ELISABETH GIBERT-SOTELO · ISABEL PUJOL PAYET
ASSUMPCIÓ ROST BAGUDANCH · TERESA DE JESÚS TRO MORATÓ
(eds.)

LINGUISTICA SINE FINIBUS

Elisabeth Gibert-Sotelo, Isabel Pujol Payet,
Assumpció Rost Bagudanch, Teresa de Jesús Tro Morató
(eds.)

LINGUISTICA SINE FINIBUS

ESTUDIS DEDICATS A MONTSERRAT BATLLORI DILLET

Dades CIP recomanades per la Biblioteca de la UdG

CIP 806.0 LIN

Linguistica sine finibus : estudis dedicats a Montserrat Batllori Dillet / Elisabeth Gibert-Sotelo, Isabel Pujol Payet, Assumpció Rost Bagudanch, Teresa de Jesús Tro Morató (eds.). – Girona : Universitat de Girona : Documenta Universitaria, 2023. – 578 pàgines : il·lustracions, taules, fotografies ; cm
ISBN 978-84-9984-671-2 (Document Universitaria).
ISBN 978-84-8458-668-5 (Universitat de Girona. Servei de Publicacions)

I. Gibert Sotelo, Elisabeth, editor literari II. Pujol Payet, Isabel, editor literari III. Rost Bagudanch, Assumpció, editor literari IV. Tro Morató, Teresa de Jesús, editor literari 1. Batllori Dillet, Montse 2. Llibres homenatge 3. Lingüística històrica

CIP 806.0 LIN

Aquesta publicació és part del projecte I+D+i PID2021-123617NB-C42, finançat per MICIU/AEI/10.13039/501100011033 i per FEDER, UE.

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN

En el seu finançament també hi han col·laborat la Facultat de Lletres i el Departament de Filologia i Comunicació de la Universitat de Girona.

Universitat de Girona
Facultat de Lletres

Universitat de Girona
Departament de Filologia i Comunicació

ISBN Servei de Publicacions de la UdG: 978-84-8458-682-1

ISBN Documenta Universitaria: 978-84-9984-616-3
DOI: 10.33115/c/9788499846163_24

Girona, 2023

ÍNDEX

LA VARIACIÓ I LA LINGÜÍSTICA DE CORPUS

JOAN TORRUELLA
ICREA – UAB
joan.torruella@uab.cat

DOI: 10.33115/C/9788499846163_24

Keywords
Textual corpora, corpus linguistics, variables, frequencies, linguistic variation.

Paraules clau
Corpus textuals, lingüística de corpus, variables, freqüències, variació lingüística.

Abstract

This paper aims to showcase the new prospects in the studies regarding the fields of linguistics in general and variation in particular that computerized textual corpora and the methodology to use them on a scientific basis stipulated by corpus linguistics can help researchers. We focus on two fundamental concepts which are often adopted in the scientific method: on the one hand, absolute and relative frequency, and on the other, dependent and independent statistic variables. These notions should be used carefully and correctly when we want to obtain data from corpora, so it helps us to interpret and understand the underlying reasons behind the data, and it enables us to develop working hypotheses. To illustrate this, we explain the use of these terms in one example referring to graphic and phonetic variation (*nuit*, *nuyt* i *nit*).

Resum

En aquest treball es volen posar en valor les noves possibilitats, en l'estudi de la lingüística en general i en el de la variació en particular, que els corpus textuals informatitzats i la metodologia per utilitzar-los amb bases científiques que estipula la lingüística de corpus obren als investigadors. En aquest cas, ens centrem en dos conceptes bàsics en el moment de fer recerca amb mètodes científics: el de freqüències absolutes i freqüències relatives, i el de variables estadístiques dependents i independents. Aquests dos conceptes cal utilitzar-los i cal fer-ho curosament i correcta quan es volen obtenir dades dels corpus que permetin ser interpretades de manera que ajudin a entendre el seu perquè i a desenvolupar hipòtesis de treball. Per exemplificar el tema es tanca el capítol amb l'explicació de la utilització d'aquests dos conceptes en el cas referit a la variació gràfico-fonètica (*nuit*, *nuyt* i *nit*).

DOI: 10.33115/C/9788499846163_24

1. INTRODUCCIÓ

Els corpus textuals[1] són recursos informàtics que faciliten enormement l'obtenció de dades, no solament perquè fan possible recollir-ne moltes més de les que es podien recollir en les lectures pacients dels textos d'un corpus, sinó, sobretot, perquè ajuden a ordenar-les, classificar-les i a quantificar-les, quelcom imprescindible per poder-les interpretar.

Inicialment, els corpus textuals servien per buscar exemples que recolzessin una teoria lingüística ja establerta, però, amb l'aparició dels nous mètodes de recerca que proposa la lingüística de corpus,[2] el procés ha canviat i és possible fer el camí a l'inrevés, primer estudiar els textos a partir de les anàlisis amb mètodes estadístics de les seves dades i, avaluant els resultat d'aquestes anàlisis, establir hipòtesis de treball i vestir teories lingüístiques. És a dir, els corpus textuals avui fan de pont entre la realitat lingüística existent en els seus textos i la teoria lingüística que es vol desenvolupar.

Com la resta de les disciplines lingüístiques, l'estudi de la variació —aquesta disciplina que tant i tan bé ha estudiat la homenatjada en aquest llibre, la professora Montserrat Batllori—,[3] ha obtingut ajudes molt notables a partir de l'aparició dels corpus textuals i de la lingüística

1 Actualment el terme *corpus* es fa servir en un sentit molt general i, també, en un sentit més restringit. Si s'observa la seva definició en el *DIEC*, es pot comprovar com de les accepcions referides a l'àmbit de la filologia, les dues primeres són de caire general: «2 *1 m.* [FL] [FLL] Col·lecció general d'escrits. 2 *2 m.* [FL] [FLL] Tota la literatura sobre una matèria» i només la tercera es refereix al que s'entén per corpus en l'àmbit de la lingüística de corpus: «2 *3 m.* [FL] Conjunt d'enunciats o de textos utilitzats en l'anàlisi i la descripció lingüística d'una llengua». En aquest cas, per diferenciar-lo de les altres accepcions, se sol parlar de *corpus textual*.

