CORPUS Y UNIVERSALES LINGÜÍSTICOS EN PRAGMÁTICA[1]

CORPUS AND UNIVERSALS IN LANGUAGE IN PRAGMATICS

Antoni Brosa Rodríguez
Universitat Rovira i Virgili

RESUMEN
Este estudio se ocupa de la interrelación entre universales lingüísticos, pragmática y lingüística de corpus. Por una parte, se analizan las principales limitaciones de los corpus existentes para la sistematización de universales pragmáticos. Por otra, se defiende el uso de estos corpus como herramienta útil para la detección de universales pragmáticos. Posteriormente, se analizan cualitativamente dos herramientas de etiquetado de actos de habla en lengua inglesa (DART y DAMSL) y dos corpus de actos de habla en esta misma lengua (SPAADIA y SWDA). Dicho análisis nos permite extraer las características esenciales (y las principales dificultades) para extrapolar dicha metodología a un conjunto de lenguas tipológicamente equilibrado. Finalmente, se propone un protocolo para la compilación de corpus de actos de habla en distintas lenguas para la extracción de universales lingüísticos.

PALABRAS CLAVE: corpus, universales lingüísticos, pragmática, actos de habla, corpus pragmatics

ABSTRACT
This study deals with the interrelationship between linguistic universals, pragmatics, and corpus linguistics. On the one hand, the main limitations of the existing corpora for the systematization of pragmatic universals are analyzed. On the other hand, the use of these corpora as useful tools for the detection of pragmatic universals is defended. Two English language speech act tagging tools (DART and DAMSL) and two English speech act corpora (SPAADIA and SWDA) are qualitatively reviewed. The analysis of these tools allows us to extract the essential characteristics (and the main difficulties) to extrapolate this methodology to a typologically balanced set of languages. A protocol is proposed for the compilation of corpora of speech acts in different languages for the extraction of linguistic universals.

KEYWORDS: corpus, universals in language, pragmatics, speech acts, corpus pragmatic

[1] Este trabajo ha sido posible gracias a la ayuda de la Universitat Rovira i Virgili (URV) (2019PFR-URV-47).

CC BY NC ND

1. INTRODUCCIÓN

El objetivo de este trabajo es analizar cómo los corpus lingüísticos pueden ayudar a formular universales lingüísticos en el ámbito de la pragmática y qué aspectos deben tenerse en cuenta para ello. Por tanto, nuestro análisis se centra en los estudios sobre universales lingüísticos, en concreto sobre universales pragmáticos, y propone el uso de la lingüística de corpus como la herramienta para el establecimiento de estos universales.

En este trabajo, nos ocupamos específicamente de los denominados universales superficiales, presentados en el siguiente apartado (véase Bickel, 2010). Para poder identificar y estudiar estos, necesitamos un gran número de datos. Para ello, el análisis a través de corpus lingüísticos puede ser una buena solución. La utilidad de los corpus en el estudio de los universales se puede ver, por ejemplo, en herramientas como *Universal Dependencies* (Nivre et al., 2016) para la sintaxis. No obstante, en el ámbito de la pragmática no se dispone de corpus ni herramientas similares para poder formular universales de este tipo. Consideramos que la pragmática de corpus puede ser de gran ayuda en el análisis de los universales pragmáticos. Por ello, en este trabajo nos proponemos revisar algunas herramientas en *corpus pragmatics* para la lengua inglesa y extraer, de ahí, las directrices necesarias para poder elaborar, por un lado, un modelo ampliamente universal y, por otro, distintos corpus en lenguas con una menor tradición en *corpus pragmatics* que permitan la comparación y, en definitiva, la extracción de datos significativos para formular universales lingüísticos, en este caso, en pragmática.

Nuestra propuesta defiende que el uso de corpus lingüísticos puede ayudar a revertir la escasez de universales pragmáticos (en comparación con otras áreas lingüísticas). Lo que presentamos aquí es una primera aproximación metodológica (en la que se hace un énfasis especial en la problemática relacionada con la selección de etiquetas para catalogar las distintas realizaciones) para resolver el problema de la identificación de universales pragmáticos a partir de la intersección de tres áreas de investigación: universales, pragmática y lingüística de corpus.

El artículo se estructura de la siguiente manera. En primer lugar, repasaremos los conceptos básicos de los universales lingüísticos y prestaremos especial atención a los universales pragmáticos. Dedicaremos un apartado a los aspectos de la lingüística de corpus que nos interesan para nuestro trabajo. Posteriormente, revisaremos cómo se pueden analizar los actos de habla desde una perspectiva universal y cómo se pueden extraer estos de un corpus. Analizaremos dos herramientas y propuestas de etiquetado de actos de habla del inglés para corpus (DART y DAMSL) y dos corpus de actos de habla en inglés (SPAADIA y SWDA). Concluiremos poniendo de manifiesto los principales escollos que se han encontrado y proponiendo una vía de actuación que ayude a compilar corpus de actos de habla en distintas lenguas.

2. UNIVERSALES LINGÜÍSTICOS

Los universales lingüísticos, normalmente, son entendidos como características que comparten todas las lenguas (Bickel 2010). Estas invariancias, tradicionalmente, pueden abordarse desde dos perspectivas distintas:

BY NC ND

1. Utilizando una metodología funcionalista, apoyada por herramientas de tipología lingüística (Greenberg, 1963). En este caso, el producto final es conocido como *universal superficial*.
2. Desde la lógica propia del generativismo, es decir, la *Universal Grammar* (Chomsky, 1965). En este caso, el producto final es conocido como *universal profundo*.

