AZCATL
Revista de divulgación en ciencias, ingeniería e innovación

¿Y si una computadora pudiera entender los dichos mexicanos?

Angeles Belém Priego Sánchez
Universidad Autónoma Metropolitana,
Unidad Azcapotzalco
abps@azc.uam.mx

Resumen

¿Qué tienen en común expresiones como «colgar los tenis» o «dejar con el ojo cuadrado»? Ambas son locuciones verbales, es decir, combinaciones de palabras con un verbo, cuyo significado no se puede deducir literalmente al analizar cada término por separado. Estas expresiones son parte del español mexicano y reflejan aspectos culturales, sociales y emocionales que a menudo se apartan de las reglas gramaticales tradicionales; pero ¿qué pasa cuando intentamos detectar estas expresiones automáticamente en textos digitales? En este artículo exploramos cómo enseñar a las computadoras a reconocerlas, con el objetivo de que comprendan mejor cómo hablamos en la vida cotidiana.

Palabras clave

Frases mexicanas, sentido figurado, aprendizaje automático e inteligencia artificial.

Abstract

What do expressions like «colgar los tenis» (kick the bucket) or «dejar con el ojo cuadrado» (knock someone's socks off) have in common? Both are verbal idioms, meaning combinations of words that include a verb, whose meaning can't be literally deduced by analyzing each term separately. These expressions are part of Mexican Spanish and reflect cultural, social, and emotional aspects that often deviate from traditional grammatical rules. But what happens when we try to detect these expressions automatically in digital texts? In this article, we explore how to teach computers to recognize them, with the goal of helping machines better understand how we speak in everyday life.

Keywords

Mexican expressions, figurative meaning, machine learning, artificial intelligent.

APA: Priego, A. (2025). ¿Y si una computadora pudiera entender los dichos mexicanos? *Azcatl, 5,* 7-11. DOI: 10.24275/AZC2025B002

De qué trata este estudio

La capacidad para comunicarnos se construye gracias a toda la información que vamos aprendiendo y guardando en la memoria a lo largo de la vida. Esa información se organiza en frases, expresiones y dichos que usamos para expresar ideas o conceptos. Estudiar estas frases y expresiones es el objetivo de la fraseología, una rama que se enfoca en entender cómo funcionan estos grupos de palabras. Esto es diferente a la lingüística tradicional, que abarca varias áreas como la estructura de las palabras, la sintaxis, el significado, el uso social del lenguaje y cómo pensamos al hablar.

En el mundo de la lingüística hay muchas formas de nombrar a las expresiones hechas que usamos en el día a día: modismos, locuciones, frases hechas, dichos, refranes, entre otras; se han registrado hasta 64 términos distintos para referirse a este tipo de combinaciones de palabras. En este trabajo se opta por llamarlas unidades fraseológicas, un término ampliamente aceptado por los especialistas tanto en el ámbito hispano como internacional (Priego, 2020).

El presente estudio se enfoca en un tipo muy particular de estas unidades: las locuciones verbales dentro del español que se habla en México. Expresiones como «colgar los tenis» (morir) o «zafarse un tornillo» (perder la razón) dicen mucho más de lo que aparentan y su sentido no se puede deducir simplemente leyendo palabra por palabra. Estas frases son parte del lenguaje local y representan un reto tanto para quienes aprenden español como para las máquinas que intentan entenderlo.

Las locuciones verbales son un tipo particular de unidad fraseológica, combinaciones de dos o más palabras que funcionan como una unidad dentro de una oración; lo interesante es que su significado no puede entenderse simplemente sumando lo que significa cada palabra por separado. En el caso de las locuciones verbales, éstas incluyen un verbo como núcleo y su sentido no es literal. Por ejemplo, «colgar los tenis» no implica realmente que alguien haya puesto calzado en un gancho, sino que significa morir. Este tipo de expresiones tienen tres propiedades fundamentales: cuentan con más de una palabra, se recuerdan como un todo, como una sola *palabra mental* y su significado no puede deducirse por partes. Detectarlas no es sencillo, ya que se puede decir de varias formas una idea con diferentes frases.

Las locuciones verbales son una parte rica y compleja del lenguaje, llenas de matices culturales y lingüísticos. Uno de los grandes aportes del estudio es demostrar que, aunque entender estas locuciones parece intuitivo para hablantes nativos, enseñarle a una computadora a entender el lenguaje humano y reconocerlas es un reto interesante que requiere métodos de inteligencia artificial.

