L'ambigüitat lèxica és útil i necessària[1]

Gemma Boleda
gemma.boleda@upf.edu

Universitat Pompeu Fabra / ICREA
Departament de Traducció i Ciències del Llenguatge

1. Introducció

Aquest treball tracta de l'ambigüitat semàntica de les paraules.[2] Comencem amb uns quants exemples dels tipus d'ambigüitat que trobem al lèxic:

(1) *ratolí*: MAMÍFER ROSEGADOR / APARELL PER FER MOURE EL CURSOR

(2) *lime* (en anglès): LLIMA / CAL VIVA

(3) *dit*: MEMBRE FINAL DE LA MÀ / DEL PEU

(4) *tawa* (en Dhudhuroa)[3]: ARBRE / FOC

(5) *estrella*: ESTEL / PERSONA FAMOSA

Pel que fa al significat de les paraules, com en molts altres àmbits de la llengua (v. altres contribucions a aquest volum), l'ambigüitat és la norma (Murphy 2002, Wasow et al. 2005). Per exemple, segons el recurs lexicogràfic WordNet, en català les paraules tenen de mitjana dos significats (Bond et al. 2016, Gonzalez-Agirre et al. 2012). A més, dins un mateix significat, hi ha vaguetat: per exemple, és impossible delimitar amb exactitud una frontera entre *alt* i *mitjà* a l'hora de qualificar l'alçada de persones (van Deemter 2010).

En aquest treball, em centro en l'ambigüitat, deixant la vaguetat de banda; i uso el terme «ambigüitat» per a cobrir tant la polisèmia com l'homonímia com la subespecificació (Lyons 1977). La polisèmia es produeix quan els significats d'una paraula estan relacionats entre si; per exemple, el ratolí animal i el ratolí que és un aparell d'ordinador (exemple (1)) comparteixen característiques visuals. L'homonímia, en canvi, es produeix quan els significats no estan relacionats entre si; per exemple, la paraula *lime* en anglès (exemple (2)) presenta homonímia (i, de fet, els dos significats provenen d'arrels diferents que han evolucionat cap a una mateixa forma fonològica de manera independent).[4] Es diu que el significat d'una paraula està subespecificat quan engloba diversos subsignificats; per exemple, el significat de *dit* (exemple (3)) cobreix tant el significat corresponent a *finger* com el significat corresponent a *toe* en anglès (és a dir, tant es pot

1 Aquest treball s'ha beneficiat dels ajuts PID2020-112602GB-I00 (Ministerio de Ciencia, Innovación y Universidades) i 2021 SGR 00470 (Generalitat de Catalunya).
2 En aquest treball, no distingiré entre ambigüitat i polisèmia; cal, però, tenir en compte que la gran majoria de casos d'ambigüitat lèxica corresponen a polisèmia.
3 El Dhudhuroa és una llengua aborigen australiana extinta.
4 V. *Online Etymology Dictionary*, entrada *lime* https://www.etymonline.com/search?q=lime).

usar per als dits de la mà com per als dits del peu). Sovint és difícil distingir entre polisèmia, homonímia i subespecificació, i diferents autors divergeixen sobre exactament què cau dins el terme *ambigüitat* (per exemple, Sennet 2023 exclou la subespecificació; v. també la contribució de Jordi Fortuny i Lluís Payrató a aquest volum). Per als objectius d'aquest treball, però, no cal fer distinció entre els diferents tipus de fenòmens.

Dèiem més amunt que l'ambigüitat lèxica és la norma, ja que la majoria de paraules són ambígües. Això sembla contravenir la funció principal del llenguatge, que és la de poder comunicar-nos: si les paraules signifiquen coses diferents, no ens fem un embolic? De fet, hi ha una llarga tradició en filosofia i lingüística (inclosa la lingüística computacional) que critica l'ambigüitat lèxica i la considera una nosa (Sennet 2023). Per exemple, Gottlob Frege va escriure:

> So long as the reference remains the same, such variations of sense may be tolerated, although they (...) ought not to occur in a perfect language. / Mentre el referent sigui el mateix, aquest tipus de variacions de sentit [referint-se a l'ambigüitat lèxica] es poden tolerar, però no haurien d'ocórrer en una llengua perfecta.[5]