Estas dos metodologías utilizan distintas herramientas para llegar a los universales (Scalise et al. 2009). No obstante, el aspecto, seguramente, más relevante es el lugar en el que focalizan la investigación: en terminología chomskiana podríamos hacer la distinción entre *e-language*, como Greenberg, e *i-language*, como Chomsky (Boeckx et al., 2009). Por lo tanto, Greenberg trabajaría con elementos tangibles, cuantificables y explícitos de distintas lenguas (teorización a partir de aspectos externos de la lengua); mientras que Chosmky buscaría la universalidad a través de formulaciones teóricas internas del funcionamiento de las lenguas (teorización a partir de lo interno de las lenguas). En este trabajo, ya que trabajamos con la unión de los universales pragmáticos con los corpus, será de mayor interés el acercamiento de Greenberg, cuya propuesta casa mejor con la metodología de corpus. Dicho de otra manera, los universales profundos surgen de formulaciones abstractas en el seno de una lengua. En cambio, los universales superficiales son el producto de una comparación de las manifestaciones de un conjunto significativo de lenguas equilibrado tipológicamente (Dryer, 1992). Por ello, si se desea trabajar a partir de corpus, tiene más sentido hacerlo en este segundo ámbito, promovido por Greenberg.

Por otro lado, también se deben destacar dos distinciones esenciales a la hora de etiquetar o clasificar universales (Moravcsik, 2010). En primer lugar, es relevante reparar en si el universal descubierto contiene excepciones o no. Si las tiene, aunque sea una, será considerado un *universal estadístico*. Si no se han constatado excepciones, un *universal absoluto*. Lo más habitual es encontrar tendencias universales con excepciones. De hecho, expertos como Dryer (1998) justifican los beneficios de los estadísticos frente a los universales absolutos, ya que estos últimos son tan generales que aportan poca información relevante en relación con las lenguas. O, dicho de otro modo, los universales absolutos suelen ser poco representativos del funcionamiento detallado de las lenguas.

En segundo lugar, también debe atenderse al modo en el que está formulado dicho universal. Se puede hacer de un modo general, aplicable a todas las lenguas del mundo ("todas las lenguas tienen verbos", por ejemplo), es decir, un *universal incondicional*. En cambio, lo más frecuente es encontrar fórmulas que restringen el ámbito de aplicación de dicho universal ("si una lengua tiene orden SOV, entonces tiene preposiciones", por ejemplo), es decir, un *universal condicional*.

Todas estas distinciones pueden hacerse en cualquier nivel lingüístico. Este es también un modo posible de clasificarlos. Así se demuestra, por ejemplo, en el repositorio *The Universals Archive* (Plank y Filimonova, 2000). En todo caso, es evidente que el interés que han suscitado los distintos niveles lingüísticos en el estudio de los universales no es el mismo. De hecho, si se accede al recurso citado anteriormente, se pueden ver claramente estas diferencias:

- 418 universales en sintaxis
- 462 universales en derivación-sintaxis
- 416 universales en morfología
- 82 universales en léxico
- 14 universales en pragmática

BY NC ND

En realidad, estos datos no son sorprendentes en absoluto. En primer lugar, los ámbitos con mayor tradición lingüística son los más profusos. La complejidad de los análisis pragmáticos puede ser la causa de la escasez de estudios sobre estos. Por ello, creemos que es interesante indagar más allá de estas primeras impresiones y adentrarnos en el ámbito de los universales pragmáticos para ver su idiosincrasia y baja representatividad.

Los universales pragmáticos son manifestaciones del uso de la lengua que están presentes en todas estas por igual. Levinson (2011) es quien mejor resume qué implica tener universales en pragmática. Según este investigador, se pueden encontrar universales pragmáticos (es decir, elementos presentes en todas las lenguas en el plano pragmático) en deixis (demostrativos, personas, tiempo-lugar, etc.), presuposiciones (inferencias a partir del contexto, el artículo, la negación, etc.), implicaturas, actos de habla (la existencia de todos los tipos en las lenguas del mundo, pero también las distintas formas de manifestarse pueden ser universales) y la organización de la conversación (reparadores conversacionales, pares adyacentes, etc.).

Desde un punto de vista más teórico, la universalidad de los distintos elementos pragmáticos se da por supuesta. En otras palabras, es un axioma que ya aparece desde los primeros postulados de Austin (1962) y Searle (1979). Estos usos del lenguaje se suponen presentes en todas las lenguas, dado que se consideran intrínsecos a la comunicación humana. En cambio, si se buscan invariancias en la forma en que las distintas cuestiones pragmáticas se manifiestan en las distintas lenguas, podemos encontrar mayores dificultades para postular una supuesta universalidad. A pesar de que existen propuestas que tratan de recoger manifestaciones pragmáticas desde un punto de vista universal, estas no han sido aceptadas unánimemente. Para verlo de un modo más claro, este proceso se puede concretar a través de la cortesía, por ejemplo. Hay numerosos estudios que formulaban universales de la cortesía y que actualmente son bastante cuestionados (Brown y Levinson, 1987) del tipo: "las lenguas usan la característica lingüística X para mostrar cortesía". Sin embargo, muchos autores, ya a partir de los años 80 a través de comparaciones entre dos o tres lenguas, como, por ejemplo, la comparación entre el japonés y el inglés en Hill et al. (1986) o el polaco y el inglés en Wierzbicka (1985), han señalado el alto nivel de influencia occidental en tales afirmaciones y han puesto de manifiesto su anglocentrismo y poca universalidad. Actualmente, esta oposición a la supuesta universalidad de Brown y Levinson se cuestiona también desde distintas lenguas centroafricanas que, cada vez, por fortuna, son más estudiadas, como el yoruba, por ejemplo.

