

Realment parlam el “polaco”?

J. Miró, F. Rosselló

Dept. de Ciències Matemàtiques i Informàtica
Universitat de les Illes Balears

Els historiadors no es posen d'acord sobre l'origen del mal costum que tenen alguns espanyols d'emprar l'adjectiu “polaco” per referir-se als catalans i, per extensió, als catalano-parlants. La teoria amb més partidaris és que es tracta d'un invent castrense, arran de la coincidència temporal entre l'ocupació de Catalunya per les tropes franquistes a començaments de l'any 1939 i l'ocupació de Polònia pels seus amics de l'exèrcit nazi el setembre d'aquell mateix any. Sembla que el paral·lelisme entre aquells dos esdeveniments portà els militars franquistes a reflexar el seu menyspreu pels catalans en l'ús de l'adjectiu “polaco” com a insult. En tot cas, l'experiència confirma que és en els quarters militars on aquest epítet està més arrelat, així que és plausible que fos en aquests on sorgí i des d'on s'encomanà a alguns altres sectors de la població mesetària. Tanmateix, era per aprendre coses com aquesta, per al que servia el servei militar!

En cas que aquest origen sigui cert, qualsevol que conegui mínimament la cruesa de l'actuació dels nazis durant l'ocupació de Polònia i llurs sentiments envers els polonesos hauria de trobar profundament desagradable l'adjectiu “polaco” adreçat als catalans. Nosaltres ens estimam més creure que, independentment del seu origen, el fet que alguns espanyols encara l'emprin no té res a veure amb desitjos subconscients d'ocupacions militars i camps de concentració. Preferim pensar que el català, molt més ric en sons que el castellà, els recorda el polonès, si és que l'han sentit mai, i empen aquest adjectiu per mostrar el seu menyspreu per una llengua que consideren pròpia de pagesos i muntanyans incultes.

Més d'una vegada ens hem demanat si la comparació entre la nostra llengua i el polonès té qualche base científica. S'assembla realment el català al polonès? La veritat és que, llevat d'algunes excepcions curioses com ara el mot polonès “benzyna” per referir-se a la benzina, molt poc. Però podríem atacar una pregunta més realista:

Aquest article és una adaptació de l'article “Czy w Unii Europejskiej mówiono po polsku?” (Ja es parlava polonès a la Unió Europea?) dels mateixos autors, que apareixerà en un número especial de la revista polonesa *Delta* dedicat a la incorporació de Polònia a la Unió Europea; agraïm als seus editors el permís per publicar-lo. La *Delta* és una revista mensual de divulgació de física i matemàtiques molt popular a Polònia. Fa poc s'ha publicat en català una selecció d'articles apareguts en aquesta revista, que recomanem a tots els aficionats i professionals de les matemàtiques: es tracta de *Llet d'Ocellet Matemàtica. 25 anys de matemàtiques a la revista Delta* (Ed. Belladonna, 2002; ISBN 84-95992-00-0).

se sembla realment el català al polonès *més que l'espanyol*? Per contestar aquesta qüestió, hem comparat la distància mitjana entre una certa quantitat de paraules poloneses i, d'una banda, la seva traducció natural espanyola i, de l'altra, la seva traducció natural catalana. Com que es tractava de comparar paraules, hem emprat per fer-ho una distància d'edició de paraules.

El padrí de les distàncies d'edició va ser Lewis Carroll, el qual va inventar l'entreteniment següent per a la secció de passatemps de la revista *Vanity Fair* l'any 1879: donades dues paraules (en un mateix idioma, anglès a l'original) de la mateixa longitud, hem de transformar-ne una en l'altra per mitjà d'una seqüència el més curta possible de paraules (en el mateix idioma), cada una de les quals ha de diferir de l'anterior en només una lletra. Per exemple, podem transformar “pera” en “figa” per mitjà de la seqüència de paraules intermèdies “fera” i “fira”. Intuïtivament, podríem dir que les dues paraules donades estan a distància 3, perquè hem sabut anar d'una a l'altra en tres passos, tres substitucions, i no hi ha cap seqüència de canvis més curta que transformi una en l'altra.

