RECURSOS DIGITALES PARA EL ANÁLISIS DEL LÉXICO DE ORIGEN LATINO EN LAS LENGUAS ROMÁNICAS[1]

SIMONA GEORGESCU, ALINA MARIA CRISTEA, ANCA DINU, LIVIU P. DINU, BOGDAN IORDACHE, ANA SABINA UBAN, LAURENȚIU ZOICAȘ

Universitatea din București
simona.georgescu@lls.unibuc.ro

Cuvinte-cheie: *lingvistică istorică, lingvistică computațională, etimologie, cuvinte moștenite vs împrumuturi lexicale.*

Keywords: *historical linguistics, computational linguistics, etymology, inherited vs borrowed words.*

INTRODUCCIÓN

Desde la creación de bases digitales de textos y tesoros lexicográficos, los recursos computacionales no han dejado de mostrar su utilidad en las investigaciones lingüísticas y filológicas. El procesamiento del lenguaje natural ha facilitado un marco propicio para una verdadera simbiosis entre la lingüística y la informática, proporcionando herramientas significativas para el avance de ambos dominios.

La mayoría de los trabajos desarrollados en este ámbito explotan el eje sincrónico de la lengua – ya se trate de traducción automática o de lingüística de corpus –, aunque el interés por el estudio diacrónico de las lenguas a través de métodos digitales se hace cada vez más visible.

Al considerar que el uso de los recursos computacionales podría resultar en perspectivas renovadas sobre la lingüística histórica, hemos propuesto el proyecto *Computational Tools for Historical Linguistics* (CoToHiLi), con el propósito general de integrar el conocimiento experto y el poder computacional en la aproximación a temas como la identificación de palabras cognadas, la discriminación entre palabras patrimoniales y préstamos lingüísticos, la reconstrucción de protoformas léxicas, o el análisis de la divergencia semántica.

El objetivo se dibuja según dos ejes: por un lado, nos proponemos automatizar ciertas etapas en el flujo de trabajo tradicional del método comparativo

[1] Esta investigación se ha desarrollado en el marco del proyecto CNCS/CCCDI – UEFISCDI, CoToHiLi, nº 108 / PNCDI III.

DOI: 10.59277/SCL.2023.1.04

(como la recopilación y selección de datos válidos, el procesamiento inicial o la alineación automática basada en reglas predefinidas) y, por otro lado, pretendemos aportar nuevos conocimientos o vías de investigación que de otro modo no serían fácilmente accesibles (por ejemplo, la identificación automática de patrones y regularidades en un gran caudal de datos).

Para contar con un fundamento científico bien definido y provisto de recursos abundantes, capaz de facilitar los datos necesarios para el desarrollo de diferentes experimentos, nos hemos centrado en el grupo principal de lenguas románicas (rumano, italiano, francés, español y portugués). Sin embargo, el propósito final rebasa los límites de la familia románica, ya que se proponen métodos para el estudio del léxico que se podrán aplicar también a grupos de lenguas menos estudiadas (cf. List *et al*., 2017), para una clasificación precisa de las relaciones que las caracterizan.

En el artículo presente se expondrán los primeras dos temas de investigación, a saber, la identificación de palabras cognadas y la discriminación entre palabras patrimoniales y préstamos lingüísticos.

LA IMPORTANCIA DE LA IDENTIFICACIÓN DE COGNADOS

La identificación y discriminación de cognados representa tanto la base de la lingüística histórica (Campbell, 1998; Mallory y Adams, 2006), como el punto de partida en la investigación sociohistórica (Mailhammer, 2015). Las implicaciones inmediatas de la identificación precisa de palabras cognadas se pueden encontrar en la filogenia lingüística (Dunn, 2015), lo que permite rastrear la relación genética entre lenguas, así como el contacto lingüístico (Epps, 2015) y el eje geográfico y cronológico de las comunidades arcaicas (Heggarty, 2015). Las series de cognados representan, de hecho, el fundamento de la gramática comparada – reconstrucción (Chambon, 2007), proporcionando también la información necesaria para lo que Epps (2015) denominaba como “socio-cultural reconstruction”.