2 La lingüística de corpus és una disciplina que tracta del procés d'emmagatzemar i recuperar dades reals i verificables, així com del de processar adequadament aquestes dades per a poder comprovar empíricament la validesa o no d'hipòtesis de treball.

3 Amb aquesta nota vull agrair-li els anys de camí professional junts i l'amistat que sempre hi ha hagut.

DOI: 10.33115/C/9788499846163_24

de corpus. Els corpus textuals informatitzats aporten a les nostres investigacions, major fiabilitat (augment considerable del volum de textos que es poden consultar), major precisió en les anàlisis (augment de les possibilitat d'interrogació) i la possibilitat d'aplicar mètodes de recerca científics (viabilitat de treballar amb diferents tipus de variables i de comptabilitzar freqüències).

En el seu moment, ja vaig fer notar que

> «sin los avances que suponen los métodos de la nueva disciplina llamada *lingüística de corpus*, a los investigadores se les pasarían por alto muchos datos e informaciones importantes para el estudio de la lengua imposibles de detectar siguiendo el método tradicional de "filología de sillón"» (Torruella 2017: 16).

2. LA RECERCA CIENTÍFICA

La recerca científica en general, i també l'aplicada a la lingüística, es pot dur a terme des de dos mètodes: el mètode experimental i el mètode observacional o comparatiu. En el mètode experimental l'investigador pot actuar, segons els seus interessos, sobre l'objecte d'estudi (per exemple, en fonètica actual, en un estudi de gravació es poden fer i refer enregistraments a diferents informants de característiques determinades per l'investigador), en canvi, en el mètode observacional o comparatiu l'investigador no pot actuar sobre l'objecte d'estudi (les dades són les que hi ha i en la forma que hi ha) i ha de basar les seves recerques en la quantificació i la comparació sistemàtica dels elements d'anàlisi per tal de poder establir les causes de les seves analogies i de les seves diferències, i poder extrapolar-ne els resultats.

És clar que quan es tracta d'estudiar la llengua des del punt de vista diacrònic no queda altre remei que utilitzar el mètode comparatiu, ja que no es pot crear i recrear la realitat lingüística que es vol analitzar en un laboratori i aplicar tècniques del mètode experimental.

Fins fa poc, especialment quan es volia estudiar la variació des del punt de vista diacrònic, els recursos per obtenir dades que l'investigador tenia al seu abast eren la consulta a atles lingüístics, principalment per a la variació diatòpica, la consulta a diccionaris, prioritàriament per a la variació

DOI: 10.33115/C/9788499846163_24

diacrònica i semàntica, i, finalment, la lectura pacient i la confecció de fitxes de les obres que es consideraven més rellevants per a cada recerca.

De fet, atles, diccionaris i corpus són eines que l'investigador té al seu abast, cada una d'elles amb característiques pròpies i finalitats diferents, però que esdevenen

> «tres fuentes de información que, a pesar de que, algunas veces, aportan datos que se contradicen (normalmente atlas y corpus *versus* diccionarios), en la mayoría de las ocasiones, ofrecen informaciones coincidentes que se complementan mutuamente. Cada recurso aporta perspectivas, matices y puntualizaciones que no pueden ofrecer los otros. Sin embargo, [cuando se trata de diacronía] se debe tener presente que los tres recursos dan cuenta de lo que existe, pero no de todo lo que existe y, mucho menos, de lo que no existe» (Torruella 2021: 45).

La lingüística de corpus facilita la recerca científica aplicant el raonament inductiu, ja que permet poder generalitzar els resultats particulars que s'obtenen a partir de les anàlisis dels exemples reals obtinguts en els corpus textuals. En aquest mètode, es parteix d'un conjunt d'exemples significatius (contextos en les concordances obtingudes dels corpus textuals informatitzats) que comparteixen una característica concreta amb la finalitat de poder inferir que aquesta característica es donarà també en elements semblants o en situacions anàlogues, o d'observar com evolucionen numèricament en el temps, en l'espai o en la situació comunicativa. En el cas de la variació lingüística, com ja indica la pròpia paraula, el que es pretén és observar com canvien en diferents àmbits les distintes variants d'una determinada unitat lingüística (ja sigui fonètica, morfològica, sintàctica, etc.) i analitzar-ne els motius.

Així mateix, la variació lingüística tracta realitzacions concretes de la llengua i aquesta com a fet determinat i social que és, es realitza en un temps concret, en un espai específic i en una situació comunicativa precisa. Per aquest motiu, l'estudi de la variació lingüística no pot tenir en compte només un factor, el diacrònic, el diatòpic, el diafàsic o el diastràtic, per exemple, sinó que ha d'observar la materialització de cada variant en els seus distints paràmetres d'àmbit i comprovar les influències que es produeixen.

Però els corpus textuals, perquè siguin eficients a l'hora d'oferir dades comparables, han d'estar estructurats en eixos de diferents àmbits (diacrònic, diatòpic, diafàsic, etc.) segons la finalitat que cada corpus

DOI: 10.33115/C/9788499846163_24

vol tenir, i cada un d'aquests eixos d'àmbit ha d'estar dividit en diferents apartats, de manera que permetin establir les diverses variables de treball i comprovar com s'interrelacionen entre elles. Així, doncs, quan es vol estudiar la variació en la llengua, cal, d'entrada, cercar el conjunt d'exemples de la variant o variants que es vol estudiar, observar com aquests es comporten numèricament dins dels diferents apartats dels àmbits de la variació (diacrònic, diatòpic, social i situacional) i extrapolar els resultats de les mostres estudiades al total de la població.