Un altre camp on tradicionalment s'ha criticat molt l'ambigüitat lèxica és la lingüística computacional. Això és perquè l'ambigüitat presenta dificultats per al tractament computacional del llenguatge, com ara la traducció automàtica; per exemple, depenent del context, *lime* s'ha de traduir al català com a *llima* o com a *cal*. Com a solució, es va crear la tasca de desambiguació semàntica lèxica (*Word Sense Disambiguation* o WSD), és a dir, la tasca d'eliminar l'ambigüitat de les paraules (Navigli 2009; v. també la contribució de Mariona Taulé a aquest volum). La tasca de WSD consisteix en definir un conjunt de sentits possibles per a les paraules i precisar quin dels sentits és l'adequat en cada context. Això no obstant, la WSD ha afrontat moltes dificultats i avui dia pràcticament no hi ha treballs que la investiguin; en els sistemes computacionals actuals, la solució s'ha trobat en representacions lingüístiques flexibles que permeten integrar la informació lèxica i la del context de manera que no calgui una desambiguació explícita (Aina et al. 2019).

Les visions de l'ambigüitat lèxica que acabo de presentar suggereixen que és una tara del llenguatge, una propietat indesitjable. En el present article, en canvi, defenso que no és una tara sinó un tret necessari, i molt útil, cosa que explica que sigui omnipresent a totes les llengües (apartat 2). També n'examinaré les propietats (apartat 3) i els límits (apartat 4).

5 Frege (1948 [1892]), pàg. 210, nota al peu 2; citada de Sennet (2023); traducció meva. Noti's que *language* també es podria traduir per *llenguatge*.

2. L'ambigüitat lèxica és útil i necessària

Per què és necessària, l'ambigüitat? Podem entendre-ho examinant la reducció a l'absurd que presenta J. L. Borges al seu conte *Funes el memorioso*:

> [Funes], no lo olvidemos, era casi incapaz de ideas generales, platónicas. No sólo le costaba comprender que el símbolo genérico *perro* abarcara tantos individuos dispares de diversos tamaños y diversa forma; le molestaba que el perro de las tres y catorce (visto de perfil) tuviera el mismo nombre que el perro de las tres y cuarto (visto de frente).[6]

És impossible parlar sense ambigüitat, perquè la realitat de la qual parlem està en etern moviment, i, si es fa servir la mateixa paraula en ocasions diferents, automàticament s'està introduint ambigüitat. Però l'ambigüitat lèxica, de fet, no només és necessària, sinó també útil, perquè ens permet **generalitzar**. Si per exemple aprenem el significat de la palabra *gos*, això ens permet parlar d'animals que no hem vist mai abans.

Aquest tipus d'ambigüitat de què parla Borges pot semblar diferent dels casos que hem discutit a la introducció; els diferents gossos, al cap i a la fi, són gossos, mentre que la intuïció ens diu que un ratolí de camp i un ratolí d'ordinador són entitats molt més diferents. Per a mi, la diferència és només de grau, no qualitativa, i els mecanismes que hi ha sota són els mateixos. Il·lustraré aquesta proposta amb un estudi recent (v. apartat 3).

Una altra raó que fa necessària l'ambigüitat lèxica és el fet que els nostres cervells són finits. Per això no podem tenir una paraula nova per a cada cosa nova que veiem, ni tan sols per a cada *tipus* de cosa nova que veiem. Les persones dominem un vocabulari d'unes 20.000 paraules (D'Anna et al. 1991; Nation i Waring 1997); és un vocabulari molt ric, però limitat.[7]

Ens trobem, doncs, amb un trencaclosques: per una banda, tenim un lèxic limitat; per altra, ens hem de referir a una infinitat de coses. Més en general, les llengües necessiten tenir prou expressivitat com per funcionar com a vehicles de comunicació, i alhora no poden gastar recursos cognitius i físics infinits (Gibson et al. 2019, Kemp et al. 2018). La pregunta, llavors, és com s'ho fan per poder satisfer les dues restriccions alhora.