El reto de mecanizar, al menos parcialmente, el proceso de detección precisa de cognados se convierte hoy día en una necesidad, dada la gran cantidad de datos lingüísticos que aún no ha sido procesada desde el punto de vista histórico (List *et al*., 2017). Visto que los métodos clásicos, aunque con resultados excelentes, son menos eficaces y exigen un gran caudal de tiempo y esfuerzo por parte de los pocos lingüistas entrenados en este dominio, se ha pasado paralelamente a una búsqueda intensa de métodos computacionales capaces de facilitar el proceso. El interés creciente por métodos automáticos utilizables en la identificación de cognados ha atraído la necesidad directamente proporcional de bases de datos fiables y consistentes, que permitan la extracción de series de cognados lo más amplias posible. Tales listados de cognados no son fácil de obtener, aun cuando se trate de lenguas conocidas, con recursos digitales bien desarrollados.

LA BASE DE DATOS ROBOCOP

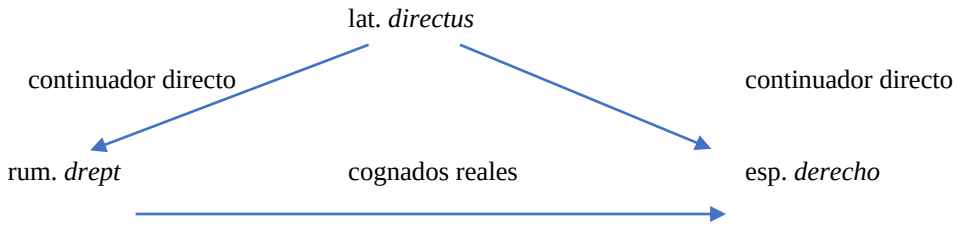
Aunque hay varios bancos de datos que comprenden los cognados románicos, estos son incompletos (como el inventario de lexemas romances basado en el listado de Swadesh, cf. Saenko y Starostin, 2015), poco fiables (como es el caso de los bancos de datos construidos a partir de Wikipedia, cf. Meloni *et al*., 2021), o basados en métodos de traducción automática (Ciobanu y Dinu, 2014). Dadas estas deficiencias, hemos decidido de construir desde cero una base de datos que cubra todas las parejas de cognados identificables entre cualesquiera dos de las principales lenguas romances (rumano, italiano, francés, español y portugués). Para la realización de esta base de datos hemos partido de los diccionarios de referencia disponibles y legibles por máquina, que contienen información etimológica[2]. Nuestra estrategia ha consistido en analizar de manera semiautomática cada uno de los diccionarios, extrayendo la información etimológica de cada lema, para luego juntar todos los datos en una única base digital, denominada RoBoCoP (*Romance Borrowings Cognates Package*). Dado que cada uno de los cinco diccionarios tenía su propia forma de presentar los datos etimológicos, las estrategias de preprocesamiento, de análisis y de posprocesamiento se han tenido que personalizar para cada idioma, lo que ha implicado un esfuerzo significativo tanto por parte de los lingüistas como de los científicos de la computación. Correlacionando toda la información etimológica obtenida, se han creado listados de cognados para cada pareja de lenguas, con el étimo mencionado cada vez, sumando un total de más de 120.000 palabras en todas las lenguas y 90.000 parejas de cognados.

Dado que la base de datos RoBoCoP cubre también los préstamos de otras lenguas aparte del latín, utilizaremos el término de *cognados* en su sentido general, a saber, palabras provenientes de un mismo étimo, ya sean heredadas o tomadas prestadas. Por ejemplo, no solo it. *madre* y esp. *madre* son cognados, sino que también it. *arancia* y esp. *naranja* entran en esta categoría, dada su proveniencia de la misma palabra árabe *nāraŋǧ*.