Per exemple, a l'estudiar el comportament de les dues variants lèxiques per al concepte '*Quercus ilex*', *alzina* i *carrasca*, no n'hi ha prou en observar la seva presència o absència en l'eix diacrònic o en l'eix diatòpic per separat, sinó que és en l'observació de conjunt del que passa a cada una de les dues variants lèxiques quan es canvia d'apartat en cada eix d'àmbit (per exemple, quan es canvia de segle en l'eix diacrònic o de territori en l'eix diatòpic) que es pot tenir una idea prou clara de la presència espacio-temporal de les dues variants en la llengua catalana.

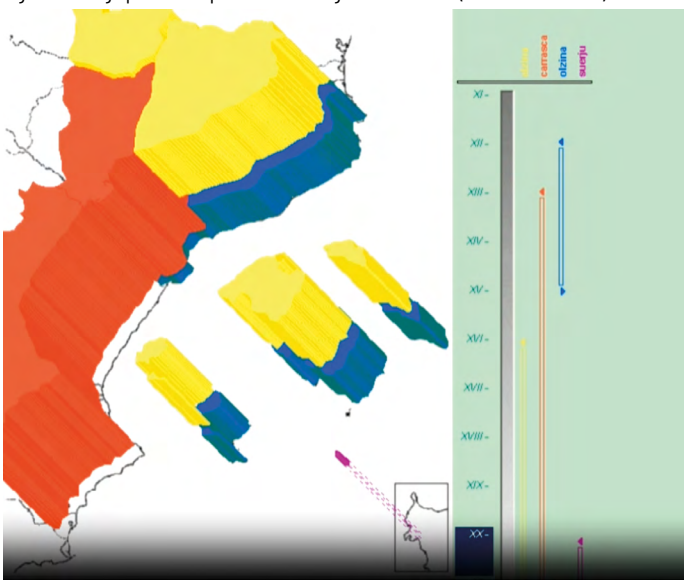
En l'aplicació informàtica *Estratigrafia dialectal* de Maria-Pilar Perea i Germà Colón (explicada a Perea i Colon 2010), es pot observar com s'interrelacionen les variables independents de temps i espai geogràfic en el conjunt de les quatre variants per al concepte 'Quercus ilex' (variables dependents): *alzina*, *carrasca*, *olzina* i *suerju*.

En aquest cas, s'ha establert un color per a cada variant lèxica (variables dependents) i a la banda dreta del mapa dels Països Catalans s'hi ha col·locat un eix vertical en el que s'hi marquen els segles amb un botó que l'usuari pot moure des del segle XI fins el XXI. Així, observant les freqüències de cada una de les variants en els diferents apartats de l'eix diacrònic es pot comprovar com, des del punt de vista cronològic, la primera variant documentada per al concepte '*Quercus ilex*' és *olzina*, que apareix a principis del segle XII, però desapareix a finals del segle XIV. La següent variant en aparèixer és *carrasca*, que es comença a documentar a principis del segle XIII i ja no desapareix de la llengua catalana. La variant *alsina*, no es comença a documentar fins a principis del segle XVI, i tampoc desapareix posteriorment. Finalment, la variant *suerju* no es comença a documentar fins a principis del segle XX i tampoc desapareix de la llengua catalana. Però el que és més revelador, és quan aquestes dades es complementen amb les de l'eix geogràfic, de manera

DOI: 10.33115/C/9788499846163_24

que es pot observar com cada una d'aquestes variants es va estenent o retraient en el territori de parla catalana segons el segle que s'analitza. Així, la variant *olzina* al segle XII ja es documenta tant a les Illes Balears com a Catalunya i mai passarà al País Valencià. Per contra, la variant *carrasca*, ja d'entrada es documenta en tot el País Valencià i en la part sud-oest de Catalunya, i aquests límits geogràfics no es mouran al llarg de tot el temps. La variant *alzina*, és una variant que geogràficament es complementa amb la de *carrasca*, i així, ja des del seu inici es troba a Catalunya, menys en la part sud-oest on s'utilitza *carrasca*, i a totes les Illes Balears, també des del seu inici fins al dia d'avui. Finalment, la variant *suerju* es documenta, a partir del segle XX i només, a l'Alguer.

Figura 1. Imatge presa de l'aplicació «Estratigrafia dialectal» (Perea i Colon 2022)



3. DOS CONCEPTES BÀSICS

Quan es tracta de fer recerca científica seguint el mètode observacional o comparatiu utilitzant corpus textuals sota les premisses de la lingüística de corpus, es necessita treballar amb dos conceptes de caire estadístic: les *freqüències* i les *variables*.