L'any 1950, R. W. Hamming formalitzà aquesta noció de distància. La *distància de Hamming* entre dues paraules de la mateixa longitud és el nombre de posicions on les dues paraules tenen lletres diferents. Per exemple, la distància de Hamming entre “pera” i “figa” és 3, perquè difereixen exactament en tres lletres. Podem entendre aquesta distància com el nombre mínim de substitucions de lletres necessari per transformar una paraula en l'altra, sense afegir ni esborrar lletres, però, contràriament a l'entreteniment de Lewis Carroll, sense necessitat que les paraules intermèdies tinguin cap mena de sentit. La distància de Hamming s'empra encara ara en el disseny de codis per a la transmissió de dades.

L'any 1966, el matemàtic rus V. I. Levenshtein introduí la *distància d'edició* entre paraules, no necessàriament de la mateixa longitud, com una generalització de la distància de Hamming. Donades dues paraules, miram de transformar una en l'altra per mitjà d'una seqüència d'operacions d'edició dels tipus següents: inserir una lletra, eliminar una lletra o substituir una lletra per una altra. Assignam a cada una d'aquestes operacions un cost, i cercam una transformació el cost total de la qual sigui mínim: aquesta cost total mínim serà la distància entre les dues paraules. Com veiem, aquesta distància d'edició generalitza la distància de Hamming en dos sentits. En primer lloc, considera operacions que no siguin la substitució rònega d'una lletra per una altra, la qual cosa permet comparar paraules de longituds diferents. Segonament, incorpora la possibilitat que unes operacions costin més que unes altres: per exemple, permet modelar el fet que el cost de substituir una lletra per una altra no tingui per què ser constant (1, a la distància de Hamming), sinó que pugui dependre de les lletres.

Les transformacions emprades en la definició de la distància d'edició se solen representar per mitjà d'*alineaments*. Un alineament entre dues paraules $X = x_1 \dots x_n$ i $Y = y_1 \dots y_m$ és una taula de dues files. La fila superior conté totes les lletres de X en l'ordre amb què hi apareixen, possiblement separades per espais en blanc, i el

mateix passa amb la fila inferior i la paraula Y ; l'única restricció és que cap columna no pot contenir un espai en blanc a les dues files. El fet que un alineament aparelli en una columna una lletra x_i de X amb una lletra y_j de Y representa la substitució de la x_i per y_j . Una columna que contingui una x_i a la fila superior i un espai en blanc a l'inferior significa que la lletra x_i s'ha esborrat i, per contra, que una columna aparelli una y_j a la fila inferior amb un espai en blanc a la superior, significa que la lletra y_j ha estat inserida. Per exemple, dos possibles alineaments entre “pera” i “figa” són (indicam els espais en blanc amb un símbol $_$)

```

p e r a      p e r _ _ _ a
f i g a      _ _ _ f i g a

```

El primer correspon a substituir les lletres “p”, “e” i “r” per “f”, “i” i “g”, respectivament, i el segon a esborrar el bocí “per” i afegir el bocí “fig”.

Si bé Levenshtein inventà la distància d'edició, no proposà cap algoritme explícit per calcular-la, i en els anys immediatament posteriors diversos autors proposaren algoritmes similars per calcular-la, tots ells basats en el paradigma de computació anomenat *programació dinàmica*. Un dels primers fou introduït per S. B. Needleman i C. D. Wunsch l'any 1970 per comparar proteïnes, en un dels primers treballs de biologia computacional, la branca teòrica de la bioinformàtica.

No és casualitat que un dels primers algoritmes per calcular distàncies d'edició es publicàs en una revista de biologia molecular. En darrera instància, les molècules bàsiques per a la vida, els àcids nucleics i les proteïnes (permeteu-nos dir-ne, per abreviar, *biomolècules*), poden ser enteses com a paraules sobre uns alfabet adients: per al DNA, aquest alfabet està format per les bases nucleiques adenina, citosina, guanina i timina, que s'abreuen A, C, G i T, respectivament; l'alfabet per al RNA és similar, només hi hem de canviar la timina per l'uracil, representat per una U: pel que fa a les proteïnes, l'alfabet és donat pels 20 aminoàcids. Llavors, la comparació d'aquestes paraules es pot emprar per deduir propietats de les molècules que representen, com ara l'homologia. Diem que dues biomolècules d'aquestes son *homòlogues* quan creiem que tenen un avantpassat en comú. Per exemple, un gen de l'home i un gen del ximpanzé (recordau que els *gens* són bocins de DNA) són homòlegs quan un gen d'un organisme del qual provenen tant els homes com els ximpanzés ha sofert, en el decurs dels mil.lennis, dos camins diferents de mutacions que han produït els dos gens actuals. Per desgràcia, mai no tenim accés a l'avantpassat en comú, i per tant hem de deduir l'homologia de dues biomolècules a partir de la similaritat de les paraules que les representen: com més se semblen dues biomolècules, més probable entenem que és la seva homologia. I atès que les operacions d'edició corresponen a mutacions elementals, les quals podem suposar que es produeixen amb velocitat constant, com més petita sigui la distància d'edició entre dues biomolècules, menys temps representa que fa que s'esbrancaren a partir de l'avantpassat en comú. Consideracions com aquestes han fet que la comparació de paraules sigui un dels temes clau en biologia molecular els darrers quaranta anys.