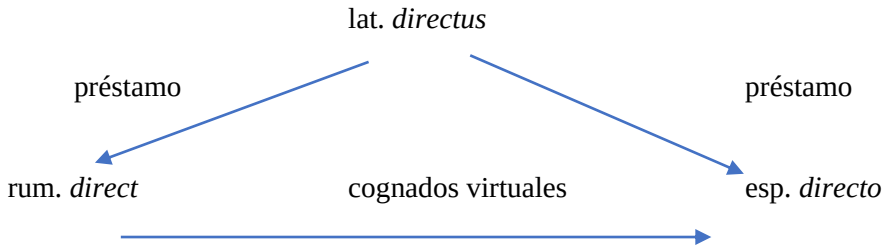
Por lo que atiende a las palabras de origen latino, la base de datos no distingue todavía entre *cognados reales* y *cognados virtuales*. Entendemos por "cognados reales" los descendientes patrimoniales (i.e. que se han conservado de la protolengua en las lenguas vernáculas a través del uso oral ininterrumpido) de un mismo étimo en más de una lengua románica. Por ejemplo, el étimo latino *directus* se ha conservado por transmisión oral en las siguientes formas románicas: rum. *drept*, it. *dritto*, fr. *droit*, esp. *derecho*, ptg. *direito*. Las palabras románicas listadas se relacionan entre ellas en cuanto cognados reales.

[2] *Dicționarul Explicativ al Limbii Române* (dexonline.ro), *Dizionario della lingua italiana De Mauro* (dizionario.internazionale.it), *Trésor de la Langue Française informatisé* (www.cnrtl.fr), *Diccionario de la lengua española*, publicado por la Real Academia Española (dle.rae.es), *Dicionario Infopedia da Lingua Portuguesa*, publicado por Porto Editora (infopedia.pt/lingua-portuguesa).

Lat. *directus*



El concepto de "cognados virtuales" se aplica a las palabras tomadas prestadas del latín (por vía culta) de un mismo étimo, en más de una lengua románica. No pocas veces se da el caso de que un étimo latino que ya tenía descendientes patrimoniales en las lenguas románicas se haya tomado también prestado, justamente porque la variante heredada había cambiado de sentido, de modo que se hizo necesario un doblete más cercano semánticamente a la palabra de origen. Es el caso del mismo lexema latino *directus*, que, al lado de la familia de cognados reales, desarrolla una serie de cognados virtuales: rum. *direct*, it. *diretto*, fr. *direct*, esp. *directo*, ptg. *directo*. Estos lexemas han penetrado en las lenguas románicas a partir del siglo XIII (en el caso del francés) hasta el siglo XIX (cuando penetró en rumano).



La posibilidad de distinción automática entre cognados reales y cognados virtuales es el desafío principal para el cual nos proponemos buscar una resolución.

EXPERIMENTOS

Con este fin, hemos llevado a cabo varios experimentos, cuyo propósito ha sido el de observar la capacidad del sistema computacional de decidir si dos palabras provenientes de dos lenguas romances son cognados o no (para una descripción detallada de los experimentos, v. Cristea *et al*., 2023 [en curso de publicación]). Este recurso supone el entrenamiento del ordenador tanto con datos positivos (parejas verdaderas de cognados) como negativos (pares de palabras sin

relación recíproca). La elección de ejemplos negativos es esencial para fomentar una interpretación correcta de los resultados del proceso automático de detección. Por ejemplo, es fácil decidir que dos palabras obviamente diferentes desde el punto de vista fonético, provenientes de dos lenguas distintas, como rum. *măr* 'manzana' y esp. *manzana* no son cognados, pero el proceso de discriminación se vuelve más difícil a la hora de distinguir entre dos palabras similares formalmente, como rum. *măr* y esp. *mar*. Para abordar este problema, hemos propuesto como método principal de generación de ejemplos negativos un sistema basado en Levenshtein (1965): se han seleccionado como ejemplos negativos pares de palabras similares desde el punto de vista gráfico, que no tienen etimología común, y cuya diferencia formal no rebasa los límites de la distancia de Levenshtein (1965). El criterio que se ha utilizado en la elección de distancias medias de Levenshtein para los pares negativos ha sido que no excedan los intervalos entre palabras de pares positivos, lo que ha garantizado que la distinción entre cognados y no cognados no se base simplemente en la similitud formal. Los experimentos se han desarrollado tanto a partir de la forma gráfica, como de la fonética, y utilizando la extracción de características a base de las alineaciones entre representaciones gráficas tal y como las genera el algoritmo Needleman-Wunsch (cf. Needleman y Wunsch, 1970).