DOI: 10.33115/C/9788499846163_24

3.1. Les variables

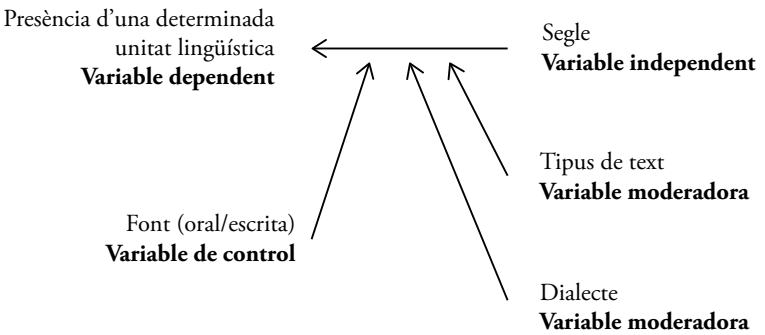
Com ja van expressar Marta Gómez i Nieves Sánchez (2010:109), «La variación es un hecho fundamental en las lenguas que se manifiesta en sus distintos aspectos y niveles, entre ellos el cronológico, el geográfico o el social», per això, per estudiar científicament la variació a partir de corpus textuals i poder aplicar les tècniques estadístiques en les seves anàlisis, aquests han d'estar estructurats en uns eixos principals que responguin a aquests «aspectos o niveles» que han esmentat les dues autores (nosaltres els anomenem «àmbits») i, dins de cada un d'aquests eixos d'àmbit, s'han de definir diferents apartats.[4] D'aquesta forma, en una investigació, els diferents eixos d'àmbit es podran utilitzar com a *variables independents*, mentre que els elements lingüístics —entre ells els referents a la variació— que es volen analitzar es podran emprar com a *variables dependents*, de manera que sigui possible observar com reacciona la *variable dependent* que s'està analitzant quan es canvien els valors de les *variables independents* implicades en la investigació. Es tracta de quantificar com canvia el nombre de casos (les *freqüències*) d'un determinat exemple dins de cada un dels apartat que conformen els eixos d'àmbit.

Així, doncs, per treballar amb corpus seguint un mètode científic i poder aplicar les tècniques estadístiques en les anàlisis de les dades que aquests proporcionen, cal disposar de *variables* que facin possible l'experimentació, de manera que, quan l'investigador canviï els apartats d'alguna de les *variables independents*, es pugui observar com reacciona la *variable dependent* que s'està analitzant. Per exemple, què succeeix numèricament amb la variant lèxica *vesprada* (*variable dependent*) quan en l'eix de l'àmbit geogràfic (*variable independent*) es canvia de dialecte. Així, els termes *variable dependent* i *variable independent* fan referència a valors que poden evolucionar de forma correlacionada.

4 Per definir els diferents apartats d'un eix principal d'un corpus cal seguir els següents criteris: 1. Els diferents apartats de cada eix classificador han d'haver estat dissenyats amb els mateixos criteris. 2. Els trets distintius de cada apartat han de ser excloents amb els dels altres apartats del mateix eix classificador. 3. Cada text del corpus ha d'encaixar en un apartat i, si és possible, només en un. Cada projecte de construcció d'un corpus ha d'establir, segons millor consideri, els paràmetres dels diferents apartats (anys, segles, èpoques històriques, etc.).

DOI: 10.33115/C/9788499846163_24

Figura 2. Variables de recerca



Les *variables dependents* són les que evolucionen (o no) en resposta als canvis dels valors de les *variables independents*. Les *variables independents* són les que es manipulen de forma deliberada per provocar els canvis de les *variables dependents*.

Per exemple, si es vol estudiar en quina època l'estructura antiga del possessiu en català, «article + possessiu + nom» (*la meva casa*), va passar a l'estructura actual de «nom + possessiu» (*casa meva*) i saber si les dues estructures van conviure en el temps i quan una estructura es va imposar a l'altra,[5] només cal utilitzar com a *variable independent* els diferents apartats de l'eix temporal i observar si les freqüències de cada un dels dos fenòmens, és a dir de les dues estructures del possessiu (*variables dependents*), tenen relació, com ara que a l'augmentar la freqüència d'una disminueixi la de l'altra.

Si un altre dels eixos d'àmbit dissenyats al construir el corpus és el tipològic i també es vol saber si el tipus de text va tenir influència en el procés de canvi entre les dues estructures del possessiu, caldrà gestionar també com a *variable independent* els diferents apartats de l'eix tipològic i esbrinar si la freqüència del fenomen estudiat (*variable dependent*) canvia entre els diferents apartats referents al tipus de text.

Si, a més, es vol saber si la zona geogràfica de procedència dels textos pot tenir rellevància en el canvi, caldrà gestionar com una altra *variable*

5 L'estudi de les dues variants sintàctiques del possessiu en català (article + possessiu + nom / nom + possessiu) es troba a Pérez Saldanya (2009).

DOI: 10.33115/C/9788499846163_24

independent els diferents apartats de l'eix de l'àmbit diatòpic i observar si la freqüència del fenomen estudiat (*variable dependent*) canvia entre els diferents apartats referents a aquest àmbit.

Finalment, serà l'anàlisi en conjunt de la influència de cada un dels eixos en la variable dependent, qui donarà una visió general del comportament del fenomen estudiat.

Però en una investigació d'aquest tipus no només intervenen les *variables dependents* i les *variables independents*, sinó que també poden (i moltes vegades deuen) intervenir-hi altres tipus de variables com les *variables de control*, les *variables moderadores*, les *variables aleatòries*, *variables intermèdies*, etc. que són variables que poden influir en l'efecte que tenen les *variables independents* sobre la *variable dependent*.

D'aquestes variables, les més utilitzades són la *variable moderadora* i la *variable de control*.

Una *variable moderadora* és un tipus de *variable independent* que pot influir en el grau de relació entre la *variable independent* i la *dependent*. Per exemple, la variable diatòpica pot influir, i a l'hora explicar, el resultat de la variable diacrònica. Les *variables moderadores*, quan convé, poden esdevenir *variables independents*.

Una *variable de control* és una variable que l'investigador vol tenir vigilada per tal que no canviï en el transcurs de la investigació perquè podria afectar els valors de les *variables dependents* i, en conseqüència, el resultat final de l'anàlisi. Per exemple, si es creu que el fet que els textos analitzats estigui escrit per un home o per una dona pot influir en el resultat, es controla la variable gènere de l'autor, analitzant només textos d'un sol gènere.

3.2. Les freqüències

L'altre concepte estadístic important a l'hora de treballar la variació a partir de corpus és el de les *freqüències*, ja que la quantificació dels elements lingüístics, la seva ordenació, classificació en apartats i la comparació numèrica entre aquests apartats és el camí per intentar resoldre hipòtesis de treball.

