Autores Gemma Bonet i Mireia Calm *Oficina de Política Lingüística, Comissionat per a Universitats i Recerca, Departament d'Innovació, Universitats i Empresa*

Tecnologia, universitats i llengua: recursos lingüístics tecnològics, en català, desenvolupats per les universitats

Abast dels recursos lingüístics desenvolupats per les universitats en el marc del conveni de col·laboració signat entre el Ministeri d'Indústria, Turisme i Comerç i l'Administració de la Generalitat de Catalunya per a l'aplicació de les noves tecnologies en els àmbits de la investigació i la docència universitàries.

L'any 2006, i arran del conveni específic de col·laboració signat l'any 2005 entre el Ministeri d'Indústria, Turisme i Comerç i l'Administració de la Generalitat de Catalunya per a l'aplicació de les noves tecnologies en els àmbits de la investigació i la docència universitàries, l'antic Departament d'Universitats, Recerca i Societat de la Informació va signar set convenis amb les universitats catalanes per un import d'1.305.460 euros, l'objectiu dels quals era obtenir aplicacions i altres productes tecnològics en català i desenvolupar nous recursos lingüístics en aquesta llengua.

Els projectes resultants, que fomenten les recerques i el desenvolupament sobre noves tecnologies aplicades al tractament de les llengües, la gestió del multilingüisme, les metodologies didàctiques i els recursos i les eines d'aprenentatge del català en línia, es descriuen a continuació.

Generació de recursos lingüístics per al desenvolupament de tecnologies de la parla en català

Impulsat per la Universitat Politècnica de Catalunya, aquest projecte crea recursos lingüístics bàsics en català a fi de desenvolupar sistemes de tecnologies de la llengua i la parla, en l'àmbit de la telefonia fixa i mòbil, el sector automobilístic i el sector del consum. Amb aquest treball es pretén posar la llengua catalana al nivell d'altres llengües europees, adaptar tecnologies desenvolupades en llengües diverses, atreure empreses tecnològiques per crear productes en català i promoure la recerca en el camp del processament de la llengua i la parla també en català.

Per al reconeixement automàtic de la parla, calen grans bases de dades orals amb l'objectiu d'entrenar els sistemes, concretament perquè aprenguin diferents formes de pronunciació. Es tracta d'enregistrar una gran quantitat de persones d'edats diverses i de distinta procedència dialectal, i en diferents entorns, que llegeixen textos o frases i responen preguntes curtes sobre temes específics. Després es fa una transcripció ortogràfica del que realment ha pronunciat l'informant i una transcripció fonètica.

Com que la tecnologia actual és molt sensible al canvi d'ambient, al soroll i als micròfons, han calgut un gran nombre d'enregistraments en un entorn concret per assegurar que el sistema de reconeixement pot treballar en condicions

òptimes en aquest entorn. Els recursos obtinguts consisteixen en quatre bases de dades referents als quatre àmbits següents:

Telefonia fixa i mòbil

S'han gravat 4.000 persones de distinta procedència dialectal (hi són representats el quatre grans dialectes de Catalunya —central, gironí, nord-occidental i tortosí— i el valencià i el balear), de diferent sexe i edat, les quals fan la trucada des de diferents llocs (al carrer, a casa, dins del cotxe...).Els informants llegeixen, durant 15 minuts, textos amb informació variada.