Aquí hem fet servir una simplificació de l'algoritme de Needleman-Wunsch pro-

posada per P. H. Sellers l'any 1974 per tractar el cas quan el cost d'afegir o eliminar una lletra és constant. Suposem, per fixar idees, que tenim un alfabet Σ . Definim una matriu de costos $(\sigma(a, b))_{a, b \in \Sigma}$: cada $\sigma(a, b)$ representa el cost de substituir una a per una b . A més, fixam un cost γ constant per a l'afegit o esborrat d'una lletra. D'aquesta manera, es defineix el *cost* d'un alineament entre dues paraules $X = x_1 \dots x_n$ i $Y = y_1 \dots y_m$ com la suma dels valors $\sigma(x_i, y_j)$ per a cada columna que aparelli una x_i amb una y_j , més γ multiplicat pel nombre de columnes que tenen qualche espai en blanc en una de les files. Diem un *alineament òptim* de X i Y a un alineament de cost mínim.

Ara, donades dues paraules $X = x_1 \dots x_n$ i $Y = y_1 \dots y_m$ escrites en l'alfabet Σ , l'algoritme de Needleman-Wunsch-Sellers calcula de manera recurrent la matriu d'ordre $(F(i, j))_{\substack{i=0, \dots, n \\ j=0, \dots, m}}$, cada entrada $F(i, j)$ de la qual dóna el cost d'un alineament òptim entre els prefixos $x_1 \dots x_i$ i $y_1 \dots y_j$ (quan $i = 0$ o $j = 0$, el prefix corresponent és la paraula buida, sense lletres). D'aquesta manera, el valor $F(n, m)$ serà la distància d'edició entre X i Y .

La idea en què es basa l'algoritme és que, de la manera additiva com hem definit el cost d'un alineament, un alineament òptim entre X i Y dóna lloc a alineaments òptims entre els prefixos que fa correspondre. Això implica que si un alineament òptim entre $x_1 \dots x_i$ i $y_1 \dots y_j$ aparellà x_i amb y_j , aleshores induiria un alineament òptim entre $x_1 \dots x_{i-1}$ i $y_1 \dots y_{j-1}$, i per tant en aquest cas tindríem que $F(i, j) = F(i-1, j-1) + \sigma(x_i, y_j)$. De manera similar, si un alineament òptim entre $x_1 \dots x_i$ i $y_1 \dots y_j$ aparellà x_i amb un espai en blanc, aleshores induiria un alineament òptim entre $x_1 \dots x_{i-1}$ i $y_1 \dots y_j$, de tal manera que tindríem que $F(i, j) = F(i-1, j) + \gamma$. I si un alineament òptim entre $x_1 \dots x_i$ i $y_1 \dots y_j$ aparellà y_j amb un espai en blanc, aleshores induiria un alineament òptim entre $x_1 \dots x_i$ i $y_1 \dots y_{j-1}$ i per consegüent $F(i, j) = F(i, j-1) + \gamma$. Aleshores, a cada moment, el valor real de $F(i, j)$ serà el mínim d'aquests tres valors possibles: això defineix una recursió on $F(i, j)$ es calcula un cop coneguts $F(i-1, j-1)$, $F(i-1, j)$ i $F(i, j-1)$. D'altra banda, els valors $F(i, 0)$ i $F(0, i)$ del cost d'un alineament òptim entre $x_1 \dots x_i$ i $y_1 \dots y_i$, respectivament, i la paraula buida són forçosament $i\gamma$. Tot plegat, això ens dóna l'algoritme senzill següent:

```

Input  $x_1 \dots x_n, y_1 \dots y_m$ 
 $F(0, 0) = 0$ 
For  $i = 0, \dots, n$ 
     $F(i, 0) = i\gamma$ 
For  $j = 0, \dots, m$ 
     $F(0, j) = j\gamma$ 
For  $i = 1, \dots, n$ 
    For  $j = 0, \dots, m$ 
         $F(i, j) = \min\{F(i-1, j-1) + \sigma(x_i, y_j), F(i-1, j) + \gamma, F(i, j-1) + \gamma\}$ 
Output  $F(n, m)$ 