No obstante, los planteamientos que señalan dicho sesgo no son suficientes, ya sea en pragmática intercultural o transcultural (Kecskes, 2016, 2017, 2019). En muchos casos, no se lleva a cabo una comparación real entre lenguas, no se trabaja con datos fiables, salvo alguna excepción próxima a un planteamiento universal, aunque no llegue a serlo, como CCSARP (Blum-kulka y Olshtain, 1984), donde se analizan bastantes lenguas (inglés, hebreo, ruso, etc.). Trosborg (2010), por ejemplo, ya señala de un modo preciso, evidente y objetivo las diferencias cualitativas y cuantitativas que hay entre estudios sobre pragmática universal, dependiendo de si se hiciera a través de corpus o sin estos.

Lamentablemente, esta opción tan interesante a través de corpus no ha sido satisfactoriamente desarrollada. Esto puede deberse a ciertas limitaciones metodológicas o de recursos en relación con la aplicación de la pragmática de corpus a lo universal. Por ello, conviene revisar qué ocurre y de ahí el interés de este estudio.

CC BY NC ND

3. CORPUS PRAGMATICS

Tanto la pragmática como la lingüística de corpus son dos disciplinas relativamente modernas. Lo es todavía más la unión de ambos conceptos, *corpus pragmatics*, que se puede entender como la ciencia que describe el uso del lenguaje en contextos reales a través de corpus, según Romero-Trillo (2017). Su fusión, tradicionalmente, se ha supuesto complicada por cuestiones metodológicas, pues lo verdaderamente interesante de un corpus es la información cuantitativa que ofrece, mientras que si algo precisa la pragmática es de un análisis cualitativo (Rühlemann y Clancy, 2018). En otras palabras, se podría decir que, de un modo abstracto, el análisis de la pragmática es horizontal, el análisis de corpus es vertical y, en cambio, el análisis de la pragmática de corpus es diagonal (más complejo, por lo tanto), ya que se deben tener en cuenta, a la vez, factores cuantitativos y cualitativos. Dicho de otro modo, el hecho de ser diagonal implica la presencia de un interés por el cálculo de las frecuencias de aparición de una manifestación concreta y, a su vez, el anhelo por plasmar las particularidades de cada realización (ya que, al trabajar con pragmática, se deben tener en cuenta los distintos condicionantes contextuales, marcándolos con etiquetas, para conocer así mejor la naturaleza de todos los *tokens* y, a su vez, poder compararlos). En todo caso, de todo ello debe destilarse la filosofía que hay en *corpus pragmatics*: hacer pragmática a partir de resultados empíricos (aquí radica el punto de unión con los investigadores de universales lingüísticos superficiales).

Ya desde sus inicios (Romero-Trillo, 2008), esta subdisciplina ha experimentado dificultades, pues es mucho más fructífero para un corpus buscar relaciones explícitas y fijadas, manifestaciones literales. Por ello, es tan interesante el uso de un corpus en léxico, sintaxis, etc. En cambio, una de las características esenciales de la mayoría de los fenómenos pragmáticos es, precisamente, que no suele haber una relación directa y manifiesta entre forma y función (Aijmer y Rühlemann, 2015). Esto se traduce, por supuesto, en la necesidad de, muchas veces, anotar manualmente los corpus.

Puesto que la pragmática de corpus es un campo muy extenso y prolífico en los últimos cinco años (especialmente desde el impulso que tomó a partir de la creación de una revista con título homónimo en inglés en 2017), debe acotarse su análisis en este estudio. Por ello, a continuación se ejemplificará su revisión en el caso de los actos de habla.

El estudio de los actos de habla puede hacerse a partir de tres metodologías distintas: desde el sillón, desde el laboratorio y desde el terreno (Trosborg, 2010). La primera de las opciones es la que goza de un menor crédito, especialmente cuando se trata de estudios empíricos como en el caso de los universales superficiales. No obstante, las teorías de sillón (teorías deductivas formuladas por el investigador) pueden adornarse con ejemplos extraídos de corpus, sin que estos sean la fuente principal de la formulación de la teoría. Por ello, no se puede considerar nunca trabajo de corpus (Jucker, 2009).

La segunda de las opciones es, seguramente, la más fecunda en el estudio de los actos de habla en comparaciones entre lenguas (ya sea universal o no), como señalan Flöck y Geluykens en Romero-Trillo (2015). Se suele desarrollar a través de DCT (*Discourse Completion Tasks*), es decir, cuestionarios donde se debe completar una parte del discurso. Los beneficios más destacables de esta metodología son, por un lado, la oportunidad que se ofrece de unificar al máximo el contexto (un hecho esencial en comparación entre lenguas dispares en pragmática) y, por otro lado, el

CC BY NC ND

bajo requerimiento de medios y tiempo, ya que las producciones y los etiquetados están mucho más acotados y los datos son más manejables.