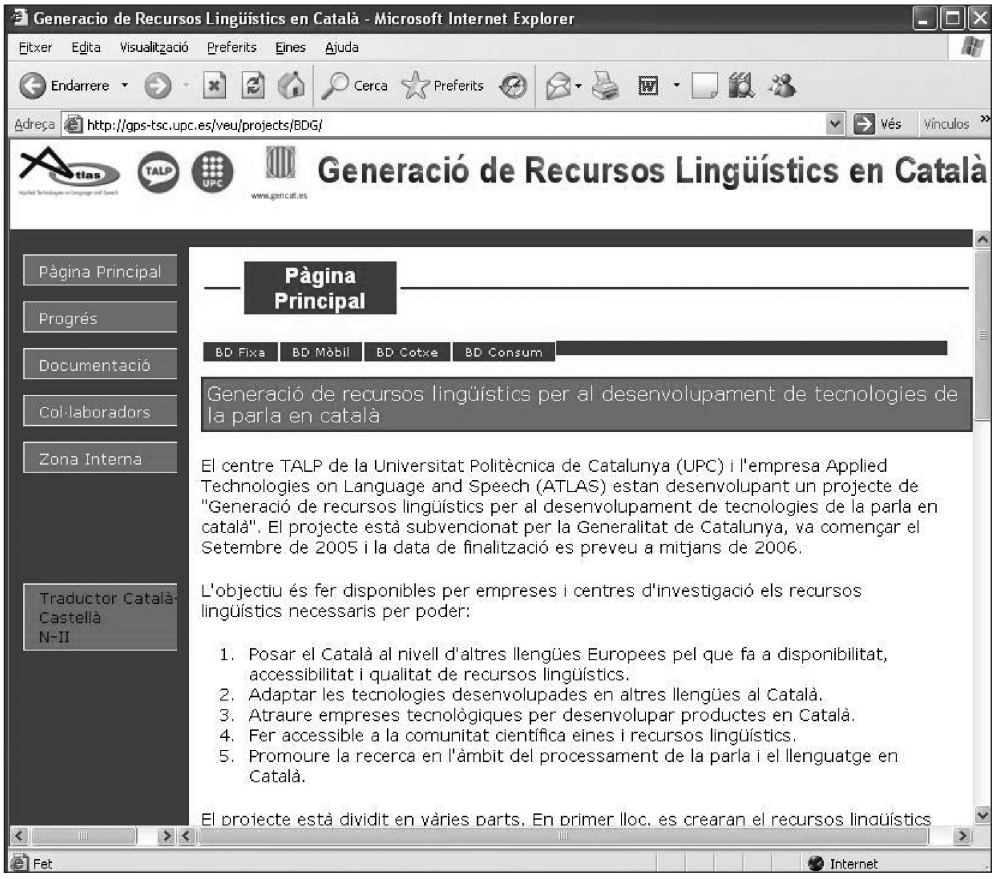
Aquestes bases de dades són útils per a punts d'informació telefònica, per a centraletes telefòniques, per fer reserves de d'hotels, trens, viatges, entrades a espectacles, etc.…

Consum

Durant aproximadament 60 minuts, 550 persones (representants dels quatre dialectes de Catalunya) llegeixen textos mitjançant un sistema equipat de micròfons i en diferents entorns (en llocs públics, a l'oficina, a casa…). Aquestes bases de dades són aplicables a quioscos d'informació, al control d'eines al treball (preferentment ordinadors) i a casa, aen l'àmbit sociosanitari (per ajudar la gent gran)….

Sector automobilístic

Uns 300 informants (delimitats entre dialecte oriental i occidental) llegeixen uns 60 minuts cadascú dins de cotxes en diverses situacions de trànsit i velocitat. Els recursos lingüístics resultants es poden usar per comandar sistemes de



GPS, controlar certs dispositius del vehicle…, i atès que el missatge reconegut es pot enviar a qualsevol lloc a través d'un telèfon mòbil, també s'hi poden incloure aplicacions de control a distància de fax, telefonia...

Totes les bases de dades s'han enregistrat d'acord amb les especificacions establertes en projectes europeus anteriors. Aquestes especificacions inclouen els criteris per al disseny dels corpus, les descripcions de les plataformes d'enregistrament, la distribució dels informants per dialecte, sexe, edat i ambient d'enregistrament, la normativa utilitzada en la transcripció del que realment s'ha pronunciat, la documentació de les bases de dades finals i els criteris per a la validació d'aquestes. Aquesta validació i homologació finals han anat a càrrec de la Universitat de Vigo, centre extern amb experiència en aquest tema i amb la reputació d'haver participat en la validació de recursos lingüístics en projectes europeus.

En definitiva, amb aquest projecte s'ha pretès posar el català al mateix nivell que altres llengües europees pel que fa a disposar de bases de dades orals adreçades a aplicacions industrials.

El producte resultant és públic i gratuït, accessible a través d'ELRA (European Language Resources Association), associació que en un futur immediat difondrà gratuïtament les bases de dades corresponents i les incorporarà en el seu catàleg general, que es distribueix a empreses i universitats d'arreu del món.