DOI: 10.33115/C/9788499846163_24

Johannes Kabatek (2006: 171) ho expressava de la següent manera: «la cuantificación de elementos nunca va a ser sustituto del análisis filológico de detalles, pero es una base objetiva para la comparación, fundamento de cualquier estudio de evolución histórica», i ho reafirmava dient:

> «La lengua no son números. Pero los números, como bien dice Joan Torruella, permiten una proyección de los datos lingüísticos que posibilitará su análisis objetivo. Sin embargo, la "objetividad" científica no reside solo en el tratamiento numérico adecuado, sino también en el paso previo: el rigor metodológico necesario para la transformación de textos en datos numéricos.» (Kabatek 2017: 13)

Està clar que la *freqüència* és l'element base que ens permet fer comparacions de la presència d'un mateix element lingüístic entre diferents apartats d'un mateix àmbit o de diferents elements lingüístics dins d'un mateix apartat, la qual cosa és el fonament per a la obtenció de dades.

Però en estadística hi ha dos tipus de *freqüències*: les absolutes i les relatives. La *freqüència absoluta* és una mesura que ens dóna informació sobre la «quantitat» de vegades que es repeteix un mateix fet, la *freqüència relativa* és una mesura que ens indica la «proporció» que la *freqüència absoluta* té en el total de la població.[6]

En el cas dels corpus textuals, si les comparacions es fan entre diferents elements lingüístics dins d'un mateix apartat, com que el nombre total de mostres és el mateix, s'utilitza la *freqüència absoluta* per expressar les vegades que succeeix un fet i la *freqüència relativa* per expressar la proporció en què aquest succeeix. Per contra, quan la comparació es fa sobre la presència d'un mateix element lingüístic en diferents apartats, com que, per molt equilibrats[7] que siguin els corpus, el nombre total de mostres que tenen els diferents apartats d'un eix d'àmbit no serà mai igual, la freqüència que s'ha d'utilitzar és la relativa, ja que no té el mateix valor que un fet lingüístic ocorri 16 vegades en un total de 100 mostres que ocorri 16 vegades en un total de 200 mostres. En el segon cas, el valor

6 La freqüència relativa (n_i) és el quocient entre la freqüència absoluta (fi) d'un valor determinat i el nombre total de dades de la població (N), i se sol expressar en tants per cent. La suma de las freqüències relatives sempre és igual a 1.

7 Equilibri referit al fet que la quantitat de paraules recollides tingui una proporció adequada respecte del total i una distribució apropiada en cada apartat del corpus. Per a més detalls, veure Torruella 2017, pàg. 129 i següents.

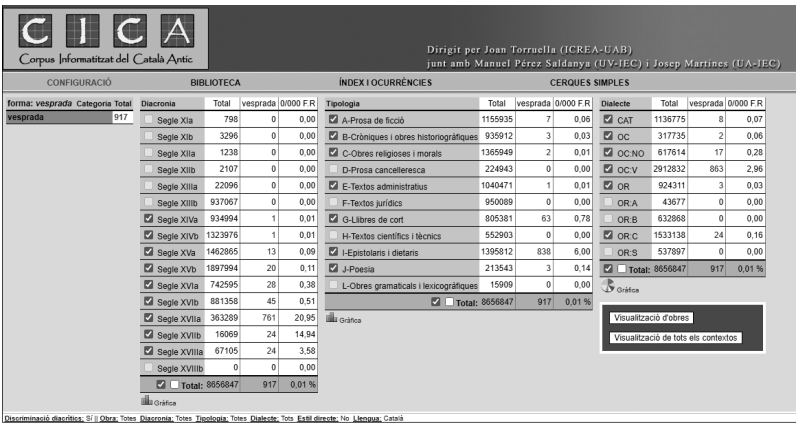
DOI: 10.33115/C/9788499846163_24

absolut és el mateix, però el valor proporcional és exactament la meitat que en el primer.

Per exemple, si en una mostra de 1000 paraules la variant A ocorre 230 vegades, la variant B ocorre 570 vegades i la variant C ocorre 200 vegades, les seves *freqüències relatives* són, respectivament, d'A = 0,23, B = 0,57 i C = 0,20 i la suma total = 1. Però si la mostra és de 2000 paraules, les *freqüències relatives* passen a ser la meitat A = 0,11, B = 0,28, C = 0,10 [ALTRES = 0,50]. Suma total = 1.

En la imatge següent, treta del *Corpus Informatitzat del Català Antic* (*CICA*), es pot apreciar com, en l'eix referent a la diacronia, la paraula *vesprada* en la segona meitat del segle xvii i en la primera meitat del segle xviii tenen el mateix valor de la *freqüència absoluta* (tercera columna), que és de 24, però el valor de la *freqüència relativa* (quarta columna) és de 14,94 per mil en la segona meitat del segle xvii i de 3,58 per mil en la primera meitat del segle xviii.[8] La diferència està en què el total de mostres (segona columna) de la segona meitat del segle xvii és de 16.069 i el de la primera meitat del segle xviii és de 67.105.

Figura 3. Distribució de les freqüències absolutes i relatives de la paraula vesprada en els tres eixos d'àmbit presents en el CICA



8 Quan hi ha molts ceros en els valors decimals és més visual donar el *tant per mil* que no pas el *tant per cent*.

DOI: 10.33115/C/9788499846163_24

L'anàlisi de la distribució de *freqüències relatives* en els diferents paràmetres del corpus no sols dóna una visió de la dimensió del fet estudiat, sinó que pot apuntar a possibles explicacions.