RESULTADOS

Para cualquier pareja de idiomas, la detección de pares de cognados a partir de listados cargados de ejemplos positivos y negativos ha alcanzado un máximo de 95% (en el caso de la pareja italiano – español) y un mínimo de 91% (en la pareja español – rumano). Se ha observado que los valores mínimos resultan de parejas que incluyen la lengua rumana, mientras que los valores máximos los producen los pares de los que forma parte el italiano. La interpretación de estos resultados puede renovar la perspectiva global sobre el grado de similitud entre lenguas (cf. Dinu y Dinu, 2005; Ciobanu y Dinu, 2014), favoreciendo un escudriño más profundo de la estructura fonética de las lenguas romances, observada tanto en relación con el latín, como en comparación recíproca.

Cuando la búsqueda se ha restringido a los cognados originados en latín, el promedio de los resultados correctos ha superado los valores obtenidos en el experimento general. Esta precisión creciente se apoya en la regularidad de la evolución fonética del latín hacia las lenguas romances, lo que también lleva a unas correspondencias bastante constantes entre cualesquiera dos lenguas románicas. Queda claro que la máquina ha sido capaz de aprender y reconocer mejor las correspondencias fonéticas entre voces originadas en latín, rasgos que no son válidos para vocablos provenientes de otras lenguas, ya sea indoeuropeas (como el inglés) o no-indoeuropeas (como el árabe).

Ha habido también casos donde el experimento ha puesto de relieve algunas limitaciones del banco de datos, o bien ha planteado ciertas preguntas, cuando la identificación de una pareja de cognados se ha categorizado como error, pese a una evidente relación genética. Por ejemplo, el esp. *cognitivo* y el rum. *cognitiv* no se han registrado como cognados en nuestra base de datos, dado que en los diccionarios figuran con etimología distinta: el lexema español se considera como creación interna (un supuesto derivado del esp. *cognición*), mientras que la voz rumana es un préstamo del fr. *cognitif*. La selección automática de tales parejas como cognados pone en duda el estatuto de creación interna de palabras como el esp. *cognitivo*, dadas las posibilidades limitadas de derivación con el sufijo *-ivo* (en este caso) en español (cf. Real Academia Española, 2010), imponiendo al mismo tiempo una reevaluación de la influencia francesa en el léxico español.

DISTINCIÓN AUTOMÁTICA ENTRE PALABRAS PATRIMONIALES Y PRÉSTAMOS DEL LATÍN EN LAS LENGUAS ROMÁNICAS

Otro experimento que hemos desarrollado plantea la posibilidad de discriminación automática entre palabras heredadas y préstamos del latín, con el propósito, a largo plazo, de poder distinguir sistemáticamente entre cognados reales y cognados virtuales.

El tópico es esencial para los estudios de filogenia lingüística, en vista de una mejor clasificación de las relaciones genéticas entre lenguas, así como para la correcta localización geográfica y cronológica de las lenguas arcaicas. El reto principal es establecer si una determinada similitud léxica entre dos lenguas provenientes del mismo ancestro es el resultado de una divergencia a partir de su origen común, o de una convergencia a través del préstamo y contacto lingüístico (Heggarty, 2012: 307). De este modo, se puede llegar finalmente a determinar si dos lenguas tienen realmente un ancestro común, o las eventuales similitudes provienen del contacto lingüístico:

> this gave way to a very different line of thought: that the method could be applied to languages not yet known to be related, and used precisely to diagnose or establish whether they were. From the information question ‘how closely related?’, one had passed to the yes/no question ‘related or not?’ (Heggarty, 2012: 308)