<http://gps-tsc.upc.es/veu/projects/BDG/>

RESTAD. Recursos de suport a la traducció automatitzada aplicats a la docència

Aquest projecte cal situar-lo en el marc del nou espai europeu d'ensenyament superior (EEES). Aquest espai significa un canvi en el model d'ensenyament-aprenentatge universitari ——homologacions de les titulacions reconegudes entre diferents països, increment de la mobilitat d'estudiants i professors, importància de la docència virtual i dels materials didàctics elaborats pel professorat, augment de la relació interpersonal professor-alumne, recerca personal de l'alumne com a part essencial del procés d'aprenentatge… ——, amb la qual cosa s'intueixen alteracions en els usos lingüístics docents i discents de la universitat. En aquest context ens cal saber preservar un espai preferent per al català com a llengua de docència, i les noves tecnologies i les tecnologies aplicades al processament del llenguatge en aquest marc del nou disseny metodològic europeu hi tenen un paper fonamental.

El projecte RESTAD[1] està desenvolupat per la Universitat Autònoma de Barcelona, la Universitat de Girona, la Universitat Oberta de Catalunya i la Universitat Politècnica de Catalunya, les quals formen part de la Subcomissió de Traducció Automatitzada de la xarxa d'universitats de tot l'àmbit català, i s'adreça a tot l'alumnat i al personal acadèmic del sistema universitari català.

L'objectiu principal d'aquest projecte consisteix a desenvolupar recursos i eines que facilitin i que millorin la traducció automatitzada al català dels documents docents i dels documents academicoadministratius. Molta docu-

1. Trobareu un article específic sobre RESTAD a la pàgina 41

mentació que el professorat posa a disposició de l'alumne l'alumnat en el campus virtual està en castellà o anglès, i cal disposar de la versió en català; d'altra banda, cal treballar també els materials acadèmics administratius, perquè gradualment augmentarà la necessitat de poder-los oferir en dues o tres llengües.

En definitiva, el projecte RESTAD serveix per gestionar el multilingüisme amb més eficàcia, contribueix a la normalització de la llengua catalana i en garanteix la presència en aquest nou marc europeu, fomenta l'autonomia del professorat en la confecció de materials, millora la qualitat final de la documentació destinada a l'estudiantat i forma en l'ús d'eines informàtiques.

Els recursos que s'han obtingut, els podem classificar en dos grans blocs:

a) Memòries de traducció i bases de dades terminològiques. S'ha sistematitzat i reaprofitat la informació lingüística de què disposen els serveis lingüístics universitaris per fer més fàcil la traducció de materials docents i d'assegurar-ne la qualitat lingüística.

b) Aplicacions per optimitzar els programes de suport a la traducció que ja utilitzen els serveis lingüístics. S'han desenvolupat les eines següents:

— Un alineador automàtic de textos, essencial per a obtenir bones memòries de traducció.
— Un extractor automàtic de terminologia a partir de textos de la memòria de traducció.
— Un cercador automàtic d'equivalents de traducció.
— Un entorn d'edició i selecció dels termes extrets amb els seus equivalents de traducció.
— Plantilles i criteris per a documents docents de postedició.

El desenvolupament d'aquesta tecnologia lingüística s'ha concebut com a programari lliure, se'n garanteix la distribució gratuïta a totes les universitats, institucions, empreses i persones interessades a través de descàrregues des d'Internet, i se'n facilitarà l'adaptació a altres necessitats.

<http://www.uoc.edu/serveilinguistic/home/restad/restad.html>

CeRes, cerca de respostes en un portal web

Promogut per la Universitat de Barcelona, aquest projecte desenvolupa un sistema automàtic de cerca de respostes (CR) basat en tècniques de processament del llenguatge natural (PLN). Els sistemes de cerques de resposta permeten localitzar informació en sistemes tancats (un portal web, una base de dades…) o en sistemes oberts (Internet), a fi d'atendre la pregunta formulada per l'usuari. Aquests sistemes donen com a resposta un fragment de text en llenguatge natural.

El sistema proposat en aquest projecte centra l'espai de cerca de la resposta en un portal web, en concret sobre VilaWeb, i permet accedir a la informació de què disposa aquest web mitjançant preguntes formulades en llenguatge natural.

Fins al moment present, el desenvolupament de la tecnologia de cerca de respostes s'ha centrat en textos escrits en anglès. En altres llengües, com ara el cata-

là i el castellà, ens movem en un terreny verge, en fase d'investigació i, lògicament i tenint en compte aquest context, l'aplicació presentada significa donar un gran impuls a les noves tecnologies en llengua catalana.