Lo que le enseñamos a las computadoras

La tecnología ha evolucionado tanto que ahora podemos pedirle a una máquina que entienda frases o dichos como si fuera una persona. En el caso del español mexicano, se han empezado a desarrollar herramientas que permitan identificar locuciones verbales en diversos tipos de textos. Cabe aclarar que cuando hablamos de locuciones verbales mexicanas, nos referimos a expresiones que se usan y entienden comúnmente en México, aunque eso no significa que no puedan aparecer en otros países de habla hispana.

Para lograr este objetivo, la comunidad de investigadores ha creado recursos lingüísticos especialmente diseñados para entrenar los sistemas de inteligencia artificial. ¿Cómo funciona esto?, a través de una técnica llamada aprendizaje automático supervisado, un área de la inteligencia artificial que capacita las máquinas para adquirir conocimientos a partir de ejemplos previamente revisados y etiquetados por personas expertas; por lo que, la detección automática se basó en este tipo de técnicas, es decir, utilizando tecnologías que aprenden observando casos concretos, así como nosotros lo hacemos: aprendiendo con ejemplos y correcciones. Supongamos que se desea que la computadora entienda la frase «andar de pata de perro», entonces primero se le *enseña* a partir de ejemplos de uso de la frase, tanto literales como figurativos; una vez que ha *aprendido* cómo se utiliza la frase, posteriormente puede emplearla en otros ejemplos y contextos diferentes.

Teniendo esto en mente, lo primero que se necesita es construir un conjunto de datos, también llamado un corpus etiquetado, con textos —algunos con locuciones verbales y otros que no— que tengan muchos ejemplos de utilización con las diversas frases mexicanas. Con motivo de ser ilustrativo, en este artículo vamos a usar un corpus supervisado en el dominio noticiero. Estos textos fueron extraídos de periódicos mexicanos que forman parte de la Organización Editorial Mexicana (OEM), cada uno fue revisado y anotado manualmente para señalar si contenía o no alguna locución verbal para entrenar modelos con los datos que le permitieran a la computadora utilizar dichas locuciones después, de manera automática y por sí sola.

Para saber qué expresiones buscar se utilizó como fuente principal el Diccionario de Mexicanismos, una obra impresa que recoge palabras y frases propias del español de México. Este diccionario es especialmente útil porque no sólo incluye términos actuales y tradicionales, sino que también compara el uso mexicano del español con el de otros países, en especial con el de España. Además, es un diccionario descriptivo: no impone reglas, sino que refleja cómo se habla realmente el español de México, incluyendo también palabras nuevas y extranjerismos que forman parte del habla cotidiana.

Con todos estos recursos —textos anotados, un diccionario especializado y técnicas de aprendizaje automático—, el objetivo es que las computadoras aprendan a identificar y entender automáticamente cuándo un texto contiene una locución verbal mexicana, ayudando así a entender mejor el lenguaje cotidiano tal como realmente se usa.

Utilizando técnicas de recuperación de información (algo así como un buscador inteligente), se identificaron 3 164 textos periodísticos provenientes de la OEM. Cada uno de estos textos contiene al menos una aparición de alguna de las locuciones verbales seleccionadas, ya sea en su forma original o en una versión ligeramente modificada. Por ejemplo, la locución «darse por vencido» puede encontrarse en otras formas como «darse por vencida», «darnos por vencidos» o «darse por vencidas». Para detectar todas estas variantes, tanto las locuciones como los textos fueron procesados con una técnica llamada lematización, que reduce las palabras a su forma base. Finalmente, se usó un método llamado validación cruzada (en este caso con 10 particiones o *pliegues*) para entrenar y probar el modelo de forma equilibrada y así asegurar que el sistema no sólo reconociera las locuciones que ya había visto, sino que también pudiera identificar otras similares en nuevos textos. La Figura 1 resume las diferentes etapas, descritas anteriormente, empleadas para que una computadora aprenda a entender locuciones verbales mexicanas.