Les llengües tenen diversos mecanismes per solucionar aquest trencaclosques. Un de molt important és la gramàtica, que permet combinar diferents paraules per crear nous significats (per exemple, si no tinguéssim la paraula *gos*, en un determinat context en què ens volem referir a un gos marró que està al costat d'un gat blanc podríem dir «l'animal marró»). Aquí, però, ens

6 «Funes el Memorioso» fou publicat originalment a: Borges, J. L. (1944) *Ficciones*. Editorial Sur.
7 D'Anna et al., com la majoria d'estudis, mesuren el vocabulari d'estudiants universitaris americans; la xifra exacta pot fluctuar entre diferents poblacions. Això no afecta el meu argument; l'important és que estem parlant de pocs milers o desenes de milers de paraules.

centrarem en un altre mecanisme també crucial: el de «reciclar» paraules, és a dir, utilitzar paraules velles per a significats nous. Un exemple molt clar és el de *ratolí* (v. exemple (1)): quan es va inventar el ratolí d'ordinador, calia anomenar-lo d'alguna manera, i es va reciclar la paraula que es referia als mamífers rosegadors. Aquest tipus de procés d'extensió semàntica condueix a l'ambigüitat lèxica, que alhora permet economitzar recursos: tot i que es podria utilitzar la gramàtica («passa'm aquell dispositiu que està connectat a l'ordinador»), és molt més econòmic capturar nous conceptes en una sola paraula («passa'm aquell ratolí»). Però, per què es va reciclar precisament *ratolí* per al ratolí d'ordinador? Més en general, quins principis regeixen aquest «reciclatge» lèxic?

3. Els principis que regeixen l'ambigüitat lèxica

S'ha observat que el que motiva l'extensió del significat d'una paraula és la **relació semàntica** entre el significat antic i les propietats de la nova entitat a la qual un es vol referir (Xu et al. 2020).[8] En el cas de *ratolí*, es tracta de similitud visual (en un moment en què els ratolins anaven tots per cable, que semblava la cua). Això no obstant, hi ha molts d'altres tipus de relació semàntica que intervenen en casos de canvi semàntic. Un és l'associació; per exemple, en el cas de *tawa* (v. exemple (4) a la introducció), hi intervé l'associació entre el foc i una cosa que se sol cremar, l'arbre. Un altre, la relació taxonòmica, com el cas de *dit* (tant els dits del peu com els de la mà són parts del cos). Els diferents tipus de relacions no són excloents: per exemple, els dits del peu i de la mà també presenten similitud visual. De fet, com més relacionats en més dimensions estan els dos significats, més fàcil és l'extensió de significat (Brochhagen et al. 2023) – fins a un cert punt, com veurem a l'apartat 4. Sovint les relacions són menys directes, com el cas de la metàfora (p. ex., la relació entre els dos significats de la paraula *estrella* a l'exemple (5) és de caire més abstracte, amb una noció de prominència sobre un fons).

Em centraré en la resta d'aquest apartat en un treball recent (Brochhagen et al. 2023) en què hem defensat que el fet que l'ambigüitat es regeixi per relacions semàntiques és part dels fonaments cognitius humans, i que va més enllà del canvi semàntic. Segons aquesta visió, l'ambigüitat lèxica i el canvi semàntic són manifestacions particulars d'una propietat general de la cognició humana tal i com es manifesta en el llenguatge: la **creativitat lèxica**. Denominem així la capacitat que tenim les persones per utilitzar les paraules del nostre vocabulari per referir-nos a noves entitats, situacions i propietats. La creativitat lèxica, doncs, se sosté sobre un fonament cognitiu bàsic: la capacitat de percebre relacions entre objectes diferents.[9]

8 La lectora atenta observarà que hem passat d'ambigüitat lèxica a canvi semàntic; la majoria de casos d'ambigüitat lèxica s'originen en processos d'extensió semàntica, resultant en polisèmia (l'homonímia és infreqüent).

9 Aquí i a sota, uso «objecte» en un sentit abstracte, que encompassa entitats, situacions, esdeveniments, proposicions, atributs, i qualsevol altre aspecte del qual es pugui parlar .

La hipòtesi del nostre treball era que la creativitat lèxica es manifesta no només en la llengua adulta, tant en la seva situació sincrònica (ambigüitat) com en la seva evolució (canvi semàntic), sinó també en el procés d'adquisició de la llengua per part dels infants. Ja Vygotsky (1962) va observar un fenomen que després s'ha anomenat sobreextensió (*overextension*): els nens d'entre aproximadament un any i dos anys i mig tenen un vocabulari molt mins, i l'utilitzen de manera creativa per poder referir-se a les coses per a les quals encara no tenen una paraula; per exemple, una nena que veu una granota pot exclamar «gos!» (o, més probablement, «bubú!»). En treballs anteriors s'havia observat que la sobreextensió ve motivada per algun tipus de similitud o relació semàntica entre l'objecte al qual l'infant es vol referir i el significat de la paraula (Clark 1978, Rescorla 1980); al capdavall, les granotes i els gossos són animals, és a dir, tenen una relació taxonòmica.