La tercera de las opciones es la que nos ocupa en esta investigación: los corpus. Esta metodología es la más interesante, pues permite un acercamiento a los actos de habla de un modo mucho más natural y real. Además, ofrece la posibilidad de tener en cuenta las distintas frecuencias de uso, que puede ser un factor determinante a la hora de establecer una relación de las estructuras más frecuentes para manifestar un acto de habla directivo, por ejemplo (Aijmer, 2020).

No obstante, se debe destacar que no se han hallado estudios a través de corpus para desentrañar universales en los distintos actos de habla. El principal motivo es la poca disponibilidad de determinadas lenguas. Por ello, principalmente, se encontrarán corpus de actos de habla en inglés (además de alguna excepción en alemán, francés o español). Esta carencia ha supuesto que la mayoría de los estudios para comparar actos de habla sea en el seno de la lengua inglesa: se comparan corpus en inglés de distintas áreas geográficas (SPICE Ireland, por ejemplo, con la versión británica).

Otra de las dificultades que aflora cuando se valora el uso de corpus para sistematizar actos de habla es la inexacta correlación entre estructura fija y acto de habla. Una misma intervención puede considerarse un acto de habla directivo o expresivo, dependiendo del contexto. Este impedimento, en el caso de los actos de habla obligará a etiquetar manualmente los corpus (o, al menos, a revisarlos). Sin embargo, si el estudio es de corte universal, además, habrá un problema añadido. Esta falta de relación entre forma y acto de habla dificulta mucho el trabajo de los investigadores, quienes deben trabajar con distintas lenguas a la vez (y que no siempre conocen bien). Este planteamiento obliga a contar con un experto en cada lengua que pueda certificar que esa producción de ese corpus específico es un acto de habla directivo indirecto, por ejemplo. En otros ámbitos lingüísticos, bastará con que el responsable de utilizar el corpus cuente con la anotación específica del significado de las distintas manifestaciones que se podrá encontrar y puede trabajar sobre estas. En los actos de habla, a pesar de que en algún caso se podría sistematizar alguna manifestación concreta (por ejemplo, los verbos *jurar* o *prometer*) para relacionarla directamente con un acto de habla, no tendría por qué ser siempre así. Aun en el caso de que se pudiera hacer, hay muchas otras manifestaciones indirectas que no aflorarían y, en definitiva, no se tendría el 100% de los actos de habla recogidos, simplemente se dispondría de una selección azarosa (Garcia McAllister, 2015).

4. PROPUESTAS DE ETIQUETADO DE LOS ACTOS DE HABLA EN CORPUS

Una vez revisadas las dificultades existentes en obtener corpus pragmáticos más allá del inglés y alguna otra excepción, hecho que permitiría extraer universales, vamos a revisar qué propuestas (a partir de una selección de 4 recursos) se hacen desde la lengua inglesa y valorar cómo podría adaptarse esto a un contexto universal, que permita la comparación libre entre cualquier grupo de lenguas. Para ello, se analizará, primero, *cómo* se hace y, luego, *qué* se hace.

En primer lugar, se deben revisar las distintas propuestas de etiquetado de actos de habla en corpus que han surgido en el ámbito del inglés para ver su utilidad en una investigación entre múltiples lenguas. El término etiqueta, en este caso, hace referencia a las distintas formas que recogen los supuestos actos de habla. Estas pueden articularse con un mayor o menor detalle y, además, responden a una

BY NC ND

evolución de la propuesta clásica (Austin, 1962; Searle, 1979), constatada en las últimas décadas como insuficiente para dar cuenta de la realidad en relación con las distintas realizaciones de los actos de habla en las distintas lenguas.

Para ello, entre las distintas opciones, se han seleccionado dos propuestas interesantes que, según sus ideólogos, tienen la voluntad de crecer hasta ser útiles en más lenguas y, además, tratan de ofrecer un gran detalle de las distintas manifestaciones. En este caso se revisarán las directrices del *Dialogue Annotation and Research Tool* (en adelante DART) y del *Dialogue Act Markup in Several Layers* (en adelante DAMSL). En otras palabras, en este apartado vamos a revisar *cómo* se está desarrollando la investigación de actos de habla en corpus.

4.1. DART

Este recurso propuesto por Weisser (2014, 2016, 2019b, 2019a), actualizado en 2019, cuenta con 162 etiquetas distintas para clasificar los actos de habla. Esto se debe a una mayor precisión a la hora de recoger las distintas manifestaciones, en algo más cercano a lo conocido como actos de diálogo. De hecho, actualmente, Weisser ha decidido ir más allá y ofrecer de manera libre un programa que etiqueta de manera semiautomática actos de habla en inglés. No obstante, requiere una supervisión exhaustiva.

