Gjornâl Furlan des Siencis **1.** 2002, 111-132

Corpus lenghistics[#]

VLADIMIR PETKEVIČ[*]

Ristret. In chest articul il concet di corpus lenghistic – un prodot dai plui impuartants de lenghistiche computazionâl resinte – si lu specifiche tant che une largje ricolte di tescj di une o plui lenghis naturâls, formalmentri struturade e descrite daûr de lenghistiche. La lenghistiche di cumò e cjale il corpus tant che un gnûf paradigme che lu puedin doprâ no dome i lenghiscj, ma ancje chei altris specialiscj di diviersis dissiplinis sientifichis, massime di chês umanis. Il pês dai corpus inte lenghistiche moderne al ven tratât e pandût a mieç dai aspiets clâf dai corpus: fate dai corpus; dimension dai corpus; il meti sù e la sielte dai tescj di un corpus, cjalant la rapresentativitât dal corpus (a rivuart di un fin dât di prime); la struture interne dal corpus e la nature e il sens des informazions lenghistichis intun corpus, che a varan une atenzion speciâl. I esemplis pai diviers aspiets dai corpus a vegnin massime dal Corpus Nazionâl Cec (CNC).

Peraulis clâf. Lenghistiche computazionâl, lenghis naturâls, tescj scrits, tescj orâi.

1. Descrizion e finalitâts di un corpus lenghistic. I ultins 15-20 agns de evoluzion de lenghistiche moderne, tal avignî, a saran viodûts tant che une rivoluzion. Biel che fintremai intal mieç dai agns Otante a'nd jere un grum di lenghiscj di orientazion tradizionâl che a cjalavin suspietôs a lis gnovis tindincis dal studi da la lenghe naturâl cul jutori de matematiche e de informatiche – ven a stâi la elaborazion de lenghe naturâl – rivâts chei che si clamin corpus lenghistics, cheste posizion no je plui corint. I corpus lenghistics si puedin considerâ une des risultis plui impuartantis

[#] Traduzion inglês/furlan dal prof. Zorç Cadorini.

[*] Istitût di Lenghistiche Teoriche e Computazionâl, Universitât "Carli IV", Praghe, Cechie

che o vin vût intai ultins agns inte lenghistiche computazionâl (matematiche). Un *corpus lenghistic* al è definît tant che une ricolte di dâts testuâi produsûts intune lenghe naturâl, che e je formalmentri struturade e lenghistichementri descrite (cun anotazions o cun segnacui - *tags*), simpri che chê ricolte e sedi metude jù intune forme computerizade (eletroniche), cfr. p.es. Čermák (1995). Une ricolte di tescj di cheste fate, che dispès si le cjate fûr intune maniere che e sedi rapresentative cjalant a un ciert obietîf (disìn par rindi il lengaç dai gjornâi todescs di une cierte ete), par solit e cjape dentri un biel toc: e rive pôc sù pôc jù a une magnitudin di disinis di milions di peraulis di test. Di vê a man ricoltis di tescj cussì grandis, lu à permetût la sveltece che, par altri avonde resinte, dal disvuluçament de informatiche, tâl che ur à proviodût ai ricercjadôrs machinis (*hardware*) e programs (*software*) sofisticâts. Cussì i lenghiscj, i informatics e i matematics a àn la pussibilitât di tirâ sot bielis butadis di tescj, di metilis jù in discs dûrs (*hard disks*) cuntune capacitât di giga-bytes e di tratâlis cun programs di ricercje e di realizazion statistiche fats di pueste.

La reson parcè che si metin dongje i corpus lenghistics e je massime chê che par un lenghist corpusist un corpus lenghistic al è la miôr mostre di une lenghe naturâl che si puedi vê a man ore presint, e in dì di vuê une lenghe e je dure di descrivi cence doprâ il materiâl dal corpus. Chest si lu viôt ben intai caratars di un corpus che culì a si ripuartin (viôt inmò Čermák 1995, e la bibliografie citade):

- i dâts intal corpus a son tescj reâi e no artificiâi e i lôr elements (grafems, peraulis, frasis e v.i.) si comparin intal lôr *contest naturâl*;
- un corpus consistent al riflet l'ûs de lenghe studiade e cussì al permet di inferî gjeneralizazions obietivis su lis proprietâts di chê lenghe (o chês lenghis). Il fat al è che la lenghe intal corpus si manifeste intal so *ûs tipic*. Chest pont achì e chel altri a la fin a rifletin il caratar empiric dai corpus; la fonde di un corpus e je propit chê che al riflet cemût che in chel moment si dopre une cierte lenghe. Chest mût empiric di procedi al contraste a clâr cu la vision mentaliste dal lengaç, che intal studiâ e met dongje frasis e costruzions lenghistichis artificiâls e lis dopre tant che fonde primarie des sôs teoriis;
- un lenghist al pues tirâ fûr dal corpus une vore svelt un grum di buinis informazions su la lenghe e cussì rivâ a *gnovis introspezions inte lenghe* intune forme compate e ben sestade che une volte e jere dal dut

impussibile, tant a dî cuant che al coventave dâ dongje tassis di sfueuts plens di dâts lenghistics (viôt Francis 1991); alore, un corpus eletronic al pues rindi *la vore di un lenghist tant svelte e produtive che mai*.

- un corpus al à pês e si lu pues doprâ daûr di cetant siôr che al è di informazions lenghistichis metudis dilunc dai dâts testuâi creis; cussì si pues recuperâ la informazion globâl di diviers modei struturâi. In grazie des diviersis pussibilitâts di ricercje interne, o podìn dî che il corpus al ten dentri plui informazions che no i dâts internis cjapâts di bessôi.

E je une brancje specifiche inte lenghistiche che e lavore cui corpus lenghistics: la *lenghistiche dai corpus*. Cheste brancje interdissiplinâr de lenghistiche (a son soredut la informatiche e la matematiche che a cooperin cu la lenghistiche) e je divierse di chês altris brancjis de lenghistiche inte metodologjie che parie e fâs sù, e processe i corpus lenghistics e e permet un bon acès e un recupar dai dâts. La lenghistiche dai corpus, po, e vierç cuistions dal dut gnovis e e disvuluce un gnûf mût di pensâ la lenghe (plui inte Sez. 3.1).

2. Aspiets primaris dai corpus. I corpus lenghistics si ju pues viodi e stazâ daûr di diviers aspiets: la fate dai corpus (Sez. 2.1), la dimension dai corpus (Sez. 2.2), i corpus e lis lenghis intal mont (Sez. 2.3), il metisi adun di un corpus (Sez. 2.4) cjalant: la acuisizion dai dâts (Sez. 2.4.1), la sielte dai tescj e la rapresentativitât dal corpus (Sez. 2.4.2), la regjistrazion dai tescj (Sez. 2.4.3), il processament computazionâl di fonde dai tescj e la struture interne dal corpus pandude da la sô marcadure (Sez. 2.4.4), lis informazions lenghistichis intun corpus (Sez. 2.4.5), la fate da lis anotazions lenghistichis (Sez. 2.4.5.1) considerant massime la anotazion des parts dal discors e dai dâts morfologjics (Sez. 2.4.5.2 e Sez. 2.4.5.3). Il mût di profitâ dai corpus inte Sez. 3.