El funcionament es basa en el fet que l'usuari formula una pregunta factual (qui, què, com, quan) en català sobre la base de dades de notícies periodístiques de VilaWeb. El sistema processa la pregunta i identifica què és allò que l'usuari demana. A continuació accedeix a la base de dades de notícies i fragmenta els documents, i selecciona els segments de text que poden contenir la resposta. A partir de la informació obtinguda en l'anàlisi de la pregunta, tria, d'entre els segments candidats, els que tenen més probabilitat de contenir la resposta. Tot seguit ofereix a l'usuari els segments que ha seleccionat. Exemples de preguntes factuals serien *qui és el president del Govern espanyol?*, o *quin és el riu més llarg del món?*

El tipus de preguntes que el sistema admet són de resposta sobre fets, persones, o entitats concrets; amb la qual cosa, davant la demanda d'informació, per part de l'usuari i amb un llenguatge quotidià, sobre un tema, esdeveniment, o acte específics, s'obté una resposta ràpida i precisa i que satisfà les necessitats prèvies. Aquest és l'objectiu clau de CeRes: proporciona una resposta i no una llista de documents sobre el tema objecte de la pregunta formulada per l'usuari.

Aquesta aplicació és de lliure accés i es preveu que en un termini relativament curt s'instal·li al servidor de VilaWeb.

TexToral, conferències universitàries multilingües

Elaborat per la Universitat de Barcelona, es tracta d'un producte multimèdia consistent en un corpus de conferències en català fetes per professorat universitari, que serveix de material d'autoaprenentatge per a la comprensió oral i escrita de la llengua catalana. Permet consultar oralment i per escrit discursos de divulgació científica i, a la vegada, també facilita versions del mateix text en altres llengües, principalment castellà i anglès, i proposa exercicis per treballar la comprensió.

Aquest projecte s'emmarca en el context de la mobilitat universitària i professional en l'àmbit europeu i internacional, i també en el marc de la mobilitat interuniversitària de caràcter estatal. Davant aquest fet, és essencial oferir eines a l'estudiantat de fora de l'àmbit lingüístic català, tant provinent de l'Estat com de l'estranger, perquè preferentment abans d'arribar a Catalunya o un cop arribat a la universitat catalana, tingui cert nivell de comprensió oral i escrita en català.

Amb un aprenentatge inicial basat en aquesta comprensió oral i escrita, habilitats lingüístiques bàsiques que li permetran entendre el professorat a classe i comprendre mínimament la documentació acadèmica, l'alumnat podrà, en una segona fase, incrementar els llindars del seu aprenentatge.

Tot i així, el destinatari últim d'aquests materials didàctics no és tan sols l'alumnat de fora de l'àmbit lingüístic català, sinó que aquest producte multimèdia s'adreça també al professorat que, procedent de la Unió Europea, s'incorpora al sistema universitari català, i als lectors, casals i càtedres de català d'arreu del món que en vulguin fer ús.