L'adquisició del llenguatge i el llenguatge adult se solen estudiar de manera separada. La nostra hipòtesi era que trobaríem el mateix tipus d'extensions semàntiques en sincronia, diacronia, i sobreextensió – és a dir, en diferents fenòmens que resulten en ambigüitat lèxica. Per posar aquesta hipòtesi a prova empíricament, vam construir models computacionals amb dades de més de 1,400 llengües (per a fenòmens adults) i de l'anglès (per a adquisició), extretes de bases de dades de lliure distribució (Ferreira Pinto i Xu 2021, Rzymski et al. 2020, Zalizniak et al. 2012). La Taula 1 conté exemples il·lustratius de les dades amb què vam treballar.[10]

SINCRONIA	DIACRONIA	ADQUISICIÓ
bakh (hunzib): HERBA-PLANTA	*trakaĩ* (lituà): MUNTANYA => BOSC	*dog*: GOS => GRANOTA
amciri (panare): FERRO-CADENA	jasiri (suahili) lleó => valent	*ball*: PILOTA => GLOBUS
ibu (kashinawa): MARE-PARE	*içmek* (turc): ESTIRAR => BEURE	*moon*: LLUNA => SOL

Taula 1. Exemples de les dades lingüístiques de Brochhagen et al. (2023). El símbol => significa «extensió a»; el guió separa significats que coexisteixen a nivell sincrònic.

La nostra metodologia es va basar en l'ús de models computacionals per simular els diferents fenòmens semàntics. Concretament, vam construir models computacionals que, donats dos significats (com ara HERBA-PLANTA), retornessin la probabilitat que fossin denotats per la mateixa paraula, basant-se en quatre paràmetres semàntics. Aquests paràmetres quantificaven el grau de relació entre els dos significats, en quatre dimensions semàntiques: visual, taxonòmica, associativa, i afectiva.[11] Les tres primeres les hem il·lustrades més amunt; l'última quantifica

10 El hunzib és una llengua caucàsica nord-oriental parlada al sud del Daguestan; el panare, una llengua carib parlada al sud de Veneçuela; el kashinawa, una llengua ameríndia pano parlada a l'Oest d'Amèrica del Sud; el suahili, una llengua bantu parlada a moltes regions de l'Àfrica Oriental.

11 Hi ha d'altres relacions semàntiques que poden tenir un rol en la creativitat lèxica. Vam seleccionar relacions que s'han mostrat rellevants per a un o altre fenomen en treballs anteriors i que podíem

diversos aspectes relacionats amb l'afectivitat, com ara la polaritat (de negativa a positiva, passant per neutra) o la intensitat. Els valors dels paràmetres per a cada parell de significats els vam extreure de diferents bases de dades i recursos computacionals (Anderson et al. 2018, De Deyne et al. 2018, Fellbaum 2015, Mohammad 2018).

Vam entrenar un model per a cada fenomen (sincronia, diacronia, adquisició), donant-los dades positives i negatives, és a dir, casos d'extensió que es troben a les bases de dades utilitzades i casos que no.[12] Per exemple, tal i com il·lustra la Figura 1, al model d'ambigüitat lèxica sincrònica li donàvem com a exemples positius dades com ara HERBA-PLANTA i FERRO-CADENA, i com a exemples negatius parells com ara GALLINA-SABATA (en total, entre 600 i 112,000 parells de significats depenent del model). El resultat del procés d'entrenament és un model computacional (una equació matemàtica implementada en codi de programació) que pot generar prediccions per a qualsevol parell de significats, si se'n tenen les dades dels quatre paràmetres estudiats; és a dir, si en sabem el grau de relació associativa, taxonòmica, visual, i afectiva.