2.1 La fate dai corpus. I corpus lenghistics eletronics si puedin classificâ daûr di diviers criteris. Une lôr panoramiche nus judarà a fâ viodi lis varietâts di corpus che a esistin in dì di vuê. Lis classificazions principâls a son chês lenghistichis e a classifichin i corpus daûr dai criteris primaris esponûts chi sot (viôt ancje Kocek et al. 2000):

- *la forme de produzion lenghistiche*: i *corpus*, chei *scrits* e chei *orâi*, a son

par so cont. I *corpus scrits* a son fats di tescj scrits, i corpus orâi a vegnin fâts fûr di tescj cu la trascrizion di discors (massime spontanis). Diferent che pai corpus scrits, lis spesis e la vore che a coventin par meti dongje un *corpus orâl* a son badiâls, soredut la trascrizion de forme fevelade a di chê testuâl e puarte vie une vore di timp e di fuarcis;

- *la ete che si dopre une lenghe*: i *corpus sincronics, diacronics* e *storics* a son par so cont. Il *corpus sincronic* (scrit, si capìs) al è chel predominant di cumò. Al riflet la lenghe contemporanie e al è fat di tescj contemporanis, ven a stâi di tescj produsûts no plui tart di une desine di agns indaûr, pa la cuâl une vore dongje dal timp che il corpus al ven metût sù. Diviers che i corpus sincronics, un *corpus scrit diacronic* al ten dentri tescj che a rifletin la evoluzion di une lenghe intune scjale temporâl: al cjape dentri tescj di diviersis fasis dal disvuluçament de lenghe e la storie da la comunitât che le dopre. Il tiermin temporâl che al dissepare il corpus sincronic di chel diacronic par solit al è arbitrari. Un *corpus storic* al atint a di une fase (intune scjale temporâl) intal disvuluçament di une lenghe;
- *la fate dai tescj intun corpus*: i corpus *lenghistics gjenerâi* e chei *specifics* a son par so cont. Un *corpus lenghistic gjenerâl* al ten dentri tescj produsûts di plusôrs autôrs e si pues dî che al riflet la lenghe (plui just: l'ûs) tant che une sô mostre, biel che un *corpus specific* al ten dentri tescj di un cjamp particolâr, tescj scrits dome che di un autôr (tant a dî il corpus di Honoré de Balzac) oben tescj publicâts di un gjornâl dentri di un ciert timp;
- *la varietât lenghistiche*: i *corpus dialetâi* a son par so cont. A rifletin diviers dialets di une lenghe;
- *plurilenghisim*: i *corpus monolengâi, paralêi* e *a confront* a son par so cont. Un *corpus monolengâl* al ten dentri tescj dome di une lenghe, biel che chei paralêi e chei a confront a tegnin dentri tescj di plui di une lenghe. Un *corpus paralêl* al ten dentri i tescj origjinâi parie cu lis relativis traduzions. Cul integrâsi de Europe e dal mont, la traduzion jenfri diviersis lenghis e cjape simpri plui pês e i ecuivalents metûts jù inta chescj corpus a judin une vore i tradutôrs, lis voris par fâ dizionaris bilengâi (magari plurilengâi) e i students di lenghis forestis, di traduzion e v.i. Chescj corpus achì par solit no son cuissà ce grancj, par vie che no'nd è tancj tescj che a vebin la relative traduzion. Intai *cor-*

pus a confront la relazion jenfri il test origjinâl e l'ecuivalent inte(s) lenghe/is foreste/is e je plui lascje: il test inta chê altre lenghe, plui che une traduzion, al è un test di une fate dongje di chê dal origjinâl. Il rapuart jenfri chescj tescj si pues pandi in tiermins di cjamp tematic (chimiche, politiche), fate gjenerâl (narative, gjornâi, tescj professionâi), contemporaneitât de produzion dai tescj e v.i.;

- *aspiets tecnics*: i *corpus di control* e chei *di prove* a son par so cont. Un *corpus di control* al è par solit un corpus ausiliari di un corpus principâl: par solit al ten dentri tescj cun fenomens lenghistics particolârs e elements che il corpus principâl ancjemò no ju rapresente. I *corpus di prove* si ju dopre par provâ diviersis tecnichis di aplicazion di informazions lenghistichis ai tescj creis (ven a stâi par meti i segnacui, viôt plui jù);
- *studi di lenghis forestis*: i *corpus dal student* a son par so cont. Chescj corpus achì a son fats di tescj scrits di students di lenghis forestis; a tegnin dentri ancje diviers erôrs, di chei tipics dai forescj inte scriture dai arlêfs, cussì a procurin materiâl di grant valôr par che i students a si sfrancjin inte lenghe foreste che a studiin.

2.2 Dimension dai corpus. La dimension di un corpus si le misure par numar di posizions intal corpus (corintis: peraulis testuâls e puntuazion). In gjenar e va simpri ben cheste massime: *plui dâts che si pues, miôr dâts che si pues*. Se cualchidun al vûl meti sù un corpus lenghistic gjenerâl rapresentatîf, disìn, al scuen tirâ sot plui dâts lenghistics che si pues, ma chescj dâts a scugnin jessi, tal stes moment, chei plui rapresentatîfs che si pues (viôt plui in jù). In dì di vuê i corpus pal plui a tegnin dentri disinis di milions di peraulis, par esempli il *Bank of English* al ten pôc sù pôc jù 300.000.000 di peraulis, il *British National Corpus* e il *Czech National Corpus* plui o mancul 100.000.000 di peraulis par om, il corpus todesc dal gjornâl *Süddeutsche Zeitung* al è grandonon: al ten plui di un miliart di peraulis testuâls! I corpus lenghistics plui grancj no son tant impuartants pal aspiet dai fenomens lenghistics tipics (i fenomens tipics si puedin cjatâ ancje intun corpus plui piçul), ma pluitost par chei margjinâi, par vie che la probabilitât di cjatâju e je par fuarce plui alte intun corpus grant.

2.3 Lis lenghis intal mont e i corpus. Intai agns resints ator pal mont si à viodût une svelte incressite dai corpus lenghistics. Biel che intal principi

dai agns '90 a esistevin dome corpus pes lenghis plui cognossudis intal mont, cuntun lûc primari pal inglês, cumò la situazion e fâs viodi un cuadri dal dut diviers. In teorie dutis lis nazions plui grandis (ma ancje la piçule nazion slovene!) a àn i lôr corpus lenghistics, par altri a son disferencis intai nivei di cooperazion e di finanziament intai diviers paîs. Il finanziament pal disvuluçament dai corpus in cualchi paîs lu proviôt il stât, tant a dî inte Republiche Ceche diviersis istituzions a àn judât, cooperant a strent, cuant che si à metût sù il Corpus Nazionâl Cec (cfr. http://unck.ff.cuni.cz), biel che une ativitât coordenade cuntun finanziament di chê istesse fate no à puartât a disvuluçâ un prodot analic pal polac. I corpus ju metin sù ancje compagniis privadis, plui che altri par vê une buine fonde di dâts (*database*) pe preparazion di dizionaris di diviersis fatis.

