjapant dentri ancje i lemis iregolârs), cun 26 pussibilitâts tal cjamp de part dal discors (in plui di preposizions, averbis, coniunzions,

interiezions, fonosimbui, adietîfs, e des diviersis sortis di sostantîfs, di verps e di pronons, a son ancje etichetis specifichis pe puntuazion e par elements tant che numars in cifris, leams web, direzions di pueste eletroniche...) e passe 1.200 pussibilitâts tal cjamp di categorie gramaticâl (il numar al è cussì alt soredut par vie de enclisi pronominâl tai verps).

Il program automatic che al etichete lis formis al è stât fat de *Serling* soc. coop.[2], su la fonde dal materiâl che al jere za stât prontât pal tradutôr automatic *Jude3*[3]. Pal moment al rive a etichetâ in maniere corete plui dal 85% des peraulis di un test: se i tescj a àn une sintassi no masse complesse, par esempli tescj gjornalistics, al pues rivâ parsore dal 90%. Ancje dome cjalant i numars des pussibilitâts di etichetadure si pues provâ a imagjinâ la complessitât dai struments di linguistiche computazionâl e si capìs che cence un program automatic o semiautomatic al sarès masse lunc e dificil etichetâ tescj.

Se cemût che si à dite il nivel di precision al è tor dal 85–90%,[4] par rivâ al 99% (la coretece di etichetadure dal 100% tai corpus e je un obietîf utopistic) al covente l'intervent manuâl di un linguist. Chest lavôr al è impegnatîf e, a diference di ce che al jere sucedût pal coretôr COF ("Coretôr Ortografic Furlan") e pal tradutôr automatic *Jude3*, nol da une pussibilitât di jentradis economichis che e justifichi un impegn privât, almancul te realtât furlane.

Par un tant la *Serling*, dopo di vê disvilupât la tecnologjie che e coventave, scuasi 10 agns indaûr, e veve presentât il progjet ae *Agjenzie Regjonâl de Lenghe Furlane* (ARLeF), che però no lu veve capît o calcolât interessant. Cussì chest progjet al è restât fer, o disvilupât in piçule misure e in pierdite economiche, fin al 2014 cuant che a son mudadis lis cundizions politichis e la ARLeF e à finanziât il progjet, puartât indevant di chel moment dal *Centri di Linguistiche Aplicade "Agnul Pitane"* (CLAAP)[5].

[2] La *Serling* e je une societât cooperative costituide tal 2001, impegnade in servizis linguistics, in particolâr traduzion, revision di tescj, insegnament, editorie. E à disvilupât il tradutôr automatic talian–furlan *Jude*. Viôt <www.serling.org>.

[3] Il tradutôr automatic *Jude3* al funzione cuntun sisteme a trasferiment superficiâl, che al analize la morfologjie taliane e al cognòs chê furlane (cf. Carrozzo/Feregot/Mistrut 2010).

[4] Tal ultin an di lavôr il sisteme al è za stât perfezionât e i algoritmis a garantissin une coretece medie dal 92% calcolade su tescj di plui gjenars e di varie complessitât.

[5] Il CLAAP al è une societât cooperative che si è costituide tal 2012 in consecuence de crisi e dissoluzion dal *Centri Friûl Lenghe* 2000 (CFL 2000), che al jere un consorzi che par agns al jere stât l'interlocutôr des struturis regjonâls pe lenghe furlane, par ce che al tocjave la lessicografie. Il CLAAP al à dât dongje i ex

Il progjet, rivât al cuart an (un di finanziament isolât e trê di progjet trienâl), al vûl dâ dongje un corpus etichetât di 500.000 peraulis, cun precision dal 99%. I timps di realizazion a son stâts rispietâts dal CLAAP, e cumò si à za un corpus che si pues consultâ in rêt di 380.000 peraulis.