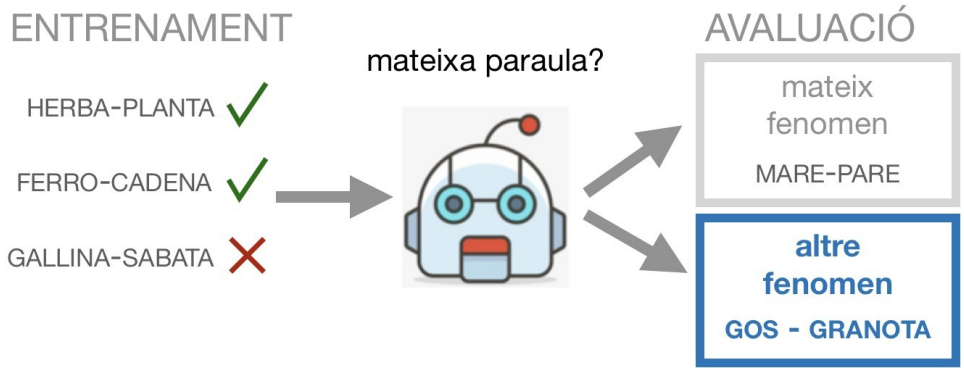

Figura 1. Representació dels models i el mètode d'avaluació.

Vam avaluar els models de dues maneres. La primera posava a prova la seva capacitat de generalització dins el mateix fenomen: cada model havia de determinar si es pot utilitzar la mateixa paraula per a dos significats que no havia vist durant el procés d'entrenament però provenen del mateix fenomen (per exemple, MARE-PARE per al model sincrònic). Això ens va permetre determinar un punt de comparació per a la prova crucial: la **generalització** a fenòmens diferents dels que els models havien vist durant el procés d'entrenament. Per exemple, tal i com il·lustra la Figura 1, vam avaluar el model entrenat amb dades sincròniques com ara HERBA-PLANTA i ara GALLINA-SABATA comprovant la seva capacitat de predir dades de sobreextensió com ara GOS-GRANOTA. La idea és que, si els models poden generalitzar d'un fenomen a un altre,

operacionalitzar per fer una exploració a gran escala amb les dades disponibles.
12 Les dades de sincronia provenien de la base de dades CLICS (Rzymski et al. 2020); les de diacronia, de DatSemShift (Zalizniak et al. 2012); les d'adquisició, de Ferreira Pinto i Xu (2021). Per construir dades negatives vam agafar significats que ocorren a les bases de dades i els vam aparellar aleatòriament.

tenim evidència que els tipus de relacions semàntiques que trobem en un cas i en l'altre són els mateixos.

Això és exactament el que vam obtenir. En l'avaluació amb dades del mateix fenomen, els models van obtenir entre un 74 i un 81% d'encert, depenent del fenomen.[13] Quan els vam avaluar amb dades d'un fenomen diferent, van donar exactament el mateix: entre un 72 i un 81% d'encert. Per exemple, el model il·lustrat a la Figura 1, que estava entrenat amb dades d'ambigüitat sincrònica en diferents llengües, va obtenir un 81% d'encert per a dades d'adquisició lèxica per part d'infants anglesos.

En aquell treball, doncs, vam concloure que hi ha indicis robustos que hi ha una base cognitiva única que regeix la creativitat lèxica, que trobem en múltiples manifestacions que divergeixen en aspectes clau. L'ambigüitat sincrònica i el canvi semàntic són fenòmens evolutius amb una escala temporal d'entre anys i milers d'anys, i són el resultat de dinàmiques poblacionals. La sobreextensió, en canvi, es dona durant l'adquisició de la llengua, a nivell individual (la majoria d'innovacions lèxiques dels nens ni perduren ni es transmeten), i desapareix al cap de pocs mesos. El substrat cognitiu que subjau aquests fenòmens i causa ambigüitat lèxica és, però, el mateix.

Resumint el que hem vist en aquest apartat, la major part de l'ambigüitat lèxica que trobem a les llengües s'origina en la identificació de relacions semàntiques entre el nou referent del quan hom vol parlar i el significat prèviament existent de les paraules que hom coneix. Les relacions semàntiques són multi-dimensionals i diverses, incloent-hi factors perceptuals i cognitius. Ras i curt, reciclem paraules per a significats que s'assemblen. Però, per què ho fem així? I, hi ha un límit a aquesta tendència?