2.4 Meti sù un corpus. Intal meti sù un corpus eletronic si passe par chestis fasis (o assums parziâi) principâls:

- acuisizion dai tescj
- classificazion dai tescj tirâts dongje
- processament computazionâl di fonde dai tescj (peât cul fin dal corpus in cuistion)
- anotazion lenghistiche dai tescj
- disvuluçament dal program pal recupar dai dâts intal corpus (ricercje intal corpus, realizazion statistichis e v.i.)

Al è clâr che cierts assums parziâi a puedin procedi a pâr.

2.4.1 Acuisizion dai tescj. I tescj di une lenghe naturâl pal corpus si puedin tirâ dongje cussì (viôt p.es. Kocek et al. 2000):

- otignint i tescj daurman inte forme eletroniche dai editôrs o dai autôrs
- burint fûr chei tescj che si cjatin intal Internet
- passant i tescj cul passadôr (*scanner*)
- copiant a man tescj scrits tal imprin su cjarte.

La prime alternative e je chê classiche: e je produtive e no (masse) laboriose. Sedi i tescj di rivistis e gjornâi sedi i libris publicâts cul jutori dal ordenadôr (*computer*) intai agns resints a son otignûts par solit in cheste maniere. I tescj ju proviodin su la fonde di un contrat (dispès sorenuie) cu la clausule che a vignaran doprâts dome che par finalitâts di studi e

sientifics; i dirits sui tescj a restin a cui che al furnìs i dâts. Une altre risultive par grandis ricoltis di dâts testuâi al è l'Internet e pardabon tancj corpus ator pal mont a tegnin dentri dâts gjavâts dal Internet. Il probleme di chescj tescj al è l'alt nivel di fai: chescj tescj a son plens di erôrs di scriture e di altre fate, par solit la lôr struture no je clare o a son scrits cence i diacritics. A passâ i tescj cul passadôr si procêt avonde planc e chest lavôr si scuen limitâ a di chei tescj che no si pues otignî inte forme eletroniche. Chest al è il câs massime pes publicazions vecjutis (soredut libris) publicadis cence doprâ il computer o che la lôr version eletroniche no je disponibil. La batidure a man, tant laboriose e lente, e je la ultime soluzion pussibil inte acuisizion dai dâts: si le dopre intal câs che il test scrit sul sfuei al vedi une cualitât tipografiche tant basse di no podê passâ il passadôr.

2.4.2 Sielte dai tescj e rapresentativitât di un corpus. Ogni corpus si lu cree par un ciert fin che si manifeste come prime robe inte sielte dai tescj. Cjalant chest fin, si dîs che chel corpus al è (o nol è) *rapresentatîf*. Disìn che se o volìn studiâ la lenghe di Honoré de Balzac par mieç di un corpus di Balzac, alore nus covente di tirâ dongje chei tescj di Balzac che a son tipics, rapresentatîfs dal so stîl. Intal câs di Balzac, ven a stâi par un corpus specific (viôt 2.1 insomp), miôr di dut al è di cjoli ducj i siei tescj, parcè che un corpus cussì, ancje grant che al sedi, al à une dimension limitade (no plui di 75 prosis). Une "mostre" rapresentative di chest gjenar e sarès identiche de vore complete di Balzac. Par altri, se o volìn studiâ une lenghe contemporanie intal so complès, o vin di meti sù un grant corpus lenghistic gjenerâl di chê lenghe, e alore il concet di rapresentativitât al è cruciâl, par vie che no podarìn mai tirâ sù ducj cuancj i tescj produsûts di une cierte comunitât lenghistiche intune cierte ete. Cussì o scugnìn sielzi chei tescj che pardabon a rifletin l'ûs di chê lenghe e i fenomens lessicâi, morfologjics, sintatics, semantics e pragmatics che e ten dentri. Se o volìn studiâ diviers fenomens, o scugnìn tignî cont des probabilitâts che si cjatin pardabon intal corpus. O podaressin, tant a dî, inmaneâ un piçul corpus che al fasi di pont di partence e contâi diviersis distribuzions probabilistichis dai fenomens di studiâ; oben doprâ di pont di partence tescj tirâts dongje a man e definîts intes lôr frecuencis e inte distribuzion di diviers fenomens, e di li tirâ sot ancjemò altri tescj

che a vebin di mantignî chês istessis frecuencis e distribuzions. Par altri, cun grandis partidis di dâts no si pues savê se a laran pardabon a rifleti almancul di dongje lis distribuzions che al ten dentri chel (par solit) piçul corpus metût dongje tal principi, par vie che la incressite de rapresentativitât dai fenomens no je liniâr tant che chê in assolût dai dâts. Dut chest resonament, par altri, chi nol va ben, parcè che par solit o volìn cjatâ fûr la distribuzion statistiche e lis frecuencis di une lenghe cence vê dacjâf une idee di chês frecuencis. Alore al è rasonevul di scrutinâ cemût che a son i tescj che a rivin a di une cierte popolazion intun ciert periodi e cemût che a son chei che ju produsin. Intal câs dal Corpus Nazionâl Cec, massime pal corpus sincronic scrit, i tescj a rifletin trop che la popolazion ceche e lei (tor il 1996) libris (soredut la narative), gjornâi e po rivistis e libris specializâts. Inta chel corpus i tescj a son dispartîts in doi grups: i tescj *di informazion* e chei *di imagjinazion*. Daûr des risultis di une ricercje sociologjiche de popolazion ceche tor il 1996 a son stâts stabilîts chescj parcentaçs (valôrs a par cent) chi sot tant che rapresentativitât ideâl dal corpus:

- *tescj di informazion* 85 % (tescj *gjornalistics*: 60 %, tescj *professionâi, tecnics* e *specializâts* in gjornâi e libris: 25 %)
- *tescj di imagjinazion* (narative) 15 %.

Ognidun di chescj doi grups al è stât ancjemò dispartît daûr di une classificazion plui fine pandude par proporzions (li dai tescj di informazion: siencis naturâls, siencis sociâls, economie, storie de art, leç, religjon e v.i.; li dai tescj di imagjinazion: prose, poesie, teatri e v.i.), cfr. Kocek et al. (2000).

Plui in gjenerâl, al esist un sisteme di classificazion dai tescj elaborât da la *Text Encoding Initiative* (TEI) e dal EAGLES (*Expert Advisory Group on Language Engineering Standards*), ma lis proporzions misuradis a parcentaç a àn di rifleti un obietîf specific di rapresentativitât che si varès di rivâi.