Heggarty pone de relieve la necesidad de distinguir entre cognados y préstamos, haciendo hincapié en que los eventuales errores, al distorsionar la situación real, pueden invalidar los resultados. Por ende, apunta el autor, la discriminación entre estas dos categorías etimológicas debe ser muy precisa, y, en consecuencia, la aproximación computacional se estima como un acercamiento muy propicio:

> What solution is there, then, if we can neither ignore the problem of distinguishing cognates from loanwords it, nor overcome it in the many cases where we do not have the necessarily linguistic knowledge to do so? There is in fact a possibility: to sidestep the question entirely at the data analysis stage, and simply to identify which forms are judged to be somehow 'correlate' with each other – whether by specialists in those languages, or more objectively by computerised approaches such as the Automated Similarity Judgment Program, or ASJP (Brown *et al.* 2008, Holman *et al.* 2008). (Heggary, 2012: 310)

Siguiendo estas sugerencias, nos hemos planteado la pregunta si sería posible hacer la distinción entre voces heredadas y préstamos de manera automática. Dado que la máquina puede procesar miles de datos en intervalos breves de tiempo, el propósito que se ha formulado en lingüística computacional ha sido el de distinguir entre las dos categorías utilizando como *input* solo la interfaz fonética de las palabras.

La hipótesis no es falta de razón, dado que el postulado esencial de la lingüística histórica es que el cambio fonético se produce – en principio – de modo regular y tiene, por lo tanto, resultados previsibles y limitados desde el punto de vista estructural: en otras palabras, queda un número limitado de combinaciones fonéticas posibles, un determinado sonido puede aparecer solo en ciertos contextos fonéticos, etc. Por ejemplo, si observamos que la evolución regular del grupo consonántico *-ct-* en las palabras heredadas en español tiene el resultado *-ch-* [tʃ] (cf. lat. *nocte* > esp. *noche*, lat. *pectus* > esp. *pecho*, lat. *octo* > esp. *ocho*, etc.), sacaremos la conclusión de que las palabras que no han experimentado el mismo cambio fonético deben de haber tenido otro trayecto histórico, a saber, deben de haber sido tomadas prestadas (p. ej. lat. *no**ct**urnus* – esp. *nocturno* y no **nochurno*; lat. *octavus* – esp. *octavo*, etc.). No sería irracional, por ende, categorizar automáticamente los vocablos que contienen el grupo *-ct-* en español como préstamos. El mismo mecanismo se puede aplicar a diferentes grupos de sonidos que representan el resultado de una evolución oral frente a la adopción culta de lexemas latinos: resulta plausible la premisa de que las palabras adoptadas por vía culta tengan una estructura fonética en cierta medida diferente de los descendientes por vía oral, incluso de sus dobletes etimológicos.

Siguiendo estas consideraciones, hemos utilizado un modelo de aprendizaje automático aplicado al léxico románico de origen latino (tanto heredado como adoptado por vía culta), con el fin de observar en qué medida la aproximación computacional permite delimitar las dos categorías etimológicas (cf. Cristea *et al*., 2021). En dicho experimento, además de las cinco lenguas ya mencionadas, se ha incluido también el catalán.

En una primera etapa, hemos ofrecido como *input* solamente la forma de las palabras, sin ninguna información suplementaria que concierna a las leyes fonéticas implicadas en las diferentes evoluciones. Más preciso, la máquina ha analizado un listado de voces heredadas y uno de préstamos del latín, reteniendo

los rasgos fonéticos recurrentes en cada grupo. Luego, se ha introducido en el modelo de aprendizaje automático una nueva lista de voces que no se hallaban en ninguna de las dos listas anteriores, con el requisito de que se clasifique cada palabra, automáticamente, en una de las dos categorías mencionadas. El grado de precisión en este proceso no ha sobrepasado el 54%, lo cual ha exigido el perfeccionamiento del método propuesto.