4. Explicacions i límits a l'extensió semàntica

Diversos treballs recents han defensat que l'ambigüitat lèxica basada en la relació semàntica entre els significats d'una paraula és beneficiosa des del punt de vista cognitiu, ja que facilita l'adquisició i l'ús del llenguatge (Ramiro et al., 2018, Srinivasan and Rabagliati, 2015, Xu et al., 2020). Per il·lustrar l'argument des del punt de vista de l'adquisició, imaginem una nena catalana que ha après el significat *dit* com a paraula que denota un dels cinc membres de la mà. Al cap d'un temps, la nena sent un adult usar *dit* per referir-se a quelcom que clarament és al peu, donada la situació comunicativa i el fet que l'adult està mirant el peu. El contingut semàntic que ja coneix l'ajudarà de dues maneres a interpretar l'adult: en primer lloc, a identificar la part del peu concreta a la qual es refereix (per similitud visual, pot inferir que és el dit); en segon lloc, en la

13 Els models són regressions logístiques, que retornen probabilitats. Seguint el que es fa al camp, prenem com a resposta positiva una probabilitat de 0,5 o més, i com a resposta negative menys de 0,5. Si el model prengués decisions a l'atzar, l'encert seria del 50%, perquè tenim el mateix nombre d'exemples negatius i positius. El model també tenia en compte paràmetres sobre relacions filogenètiques i de contacte entre llengües –v. Brochhagen et al 2023 per a més informació.

construcció del lèxic: a l'hora d'eixamplar l'entrada lèxica de *dit*, podrà reciclar molt de material semàntic, com ara el fet que un dit té ungles o falanges, resultant en una representació compacta. Al seu torn, aquesta entrada lèxica, amb una representació compacta i amb aspectes semàntics compartits, serà fàcil d'usar un cop adquirida. L'ambigüitat homònima, com la de *lime* en anglès, en canvi, no ofereix cap dels dos avantatges.

Això no obstant, el fet d'usar la mateixa paraula per a dos significats relacionats té el perill que introdueix ambigüitat que pot portar a confusió en interaccions comunicatives. Per exemple, els dits dels peus i els de les mans solen aparèixer en contextos similars, si més no a nivell físic, i això pot crear confusió, com en el següent diàleg:

(6) Àlex (per telèfon): m'he trencat un dit!
Bruna: del peu?
Àlex: no, de la mà.

En molts casos, l'ambigüitat no causarà problemes; si el mateix diàleg tingués lloc en persona, la interlocutora tindria informació visual sobre quin és el dit que està trencat. Hi ha diversos treballs que mostren que la possibilitat de desambiguar en context és un factor clau a l'hora de determinar si una ambigüitat sobreviu al llarg del temps en una llengua determinada (Brochhagen 2020, Santana 2014, Zipf 1949). Per això, depenent de quant de sovint necessitem distingir entre els dos significats en un context comunicatiu, podem esperar que el lèxic «es resisteixi» a posar-los junts a la mateixa paraula, encara que estiguin molt relacionats semànticament. Pensem per exemple en el cas d'*esquerra* i *dreta*. Els significats d'aquestes dues paraules estan estretament relacionats; de fet, són idèntics excepte en una dimensió (el sentit). A diferència de *dit*, però, en la majoria de situacions en què es fa referència a un dels dos, hi ha alguna alternativa contextualment rellevant que involucra l'altre (per exemple, anant per un camí que es bifurca, o en una conversa sobre política). Per això, en un altre treball recent (Brochhagen i Boleda 2022) vam hipotetitzar que trobaríem un Principi Rínxols d'Or al lèxic: com a tendència, esperàvem que la relació semàntica entre dos significats facilitaria l'ambigüitat lèxica, fins a un cert punt, en què la frenaria. En altres paraules, esperàvem trobar que diferents significats podran conviure dins la mateixa paraula si no estan ni *massa poc* relacionats, ni *massa* relacionats, sinó *al punt just* (en analogia amb el conte de la Rínxols d'Or).

Vam trobar suport per al Principi Rínxols d'Or en l'anàlisi que vam fer, utilitzant mètodes semblants a l'estudi que he descrit a l'apartat anterior i dades de més de 1200 llengües i 1400 significats.[14] Les dades provenen de la base de dades CLICS (Rzymski et al. 2020), que

14 La metodologia d'aquest estudi també es basa sobre la construcció de models computacionals, amb una diferència clau: el model no és lineal, sinó que és un model logístic additiu generalitzat. Els models additius generalitzats (GAM, en anglès) permeten detectar relacions curvilínies com la del Principi Rínxols d'Or, en què una variable exerceix una força primer positiva i després negativa.