2.4.3 Regjistrazion aministrative dai tescj. Ducj i tescj che a jentrin intal corpus a coventin notâts, ven a stâi che si scuen fâ une esate regjistrazion de divignince di ogni test. Il lûc cu lis informazions aministrativis di un test T al è par solit une cjavece (*header*) di anotazion e chê e inclût il co-

diç ugnul che al identifiche il test intal corpus, il titul di T o il titul dal test che al ten dentri T, l'autôr di T e il so ses, l'ISBN o l'ISSN, la identificazion dal editôr, la identificazion di cui che al à proviodût T, la fate e il gjenar di T, l'an di publicazion di T, la lenghe di T, la lenghe origjinâl (se T al è une traduzion) e altris carateristichis. Lis informazions aministrativis a àn pês, parcè che si pues tacâlis ae rie di concuardance (viôt plui in jù) tant che risulte di une ricercje pal corpus e par viodi la divignince di chê concuardance.

2.4.4 Il processament computazionâl di fonde dai tescj e la marcadure struturâl interne dal corpus. I tescj di jentrade (*input*), clamâts *creis* cjalant il processament di daspò, a son un grum eterogjenis. I tescj otignûts daurman inte forme eletroniche o chei tirâts dongje cul passadôr o copiâts a man a son scrits par mieç di diviers editôrs testuâi (*text editors*), e i lôr formâts a puedin jessi une vore lontans un cul altri. Par judâ il processament di daspò, ducj cuancj i tescj a coventin unificâts intune cierte maniere. Cumò o fasarai viodi intune forme semplificade il processament di fonde dai tescj e la struture interne di un corpus doprant il Corpus Nazionâl Cec (viôt Kocek et al. 2000).

Ducj i tescj di jentrade creis a vegnin convertîts intun formât speciâl SGML (Lenghe di Marcadure Gjeneralizade Standart – cfr. Burnard 1993, Bryan 1988 – viôt plui in jù) oben inta chel dal so clon (*clone*) XML (Lenghe di Marcadure Slargjade) e i tescj convertîts cussì si clamin *cjartolârs*. Ogni cjartolâr par solit al corispuint a un libri o a un numar dal gjornâl, ma no je une corispuindince obligatorie. Ogni cjartolâr al è componût di *documents* (par solit i cjapitui di un libri a son identificâts tant che documents par so cont). Ogni document al ten dentri une *cjavece* aministrative e il so contignût al è fat di *paragrafs*, ognidun al è formât di *frasis* fatis di peraulis corintis, ven a stâi *posizions intal corpus* (peraulis testuâls e puntuazion). Si trate, in sumis, di un sisteme gjerarchic:

cjartolârs
documents
paragrafs
frasis
peraulis

Ogni document al è marcât struturalmentri intun formât SGML, e la sô struture e je predefinide intune filze sacume (*template file*) a non DTD (Definizion de Fate dal Document) che e permet di analizâ automatichementri il document in cuistion e viodi se al cumbine cu la strutture definide intal DTD. Chê analisi le fâs un algoritmi analizadôr (*parser*) pal SGML speciâl.

La forme de marcadure di paragrafs e sentencis par SGML si le viôt chi sot (cfr. Kocek et al. 2000, p. 27):

<p n=3> ... III paragraf ...
<s id="S/J/1992/vesm9211:001-p3s1">
... I frase ...
<s id="S/J/1992/vesm9211:001-p3s2">
..II frase ...

...

<p n=3> ..III paragraf ...

<s id="S/J/1992/vesm9211:001-p3s6">
...penultime frase ...
<s id="S/J/1992/vesm9211:001-p3s7">
...ultime frase ...

<p n=4> ..IV paragraf ...

Cemût che si viôt, ogni paragraf e ogni frase intun document a son identificadis in maniere ugnule. Ogni segnacul, cjapât dentri intai lincins "<" e ">", al specifiche dulà che tache e dulà che e finìs la part corispuindint dal test. Par solit, un segnacul SGML *t* cul so contignût a àn la forme:

<t atribût=valôr ...>contignût_dal_segnacul_t</t>

indulà che <t atribût=valôr ...> al è il *vierzisegnacul*, che lu pues specificâ indenant la schirie dai atribûts cui lôr valôrs. Daspò al ven il contignût dal segnacul che al è delimitât dal *sieresegnacul* (facoltatîf).

Fintremai al nivel dai paragrafs il processament dai documents al è facilut, par vie che par solit i paragrafs a son identificâts a clâr intune maniere o in chê altre za intal test di jentrade (ancje se intal SGLM no diretementri). La segmentazion di un paragraf par frasis e la segmentazion

di une frase par posizions intal corpus (ven a stâi par peraulis ugnulis e puntuazion) a son ben lontanis di jessi fatis daurman. Chi al covente un program lenghistic sofisticât, parcè che i elements lenghistics (scrits e fevelâts) par solit a son ambiguis. Intal segmentâ lis frasis no si pues simpri fâ cont sul fat che lis frasis a finissin cuntun complès di delimitadôrs univocs tant che pont (.), pont di domande (?), pont esclamatîf (!) e ponts di sospension (...), parcè che chei grafems achì a puedin vê ancje altris interpretazions. O podìn viodilu intes frasis chi sot:

(1) <s n=1>The tragedy of the Jews reached its climax in 70 B.C.</s>
<s n=2>Titus destroyed Jerusalem and the Jews had to flee to the diasporas.</s>

Chest test achì al presente, difat, lis dôs frasis che a son mostradis, ma no je une robe banâl meti il prin sieresegnacul </s> scrit gruessut intal puest just, parcè che il pont che i ven daûr a C al rapresente sedi l'ultin element de imbreviadure B.C. sedi il pont fer che al siere la frase. Chi al covente doprât un program lenghistic sofisticât, bon di cerni i diviers câs daûr dai contescj lenghistics.

Prime de anotazion lenghistiche rude, i tescj a vegnin ancje *curâts* e *corezûts*. Un dai modui impuartants di curâ al identifiche chei tocs dal test di jentrade che a son scrits in(tune) lenghe/is foreste/is, e, co a no son integrâts in mût indisseparabil dal test intorsi, i tocs a vegnin taiâts fûr dal rest dal processament. Un altri modul clamât *intivadôr* (*guesser*) al identifiche lis peraulis discognossudis. Cjatadis lis peraulis discognossudis, chest modul al prove, disìn, a metilis in cjadene cun chês altris ator, se no lis cognòs nancje chês, e al viôt se la risulte di une cuncjadenazion cussì e je une forme di peraule che e esist: inta chest câs e je chê peraule cuncjadenade che e passe al rest dal processament. L'intivadôr al prove ancje, su la fonde de morfologjie des peraulis discognossudis, di assegnâur un ciert paradigme (daûr des carateristichis morfologjichis specifichis de lenghe in cuistion) cussì che almancul lis informazions su la part dal discors e su la morfologjie, plui che no chês lessicâls, a saran tignudis di cont pal rest dal processament.