En la etapa subsiguiente, hemos enriquecido la información básica proporcionada al programa con un núcleo de datos lingüísticos significativos: además de ofrecer el étimo de cada palabra, hemos introducido una serie de rasgos generales que caracterizan la evolución fonética en las voces románicas heredadas. Para simplificar el papel de la máquina, por un lado, y, por otro lado, con el fin de observar empíricamente cuál sería la cantidad mínima de datos lingüísticos necesaria para obtener resultados válidos en el procesamiento automático del lenguaje, nos hemos enfocado en las leyes fonéticas que intervinieron en el cambio consonántico, dejando de lado el comportamiento de las vocales[3].

Con este propósito, hemos intentado simplificar cuanto más posible el *input* para un procesamiento correcto. Hemos sistematizado los cambios consonánticos interpretándolos como parte de un proceso general de lenición, que engloba la mayoría de las leyes fonéticas particulares identificables en lenguas románicas diferentes. En breve, la trayectoria recurrente de una consonante en su evolución del latín hacia las lenguas románicas se podría sintetizar de la manera siguiente: oclusiva sorda → oclusiva sonora → africada sorda → africada sonora → fricativa sorda → fricativa sonora → semivocal → pérdida. Se entiende de por sí que un sonido no debe recorrer todas las etapas (o que no todas son visibles en diacronía), ni alcanzar la última etapa, que sería la pérdida. Por ejemplo, p → b → v, como en las ecuaciones lat. *ri**p**a* > esp. *ri**b**a* // fr. *ri**v**e*; lat. *ca**p**ra* > esp. *ca**b**ra* // fr. *chè**v**re*; o b → ø, como en lat. *canta**b**am* > rum. *cântam*.

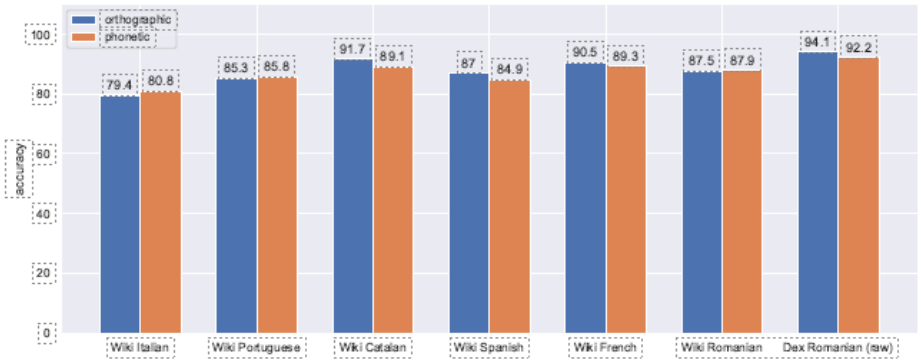
El cambio del lugar de articulación, conocido como *palatalización*, se enmarca, de hecho, en la misma pauta evolutiva que supone una debilitación articulatoria: por ejemplo, k → tʃ (→ ʃ), cf. lat. *caput* ‘cabeza’ > fr. *chef* [ʃef]; o bien, k → ts → s, cf. lat. *caelum* ['kelum] > fr. *ciel* [siel].

Según se puede observar en los ejemplos proporcionados, en ciertos casos la grafía oculta la pronunciación (cf. fr. *ciel* [siel]). Por ende, para que se puedan aplicar de manera incorrupta las reglas predefinidas, hemos construido dos bases de datos paralelas: la primera contiene las palabras románicas y los étimos latinos en su forma gráfica, mientras que la segunda se constituye de las transcripciones fonéticas de las mismas voces.

[3] Es consabido que la evolución vocálica comporta un gran número de variables difícilmente sistematizables. En ausencia de una perspectiva completa y perfectamente ordenada de todos los casos de cambio vocálico, el procesamiento automático sería viciado y no podría ofrecer resultados válidos.

Tomando como punto de partida la trayectoria recurrente dirigida por la lenición, hemos extraído todos los cambios posibles y los hemos codificado como rasgos binarios, marcando con 0 su ausencia y con 1 su presencia en las palabras dadas.