també vam usar en l'estudi exposat a l'apartat anterior. CLICS té informació sobre com es realitzen diferents significats en diferents llengües, la qual cosa es pot usar per examinar quines ambigüitats ocorren (v. columna esquerra de la Taula 1). No entraré en detalls metodològics perquè no cal per als objectius d'aquest article (v. Brochhagen i Boleda 2022 per a més informació), però sí que vull il·lustrar-ne els resultats. La Taula 2 conté exemples de prediccions del model. Recordem que, un cop induït el model a partir de les dades, aquest pot generar prediccions per a qualsevol parell de significats nou, si li podem dir el grau de relació semàntica entre tots dos. En aquest cas, només vam utilitzar un predictor per al grau de relació semàntica, és a dir que les prediccions del model depenen només d'un paràmetre.[15] Aquesta informació la vam treure d'un recurs psicolingüístic, la *Small World of Words* (De Deyne et al. 2018), que ens permet determinar per exemple que els significats TRES i SÍ estan molt poc relacionats; els significats CALENT i FOC, força relacionats; i els significats DIMARTS i DIJOUS, moltíssim.[16] A partir d'aquesta informació, el model prediu que no és gens probable que TRES i SÍ s'expressin amb la mateixa paraula; en canvi, és força probable que això passi en el cas de CALENT i FOC. Per a DIMARTS i DIJOUS, que tenen una relació semàntica molt més alta que no pas CALENT i FOC, en canvi, la probabilitat és més baixa que en cas de CALENT i FOC. Segons el model, doncs, DIMARTS i DIJOUS estan *massa* relacionats.

SIGNIFICATS	GRAU DE RELACIÓ	PROBABILITAT MATEIXA PARAULA
TRES, SÍ	Molt baix	Molt baixa
CALENT, FOC	Força alta	Força alta
DIMARTS, DIJOUS	Molt alta	Mitjana

*Taula 2. Exemples de predicció del model a Brochhagen i Boleda (2022). El grau de relació (columna 2) és l'*input *al model; la probabilitat que les llengües tinguin una mateixa paraula per a ambdós significats n'és l'*output*.*

Fixem-nos que *dimarts* i *dijous*, tot i no ser estrictament parlant antònims (com en canvi ho són *esquerra* i *dreta*), corresponen al perfil de ser significats que es poden confondre sovint en un context comunicatiu: part del que estic dient quan dic *dimarts* és «no dijous (ni tampoc dilluns, dimecres, etc.)». Per comprovar que el que hi ha darrere dels resultats és realment la possibilitat de confusió comunicativa, i no simplement el grau de relació semàntica, vam fer una altra anàlisi que

15 De fet, com en el cas anterior, el model també tenia en compte dades de relació filogenètica i de contacte, com a paràmetres de control; però la variable d'interès era la relació semàntica.
16 Els valors que utilitzem per al model són numèrics, derivats amb un procediment automàtic a partir de les dades de la *Small World of Words*.

va demostrar que la relació de meronímia (part-tot, com ara COLZE-BRAÇ) i la d'hiperonímia (subsumpció, com ara GOS-MAMÍFER) faciliten l'ambigüitat molt més que no pas l'antonímia (com ara ESQUERRA-DRETA).

Noti's que, segons el fil argumental que hem seguit fins ara, pot semblar contraintuïtiu que la probabilitat assignada pel model a DIMARTS-DIJOUS sigui mitjana, i no baixa o molt baixa. El que passa és que les paraules que estan *massa* relacionades segueixen estant *molt* relacionades, i de fet trobem molts casos de paraules que engloben significats oposats. Per exemple, 117 llengües de la base de dades CLICS tenen una sola paraula per a MARIT i MULLER; i 25, per a COMPRAR i VENDRE. En català, el verb *llogar* té una ambigüitat semblant a aquesta última: tant el propietari com el llogater poden «llogar un pis».

El que no hem trobat en cap llengua de la base de dades, en canvi, és una paraula que tingui la denotació d'antònims adjectivals com ara *esquerre* i *dret* o *fred* i *calent*. La raó de ser d'aquest tipus d'antònims és poder expressar diferents punts en la mateixa dimensió de significat, i combinar-los en la mateixa paraula ho impediria.

Resumint el que hem vist en aquest apartat, basar l'ambigüitat en relacions semàntiques és beneficiós a nivell cognitiu però pot portar a problemes comunicatius; i la tensió entre aquests dos factors deixa una petja clara en el lèxic de les llengües del món.