Frasis tant che (1) e la lôr marcadure struturâl nus àn fat viodi che il processament lenghistic al è dut fûr che facil e che nus puarte di considerazions a rivuart de marcadure struturâl dai elements dal corpus al

component plui intrigât e ancje plui interessant dal processament dai tescj intal corpus: la notazion lenghistiche dal corpus.

2.4.5 Informazions lenghistichis intun corpus. A son corpus che a son marcâts dome struturalmentri, ven a stâi che i documents a son, come che o vin viodût, segmentâts gjerarchichementri par paragrafs, frasis e peraulis (posizion intal corpus). La segmentazion in jù fintremai al nivel dai paragrafs si pues considerâ tant che marcadure struturâl rude, là che nissune informazion lenghistiche no ven inzontade al materiâl testuâl. Par altri, se o volìn rivâ a di une segmentazion juste intai elements plui in jù intes struturis tant che frasis, peraulis o ancje multiperaulis, e covente impleâ une cognossince lenghistiche. Cundiplui che un ricercjadôr, par solit, al cîr no dome formis di peraule intal corpus, ma intant de ricercje al vûl ancje doprâ lis informazions lenghistichis (massime su lis parts dal discors, su la morfologjie, su la sintàs, su la semantiche e magari ancjemò altri). Par apaiâ lis dibisugnis dal ricercjadôr, aes peraulis o ai grups di peraulis a tocje assegnâ informazions lenghistichis, par chest si dîs che un corpus al è *lenghistichementri notât*. In tancj a volaressin vê a disposizion corpus notâts, ma par vie che si brame che la dimension di un corpus cussì e sedi plui grande che si pues, la notazion si scuen fâle cence ecezions suntune fonde automatiche. Une notazion lenghistiche automatiche juste di un corpus e je, par altri, un assum une vore dûr che al met fûr une grande disfide pe lenghistiche computazionâl. O fasarai viodi (inte Sez. 2.4.5.2) ce complès che al è l'assum cul meti i segnacui pes parts dal discors e pe morfologjie.

2.4.5.1 Nivei di notazion lenghistiche. Intal so articul cognossût (Leech 1993), l'autôr al classifiche diviers nivei di notazion lenghistiche:

- notazion ortografiche
- notazion fonetiche/fonologjiche
- notazion prosodiche
- notazion des parts dal discors (*part of speech* - POS) e de morfologjie
- notazion sintatiche
- notazion semantiche
- notazion pragmatiche/dal discors.

La *notazion ortografiche* e consist inte disambiguazion grafematiche,

che e esplicite i elements ortografics ambiguis di un test, tant che lis virgulutis viertis e chês sieradis, che cualchi volte si pandin tal stes grafem. Ancje la disambiguazion de funzion di grafems tant che l'apostrof, il pont e lis maiusculis (a puedin mostrâ il cjaveç tant di un non propri che chel di une frase) e je intal assum de notazion ortografiche.

La *notazion fonetiche/fonologjiche* si le dopre intal segmentâ fonems e e torne buine intal studi su la lenghe fevelade e il so processament.

La *notazion prosodiche* si interesse de marcadure de prosodie dal test fevelât cun tant di acent, intonazion, pausis e v.i. La notazion di cheste fate e je tant laboriose che mai, compagn che ducj i processaments de lenghe fevelade.

Des POS e de notazion morfologjiche si tratarà plui in detai inte Sez. 2.4.5.2.

La *notazion sintatiche* e consist intal fâ fûr des frasis intal corpus *struturis sintatichis* (par solit cu la forme di un arbul sintatic). Un bon spieli di une notazion di cheste fate e je une des bancjis di arbui sintatics fatis plui di resint, la Bancje di Arbui Dipindinçâi di Praghe (PDT) (viôt Hajič et al. 2001), cun dentri notadis plui di 100.000 frasis dai gjornâi cecs.

La *notazion semantiche* e assegne aes peraulis e aes struturis sintatichis intes frasis i segnacui semantics daûr di diviers criteris semantics e di cjamps dal discors. Un dai assums cruciâi de notazion di cheste fate e je la *disambiguazion dal sens de peraule*, che e identifiche la interpretazion juste pes peraulis lessicalmentri ambiguis (tant che il todesc *Kiefer*, che al vûl dî tant *pin* che *massele* oben il cec *strana* che al vûl dî tant *bande* che *partît* e v.i.).

La *notazion dal discors* si interesse di dutis chês notazions dopradis par marcâ dutis lis relazions e i elements che a van di là dai confins de frase. Un toc impuartant di cheste notazion al lavore daûr des relazions anaforichis.

In dì di vuê lis notazions che si doprin di plui a son:

- la notazion des parts dal discors (POS) e de morfologjie
- la notazion sintatiche.

Achì sot o tratarìn plui in detai la notazion des POS e de morfologjie tant che notazion lenghistiche classiche. La notazion di cheste fate e varès di cjatâsi dentri di ducj i corpus notâts lenghistichementri tant che fonde

dal rest dai processaments (soredut pe analisi sintatiche e pal disvuluçament di bancjis di arbui).

2.4.5.2 Notazion des parts dal discors e de morfologjie intai corpus. La notazion des POS e de morfologjie (marcadure) intun corpus e consist intal assegnâ lis informazions su lis parts dal discors e su la morfologjie a ognidune des formis di peraule (o dai grups di formis di peraule) intai tescj dal corpus. Se o vin une tassonomie di segnacui di POS tant che complès di valôrs cun dentri elements de fate di *non*, *adietîf*, *verp* e v.i. e il repertori di valôrs fat des categoriis morfologjichis[1], o podìn assegnâi a ogni forme di peraule intal test corint i valôrs juscj des categoriis che o vin.

Se o fasìn la marcadure a man su la fonde de nestre cognossince, abilitât e esperience de lenghistiche, o cjatìn problemis avonde da râr (chest al podarès jessi il câs di elements che no son avonde clârs di identificâ tant che POS). Marcant a man, dispès no si rive a capî il cûr dal probleme clâf da la lenghistiche computazionâl: ven a stâi di risolvi il probleme de ambiguitât de lenghe *automatichementri*. Marcant a man la frase

(2) We *must* know *that* there are ambiguities in language.

o identifichìn a colp la peraule *must* tant che verp (ancje se par inglês al pues jessi ancje un non cul sens di *most*) e la peraule *that* tant che coniunzion (ancje se chê peraule cence contâ il contest e à almancul ancjemò une altre interpretazion di POS, ven a stâi chê di jessi un pronon). I corpus di milions o magari di disinis di milions di formis di peraule par fuarce no puedin vignî notâts a man. Chest al vûl dî che lis proceduris di notazion dal corpus a scuegnin sei automatichis, ma e sarès dure peâ chestis proceduris a cualchi esperience di cui che al note a man. Alore la realizazion di une notazion lenghistiche automatiche juste e je un assum intrigât, un montafin che difat par tantis lenghis (inglês, todesc, cec e v.i.) no i àn ancjemò cjatât une soluzion apaiadôre.