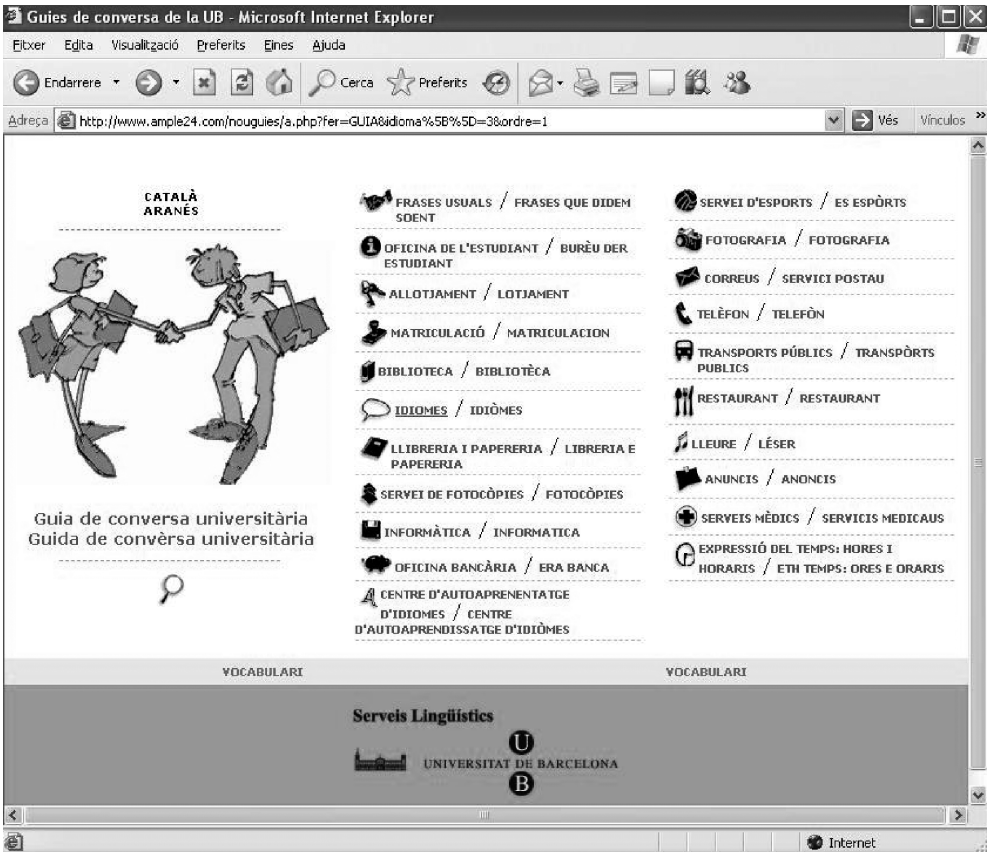
Actuació pensada dins del marc d' Interc@t, programa d'acollida lingüística, és d'accés lliure a través d'Internet.

Guia de conversa català-castellà-aranès-euskera-gallec

Elaborada per la Universitat de Barcelona, aquesta guia forma part de la col·lecció *Guies de conversa universitària*, adreçada a la població estudiant universitària estrangera que fa part dels seus estudis a les universitats de l'àrea lingüística catalana.

La Universitat de Barcelona, en la seva preocupació per acollir amb qualitat aquest públic i afavorir-ne una estada profitosa tant en l'aspecte acadèmic i cultural com en el personal, ha elaborat aquestes guies, l'objectiu de les quals és proporcionar recursos lingüístics per comunicar-se i facilitar l'aprofitament dels serveis que ofereix la universitat.

La *Guia de conversa català-castellà-aranès-euskera-gallec* inclou, en una mateixa edició —i amb els arxius de veu corresponents—, totes les llengües amb rang d'oficialitat a l'Estat espanyol. El seu públic principal, el constitueix l'alumnat



que, des del Districte obert (universitats espanyoles de fora de l'àmbit territorial català), s'incorpora a les universitats catalanes.

La guia conté més de 2.000 frases, organitzades per àmbits —matriculació, allotjament, transport, restaurant, etc…—, que ajuden la persona usuària d'aquests entorns en les situacions de relació personal. També inclou nocions de gramàtica i un vocabulari d'equivalències en totes les llengües.

Es tracta d'una eina que permet a l'alumnat utilitzar les frases més encertades i més eficaces en les situacions comunicatives en queuè s'anirà trobant des que arribi al nostre territori. I si, a més, contribueix a augmentar el seu coneixement i el seu domini del català, podrem avaluar-ne l'èxit i afirmar que s'ha assolit l'objectiu final.

<http://www.ub.edu/sl/guia/>

SOLC. Servidor ortològic de la llengua catalana

Aquest projecte de gran interès lingüístic impulsat per la Universitat Pompeu Fabra, té com a objecte l'elaboració d'un diccionari ortològic per resoldre els dubtes dels professionals de la llengua oral i que dóna cobertura a totes les variants considerades en la proposta d'estàndard oral de la Secció Filològica de l'Institut d'Estudis Catalans.

Evidentment, tractant-se com que es tracta d'un diccionari que havia de cobrir un lèxic que no para de créixer amb una multiplicitat de formes i amb nombroses entrades enciclopèdiques, aquesta publicació hauria estat inviable si s'hagués hagut d'elaborar en edició paper, principalment per la seva condició efímera i per la seva caducitat. Per aquest motiu, s'ha concebut com una base de dades accessible per Internet, com un diccionari en línia. A partir d'aquí, destaquen una sèrie de valors afegits: que respon millor a la immediatesa del treball quotidià als mitjans de comunicació orals, i que constitueix una formalització computacional de la proposta d'estàndard oral de la Secció Filològica de l'Institut d'Estudis Catalans, amb la qual cosa, a més d'alimentar el diccionari en línia, permet als ortòlegs estudiar els límits, les incongruències i les assistemacitats que pugui tenir, o no, la proposta esmentada.