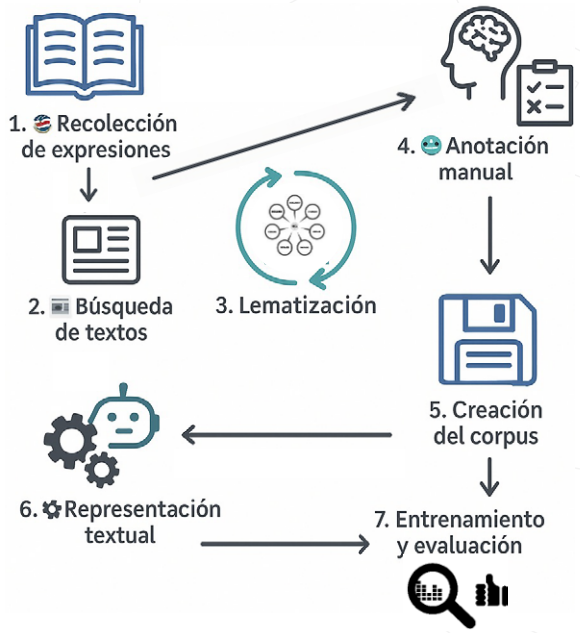

Figura 1. Cómo una computadora aprende a entender locuciones verbales mexicanas.

¿Qué se ha obtenido?

Tras enseñar a las computadoras con varios programas, las cuales aprenden del lenguaje con ejemplos reales, comprobamos si podían detectar locuciones verbales mexicanas. Entonces ¿puede una máquina entender expresiones como «meter la pata»? Aunque parezca increíble, sí; gracias a técnicas de inteligencia artificial, las

computadoras pueden aprender a reconocerlas. Para lograrlo, se entrenaron las máquinas con alrededor de 3 000 frases tomadas del uso cotidiano del español mexicano extraídas de noticias, marcadas previamente por personas expertas. Seguidamente, se pusieron a prueba cuatro algoritmos distintos —árboles de decisión, J48; vectores de soporte, SMO; modelo probabilístico, Naïve Bayes; y vecinos cercanos, K-Star— para ver cuál entendía mejor cuándo una expresión era una locución verbal (Priego y Pinto, 2023).

De los cuatro algoritmos, el más *inteligente* resultó ser J48, una técnica que aprende a tomar decisiones siguiendo patrones en el lenguaje. En el caso específico de este corpus, el modelo logró identificar correctamente estas expresiones con resultados que son cercanos al 77 % de precisión (exactamente 76.74 %), es decir, de cada 10 casos reales del lenguaje con dichos mexicanos la máquina entiende 7, en uno tiene duda y en dos se equivoca. El éxito de los árboles de decisión se debe a que entienden las palabras y el orden en que aparecen, algo clave en este tipo de expresiones idiomáticas.

Este estudio es importante porque estas frases no son sólo adornos del lenguaje, son parte de nuestra cultura, nuestra forma de hablar y de sentir. Si queremos que la tecnología nos entienda realmente —como hablamos en la calle, en las noticias o en redes sociales— necesitamos que también comprenda nuestras expresiones más auténticas. Además, detectar automáticamente estas expresiones es una tarea compleja incluso para humanos, porque muchas veces su significado no se entiende literalmente, así que el hecho de que una máquina pueda hacerlo con una precisión aceptable es un avance enorme.

Los beneficios se observan en tareas como la traducción automática, al permitir encontrar equivalencias idiomáticas más precisas en otros idiomas; el análisis de sentimientos, al facilitar la detección de emociones implícitas en expresiones coloquiales; y en la extracción de información y la generación de resúmenes automáticos, al ayudar a captar mejor el sentido global de un texto. Por otra parte, se mejora la comprensión semántica y sintáctica al tratar estas unidades como bloques coherentes y resulta especialmente útil en sistemas de diálogo, donde permite interpretar y generar respuestas más naturales y contextualmente apropiadas.

Los resultados son alentadores y se espera que, con el apoyo de enfoques lingüísticos más profundos, el rendimiento mejore en futuras investigaciones. Éste es sólo el primer paso. El siguiente reto es ampliar esta investigación a otros tipos de textos, otros acentos, otras regiones y, por supuesto, a las lenguas originarias de México, para que la tecnología también hable nuestra diversidad.