La marcadure automatiche des POS e de morfologjie intun corpus e consist intai assums chi sot che a coventin fats automatichementri:

- lematizazion
- analisi des POS e de morfologjie
- disambiguazion des POS e de morfologjie.

Chescj prins doi assums i tocjin a di un analizadôr morfologjic, che al è un imprest specializât cuntun program bon di:

- complî la *lematizazion*, ven a stâi che al assegne a di une forme di peraule intal test corint il so *leme*, ven a stâi une forme di fonde rapresentative oben plui formis se la forme di peraule e je ambigue intal leme. Par solit il leme di une forme nominâl e je la forme dal nominatîf singolâr; il leme di une forme verbâl al è l'infinît dal verp in cuistion e v.i. Tant a dî, la forme dal locatîf plurâl intal cec *tancích* si le pues assegnâ a dôs interpretazions pal leme:
  (a) leme *tanec* (fur. *danze*)
  (b) leme *tank* (fur. *tanc*)
- complî la *analisi des POS e de morfologjie*, ven a stâi che al assegne a di une forme di peraule dutis lis sôs pussibilis interpretazions di POS e morfologjiche cence contâ il contest particolâr dulà che la peraule si cjate. Tant a dî, la peraule francese *pas* e je almancul trê voltis ambigue e i coventaressin assegnadis almancul trê interpretazions par so cont:
  (a) *pas* tant che particule negative
  (b) *pas* tant che non singolâr (fur. *pas*)
  (c) *pas* tant che non plurâl (fur. *pas*).

Un altri esempli al è chel de peraule francese *chante* che e je ambigue almancul cinc voltis e cussì i coventaressin assegnadis almancul cinc interpretazions par so cont:

(a) la II persone singolâr imperative dal verp *chanter*
(b) la I persone singolâr indicative dal verp *chanter*
(c) la III persone singolâr indicative dal verp *chanter*
(d) la I persone singolâr coniuntive dal verp *chanter*
(e) la III persone singolâr coniuntive dal verp *chanter*.

Formalmentri, la interpretazion des POS e de morfologjie par solit le pant une stringhe (*string*) di valôrs di une letare e ognidun di lôr al parten a di une categorie specifiche. La categorie di POS e je par solit la prime tant che il caratar plui di pês de peraule in cuistion cjalant la sô distribuzion sintatiche e semantiche inte lenghe; chês altris categoriis morfologjichis a vegnin identificadis de posizions inte stringhe e a vegnin daûr de posizion di POS. Tant a dî, inte tassonomie che e marche il Corpus Nazionâl Cec ae forme di peraule *prezidentem* i si assegne il leme: *prezident* (di fonde, ven a stâi la forme dal nominatîf singolâr)

e il segnacul di POS e de morfologjie: NNMS7 -------A--------
Achì lis letaris e il numar si ju interpretin cussì:
I posizion: POS primarie: N = non
II posizion: altre diferenziazion dentri des POS: N = non comun
III posizion: gjenar: M = masculin (animât)
IV posizion: numar: S = singolâr
V posizion: câs: 7 = istrumentâl
VI-X posizion: irilevantis pai nons
XI posizion: negazion: A = afermative (forme dal non no negative)
XII-XV posizion: irilevantis pai nons.
Duncje, passade la analisi des POS e de morfologjie, la peraule e à assegnadis dutis lis interpretazions des POS e de morfologjie pussibilis.

2.4.5.3 Disambiguazion automatiche des POS e de morfologjie. L'analisi des POS e de morfologjie e je dome che un pas di chei che a coventin par rivâ a di un obietîf naturâl de analisi: la identificazion de interpretazion des POS e morfologjiche juste daûr dal contest, parcè che intai tescj di lenghis naturâls salacor ogni peraule e à une interpretazion di lessic e di POS e di morfologjie ugnule: cheste e je la informazion che i interesse di plui al ricercjadôr. Alore la selezion dal segnacul just e à un pês cruciâl e al è un modul di une procedure di disambiguazion automatiche a rispuindi dal bon fin di chest assum.

La disambiguazion automatiche des POS e de morfologjie pes peraulis intai tescj di un corpus e je cence confront plui dure che no la analisi des POS e de morfologjie. E je tant dure par vie che la peraule di disambiguâ e sta in diviers contescj. No son dome lis informazions morfologjichis sul contest a coventâ pe rude disambiguazion di une peraule cussì, si ben ancje i fatôrs sintatics (pal plui) e chei semantics a àn une part clâf. Si puedin definî i fatôrs che a àn un impat diret su la cualitât de disambiguazion:

- l'ambiguitât morfologjiche e di POS inerente a la lenghe processade
- il repertori e la dimension dai segnacui di POS e de morfologjie doprâts, che al è peât a strent cul pont insomp
- il metodi di disambiguazion doprât
- la cualitât dai dâts di jentrade (il nivel di fai, chel dai erôrs di scriture e v.i.).

Lis lenghis ator pal mont a son un grum diviersis inta chest prin aspiet

menzonât. Tant a dî, l'inglês e je une lenghe cuntune morfologjie flessionâl puarute, cuntun ordin des peraulis fer (par esempli, intes frasis declarativis il subiet al ven prime dal verp predicatîf) e cu la partignince des formis di peraule aes sôs classis di peraulis (parts dal discors) che pal plui no je mostrade a viert intal aspiet morfologjic (par esempli la peraule *love* e pues fâ di non, di adietîf o di verp). Il talian in pratiche nol à morfologjie pe declinazion (formalmentri al ten par so cont par nons e adietîfs dome il numar e il gjenar), ma la morfologjie verbâl e je siorone; l'ordin des peraulis al è plui libar che no intal inglês. Il cec, membri de famee des lenghis slavis, al è caraterizât di un ordin des peraulis un grum libar e di une morfologjie pe declinazion siorone; la morfologjie verbâl, par altri, e je tant plui puare di chê taliane, jessint che pe pluipart dai mûts e dai timps i sens verbâi a son di formazion analitiche, ven a stâi che a son espressions verbâls compostis. Po dopo, lis lenghis a son diviersis ancje inte lôr ambiguitât morfologjiche paradigmatiche e casuâl inerente. Chestis robis des lenghis si lis cognòs ben, ma al è interessant viodi ce cruciâl che al è l'influs che a esercitin sul grât di sucès de disambiguazion des POS e de morfologjie pes lenghis rispetivis.