5. Conclusió

En aquest article he defensat l'ambigüitat lèxica com a propietat essencial del llenguatge. L'ambigüitat lèxica és omnipresent a les llengües, i els principis que la subjauen els trobem en fenòmens i escales temporals molt diverses, des de l'adquisició fins a l'evolució del llenguatge. Segons la tesi que he defensat, això és perquè de fet és beneficiosa a nivell cognitiu: ens permet generalitzar i ens permet tenir un sistema lingüístic compacte, cosa que en facilita l'adquisició i l'ús. Això no vol dir que no causi problemes a nivell comunicatiu, tal i com assenyalen els seus detractors. Les llengües han de gestionar la tensió entre la necessitat de ser cognitivament simples (fàcils d'aprendre i usar) i comunicativament adequades (prou expressives; Brochhagen et al. 2018, Carr et al. 2020, Gibson et al. 2019). En llengua rere llengua, trobem un Principi Rínxols d'Or que regula aquestes dues pressions contraposades. El fet que trobem tanta ambigüitat lèxica al llenguatge malgrat els problemes comunicatius que causa suggereix que els seus avantatges són molt substancials.

Agraïments

Aquest article ha estat finançat en part pel Ministeri de Ciència i Innovació i l'AEI (projecte PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033). L'autora agraeix els autors de les bases de dades i recursos utilitzats que els hagin posat a disposició pública.

Bibliografia

Aina, L., K. Gulordava, Boleda, G. (2019). Putting words in context: LSTM language models and lexical ambiguity. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3342-3348.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077-6086.

Bond, F., Vossen, P., McCrae, J. P., Fellbaum, C. (2016). CILI: the collaborative interlingual index. *Proceedings of the 8th Global WordNet Conference*, 50-57.

Brochhagen, T., Boleda, G. (2022). When do languages use the same word for different meanings? The Goldilocks principle in colexification. *Cognition*, 226, 105179.

Brochhagen, T. (2020). Signaling under uncertainty: Interpretative alignment without a common prior. *The British Journal for the Philosophy of Science*, 71(2), 471–496.

Brochhagen, T., Boleda, G., Gualdoni, E., Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656), 431–436.

Carr, J. W., Smith, K., Culbertson, J., Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, 202, 104289.

Clark, E. V. (1978). Strategies for communicating. *Child Development*, 953-959.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., Storms, G. (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3), 987–1006.

D'Anna, C. A., Zechmeister, E. B., Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior*, 23(1), 109-122.

C. Fellbaum (2015). *WordNet*. MIT Press.

Ferreira Pinto, R., Xu, Y. (2021). A computational theory of child overextension. *Cognition*, 206, 104472.

Frege, G., 1948 [1892]. Sense and Reference. *The Philosophical Review*, 53: 209–230.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), 389–407.

Gonzalez-Agirre, A., Laparra, E., Rigau, G. (2012). Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. *Proceedings of the 6th Global WordNet Conference*.

Kemp, C., Xu, Y., Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1), 109–128.

Lyons, J. (1977). *Semantics*. Cambridge University Press.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual Meeting of the Association forComputational Linguistics*.

Nation, P., Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, 14(1), 6-19.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys* 41(2), 1-69.

Ramiro, C., Srinivasan, M., Malt, B. C., Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 2323–2328.

Rescorla, L. A. (1980). Overextension in early language development. *Journal of child language*, 7(2), 321-335.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., +++ Kaiping, Gereon A.; Chang, Sophie; Lai, Yunfan; Morozova, Natalia; Arjava, Heini; Hübler, Nataliia; Koile, Ezequiel; Pepper, Steve; Proos, Mariann; Van Epps, Briana; Blanco, Ingrid; Hundt, Carolin; Monakhov, Sergei; Pianykh,

Kristina; Ramesh, Sallona; Gray, Russell D.; Forkel, Robert; List, Johann-Mattis. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7, 13.

Santana, C. (2014). Ambiguity in cooperative signaling. *Philosophy of Science*, 81(3), 398–422.

Sennet, A. (2023). Ambiguity. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta i Uri Nodelman (eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/ambiguity/

Srinivasan, M., Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157, 124–152.

van Deemter, K. (2010). *Not exactly: In praise of vagueness*. OUP Oxford.

Vygotsky, L. (1962). *Language and Thought*. MIT Press.

Wasow, T., Perfors, A., Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, 265–282. CSLI Publications.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, 201, 104280.

Zalizniak, A. A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak, T., Russo, M. (2012). The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50(3), 633-669.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, New York.