Per exemple, una visió de conjunt de les *freqüències relatives* en els diferents apartats dels eixos diacrònic, tipològic i dialectal del corpus *CICA* ens apunta a què la variant *vesprada* (variant lèxica per al concepte 'tarda' o per al de 'crepuscle'), no apareix fins al segle XIV, que pren força al segle XVII, que s'utilitzava principalment en el *català occidental: Valencià*, i que s'emprava majoritàriament en textos de formalitat baixa i d'oralitat escrita (dietaris i epistolaris). La distribució de les *freqüències relatives* d'aquesta variant es podria comparar amb la distribució en altres variants del mateix concepte, com ara *vespre*, *horabaixa*, *vesprada* i, així, es podria dibuixar la vida de cada una d'elles i les seves interaccions.

De totes maneres, a vegades, el valor de les *freqüències* s'ha de relativitzar, especialment quan es tracta de corpus històrics, ja que quan el nombre de mostres és petit,[9] el fet que el valor de la *freqüència* d'un cas determinat sigui 0 no permet afirmar, amb un marge d'error acceptable, que el cas no hagi existit, només permet certificar que no apareix als documents que configuren el corpus o en els documents que han pervingut fins al dia d'avui. En canvi, si el valor és 1 o superior, ja es pot assegurar que el cas ha existit, encara que no es pugui inferir en quina proporció.

També s'ha de tenir en compte la teoria de l' N+1 Text, que diu que per molt gran que sigui un corpus (i quan més petit més probable és) sempre és possible que aparegui un nou text amb dades que rebatin les teories desenvolupades fins aquell moment.

4. UN EXEMPLE. VARIANTS: *NUIT*, *NUYT* I *NIT*

Les variants per al concepte 'nit' es poden explicar a partir de l'evolució de l'ètim llatí NOCTEM. Com és ben sabut, en posició de coda la consonant -C- es converteix en iod [j] i aquesta semivocal

9 En molts casos el nombre de mostres és obligatòriament petit, ja que no n'hi ha més. En les llengües romàniques, per exemple, el nombre de mostres dels segles X o XI existents és força limitat, tot i sumant totes les documentades.

DOI: 10.33115/C/9788499846163_24

provoca el tancament en *u* de la *o* oberta tònica del llatí tardà, sigui per diftongació i monoftongació posterior ([ɔ́j] > wój] > [új]) sigui per tancament directe de dos graus ([ɔ́j] > [új]). A partir de l'aplicació de les regles fonològiques històriques, el resultat esperable és, per tant, la forma *nuit* amb diftong decreixent ([nújt]).

Joan Coromines (*DECat*, sv: nit) explica que *nuit* és la forma més antiga, present en el català des dels orígens de l'idioma (ja documentada en les Homilies d'Organyà), i que és freqüent fins a finals del segle XIV però rara ja en el segle XV. Mentre que *nit*, tot i que ja es documenta algun cop a finals del segle XIII (apareix alguna vegada en Desclot), no predomina fins l'últim terç del segle XIV. Com apunta també Coromines, el canvi de *nújt* a *nit* suposa una etapa prèvia amb trasllat de l'accent *nwít* i una elisió posterior de la semivocal *w* per les dificultats que implicava l'obertura sil·làbica complexa *nw*, formada per dos sons amb punts d'articulació diferents, dentoalveolar en el cas de la *n* i labiovelar en el cas de la *w*. D'acord amb el que hem apuntat la formació de *nit* respondria a la derivació següent: NOCTE > nɔ́jte > nújt > nwít > nit.

En el *CICA*, per al concepte 'nit' es documenten les variants fonètiques *nit*, *nuit* i *nuyt*. La primera variant respon, evidentment, a la forma monoftongada final i la tercera a la forma amb el diftong decreixent [nújt]. Més problemàtica és la forma gràfica *nuit*, que es pot interpretar com una variant gràfica de *nuyt* (explicació més probable) o bé com la representació de la forma amb diftong creixent ([nwít]).

Quan s'observa la distribució de les freqüències d'aquestes variants en els tres eixos d'àmbit en què està estructurat el corpus (diacrònic, tipològic i dialectal), hom s'adona que es tracta de tres variants que en l'àmbit dialectal es documenten en quasi tots els dialectes i que en l'àmbit tipològic es documenten en quasi tots els tipus textuals. Només es troben diferències significatives en la distribució de les freqüències en l'eix de l'àmbit temporal.

Utilitzant els apartats de la variable independent referent a l'àmbit de la diacronia, s'observa que de les tres variants la primera que es documenta és la forma diftongada *nuit*, que es comença a trobar en textos de principis dels segle XIII i desapareix a finals del segle XIV (també hi ha un cas de *nit* al segle XIII). Però quan s'observen les freqüències relatives hom s'adona que la seva freqüència, tant en la primera meitat del segle XIII

DOI: 10.33115/C/9788499846163_24

com en la segona meitat, és força alta respecte de les dels altres apartats diacrònics (entre un 0,36 i un 0,45 per mil), en canvi, en les dues meitats del segle XIV la seva freqüència relativa es ja força residual (entre un 0,01 i un 0,03 per mil). La segona variant que des del punt de vista temporal es documenta amb consistència és la variant gràfica de la forma diftongada anterior *nuyt*,[10] que es registra per primera vegada en la segona meitat del segle XIII i desapareix a finals del segle XV, tot i que en aquest segle ja té molt poca presència en els textos (entre 0,02 i 0,03 per mil). Finalment, la variant monosil·làbica *nit* malgrat que es documenta per primera vegada en la primera meitat del segle XIII, no és fins el segle XIV que es registre amb una certa freqüència i no serà fins a partir del segle XV que la variant tindrà una freqüència considerable (entre 4,03 i 4,27 per mil).