2. Sisteme di lavôr e carateristichis dal corpus

2.1 Selezion dai tescj

Il corpus etichetât furlan al è stât metût adun cun chestis carateristichis:

1. dome tescj di prose;
2. belançament jenfri tescj cun finalitâts artistichis e cence finalitâts artistichis, e jenfri diviers gjenars, argoments e autôrs;
3. in consecuence dal pont 2., pal plui tescj curts;
4. tescj cuntune buine coretece de lenghe furlane;
5. tescj de fase contemporanie de lenghe furlane (tacant de fin dal secul XIX a vuê, cun prevalence di tescj dai ultins 60 agns);
6. tescj scrits doprant la coinè;
7. tescj origjinâi (no traduzions).

Chescj ponts a van daûr di chês che a son liniis ricognossudis a nivel gjenerâl[6] par vê corpus che a sedin "rapresentatîfs", cundut che a àn vût di jessi justâts ae realtât tipiche de produzion scrite par furlan.

Il belançament jenfri tescj cun finalitât artistiche (narative, teatri...) cul 37% e no artistics (gjornalisim, comunicazion, divulgazion sientifiche...) cul 63% no à dal sigûr chês stessis proporzions di corpus di altris lenghis europeanis, dulà che i tescj leteraris a son tor dal 15%, o ancje di mancul. Un tant al derive dal fat che la produzion scrite par furlan e je limitade e che te informazion, divulgazion e ricercje su argoments diferents e je plui scjarse di ce che e je in altris lenghis, massime se si cîr tescj cun buine cualitât espressive e che no sedin traduzions. Par altri al è di tignî cont che al è pussibil che cierts tocs jentrâts tal corpus a sedin traduzions dal talian, se te publicazion cheste carateristiche no jere segnalade.

colaboradôrs dal CFL 2000 ancjemò disponibii a meti lis lôr competencis a disposizion di progjets di linguistiche aplicade pe lenghe furlane.

6 Cf. ancjemò une volte Petkevič 2002.

La sielte di tescj curts e scugnive jessi se si voleve mantignî un ecuilibri tra gjenars e autôrs[7]. Par esempli si varès podût etichetâ dut il romanç *L'aghe dapît la cleve* di Dino Virgili (1979[2]), che al à une cualitât di espression par furlan une vore alte, ma la sô dimension di scuasi 120.000 peraulis e varès scuilibrât dal dut, suntun gjenar, suntun test e suntun autôr, un corpus che al varès vût di vê "dome" 500.000 peraulis.

Par altri la pocje fiducie te continuitât dal finanziament (mâl cronic de politiche linguistiche pe lenghe furlane) e conseave di resonâ in principi no suntun corpus che al fos rapresentatîf dome te sô completece di 500.000 peraulis, ma za in ognidune des fasis intermediis. In cheste maniere ancje se al fos restât interot nol varès vût problemis di belançament. Cussì cuasi il 74% dai tescj che cumò a son tal corpus a son di mancul di 1.000 peraulis, il 52% di mancul di 500 peraulis.

Scuasi ducj i tescj a son di dopo de metât dal secul XX: al è di chest moment che la coinè furlane e tache a cjapâ plui uniformitât, cuntune progression che e je deventade simpri plui svelte e massive tai ultins 20 agns.

2.2 Tratament dai tescj

In gjenerâl lis contis che a fasin part di racueltis publicadis intun volum sôl a son stadis tratadis tant che tescj autonoms, un par document: il titul dal volum dulà che a son stadis publicadis al è segnât intal cjamp notis; al contrari, scrits luncs ma cun continuitât narative o logjiche, ancje se dividûts in plui cjapitui, a son stâts mantignûts tun document sôl.