Lo que nos deja esta investigación

Durante los últimos tiempos, el interés por el estudio de las frases o dichos —también conocidos como unidades fraseológicas— ha crecido dentro del campo de la lingüística aplicada y computacional. Esta rama, llamada fraseología, es clave para entender cómo usamos el lenguaje en la vida real y tiene aplicaciones prácticas muy importantes, por ejemplo, en traducción automática, en análisis de opiniones o sentimientos en redes sociales, en procesamiento de lenguaje natural y en educación lingüística, por citar algunos.

Las frases son palabras que hacen equipo para expresar una idea que, aunque puede admitir algunas variaciones, no funciona si la frase se separa o se interpreta palabra por palabra. Un ejemplo claro es la expresión «tirar la toalla», que no significa literalmente lanzar una toalla, sino rendirse. Este tipo de expresiones refleja no sólo la estructura del idioma, sino también aspectos culturales y sociales.

En este trabajo, el foco estuvo en un tipo específico de estas unidades: las locuciones verbales mexicanas, es decir, frases comunes del español de México que giran en torno a un verbo. El objetivo fue analizar estas expresiones y, sobre todo, explorar cómo una computadora puede aprender a identificarlas en textos escritos. Asimismo, permitió demostrar que el español mexicano, con toda su riqueza y variabilidad, puede ser analizado con herramientas computacionales avanzadas.

La investigación destaca la importancia de las locuciones verbales como parte vital de nuestra lengua y

muestra cómo la inteligencia artificial puede acercarse —aunque con limitaciones— al modo en que los humanos entendemos expresiones idiomáticas. Éste es sólo el comienzo, pero ya estamos enseñando a las máquinas a entender cómo hablamos realmente en México. El próximo reto es extender este estudio a otros géneros textuales, a otras lenguas e incluso a las lenguas originarias de México, para que la tecnología refleje la riqueza y diversidad lingüística del país.

Referencias

Alves, D., Fischer, S. y Teich, E. (2025). Syntagmatic productivity of MWEs in scientific english, proceedings of the 21st workshop on multiword expressions. *Association for Computational Linguistics*, pp. 1-6, doi: 10.18653/v1/2025.mwe-1.1

Baldwin, T., Bannard, C., Tanaka, T. y Widdows, D. (2003). *An empirical model of multiword expression decomposability* (pp. 89-96). Association for Computational Linguistics.

Casares, J. (1992). *Introducción a la lexicología moderna*. Consejo Superior de Investigaciones Científicas.

Fakharian, S. y Cook, P. (2021). *Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity*. (pp. 23-32). Association for Computational Linguistics.

González, M. I. y Ortega, G. D. (2005). En torno a la variación de las unidades fraseológicas. En R. Almela, Ramón, E. y Wotjak, G. *Fraseología contrastiva: con ejemplos del alemán, español, francés e italiano*. Universidad de Murcia.

Gramley, S., Gramley, V. y Patzold, K. M. (1992). *A survey of modern english*. Routledge.

Lamiroy, B. (2005). *Le problème central du figement est le semi-figement*. Linx.

Manning, C.D., Raghavan, P. y Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Pinto, D. y Priego, B. (2020). Using automatic constructed thesauri instead of dictionaries in the verbal phraseological units validation task. *Journal of Intelligent & Fuzzy Systems*, *39*(2), pp. 2061-2070. https://doi.org/10.3233/JIFS-179.

Priego, B. (2020). *A new lexical resource for evaluating polarity in spanish verbal phrases*. *Computación y Sistemas*, *24*(2), pp. 725-732. doi: 10.13053/CyS-24-2-3409

Priego, B. y Pinto, D. (2023). *Locuciones verbales del español mexicano: un análisis desde la lingüística computacional*. Libros BUAP.

Ramisch, C., Walsh, A., Blanchard, T. y Taslimipoor, S. (2023). *A survey of MWE identification experiments: the devil is in the details* (pp. 106-120). Association for Computational Linguistics. doi: 10.18653/v1/2023.mwe-1.15

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. y Flickinger, D. (2002). Multiword expressions: a pain in the neck for NLP. En A. Gelbukh (Ed.). *Computational linguistics and intelligent text processing. CICLing 2002. Lecture notes in computer science. Volumen 2276*. Springer. https://doi.org/10.1007/3-540-45715-1_1

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433-460. doi:10.1093/mind/LIX.236.433

Zuloaga, A. (1980). *Introducción al estudio de las expresiones fijas*. Verlag Peter Lang.