Este alto nivel de detalle apreciado (por ejemplo, hay 22 tipos diferentes de "expresar") complica el análisis universal netamente. En primer lugar, cuantas más categorías se incluyan, mucho más tiempo (y recursos) debe dedicarse a que aparezcan todos los elementos. Esto, en el caso de lenguas con una tradición filológica escasa o, incluso, nula, puede hacer la investigación irrealizable. Además, este nivel de detalle también es conflictivo a la hora de trazar equivalencias claras. A pesar de que pueda resultar sorprendente, a mayor nivel de detalle, mayor posibilidad de etiquetar erróneamente. Esto supondría que no se podría crear una equivalencia y que la comparación sería falsa. El error en este tipo de etiquetados es altamente probable si se tiene en cuenta que se está trabajando con lenguas que tienen formas de expresar algo muy distintas a lo conocido (o directamente no lo tienen). Basta con elegir un ejemplo concreto de los 22 tipos de expresiones mencionadas anteriormente. Destaquemos las categorías *expressDisappointment, expressDisgust* y *expressDislike*. Muchas veces, un único hablante podría dudar en cuál de dichas categorías colocar una determinada manifestación lingüística. También es posible que dos hablantes de una misma lengua tengan percepciones distintas de un mismo hecho entre estas tres categorías. Esta disparidad se puede manifestar de una manera mucho más acentuada si estamos hablando de lenguas, no solo con estructuras totalmente opuestas, sino de contextos totalmente ajenos, donde, quizá, para una, un hecho encaje mejor en una u otra categoría. Si se quiere utilizar un mismo acto de habla en varias lenguas para ver si ofrece el mismo recorrido sintáctico-semántico, con esta serie de dificultades, no se podría hacer.

4.2. DAMSL

Como en el caso anterior, la propuesta de Allen y Core (1997) va más allá de los simples actos de habla y ofrece una clasificación más detallada. En este caso no hay una clasificación tan clara y ordenada como en DART, pero la clasificación estaría cerca de las 14 categorías (escalables hasta 220). Por ello, aparentemente, existirían menos posibilidades de disparidad. Además, la organización de detalle que se hace es jerárquica (en DART las 162 categorías están al mismo nivel), así que es posible

CC BY NC ND

reducir o aumentar estas 14 etiquetas, dependiendo de las necesidades del investigador o de la realidad de la lengua que se esté investigando.

Otro aspecto interesante que se ofrece en dicha clasificación es la forma de sistematizar las distintas dificultades que se puede encontrar un investigador cuando está trasladando un texto oral a un corpus escrito. Esta serie de anotaciones deben estar reguladas y ser lo más claras y abiertas posible, para que así su aplicación pueda resultar efectiva en cualesquiera lenguas. Por lo tanto, esta propuesta es interesante en cuanto a la posible utilidad como punto de partida para alcanzar un consenso general en distintas cuestiones (por ejemplo, cómo reflejar superposiciones) que no han sido fijadas en lenguas poco estudiadas. Desgraciadamente, DAMSL no ha dado el paso de DART y no ofrece un programa de etiquetado semiautomático.

5. CORPUS DE ACTOS DE HABLA

Una vez revisadas dos propuestas sobre *cómo* etiquetar actos de habla en corpus, es interesante ver *qué* corpus existen actualmente en este mismo sentido. Para ello, hemos procurado ser congruentes y, del mismo modo, presentaremos los corpus que emanan de las propuestas presentadas anteriormente. Por lo tanto, en primer lugar, se revisará el *Speech Act Annotated Dialogues* (en adelante SPAADIA) y, en segundo lugar, se recogerá el *Switchboard Dialog Act Corpus* (en adelante SWDA).

5.1. SPAADIA

El corpus de Leech y Weisser (2014) en la versión *Trainline Corpus* consta de 35 diálogos telefónicos en inglés entre una operadora y clientes de un servicio de ferrocarriles. Las llamadas se producen para pedir información sobre trayectos o para adquirir billetes. En este caso se ofrecen 41 etiquetas distintas para clasificar los actos de habla. Retomando el ejemplo citado anteriormente, en este caso, en la categoría expresiva se pasa de 22 opciones a 4. Las tres modalidades mencionadas, ahora, convergen en una sola denominada *expressOpinion*. Toda la información extra, en este caso, no condiciona la categoría de acto de habla, sino que aparece reflejada en categorías independientes (ya que también se mencionan *polarity*, *topic* y *mode*).

Esta propuesta es interesante, ya que no se queda en el mero hecho de expresar, ofrece más información, pero no plantea problemas de interpretación. En las notas del corpus se recoge de manera pormenorizada las características de aparición que se verán en el corpus. En el caso de *expressOpinion*, por ejemplo, habla de la frecuencia de aparición en otros corpus menos marcados, la frecuencia de aparición en este, los motivos por los que ocurre esto y, además, menciona los posibles conflictos de etiquetado que pueden producirse en determinadas situaciones entre declaraciones y expresiones, por ejemplo.

Si hay que mencionar algún aspecto que no sea positivo para la investigación de universales es que no hay prácticamente información sintáctico-semántica de la estructura de los distintos actos de habla. Esta información, infrecuente en los corpus específicamente de pragmática, permitiría un mayor conocimiento del funcionamiento de la pragmática y sus distintos recursos. Si bien es cierto que esto podría solucionarse integrando un corpus como este a corpus anotados universalmente en otros niveles de la lengua, todavía estamos lejos de alcanzar una

BY NC ND

especie de corpus como *Universal Dependencies* (Nivre et al. 2016) que sea aplicable a todos los niveles y todas las lenguas.