Esta vez se ha obtenido un promedio de 90% de clasificaciones correctas de las palabras dadas en las dos categorías (voz heredada *vs* préstamo léxico), un resultado visiblemente superior al que se había conseguido en la primera prueba. El diagrama que viene representado abajo recoge los resultados obtenidos para todas las lenguas románicas, distinguiendo entre la aplicación del algoritmo a las variantes gráficas y el empleo del mismo a las formas transcritas en el alfabeto fonético internacional[4].



Los resultados del experimento con información lingüística.

El uso de dos métodos paralelos ha resultado en una leve oscilación en el porcentaje obtenido según la correctitud de los resultados. La relación entre la forma gráfica y la fonética difiere de una lengua a otra, según se puede notar fácilmente: en tres de las seis lenguas analizadas, a saber, en catalán, español y francés, se observa que el resultado mejor (con más del 1%) se ha obtenido cuando se han analizado las variantes gráficas, mientras que en la otra mitad una correctitud levemente mayor la han ofrecido las variantes fonéticas (aunque, salvo en italiano, la mejora no ha sobrepasado el 1%).

Este experimento nos deja en claro la importancia de los datos lingüísticos tradicionales en el procesamiento automático del lenguaje. Una exploración exclusivamente computacional del lenguaje no ha proporcionado resultados significativos, ya que el porcentaje de 54% que describe la correctitud obtenida en la primera etapa del experimento corresponde a lo que hubiera proporcionado la mera aplicación de la teoría de la probabilidad. La integración de reglas lingüísticas predefinidas ha mejorado visiblemente la calidad del proceso.

[4] Para el rumano hemos tenido a disposición dos listados paralelos, uno que recogía las palabras del diccionario *Wiki*, el otro creado a partir de www.dexonline.ro, que junta la mayor parte de los diccionarios rumanos. Sin embargo, para asegurar la unidad del método aplicado a todas las lenguas románicas, vamos a centrarnos en los resultados relevados por el análisis del listado Wiki.

CONCLUSIONES

Las investigaciones de las últimas décadas en el ámbito de la lingüística han resaltado una necesidad creciente de métodos y herramientas que faciliten el recurso científico tradicional. Las herramientas digitales que proponemos podrían abrir el camino hacia una nueva perspectiva sobre el procesamiento del léxico. Aunque los métodos dibujados se derivan del estudio del léxico románico, los algoritmos resultantes podrían tener un área mucho más amplia de aplicación, así como permitir inferencias generales sobre el cambio lingüístico.

En el artículo presente hemos descrito dos de los principales trabajos llevados a cabo en el marco del proyecto CoToHiLi: la constitución de una base de datos que comprende las series de cognados románicos y la investigación de la posibilidad de distinguir (semi)automáticamente entre voces heredadas y préstamos léxicos. Los resultados han hecho patente la importancia de los datos lingüísticos integrados en el procesamiento automático, poniendo de relieve que la mejor aproximación a la lingüística histórica junta el saber de los lingüistas clásicos con el poder de los recursos computacionales.

BIBLIOGRAFÍA

Campbell, Lyle, 2013, *Historical Linguistics. An Introduction* (1ª ed. 1998), 3ª edición, Edinburgh University Press.

Chambon, Jean-Pierre, 2007, "Remarques sur la grammaire comparée-reconstruction en linguistique romane (situation, perspectives)", *Mémoires de la Société de linguistique de Paris*, n. s., 15 (*Tradition et rupture dans les grammaires comparées de différentes familles de langues*), Lovaina, Peeters, 57–72.

Ciobanu, Alina Maria / Dinu, Liviu P., 2014, "Building a Dataset of Multilingual Cognates for the Romanian Lexicon", in *Proceedings of the language resources and evaluation conference 2014*, p. 1038–1043.

Cristea, Alina Maria *et al*., 2021, "Automatic discrimination between inherited and borrowed Latin words in Romance languages", in: *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNPL) (Findings)*.

Cristea, Alina Maria *et al*., 2023 [en curso de publicación], "RoBoCoP: A comprehensive Romance Borrowing Cognate Package and Benchmark for Multilingual Cognate Identification".