Di chê altre bande il grât di sucès al incrès ancje par un secont fatôr, ven a stâi il repertori dai segnacui doprâts, che lu clamìn *segnacolari* (*tagset*). Tant a dî, i metisegnacui (*taggers*) pal inglês a doprin un segnacolari di disinis di segnacui (stant a la la morfologjie puare dal inglês); i metisegnacui pal cec, impen, che a àn ce fâ cuntune flession plui grande, a doprin un segnacolari di fin 2000 segnacui, par vie che ogni cumbinazion dai valôrs des variis categoriis e ven rapresentade formalmentri di un segnacul. Al è naturâl che e sedi simetriche la relazion jenfri la dimension dal segnacolari e la bondance di informazions lenghistichis gjavadis fûr di un corpus notât. Se i metisegnacui pal cec a doprassin dome segnacui pes POS e no chei morfologjics, il grât di sucès al sarès une vore plui alt (difat al è plui che il 99%); se si à di rifleti intal segnacolari la morfologjie adimplen (alore cuntuns 2000 segnacui), il grât di sucès, doprant i metodis stocastics pe disambiguazion di cumò (viôt plui in jù), al cole (plui o mancul il 94,5%).

Il grât di sucès al è peât a strent ancje cul metodi di disambiguazion doprât. I metodis corints a son di trê fatis:

- chel stocastic (statistic, probabilistic)
- chel par regulis (*rule-based*)
- un metodi cumbinât dulà che ducj i doi metodis a cooperin.

La disambiguazion stocastiche. Chest metodi di disambiguazion al è doprât cumò par dut il mont. Si fonde, disìn, su lis probabilitâts di transizion jenfri peraulis adiacentis intal test, o miôr jenfri i lôr segnacui. Il principi dal metodi al è chel di meti dongje tal imprin un spielut, clamât *corpus sfrancjadôr*, dulà che lis formis di peraule a son marcadis a man (la dimension di un corpus cussì e je di uns centenâr di miârs di peraulis). Il metisegnacui statistic daspò al "impare" chê marcadure alì, ven a stâi che al note intes sôs tabelis internis lis probabilitâts di transizion jenfri segnacui (o magari peraulis e/o lemis) adiacents e lis lôr frecuencis. Cjapât sù chest aventari, il metisegnacui lu apliche su di un grant corpus che al è stât analizât (ma no disambiguât) inte morfologjie e intes POS.

Il grât di sucès di chescj metisegnacui achì al dipent crucialmentri di cetant grant che al è il segnacolari e di cetant fer che al è l'ordin des peraulis, plui esat de distance intal ordin des peraulis che a puedin vê lis relazions sintagmatichis inte lenghe che i tescj a son daûr a vignî marcâts. Dacjâf di chest fatôr a son i balcons doprâts par solit di chei metisegnacui par lumâ i dâts, che a son, disìn, strentuts (par solit no cjapin dentri plui che trê peraulis vicinis). Il grât di sucès dai metodis di cheste fate pal inglês e pal cec lu vin scrit prime e al cuantifiche a clâr lis evidentis diviersitâts tipologjichis jenfri inglês e cec. Intai tescj inglês si cjatin une vore frecuentementri secuencis di segnacui tipichis (l'ordin des peraulis fer dal inglês al causione che a sedin plui ponts fers inte struture de frase inglese), dulà che intune frase ceche di ponts fers sintatics no'nd è tancj e il numar des configurazions sintatichis di pôc sù pôc jù chê stesse frecuence al è avonde plui alt e, cussì, la selezion dal segnacul just e je une vore mancul fortunade.

Un dai problemis primaris pai metisegnacui stocastics e je chel batiât "probleme dai dâts sparniçâts": di grancj corpus sfrancjadôrs notâts juscj di podêju doprâ pai metisegnacui stocastics no'nd è. Il probleme al è che intun corpus di marcâ a puedin cjatâsi secuencis di segnacui che a mancjavin intai dâts sfrancjadôrs (il metisegnacui "alì no ju à viodûts") e alore il metodi al scuen doprâ altris mecanisims par risolvi il probleme

(chest lavôr si lu clame *pulidure*). Al reste un probleme viert chel di capî se, a fâ incressi sostanzialmentri la dimension dai dâts sfrancjadôrs, si rive a une cualitât *sostanzialmentri* miôr pe disambiguazion interie, ven a stâi che al sarès di capî trops dâts sfrancjadôrs che a coventin par che il metisegnacui al passi il nivel dal, disìn, 98% pal cec.

La disambiguazion par regulis. Dongje dai metodis di disambiguzions de fate stocastiche, e esist une fate di disambiguazion che e consist inte formulazion di *regulis sintatichis* che a identifichin i segnacui juscj e a parin vie chei falâts su la fonde di un contest descrit lenghistichementri. Regulis cussì par fuarce no si fondin su considerazions lenghistichis superficiâls e a cirin pluitost di fâ sù un model dal sisteme (sintatic) di une lenghe stabilide. Meti dongje regulis cussì al è une vore laboriôs e une disambiguazion di sucès e covente che i lôr autôrs a sedin sintaticiscj brilants, parcè che chês regulis a domandin introspezions lenghistichis (e massime sintatichis) intal sisteme de lenghe studiade tant sofisticadis che mai. Par altri, cun chest metodi nancje nol covente un corpus sfrancjadôr, parcè che l'autôr des regulis si fonde dome su lis sôs cognossincis e la sô intuizion lenghistichis, ma al è un grant jutori pal disegnadôr des regulis chel di verificâ lis sôs intuizions intun corpus (miôr se piçul) ben notât di prime, ven a stâi un corpus "sfrancjadôr" intun altri sens. Se si lu fâs suntun model dal sisteme di chê lenghe si à in plui ancjemò un costrut: lis regulis metudis dongje par une cierte lenghe contemporanie a si puedin doprâ in maniere diacroniche: si puedin aplicâ a tescj plui vieris (di corpus diacronics o storics) o a tescj plui resints (passâts, disìn, di disinis di agns) e cussì si viôt a colp se lis regulis metudis dongje par sagomâ il sistem de lenghe intune cierte fase evolutive a passin ancje pai tescj produsûts in altris etis dal disvuluçament de lenghe in cuistion.

Il metodi cumbinât. I doi metodis plui in sù a puedin sei cumbinâts. Un mût al è chel di fâ partî il component par regulis che al cure i dâts, ven a stâi che al pare vie cualchi interpretazion morfologjiche e di POS a prin tîr falade, cussì il component stocastic che al ven daûr al pues lavorâ cuntune jentrade mancul ambigue. La disambiguazion si pues fâle ancje par fasis intune altre maniere. Il component stocastic al fâs une disambiguazion a stroz des POS (e magari de morfologjie): chi al varès di cometi

pardabon pôcs fai. La disambiguazion plui fine le pues fâ parsore de risulte de disambiguazion a stroz il metisegnacui par regulis. In ducj i doi câs i components diviers par metodologjie a son cubiâts. A son, si sa, ancje altris metodis cumbinâts che chei menzonâts, ma si pues dî che i metodis ibridis a somein prometi ben.