5.2. SWDA

SWDA (Jurafsky et al. 1997) es un corpus que contiene 1155 diálogos telefónicos. En este caso, se dispone de 42–43 etiquetas para los actos de habla (1 o 2 más que en SPAADIA), dependiendo de la versión. Por lo tanto, no es una clasificación tan exhaustiva como DART (que era contraproducente por los riesgos que conlleva el alto nivel de detalle) ni escueta como DAMSL. Hay alguna referencia sintáctico-semántica englobada dentro de las etiquetas que ofrecen. No obstante, podría haber más detalles para una mayor utilidad. Para que el proceso de etiquetado no se viera afectado por los problemas descritos en DART, seguramente, sería interesante mantener una perspectiva bastante general en los actos de habla y, posteriormente, de manera aislada, añadir la caracterización desde el plano de la sintaxis o la semántica. De este modo, no solo se superan los problemas mencionados anteriormente, sino que, además, permite la sustitución de estas notas sintácticas en el futuro por un análisis integral. El reto verdaderamente interesante sería alcanzar una especie de *Universal Dependencies* que contenga la parte pragmática.

Este corpus es también muy interesante por la gran cantidad de datos sociolingüísticos que se incluyen. La información puede usarse para certificar muchos otros aspectos de la sociolingüística que aquí no se revisan pero que, sin duda, son relevantes como, por ejemplo, la edad o el sexo de los emisores.

Finalmente, debe destacarse que, del mismo modo que ha ocurrido con el resto de las propuestas, a pesar del gran interés que suscitan los grupos de etiquetas que han fijado, estos van perdiendo la potencial universalidad si uno va alejándose del inglés. Por supuesto, las 43 etiquetas son aplicables al español, por ejemplo. Incluso es posible que se pueda afirmar que todas las lenguas del mundo puedan clasificarse bajo estas mismas reglas. No obstante, es también probable que alguna de las lenguas del mundo demuestre que alguna de esas categorías no es aplicable en su caso. Por ejemplo, quizá hay alguna lengua que no permite la aparición de *self-talk*, o que no se pueden dar respuestas del tipo *dispreferred answers*. Es bien conocido que lenguas con culturas bastante alejadas a la nuestra obligan a responder afirmativamente incluso si lo que se afirma no se terminará realizando. Si este tipo de situaciones ocurren en culturas como Japón, es también posible que esto sea más extremo todavía en alguna isla Papunesia, por ejemplo. En definitiva, creemos que, antes de aplicar la clasificación de actos de habla planteada aquí, se debe hacer una comprobación de que es válido universalmente y no se incurre en algo que ya denunció Wierzbicka (2003): la falta de universalidad en las herramientas de análisis y el anglocentrismo de las teorías que engloban dichos análisis (algo mucho más evidente en sintaxis o morfología, por ejemplo).

6. PROPUESTA DE PROTOCOLO PARA LA CREACIÓN DE CORPUS DE ACTOS DE HABLA UNIVERSALES

Una vez revisados cuatro recursos útiles para etiquetar actos de habla en corpus, se puede afirmar que las distintas propuestas no son satisfactorias a la hora de extender el estudio a cualquier lengua.

Creemos que, inicialmente, se deben dar unos pasos previos para asegurar el correcto funcionamiento de una clasificación como la planteada:

BY NC ND

1. Sería positivo un inicio de la investigación mediante DCT, que permitiría, si se hiciera un análisis verdaderamente universal en pragmática, obtener datos fiables, concretos y significativos sobre las distintas realizaciones de los actos de habla en lenguas todavía no analizadas. Para poder unificar los criterios, sería interesante crear situaciones para completar unificadas. En este caso, se puede seguir la propuesta de universales pragmáticos de Kasper (1992) y analizar los actos de habla que propone y se tendrían en cuenta las distintas variables sociales que ya tienen en cuenta en el CCSARP (Blum-kulka y Olshtain, 1984). Lo más razonable también es seguir esta propuesta en cuanto a los informantes, alumnos de filología que, sin llegar a ser expertos, presentan un conocimiento consciente de la lengua. En cuanto al set de lenguas, es necesario tener en cuenta equilibrios de *genus*, familia, área, y tipología lingüística. Por ello, y teniendo en cuenta las dificultades para encontrar datos, creemos que un buen inicio se podría hacer con el siguiente set lingüístico, seleccionado con dicho equilibrio: alemán, árabe, vasco, estonio, español, turco, indonesio, coreano y ruso.
2. Lo dicho permitiría hacer un análisis pormenorizado de las etiquetas que se querrían utilizar en el corpus (que recordemos que no deberían ser demasiadas), para poder validar su efectividad y universalidad.
3. Posteriormente, una vez se dispone de estos datos, ya se podrían confeccionar corpus reducidos, con un contexto muy controlado, para poder analizar cómo se comportan estos actos de habla en un contexto natural, lejos del laboratorio. Pero es importante recordar que se ha demostrado que se debe disponer previamente de este informe (a partir de un DCT) que ayude al investigador de universales a conocer qué posibles manifestaciones tiene aquello que investigará en los corpus. En otras palabras, los DCT permiten la comprobación de la adecuación y legitimidad del uso de las distintas etiquetas en lenguas en las que todavía no existe un análisis pormenorizado de dichos actos de habla.
4. Estos corpus reducidos deben entenderse como un paso intermedio. Los beneficios de disponer de un corpus en este momento de la investigación, a pesar de que sea bastante controlado y no demasiado extenso, son múltiples. Estas muestras de textos, al no ser muy numerosas y estar hechas a partir de los DCT que se proponen como primer paso, permitirían una unificación de los contextos y de los elementos que aparecerían. Un investigador sobre universales lingüísticos agradecería mucho este hecho, ya que facilita el afloramiento de posibles tendencias compartidas por las lenguas que se revisan.
5. La presencia de distintos corpus pequeños de distintas lenguas permitiría, posteriormente, la recopilación de corpus más grandes y que fueran significativos en cuanto a la frecuencia de uso (que también aparece en un corpus reducido pero su valor es más relativo, puesto que está ligeramente dirigido y controlado). Estos corpus más grandes se obtendrían sin ningún tipo de condicionante ni premisa, en un contexto libre y sin dirección. En este caso el valor fundamental de este corpus sería que ya se podrían valorar cuestiones relacionadas con la frecuencia y con otros aspectos que son tenidos en cuenta en un corpus en otros ámbitos lingüísticos que no presentan tantas dificultades y con lenguas que ya conocemos hoy en día.
6. Finalmente, el paso ideal sería, una vez comprobada la eficacia del etiquetado y una vez se dispone de un corpus significativo, integrar el análisis de actos