Dunn, Michael, 2015, "Language phylogenies", in Bowern, Claire / Evans, Bethwyn, *The Routledge handbook of historical linguistics*, London/New York, Routledge, 190–211.

Dinu, Anca / Dinu, Liviu P., 2005, "On the syllabic similarities of Romance languages, in *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing* 2005, Mexico City, 785–788, Springer.

Epps, Patience, 2015, "Historical linguistics and socio-cultural reconstruction", in Bowern, Claire / Evans, Bethwyn, *The Routledge handbook of historical linguistics*, London/New York, Routledge, 579–597.

Heggarty, Paul, 2012, "Beyond lexicostatistics: how to get more out of 'word list' comparisons", *Diachronica* 27(2), 301–324.

Heggarty, Paul, 2015, "Prehistory through language and archaeology", in Bowern, Claire / Evans, Bethwyn, *The Routledge handbook of historical linguistics*, London/New York, Routledge, 598–626.

Levenshtein, Vladimir I., 1965, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", *Soviet Physics Doklady* 10, 707–710.
List, Johann-Mattis *et al*., 2017, "The Potential of Automatic Word Comparison for Historical Linguistics", PLOS ONE, 12(1), 1–18.
Mailhammer, R. 2015, "Etymology", in Bowern, C./Evans, B., *The Routledge handbook of historical linguistics*, London/New York, Routledge, 423–441.
Mallory, J.P. / Adams, D.Q. 2006, *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*, Oxford University Press.
Meloni, Carlo / Ravfogel, Shauli / Goldberg, Yoav, 2021, "*Ab antiquo*: Neural proto-language reconstruction", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2021, 4460–4473.
Needleman, Saul B. / Wunsch, Christian D., 1970, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *Journal of molecular biology* 48(3), 443–453.
Real Academia Española, 2010, *Nueva gramática de la lengua española. Manual*, Espasa Calpe.
Saenko, Mikhail / Starostin, G.S., 2015, "Annotated Swadesh wordlists for the romance group (Indo-European family)", *The Global Lexicostatistical Database*, RGU.
Uban, Ana Sabina *et al.*, 2021, "Tracking semantic change in cognate sets for English and Romance languages", in *Proceedings of the 2nd international workshop on computational approaches to historical language change*.

RESURSE DIGITALE PENTRU ANALIZA LEXICULUI DE ORIGINE LATINĂ ÎN LIMBILE ROMANICE

Rezumat

În cadrul proiectului CoToHiLi (*Computational Tools for Historical Linguistics*), constituit cu scopul de a oferi instrumente computaționale pentru cercetarea schimbării lingvistice, am creat o bază de date care cuprinde seriile de cognați din principalele limbi romanice. Pe baza acesteia, am desfășurat experimente care testează capacitatea mașinii de a detecta automat cuvintele cognate în oricare pereche de limbi romanice. Totodată, am explorat posibilitatea de a face distincția în mod semiautomat între cuvinte moștenite în limbile romanice și cuvinte împrumutate din latină sau dintr-o altă limbă romanică. Articolul descrie dificultățile generate de fiecare experiment, modul în care au fost aplicate metodele digitale și integrarea datelor lingvistice cunoscute, propunând de asemenea o interpretare a rezultatelor.

DIGITAL RESOURCES FOR THE ANALYSIS OF THE LEXICON OF LATIN ORIGIN IN ROMANCE LANGUAGES

Abstract

Within the framework of the CoToHiLi project (*Computational Tools for Historical Linguistics*), set up with the aim of providing computational tools for the research of linguistic change, we created a database containing the cognate chains from the main Romance languages. Starting from this database, we conducted experiments that test the machine's ability to automatically detect cognate sets in any pair of Romance languages. We also explored the possibility of distinguishing semi-automatically between inherited and borrowed words in the Romance languages. The article describes the challenges of each experiment, the application of digital methods, as well as the importance of integrating linguistic data. We also propose an interpretation of the results.