3. Il mût di profitâ di un corpus

3.1 Imprescj di program, concuardancis, statistichis. Il sens primari dal corpus al è chel che si pues cirî fûr i elements dal corpus (formis di peraule, notazions lenghistichis) cun diviers imprescj di program. Il program al consist di doi components primaris:

- il *motôr ciridôr* che al è responsabil dal cirî a svelt i dâts domandâts
- il component dal *program di denant* che al permet al ricercjadôr di:

  viodi i dâts cirûts inte forme di concuardancis
  specificâ domandis e condizions statistichis;
  daûr dai nivei di notazion lenghistiche si pues cirî,
  dongje des (cumbinazions di) peraulis corintis,
  ancje i dâts relatîfs ai segnacui morfologjics, ai lemis,
  ai segnacui sintatics e v.i.
  cernî e meti vie riis di concuardance
  otignî informazions statistichis di diviersis fatis.

Une *concuardance* e je une secuence di riis cun dentri ognidune lis formis di peraulis che si cirive intal lôr contest naturâl (e parametrizabil). Il standart stabilît, achì al è il cussì clamât formât KWIC (*Peraulis Clâf intal Contest*) che al scrîf la risulte di une domande (viôt chi sot) di une maniere lindule, cul/i element(s) cirût(s) intal mieç des riis e il contest rispetîf dulintor di chescj elements. Lis riis di concuardance a puedin sei cernidis in diviersis manieris. Un spieli di une concuardance cussì al è chel chi sot:

Sunday, calling for greater economic reforms *save* China from poverty.
ion asserted that “the Postal Service could *save* enormous sums of money in contractin
Then, she said, the family hopes to *save* enough for a down payment in a ghome
t-of-work steelworker, “because that doesn’t *save* jobs, that costs jobs.”
We suspend reality when we say we’ll *save* money by spending $10,000 in wages
won the first round in an effort to *save* one of Egypt’s great treasures, the

three children in a mining town who plot to *save* the "pit ponies" doomed to be
"Basically we could *save* the operating costs of the Pershings
(Part de concuardance de peraule *save* cjapade dal corpus de AP [Associated Press] 1987)

O viodìn che la peraule cirude (ven a stâi *save*) e je intal mieç des riis cui rispietîfs contescj intorsi. Cussì o podìn viodi a lindul cemût che la peraule si puarte in diviers contescj.

Il motôr ciridôr al covente inmaneât cussì che al puedi cirî sveltissim ancje intun corpus grandon. La rispueste e je peade a strent cu la nature des dibisugnis pandudis dal ricercjadôr: une domande cun dome une forme di peraule e ven apaiade daurman par vie de indicizazion interne dai dâts, domandis plui intrigadis fatis di cumbinazions di diviersis fatis (magari di diviers nivei di notazion) a cjatin la rispueste plui tart.

Duncje, i imprescj di cirî par domandâ fûr i dâts a proviodin pussibilitâts di svelt recupar dai dâts cumbinant diviersis condizions e cussì a dan introspezions tant gnovis di no crodi inte struture di une cierte lenghe. Compagn par impuartance al è par altri il fat che il corpus e i siei servizis par cirî a permetin di recuperâ informazions sofisticadonis su la lenghe. Se, disìn, si vûl recuperâ une informazion une vore specifiche intal corpus, si scuen formulâ une domande un grum sofisticade. Di spieli o podìn formulâ i doi problemis chi sot:

- cjatâ fûr ducj i verps cu la valence gjenitivâl
- cjatâ fûr ducj i verps riflessîfs.

Si puedin formulâ sedi domandis une vore gjenerâls che a finissin par ingrumâ un mont di dâts che no coventin e che po si scuen butâju vie a man, sedi domandis sofisticadis che a puedin eliminâ un biel toc dal lavôr di postedizion.

Ancjemò, si puedin cjatâ gnovis relazions inte lenghe parcè che concuardancis cussì perspicuis a puedin puartâ un ricercjadôr a definizions di problemis gnûfs adimplen e magari ancje aes soluzions.

Dongje des concuardancis, il ricercjadôr al pues recuperâ ancje *informazions statistichis* di diviersis fatis. Une funzion statistiche elementâr e rigjave la frecuence assolude di un ciert fenomen. Ancjemò plui interessantis a son lis diviersis misurazions che a misurin la colocabilitât des peraulis,

ven a stâi che a cuantifichin trop che dôs (o plui) formis di peraule a fasin une associazion di peraulis stabil, inclapide (*funzions di informazion mutuâl* e *t-score*). A si doprin ancje diviersis distribuzions di frecuence di un ciert fenomen (par esempli rivuart a diviersis risultivis di tescj).

3.2 Il profit dai corpus. Il corpus al proviôt al ricercjadôr, cemût che si pues inferî de descrizion chi insomp, une vore di mûts diviers di profit, massime in lenghistiche. I corpus a puedin fâ di fonde par diviersis fatis di studi lenghistic teoric e di aplicazions tant che dizionaris, gramatichis, manuâi e v.i. Il profit dai corpus plui di pês pal grant public al consist inte *lessicografie fondade sui corpus*. I dizionaris modernis in dì di vuê a vegnin metûts sù su la fonde di corpus, parcè che lis peraulis presentadis intal lôr contest a dan al lessicograf un grum di informazions (a voltis ancje masse!) sui diviers contescj là che la peraule e podarès saltâ fûr e cussì i da al lessicograf indicazions par tratâ ducj i ats di sens de peraule studiade. In dì di vuê i corpus a tegnin dentri materiâl indispensabil tal meti dongje dizionaris crêts.

Il corpus no lu doprin dome i lenghiscj. Ancje studiôs di altris dissiplinis, soredut di chês umanis (sociolics, storics e v.i.), a puedin recuperâ dal corpus materiâl di doprâ, soredut informazions lessicâls, colocazions tipichis des lôr dissiplinis e v.i. Par dîle cun bielis peraulis, il corpus (massime un corpus lenghistic gjenerâl) al riflet a mieç dai siei tescj la vite di une cierte ete inte storie di une comunitât lenghistiche specifiche.

4. Conclusion. In struc, o ai cirût di dâ al letôr une idee almancul superficiâl di ce che al è un corpus lenghistic, ce obietîfs che al à e cemût che si pues impleâlu. O ai ancje informât il letôr sui siei diviers aspiets e dimensions. Se cualchidun al vûl vê plui informazions sui corpus lenghistics, a esistin tassis di leterature juste su chest teme. Une piçule part de bibliografie sui corpus lenghistics si cjate chi sot.

Note

[1] Par nons, adietîfs, numerâi e pronons o podìn disseparâ il *câs* (cun valôrs tant che nominatîf, gjenitîf, ...); il *numar* cun valôrs tant che singulâr e plurâl; il *gjenar* cun valôrs tant che masculin, feminin, neutri. Lis categoriis di *persone, numar, timp, vôs* e v.i. a partegnin ai verps (cu la categorie de persone di valôr I, II e III; jessint i valôrs dal timp passât, presint, futûr; cognossint pe vôs i valôrs atîf e passîf, e v.i.)); la categorie dal *grât* e parten ai adietîfs e ai averbis (positîf, comparatîf e superlatîf).