Figura 4. Freqüències absolutes i relatives en l'eix d'àmbit diacrònic de les variants nuit, nuyt i nit

Diacronia	Total	nuit	0/000 F.R
Segle XIa	798	0	0,00
Segle XIb	3296	0	0,00
Segle XIIa	1238	0	0,00
Segle XIIb	2107	0	0,00
Segle XIIIa	22096	1	0,45
Segle XIIIb	937067	34	0,36
Segle XIVa	934994	3	0,03
Segle XIVb	1323976	1	0,01
Segle XVa	1462865	0	0,00
Segle XVb	1897994	0	0,00
Segle XVIa	742595	0	0,00
Segle XVIb	881358	0	0,00
Segle XVIIa	363289	0	0,00
Segle XVIIb	16069	0	0,00
Segle XVIIIa	67105	0	0,00
Segle XVIIIb	0	0	0,00
Total:	8656847	39	0,00 %

nuit

Diacronia	Total	nuyt	0/000 F.R
Segle XIa	798	0	0,00
Segle XIb	3296	0	0,00
Segle XIIa	1238	0	0,00
Segle XIIb	2107	0	0,00
Segle XIIIa	22096	0	0,00
Segle XIIIb	937067	179	1,91
Segle XIVa	934994	175	1,87
Segle XIVb	1323976	182	1,37
Segle XVa	1462865	4	0,03
Segle XVb	1897994	4	0,02
Segle XVIa	742595	0	0,00
Segle XVIb	881358	0	0,00
Segle XVIIa	363289	0	0,00
Segle XVIIb	16069	0	0,00
Segle XVIIIa	67105	0	0,00
Segle XVIIIb	0	0	0,00
Total:	8656847	544	0,01 %

nuyt

Diacronia	Total	nit	0/000 F.R
Segle XIa	798	0	0,00
Segle XIb	3296	0	0,00
Segle XIIa	1238	0	0,00
Segle XIIb	2107	0	0,00
Segle XIIIa	22096	1	0,45
Segle XIIIb	937067	51	0,54
Segle XIVa	934994	101	1,08
Segle XIVb	1323976	237	1,79
Segle XVa	1462865	590	4,03
Segle XVb	1897994	811	4,27
Segle XVIa	742595	267	3,60
Segle XVIb	881358	211	2,39
Segle XVIIa	363289	593	16,32
Segle XVIIb	16069	29	18,05
Segle XVIIIa	67105	20	2,98
Segle XVIIIb	0	0	0,00
Total:	8656847	2911	0,03 %

nit

El resultat obtingut de les freqüències que ens proporciona el *CICA* està d'acord amb les dades que ofereix Joan Coromines, però utilitzant un

10 D'acord amb les observacions que m'ha fet en Manuel Pérez Saldanya, interpreto que el diftong és decreixent, ja que la «y» representa la semivocal, no la vocal. La qüestió és si *nuit* i *nuyt* són variants gràfiques o representen pronúncies diferents. No es pot descartar que siguin variants gràfiques com passa amb *feit/feyt*. En aquest cas el diftong sempre és decreixent.

DOI: 10.33115/C/9788499846163_24

corpus estructurat i observant la distribució de freqüències les anàlisis que aquesta ens permet són molt més precises.[11]

D'aquesta manera, si es transporten les dades de les freqüències sobre uns gràfics i aquests se sobreposen, observant la corba de les freqüències de cada una de les variants es comprova que es produeix un desplaçament del moment àlgic de cada una de les tres variants que sembla indicar un procés de canvi (primer simplement gràfic o, menys probablement, de tipus de diftong, i, posteriorment, a monoftongació), que va des de les variants diftongades (*nuit* i *nuyt*), i arriba a la variant monosil·làbica (*nit*). Aquest procés es produeix entre finals del segle XIII i principis del segle XV. En el gràfic, també es posa en evidència que quan comença a aparèixer una nova variant, la freqüència de la variant anterior comença a baixar, i que el punt màxim de la nova variant coincideix amb el punt de desaparició de l'antiga variant.

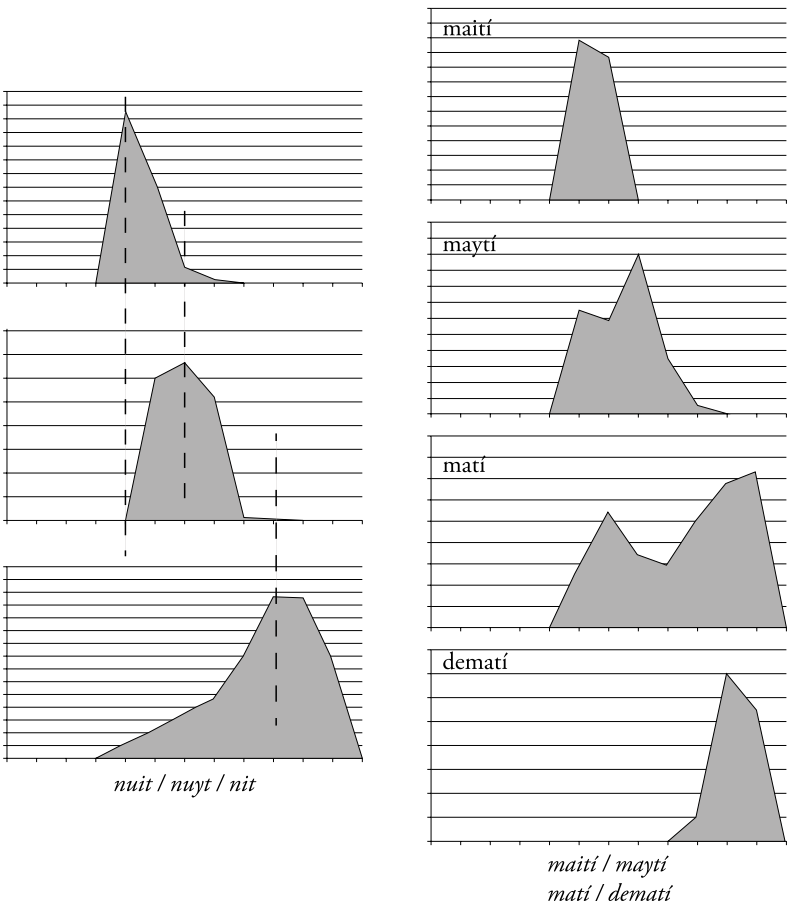
Un cas similar passa amb les variants del concepte 'matí'. A partir de la distribució de freqüències de les distintes variants (*maitís*, *maytí*, *matí* i *dematí*), es pot observar com els gràfics presenten una distribució força similar a la establerta en el cas de les variants per al concepte 'nit'.