Els objectius específics del projecte han estat implementar la proposta d'estàndard oral del l'IEC; sistematitzar-la i aplicar-la exhaustivament, modelitzar-la per preveure'n les evolucions; recopilar tota la casuística que s'avé a un tractament regular, i separar l'ortologia sintàctica de la lèxica.

Aquest producte serà accessible a través d'Internet i estarà a l'abast de tots els professionals de la llengua oral que en vulguin fer ús.

Desenvolupament d'una plataforma de treball en xarxa via Internet en neologia catalana i millora de les estratègies de detecció automàtica de neologismes

Aquest darrer projecte, impulsat també per la Universitat Pompeu Fabra, es concreta en la creació d'una plataforma de treball a través d'Internet, que permeti, d'una banda, col·laborar activament amb l'Institut d'Estudis Catalans, en el recull i l'anàlisi de neologismes procedents de les diverses varietats dialectals, i, de l'altra, facilitar l'accés a les dades a través d'una interfície web.

Això implica el desenvolupament d'una aplicació web d'accés a les bases de dades, la creació d'una plataforma de treball de detecció i anàlisi de neologismes en català en xarxa i a distància mitjançant Internet i que permeti introduir informació recollida pels observatoris de neologia d'altres llengües romàniques, i el perfeccionament del procés de detecció automàtica de neologia.

Actualment s'han assolit els dos primers objectius d'aquest projecte complex:

-S'ha desenvolupat la plataforma d'accés a la informació multilingüe que permet accedir, via web, a la base de dades de l'Observatori de Neologia de la Universitat Pompeu Fabra, i que permet incorporar tots els buidatges d'altres grups que treballen en neologia i que utilitzen les mateixes eines metodològiques. El recurs permet fer cerques bàsiques i cerques avançades.

La plataforma també permet registrar neologismes dels diferents dialectes del català. A la vegada s'ha creat una bústia neològica que permet a qualsevol usuari enregistrar neologismes procedents dels mitjans de comunicació. El neologisme, un cop revisat, forma part de la base de dades de l'Observatori i, consegüentment, es pot consultar a través de la plataforma d'accés a la informació BOBNEO.

-S'ha creat la xarxa d'observatoris de neologia per a la llengua catalana (NEOXOC), integrada per sis grups representatius de diverses variants diatòpiques del català (la Universitat de Perpinyà, la Universitat de les Illes Balears, la Universitat Rovira i Virgili, la Universitat d'Alacant, la Universitat de Lleida, la Universitat d'Andorra i la Universitat Pompeu Fabra).

D'aquesta interfície web d'accés obert, se'n pot beneficiar tota la comunitat lingüística, ja que posa a disposició de qualsevol ciutadà o ciutadana i de tots els professionals que es dediquen professionalment a la llengua —en el camp de la traducció, correcció, ensenyament de llengües, edició, periodisme...—, un corpus de paraules noves permanentment actualitzat i amb informació sobre el seu ús real.

<http://obneo.iula.upf.edu/bobneo/>

En resum, tots aquests projectes són fruit de l'interès i la recerca de les universitats per aplicar noves tecnologies de la informació i la comunicació a la llengua catalana, i que ha suposat, en aquest cas concret, l'obtenció de noves eines d'aprenentatge en línia, la creació de nous recursos terminològics, l'elaboració i la traducció automàtica de materials universitaris de suport a la docència, el desenvolupament de productes tecnològics i industrials...; això sí, tot en llengua catalana, i amb el resultat evident que no tan sols se'n beneficia la comunitat universitària, sinó també els col·lectius professionals relacionats amb aquests àmbits i, en general, tota aquella ciutadania que vol viure en català amb normalitat.

Es tracta, doncs, d'un conjunt de projectes que interrelacionen, d'igual manera, la recerca, la tecnologia i la llengua catalana: tres factors que constitueixen el denominador comú de tots els productes resultants.

■