CC BY NC ND

de habla a otros corpus para, así, tener una fuente de datos mucho mayor (en relación con otros aspectos lingüísticos), que permita hacer correlaciones entre distintos ámbitos lingüísticos y que permita formular universales condicionales que ayuden a comprender mejor el funcionamiento de las lenguas del mundo. Por lo tanto, el objetivo final conlleva la conjunción de los datos pragmáticos obtenidos con otros para, así, tener un conocimiento total de lenguas que, en muchos casos, hoy en día todavía desconocemos, desgraciadamente.

7. CONCLUSIONES

En definitiva, hemos podido comprobar cómo la novedad de la unión de los tres ámbitos tratados en esta investigación (universales, pragmática y corpus) supone un estado incipiente de su estudio. A pesar de ello, creemos que su estudio sería altamente beneficioso para el conocimiento general del funcionamiento de las lenguas del mundo y su comparación.

El principal beneficio de la propuesta, fundamentalmente, radica en la combinación de las distintas metodologías propias de cada ámbito. En primer lugar, se han podido destacar que hay una buena complementación de aspectos diferenciados. La relación de la pragmática (cualitativa) y la lingüística de corpus (cuantitativa) origina estudios muy ricos y completos (*corpus pragmatics*). Asimismo, los corpus, si contienen información adecuada y relevante, ofrecen todos los datos contextuales necesarios para trabajar de forma satisfactoria en pragmática. La principal dificultad latente en este binomio es lo complejo que resulta, en algunos casos, sistematizar y cuantificar hechos pragmáticos en un corpus (especialmente si se desea hacer de un modo automático).

En segundo lugar, también es interesante destacar que la metodología de corpus casa a la perfección con el estudio de universales en el lenguaje desde una perspectiva superficial. En ambos casos el objetivo es tratar de acercarse al conocimiento del hecho lingüístico (en este caso pragmático) a partir de datos empíricos y gracias a cifras cuantitativas relevantes. Por lo tanto, la estadística será fundamental en ambos casos.

En el último de los vértices del triángulo que proponemos, la relación entre universales y pragmática, podemos mencionar que hay mucho recorrido pendiente todavía. Como hemos comentado, hay una baja producción de universales en este ámbito por las dificultades que entraña analizar aspectos pragmáticos en lenguas con poca información o que, directamente, se desconocen. La propuesta más solida, como se ha recogido, es el CCSARP, con casi una decena de lenguas relativamente distintas, pero solo con dos actos de habla muy determinados. Por ello, el corpus lingüístico sería una solución óptima para dicho escollo.

Por consiguiente, creemos que, si realmente se siguieran los pasos que aquí proponemos y se obtuvieran corpus de actos de habla en distintas lenguas (más y menos conocidas), se podrían extraer universales lingüísticos interesantes y significativos. Estos, a su vez, permitirían adaptar muchas de las herramientas existentes en NLP (*Natural Language Processing*) a lenguas de las que no se dispone una gramática completa y, por lo tanto, se podrían hacer extensivos todos estos avances a muchas más lenguas con bajos recursos (o todas). Por otro lado, otra de las líneas de investigación futuras, fruto de la formulación de universales pragmáticos a partir de corpus, sería la mayor precisión a la hora de enseñar segundas lenguas. Si se conocieran las estructuras pragmáticas (en este caso actos de habla) y sus

CC BY NC ND

realizaciones concretas que estuvieran presentes en todas las lenguas del mundo (o en su amplia mayoría) se podría aplicar dicha información a la enseñanza de L2 para, así, facilitar el proceso de enseñanza y aprendizaje.

REFERENCIAS

Aijmer, J. 2020. Contrastive pragmatics and corpora. *Contrastive Pragmatics* 1: 1–30.