I tescj che a jerin dome stampâts te cjarte (21% dal volum totâl) si à vût di scansionâju, passâ lis imagjins par un program di ricognossiment di test che al sveltìs la trascrizion e po dopo a son stâts corets, par vie che i programs di cheste sorte a zovin, ma no son perfets, e dispès i tescj furlans a puedin jessi in grafiis diviersis di chê uficiâl o vê erôrs di batidure. Altris tescj a son stâts convertîts dai formâts pdf (12% dal volum totâl): in chest câs il tratament al è plui lizêr che no la scansion de cjarte, ma in ogni câs al covente un passaç par recuperâ la struture dal test, par vie che tal pdf al pues sucedi che ogni rie e vegni dividude di chê altre e cierts caratars a puedin pierdisi.

[7] Cf. Carrozzo/Feregot 2012, 241–262.

La conversion di tescj cirûts tal web o za in formât .doc, .rtf, .odt o altris (67% dal volum totâl) invezit e je chê plui svelte. Par jentrâ tal program di etichetadure ducj i scrits a àn di jessi metûts in formât .txt, po a vegnin impuartâts te base di dâts e convertîts in documents cun formât .ann (test cun anotazions). Ogni document txt al à une intestazion là che si segne autôr, curadôr, titul, an, edizion, varietât e si puedin zontâ notis. Ducj i tescj a son stâts voltâts te grafie uficiâl: tancj dai tescj plui resints a jerin za scrits te grafie uficiâl, ma a podevin vê fai di batidure, duncje a àn vût bisugne ancje chei di une revision ortografiche. I tescj a son stâts selezionâts jenfri chei za scrits te coinè (furlan standard) ma chest nol gjave che a sedin elements di variabilitât, plui di dut tai tescj plui vecjos o di autôrs mancul precîs. Si puedin fâ trê esemplis:

- daûr di ce che al sucêt in diviersis varietâts, si pues cjatâ alternance di formis tal pronon aton subiet di tierce persone plurâl *(lôr a son/lôr e son)*, li che il standard al à finît par favorî la forme *a*;
- il standard definît di Xavier Lamuela (1987, 36) tai agns ’80 al acetave dutis dôs lis formis interogativis par *al è = esal?/isal?*, po te scriture de coinè esal al è sparît in maniere spontanie, ma al jere la forme plui doprade di diviers grancj autôrs dai agns ’50 – ’60 – ’70 dal secul passât, tra chei Josef Marchet, che al è calcolât un pôc il pari de coinè;
- il diftonc *ie/je* derivât di Ĕ latin, a pene prime di un grup consonantic che a tacave par r, te coinè al risultave pal plui *ia/ja (jarbe, fiar, piardi, tiare...)* daûr dal model centrimeridionâl. Al jere doprât in maniere sistematiche tant che de forme de coinè ancje di scritôrs che dal sigûr te lôr varietât a varessin scrit ie *(jerbe, fier, pierdi, tiere...)*. Po cu la normalizazion di Lamuela, fate buine par leç tal 1996, la forme *ie/je* e *je* deventade chê normative.

Par che l’etichetadôr al puedi funzionâ in maniere automatiche, il fat che tai tescj a sedin alomorfs, massime chei no acetâts dal standard, al è un probleme. Par un tant, scuasi ducj i tescj a son stâts tratâts, a nivei diferents daûr di ogni câs.

Par tescj di autôrs che a dopravin cun cussience la coinè, pe plui part cu lis formis che a corispuindin al standard di cumò, ma cun pôcs elements di diference (par esempli, cemût che si à viodût parsore, pronon aton subiet plurâl *e* invezit di *a*, o articul definît feminin *le* invezit di *la* o altris) e je stade fate une uniformazion su lis formis dal standard: chest intervent al è segnalât tal cjamp notis de intestazion dal document.

Si previôt un altri tratament pai tescj li che e je une fase de coinè pardabon divierse di chê di cumò e pai tescj che a sedin scrits in varietâts locâls: si fasarà

une version, che e sarà chê viodude dal utent, li che si intervignarà dome a nivel di ortografie (es. *l'âga* > *la aga*); po si fasarà une altre version, che e sarà chê doprade dal program pal processament, li che ogni peraule furlane e varà la forme standard (es. *la aga* > *la aghe*), ma cence intervents a altris nivei, par esempli a chel sintatic. Al è clâr che cheste version dople e compuarte ancje un lavôr dopli e une competence specifiche tal cognossi lis formis des varietâts locâls.