```

A més, si recordam en cada càlcul de $F(i, j)$ per a $i, j > 0$ quina de les tres opcions hem aplicat, podem obtenir fàcilment un alineament òptim entre les dues paraules entrades.

En la nostra aplicació, hem pres com a alfabet Σ el conjunt de totes les lletres, possiblement amb accents, que es fan servir en polonès, espanyol o català: en total 46 caràcters. Pel que fa a la matriu de costos, per simplificar i sense entrar ara en detall, l'hem definida de la manera següent:

$$\sigma(a, b) = \begin{cases} 0 & \text{si } a = b, \text{ llevat d'accents (és a dir, els accents no compten)} \\ 1 & \text{si } a \text{ i } b \text{ són lletres diferents, però el seu sò és similar} \\ 2 & \text{en tot altre cas} \end{cases}$$

La matriu detallada la trobareu a la pàgina web bioinfo.uib.es/~joemiro/polcat. Finalment, hem pres $\gamma = 3$.

Així, per exemple, si empram aquest algoritme per calcular la distància d'edició entre la paraula catalana “anglès” i la seva traducció polonesa “angielski”, construïm la matriu següent (totes les lletres en aquest exemple o bé són iguals, o bé sonen diferent):

		a	n	g	i	e	l	s	k	i
	0	1	2	3	4	5	6	7	8	9
0	<u>0</u>	3	6	9	12	15	18	21	24	27
a	1	3	<u>0</u>	3	6	9	12	15	18	21
n	2	6	3	<u>0</u>	3	6	9	12	15	18
g	3	9	6	3	<u>0</u>	3	6	9	12	15
l	4	12	9	6	3	<u>2</u>	5	6	9	12
è	5	15	12	9	6	5	<u>2</u>	<u>5</u>	8	11
s	6	18	15	12	9	8	5	5	<u>5</u>	<u>8</u>
i									<u>8</u>	<u>11</u>

Això mostra que la distància d'edició entre aquestes dues paraules és 11. A més, hem subratllat el camí de les entrades que han produït la darrera entrada: l'11 del racó inferior a la dreta prové del 8 a la seva esquerra, el qual prové del 5 a la seva esquerra, i així successivament. Si traduïm aquest camí en un alineament, obtenim l'alineament òptim següent:

a n g i e l s k i
a n g l è _ s _ _

Us deixam com a exercici que calculeu la distància d'edició i un alineament òptim entre “angielski” i el corresponent espanyol “inglés”.

Bé doncs, hem aplicat aquest algoritme a 427 paraules poloneses extretes a l'atzar d'un diccionari polonès-anglès que tenim; els càlculs detallats els trobareu a la pàgina web esmentada abans. El resultat ha estat que la distància mitjana d'aquestes paraules a la seva traducció catalana ha estat 12.43, mentre que la distància a la seva

traducció espanyola ha estat 12.31. Com veieu, els resultats són similars (què esperàveu?), però la distància mitjana de les paraules poloneses a les espanyoles ha resultat ser una mica inferior a la seva distància mitjana a les catalanes!

És clar que la mostra ha estat molt petita (el nostre temps lliure és escàs), i que la matriu de costos és millorable. A més, segurament aquesta anàlisi seria més acurada si la basàssim en fonemes, i no en lletres escrites: així podríem estudiar quin d'entre el català i el castellà “sona més a polonès”. Qualcú s'anima a portar a terme de manera seriosa aquest estudi? Per ventura l'Ig Nobel us està trucant a la porta!

Ah, i com que som molt agraïts, no podem acabar sense esmentar que aquest treball ha estat finançat en part pel projecte BFM2003-00771 ALBIOM de la DGES espanyola.