A partir d'aquestes dades, en una recerca de caràcter científic, utilitzant les distintes variants d'una mateixa família lèxica com a variables dependents i els diferents apartats de l'eix d'àmbit diacrònic com a variable independent, es podrà començar a estudiar si entre finals del segle XIII i finals del segle XIV hi ha un procés de monoftongació generalitzat en la llengua catalana i, també, si hi ha una evolució de diftong decreixent a diftong creixent o es tracta només de variants formals. Si aquest mateix procediment s'observa en altres casos de variants fonètiques que comporten un procés de monoftongació i el procés coincideix en el temps, és quan serà possible començar a plantejar una hipòtesi de treball que es podrà validar o no a partir dels resultats de diversos casos similars.

DOI: 10.33115/C/9788499846163_24

11 Un dels inconvenients de les dades que ofereix Coromines, i que ell ja esmenta alguna vegada, és que utilitza molts textos literaris, la font dels quals no és l'original, amb la qual cosa es passen per alt les possibles intervencions dels diferents copistes.

Figura 5. Desplaçament cronològic de variants



5. CONCLUSIONS

Els corpus i la seva explotació sota les premisses de la lingüística de corpus poden resultar de gran ajuda a la hora de fer recerca amb bases científiques sobre temes lingüístics en general i sobre temes de variació en particular. Per poder-ho fer de manera solvent i correcta cal tenir en compte dos conceptes que s'han d'aplicar de forma convenient:

DOI: 10.33115/C/9788499846163_24

el de les variables estadístiques i el de les freqüències. Pel que fa a les variables, en la formalització de la nostra recerca s'ha d'establir quines seran les variables dependents i quines les independents, i observar com les freqüències dels elements lingüístics que es volen estudiar (variable dependent) es comporten quan es canvien els paràmetres dels eixos dels diferents àmbits en què ha d'estar estructurat el corpus (variables independents). Pel que fa a les freqüències, absolutes i relatives, cal distingir el valor de cada una d'elles i tenir criteri per saber quan s'ha de tenir en compte una o l'altra, depenent de si es vol treballar amb quantitats absolutes o amb quantitats proporcionals. Tenint en compte que els corpus difícilment poden estar equilibrats en tots els seus apartats de cada eix d'àmbit, és més probable que convingui utilitzar el valor de les freqüències relatives, altrament s'estarà passant per alt la diferent magnitud del nombre de mostres entre els diversos apartats, quelcom important a l'hora de valorar correctament els resultats.

REFERÈNCIES BIBLIOGRÀFIQUES

[*CICA*] Torruella, Joan, Pérez Saldanya, Manuel i Martines, Josep (dirs.). *Corpus Informatitzat del Català Antic*. http://www.cica.cat

DECat = Coromines, Joan. 1980-2001. *Diccionari etimològic i complementari de la llengua catalana*. Barcelona: Curial.

DIEC = Institut d'Estudis Catalans. Diccionari de la llengua catalana. https://dlc.iec.cat/

Gómez, Marta y Nieves Sánchez. 2010. La marcación diatópica. Dins Marta Gómez y José Ramón Carriazo (eds.), *La marcación en lexicografía histórica,* 109-169. San Millán de la Cogolla: Cilengua.

Kabatek, Johannes. 2006. Tradiciones discursivas y cambio lingüístico. Dins Guiomar Ciapuscio, Konstanze Jungbluth, Dorothee Kaiser i Célia Lopes, (eds.), *Sincronía y diacronía de tradiciones discursivas en Latinoamérica*, 151-173. Madrid / Frankfurt am Main: Iberoamericana / Vervuert.

DOI: 10.33115/C/9788499846163_24

Kabatek, Johannes. 2017. Prólogo. Dins Joan Torruella, *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. New York: Peter Lang.

Perea, Maria-Pilar i Germà Colón. 2010. Cronoestratigrafía dialectal. Dins Maria Iliescu, Heidi Siller-Runggaldier, Paul Danler (eds.), *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes*, vol. IV, 199-211. Berlin/New York: De Gruyter.

Perea, Maria-Pilar i Germà Colon. 2022. *Estratigrafia dialectal*. http://www.ub.edu/lexdialgram/estratigrafia/html/pagina2.html?inputIdDiatopisme=26

Pérez Saldanya, Manuel. 2009. Si per la tua gràcia podia eu conservar ma vida 'If by your grace I could preservem y life?: pronominal possessive constructionsin Old Catalan. Dins Joan Rafel Cufí (ed.), *Diachronic Linguistic*, 275 – 298. Girona: Documenta Universitaria.

Sánchez-Prieto Borja, Pedro. 2012. Un corpus para el estudio integral de fuentes documentales (CODEA). Dins Emilio Montero Cartelle i Carmen Manzano Rovira, (coord.), *Actas del VIII Congreso Internacional de Historia de la lengua española (Santiago de Compostela, 14-18 de setiembre de 2009)*, vol. I, 445-466. Santiago de Compostela: Meubook.

Torruella, Joan. 2017. *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. New York: Peter Lang.

Torruella, Joan. 2021. Atlas, diccionarios y corpus: tres recursos lingüísticos en contraste. Dins Matteo de Beni i Dunia Hourani-Martín (eds.), *Corpus y estudio diacrónico del discurso especializado en español*, 25-53. Berlín: Peter Lang.

DOI: 10.33115/C/9788499846163_24