Aijmer, J. y Rühlemann, C. 2015. *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.

Allen, J. y Core, M. 1997. Draft of DMSL: Dialogue Act Markup in Several Layers. <http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html> (acceso: 11/12/2020)

Austin, J. L. 1962. *How to do things with words*. Oxford: Oxford University Press.

Bickel, B. 2010. Absolute and statistical universals. En P. C. Hogan ed. *The Cambridge encyclopedia of the language sciences*. Cambridge: Cambridge University Press, 77–79.

Blum-kulka, S. y Olshtain, E. 1984. Requests and apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP). *Applied Linguistics* 5(3): 196–213.

Boeckx, C., J. D. Fodor y Glertman, L. 2009. Round table 'Language universals: Yesterday, today, and tomorrow'. En M. Piattelli-Palmarini, J. Uriagereka y P. Salaburu eds. *Of minds and language: A dialogue with Noam Chomsky in the Basque Country*. Oxford: Oxford University Press, 195–220.

Brown, P. y Levinson, S. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, Mass.: The MIT Press.

Dryer, M. S. 1992. The Greenbergian word order correlations. *Language* 68(1): 81–138.

Dryer, M. S. 1998. Why statistical universals are better than absolute universals. En *Papers from the 33rd Annual Meeting of the Chicago Linguistics Society*, 123–146.

Garcia McAllister, P. 2015. Speech acts: A synchronic perspective. En K. Aijmer y C. Rühlemann eds. *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press, 29–51.

Greenberg, Joseph H. 1963. *Universals of language*. Cambridge, Mass.: The MIT Press.

Hill, B., S. Ide, S. Ikuta, A. Kawasaki y Ogino, T. 1986. Universals of linguistic politeness. Quantitative evidence from Japanese and American English. *Journal of Pragmatics* 10(3): 347–371.

Jucker, B. 2009. Speech act research between armchair, field and laboratory. *Journal of Pragmatics* 41(8): 1611–1635.

Jurafsky, D., E. Shriberg y Biasca, D. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Draft 13. <https://web.stanford.edu/~jurafsky/ws97/manual.august1.html> (acceso: 11/12/2020)

Kasper, G. 1992. Pragmatic transfer. *Second Language Research* 8(3): 203–231.

Kecskes, I. 2016. Cross-cultural and intercultural pragmatics. En Y. Huang ed. *The Oxford handbook of pragmatics*. Oxford: Oxford University Press, 1–19.

Kecskes, I. 2017. *Intercultural pragmatics*. Oxford: Oxford University Press.

Kecskes, I. 2019. Editorial to the special issue 'Current developments in intercultural Pragmatics.' *Journal of Pragmatics* 151: 74–75.

Leech, G. y Weisser, M. 2014. The SPAADIA annotation scheme. <http://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf> (acceso: 08/12/2020)

Levinson, S. C. 2011. Pragmatics, Universals. <https://pure.mpg.de/rest/items/item_67085_12/component/file_1451158/content> (acceso: 08/12/2020)

Moravcsik, E. A. 2010. Explaining language universals. En J. J. Song ed. *The Oxford handbook of linguistic typology*. Oxford: Oxford University Press, 69–89.

BY NC ND

Nivre, J., M. Catherine De Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty y Zeman, D. 2016. Universal dependencies v1: A multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 1659–1666.

Plank, F. y Filimonova, E. 2000. The universals archive: A brief introduction for prospective users. *STUF – Language Typology and Universals* 53(1): 109–23.

Romero-Trillo, J. 2008. *Pragmatics and corpus linguistics*. Berlin: De Gruyter Mouton.

Romero-Trillo, J. 2015. *Yearbook of corpus linguistics and pragmatics 2015*. Heidelberg: Springer.

Romero-Trillo, J. 2017. Corpus pragmatics. *Corpus Pragmatics* 1(1): 1–2.

Rühlemann, C. y Clancy, B. 2018. Corpus linguistics and pragmatics. En N. Norrick y C. Ilie eds. *Pragmatics and its interfaces*. Amsterdam: John Benjamins, 241–266.

Scalise, S., E. Magni y Bisetto, A. 2009. *Universals of language today*. Heilderberg: Springer.

Searle, J. 1979. A classification of illocutionary acts. *Language in Society* 5:1–24.

Trosborg, A. 2010. *Pragmatics across languages and cultures*. Berlin: Walter de Gruyter.

Weisser, M. 2014. Speech act annotation. En K. Aijmer y C. Rühlemann eds. *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press, 84–114.

Weisser, M. 2016. DART – The Dialogue Annotation and Research Tool. *Corpus Linguistics and Linguistic Theory* 12(2): 355–388.

Weisser, M. 2019a. Manual for the Dialogue Annotation & Research Tool (DART). <http://martinweisser.org/publications/DART_manual.pdf> (acceso: 09/12/2020)

Weisser, M. 2019b. The DART Taxonomy v.3. <http://martinweisser.org/publications/DART_taxonomy_v3.pdf> (acceso: 09/12/2020)

Wierzbicka, A. 1985. Different cultures, different languages, different speech acts. Polish vs. English. *Journal of Pragmatics* 9: 145–178.

Wierzbicka, A. 2003. *Cross-cultural pragmatics*. Berlin: Mouton de Gruyter.

CC BY NC ND