Soredut pai tescj di dopo dal 1996, li che l'autôr al à sielzût di doprâ grafie uficiâl e coinè, a son stâts fats intervents no dome di revision ortografiche, ma ancje di uniformazion dal lessic tal câs che l'autôr al dopri in maniere no cussiente formis locâls o formis cuntune standardizazion imperfete: la stesse linie tignude in cheste azion e je chê di une normâl corezion di stampons, che dispès e mancje tal mont editoriâl furlan. Si pues duncje fâ une liste dai intervents fats e no fats sui tescj.

- E je stade fate une uniformazion sul standard di variantis locâls tant che *cjariesie, çariese, çariesie, cjarese... > cjariese; Unvier > Invier; Astât, Estât > Istât.*
- E je stade fate une uniformazion al standard in peraulis li che in timps plui o mancul resints si è slargjade une forme contaminade dal talian: *treno > tren; freno > fren; moto* (tal sens di moviment des mans, dal cjâf e v.i.) > *mot; estro > estri* ...
- Tal câs dai averbis tant che *veramentri > verementri* la norme e recupere un compuartament arcaic che si jere pierdût: tai tescj plui resints cheste uniformazion al standard e je fate tant che revision normâl, cence nancje segnalazion in note.
- E je stade fate une uniformazion sul standard di variantis morfologjichis tant che *vorès > volarès; glutì > glotè* ...
- A son stadis mantignudis lis sieltis puntuâls dai autôrs, fatis in maniere coerente o volontarie, di slontanâsi des formis ortografichis o lessicâls uficiâls, par vê cierts efiets: par esempli doprant singulis formis locâls par dâ un caratar dialetâl o scrivint peraulis in grafie particolâr (par esempli cun latinisims, esotisims, arcaisims...).
- No son stâts fats intervents di corezion su la sielte lessicâl jenfri plui alternativis, cemût che al è stât fat invezit in pocjis edizions ipercoretivis (par esempli *vecjo > vieri; curnîs > suaze; stranîr > forest; dopo > daspò*).
- No si à coret nancje contaminazions percepidis tant che faladis ancje dal locutôr comun (*azûr* par *celest*, *talpe* par *farc*...), ma in cheste fase si à cirût di selezionâ tescj cuntune cualitât linguistiche avonde alte di no vê dentri câs dal gjenar.
- No e je stade mudade la sielte dai autôrs di doprâ prestits arûts dal talian o di altris lenghis, tant che element espressîf o popolâr, massime tai discors direts.
- No son stâts fats intervents di corezion su la sintassi.

- Te conversion dai tescj intal formât .txt al è pussibil che a sedin stâts deformâts o pierdûts caratars modificâts tipics di lenghis forestis (*æ, ä, å, č, ñ*...), doprâts in citazions, elements onomastics o singulis peraulis.
- Te riproduzion dai tescj in gjenerâl si à trasformât o eliminât i elements no testuâi no compatibii cul formât .txt, tant che grafics, tabelis, imagjins e didascaliis. E je stade eliminade ancje la sintesi introdutive o finâl, se e jere, intai articui specialistics.

In conclusion di chest paragraf si pues dî a clâr che l'intindiment dal corpus, in cheste fase, al è chel di dâ un imprest pe analisi de lenghe furlane e no une riproduzion filologjiche dai tescj. A cui che al à interès specific par chescj elements si racomande di consultâ i origjinâi e no la lôr riproduzion tal corpus.

3. Resons e finalitâts dal corpus etichetât

Lis sieltis di selezion e di tratament dai tescj no son assoludis e indiscutibilis, ma a van daûr dai imprescj che si veve za in man e des finalitâts che si previodeve. Za dal 2001 la lenghe furlane e à a disposizion un coretôr ortografic[8], prodot par iniziative privade, ma daûr dal standard uficiâl: chest program al à vût diviersis evoluzions che infin a àn puartât a fâ ancje, tant che efiet colaterâl no previodût, i cuadris flessionâi dal *Grant Dizionari Bilengâl Talian–Furlan*[9] e il tradutôr automatic talian–furlan *Jude3*.

Su la stesse evoluzion di materiâl e competencis al è nassût l'etichetadôr dal corpus, che alore, pe sô divignince, al è un imprest adatât a tratâ la coinè, il furlan standard. Si à di notâ par altri che la coinè e je fissade (de norme legjislative, dal ûs, de elaborazion...) a nivel di grafie e morfologjie; a son za grancj repertoris lessicâi, dal secul XIX incà; invezit e je une vore mancul studiade e definide a nivel di sintassi, regjistris e stîi.

Une des primis finalitâts alore e je chê di servî la coinè tai cjamps che no son ancjemò ben studiâts e codificâts, cjapant fuarce dai ponts za salts. Dâ adun e ordenâ cui criteris che o vin viodût une grande racuelte di tescj al permetarà di:

[8] Il COF, prodot di *Informazione Friulana* soc. coop., al è stât cuistât in plui versions des aministrazions publichis e cumò al è dât fûr de ARLeF, viôt <www.arlef.it/struments/coretor-ortografic-furlan>.

[9] I cuadris flessionâi a jerin une funzion, in principi no previodude e zontâde fûr di progjet, in dotazion tal CD-Rom dal CFL 2000. Vuê si puedin consultâ in <www.claap.org>.

- cjatâ lemis patrimoniâi che a jerin scjampâts des racueltis fatis cun sistemis no automatizâts;
- individuâ une buine cuantitât di neologjisims;
- judâ te ricercje su la semantiche, cun risultâts une vore svelts sui ûs di ogni singul leme o di secuencis di lemis;
- fâ di miniere fraseologjiche par oparis lessicografichis;
- judâ la ricercje su la sintassi furlane e la sô codificazion;
- fâ di miniere par eserciziaris, imprescj didatics, antologjiis...

Se si decidarà di invistî sul tratament ancje di tescj no in coinè cul sisteme de dople version, visualizazion in origjinâl e processament cun formis standard, il corpus al podarà ancje:

- judâ il studi leterari, ancje individuant lis frecuencis lessicâls intai diviers autôrs;
- judâ il studi dai ûs scrits, ancje individuant lis primis atestazions di un leme, o la fase storiche li che al è stât doprât;
- fâ di sisteme di gjestion di bibliotechis digjitâls;
- cun gnovis funzions judâ la ricercje su la dialetologjie furlane.

Cun di plui un imprest di linguistiche computazionâl di chest pês al varà efiets su la incressite des competencis e su la realizazion di altris prodots, che no si rive nancjemò a previodi.

La cjadene di efiets colaterâi positîfs, suntun progjet di cheste puartade, nol è di lassâ di bande: al baste pensâ che, simpri te linguistiche computazionâl furlane, in bande dai lavôrs leâts al *Grant Dizionari Bilengâl Talian–Furlan* a son nassûts par iniziative privade COF, DOF,[10] cuadris flessionâi, *Jude3* e l'etichetadôr stes.

4. Fase di cumò e prospetivis

4.1 Suaze economiche e politiche

Te fase di cumò si è rivâts a un corpus, belançât e etichetât cuntun alt nivel di coretece, formât di 378 tescj par un totâl di 380.000 peraulis. L'obietîf al è di rivâ tal 2017 a 500.000 peraulis ma son stâts ritarts burocratics e aministratîfs di 9 mês.

[10] Il DOF al è il *Dizionari Ortografic Furlan–Talian/Talian–Furlan*, prime realizât di *Informazione Friulana* soc. coop. In version digjitâl, po rielaborât e publiât ancje in cjarte (Carrozzo 2008).

Un dai problemis principâi dal progjet al è chel di dipendi de politiche linguistiche pe lenghe furlane, che e à patît une involuzion une vore negative. E je une complicazion burocratiche simpri plui alte, gjenerâl des istituzions te realtât taliane, ma ancje cun criticitâts specifichis de *Agjenzie Regjonâl pe Lenghe Furlane*.

A son scuasi 10 agns che la ARLeF e à rinunziât a jessi propositive e a cjapâ strategjiis claris, rivant al massim a tignî une posizion ricetive, che invezit di indreçâ e acete a posteriori chê che e je la capacitât produtive private. Il sisteme di finanziaments publics su la lenghe furlane nol rive a garantî une continuitât serene a nissun progjet plurienâl: la consecuence sul progjet dal corpus e je che ancje se al è stât finanziât, tai ultins trê agns, al à vût i fonts cun intervai e ritarts luncs: par un tant i lavoradôrs impegnâts a àn vût lungjis dadis di timp li che o no àn lavorât o a àn lavorât cence paie.

Di là de situazion politiche e burocratiche, al reste in ogni câs dificil fâ capî la impuartance dal progjet: cu la cressite tecniche, i lavoradôrs specializâts a son rivâts a produsi imprescj che a van di là de capacitât di comprension imediade o dal interès de plui part dai responsabii de (mancjade) planificazion linguistiche. La dificoltât comunicative jenfri tecnics e ent finanziadôr e cres ancje par vie che al è dificil o impussibil fâ previsions cuantitativis su la utilitât dal progjet. I finanziadôrs a podaressin domandâ cun dut il dirit: trops lemis patrimoniâi che a jerin restâts fûr dai repertoris lessicâi si cjatarano ogni 100.000 peraulis etichetadis? Trops neologjisims? In ce percentuâl di timp sarano plui sveltis lis ricercjis tai diviers cjamps? Parcè rivâ a une prime fase di 500.000 peraulis etichetadis e no 400.000 o 600.000? Chestis a son domandis che no àn e no puedin vê nissune rispueste tecniche prime che il lavôr nol sedi finît e scrutinât.

Alore se la cuistion de funzionalitât strategjiche dal progjet e de sostignibilitât dai siei coscj no cjate rispuestis tecnichis, la sielte e devente politiche: valie la pene di invistî ancjemò su chest setôr, puartant il furlan tra lis lenghis che a àn ancje un futûr digjitâl, o isal miôr fermâsi e rinunziâ?

Al reste il fat che nissun element tecnic al pues solevâ la responsabilitât personâl di cui che al à di cjapâ decisions politichis. Magari cussì no, si è intun moment (par altri za miôr cumò che no agns indaûr) dulà che la zonte regjonâl e à disinterès o contrarietât rispiet ae lenghe furlane, inte ARLeF a mancjin la fuarce politiche, la capacitât di cjapâsi responsabilitâts, e al è plui facil cjatâ problemis che no soluzions.

4.2 Prospetivis tecnichis

Se l'avignî dal progjet, par ce che al dipent de politiche, al è ancjemò te fumate, a son sperancis une vore buinis a nivel tecnic. Lis prospetivis concretis a son chês di mantignî a 500.000 peraulis il corpus cu lis carateristichis di cumò, che duncje al fasarà di "corpus di alenament" par miorâ lis regulis di disambiguazion pe etichetadure automatiche e produsi, po dopo, un corpus di cualchi milion di peraulis etichetât dome in automatic, cence revision umane. Une racuelte di tescj di passe 2.500.000 peraulis e je za stade realizade fûr contrat e cence paie, in diviers mês di lavôr. Par cheste fase, lis dimensions plui grandis a permetin sieltis mancul vincoladis sui tescj. In particolâr:

- si à dât plui atenzion ai grancj autôrs de corint leterarie La Risultive (Josef Marchet, Dino Virgili, Maria Forte, Riedo Pup, Meni Ucel ...) e a altris autôrs di romançs;
- si à zontât plui tescj teatrâi;
- a son stâts tratâts tescj sientifics e divulgatîfs ancje di buine lungjece e ancje se a jerin traduzions dal talian o di altris lenghis, calcolant il grant valôr che a àn par cjatâ fûr neologjisims;
- a son stâts tratâts cul sisteme de dople version (origjinâl pe visualizazion e standard pal processament) diviers tescj in varietât locâl o in fasis de coinè diferente di chê di cumò.

La etichetadure di chest corpus, che al restarà dividût dal prin, e rive dongje dal 95%. Par otignî chest risultât a son doi nivei di intervent:

- perfezionament des regulis di disambiguazion dal etichetadôr
- riduzion de profonditât te pussibilitât di ricercje (par esempli individuant tant che part dal discors dome "verp" cence distingui jenfri ausiliâr, modâl, transitîf, intransitîf...).

Un imprest dal gjenar al podarà cressi une vore svelt in cuantitât di tescj e fâ un grant servizi par dutis lis funzions metudis in liste tal paragraf 3 e ancje par altris.

5. Bibliografie

Brill, Eric: *A Simple Rule-Based Part of Speech Tagger*, in: Association for Computational Linguistics (ed.), Proceedings of the Third Conference on Applied Natural Language Processing, Trento 1992, 112–116.

Burelli, Alessandra/Miculan, Marino: *Frecuencis lessicâls dal furlan scrit*, in: "Gjornâl Furlan des Siencis", 1, 2002, 167–207.

Calzolari, Nicoletta: *Linguistica Computazionale e Risorse Linguistiche*, in: Cicchese, Gennaro et al. (eds.), Scienze informatiche e biologiche. Espistemologia e ontologia, Roma 2011, 32–64.

Carrozzo, Alessandro: *Dizionari Ortografic Furlan–Talian/Talian–Furlan*, Udin 2008.

Carrozzo, Sandri/Feregot, Franz: *La elaborazion di un corpus etichetât de lenghe furlane scrite: esperiments e prospetivis*, in: "Ce fastu?", 2, 2012, 241–262.

Carrozzo, Sandri/Feregot, Franz/Mistrut, Dree: *Jude: la realizazion di un program di traduzion automatiche dal talian al furlan*, in: "Ce fastu?", 2, 2010, 269–280.

CFL 2000 = Centri Friûl Lenghe 2000 (ed.), *Grant Dizionari Bilengâl Talian–Furlan*, Udin 2004.

Lamuela, Xavier: *La grafie furlane normalizade*, Udin 1987.

Marinelli, Rita et al.: *The Italian PAROLE corpus: an overview*, in: Zampolli, Antonio/Calzolari, Nicoletta/Cignoni, Laura (eds.), "Linguistica Computazionale", Special Issue, XVI–XVII/I, 2003, 401–421.

Petkevič, Vladimir: *Corpus lenghistics*, in: "Gjornâl Furlan des Siencis", 1, 2002, 111–132.

Virgili, Dino: *L'aghe dapît la cleve*, Udine 1979[2].

Riferiment web

<www.arlef.it>, [11.12.2017].

<www.claap.org>, [11.12.2017].

<www.serling.org>, [11.12.2017].

Ressumé

Tres la linguistica computazionala él vegnù a se l dé aplicazions con de bogn resultac en cont dl furlan: ti ultims doi dejenés él vegnù laouré fora dizionars digitai, n coretour ortografich y n tradutour automatich talian–furlan. Dilan a chisc stromenc y a la esperienza che an à abiné adum pòn jì inant con la realisazion de n corpus eticheté. Ala nen va de na basa de dac metuda adum da tesć te chi che vigni parola vegn colieda a na anotazion con n lema, na conversazion o la morfologia, te chest cajo. Te chest contribut végnel descrit les carateristiches y les fases de laour coche ence les prospetives de chest projet.