Estructura de consultas para la selección automática de formas gramaticales analíticas del tiempo futuro en lenguas eslavas

Query Structuring for Automatic Selection of Analytical Future Tense in Slavic Languages

Serhii Fokin, *Taras Shevchenko National University of Kyiv*
sergiyborysovych@ukr.net

Received: July, 25 2020.
Accepted: October, 30 2020.

RESUMEN
Los corpus de textos actuales permiten a los investigadores llevar a cabo un amplio rango de análisis, así como automatizar la selección del material empírico gracias a una anotación detallada del texto. Ya que las anotaciones caracterizan mayoritariamente tales unidades como palabras, algunas categorías gramaticales expresadas por medio de formas analíticas no pueden ser anotadas aprovechando este enfoque, razón por la cual su selección se ve dificultada o produce resultados erróneos. Con el fin de resolver dicho problema proponemos utilizar consultas específicas mediante el lenguaje técnico *CQL* (*Context Query Language*) o similares, que puede compaginar todos los parámetros y condiciones que el investigador necesite: tanto lexemas con sus características como combinaciones enteras de ellos. Para ilustrar el uso, los resultados y potencialidades de dicha herramienta, describimos las particularidades destinadas a la selección de formas analíticas del tiempo futuro en 6 lenguas eslavas: bielorruso, checo, polaco, eslovaco, ruso y ucraniano. Las consultas *CQL* han de ser adaptadas para cada corpus modificando los nombres de las etiquetas, quedando invariable su estructura. Algunos corpus están provistos de una interfaz particular para la selección de dichas categorías. No obstante, el uso de *CQL r*esulta ser una solución más universal puesto que incluye la posibilidad de modificar parcialmente las demandas más específicas del usuario, p. ej., selección de las formas futuras en una voz y aspecto determinados.

Palabras clave: anotación del corpus, tiempo futuro, formas gramaticales analíticas, lengua de consulta, muestreo automatizado.

ABSTRACT
Modern corpora tools allow language researchers to perform a wide range and type of analysis as well as automatize the selection of empiric material thanks to a detailed annotation of the texts. Since annotation is mainly word-oriented, some grammar categories which are expressed by means of analytical forms cannot be tagged as such in a word-oriented approach, which is why in many corpora their selection is hampered or produces mistaken results. To overcome this problem, we propose to use specific queries in Context Query Language (CQL), which can combine as many parameters and conditions as the researcher might need to indicate: lexemes with their characteristics or sequence of them. To illustrate the usage, possible results, and potentiality of this tool, we make an overview of the queries aimed at selecting analytical forms of future in 6 Slavic languages: Belorussian, Check, Polish, Slovak, Russian, and Ukrainian. CQL-queries should be adapted for each corpus by modifying the tagname, though the structure of the query remains identical. Some annotated corpora do not accept CQL-query, though they provide a specific user interface for selecting grammar categories. Nevertheless, CQL appears to be a more universal solution (though, more complicated to use) because it includes the possibility of modifying partially the parameters of the search, for example, selecting future forms in a certain voice or aspect.

Keywords: corpus annotation, future tense, analytical grammatical forms, query language, automatic sampling.

Introducción

Los avances de la lingüística del corpus se dejan notar en la dinámica evolución de las herramientas de búsqueda cada vez más sofisticadas y precisas. Los usuarios pueden llevar a cabo todo tipo de búsqueda: desde las consultas más triviales (palabras sueltas, secuencias de palabras, frases), pasando por lexemas en su paradigma completo, formas gramaticales particulares, hasta las consultas más específicas basadas en expresiones regulares complejas con caracteres comodines y operadores relacionales. Además de consultas prácticas, uno de los aprovechamientos más eficientes de los corpus actuales consiste en la posibilidad de efectuar el muestreo con fines de investigación.

Cualquier fenómeno lingüístico formalmente segmentable, gracias a las emergentes herramientas del corpus, se vuelve un cebo seductor irresistible para el investigador. Mientras el mayor atractivo de los primeros corpus fue la posibilidad de manejar volúmenes, inabarcables en muestreo convencional hecho a mano, los corpus actuales son particularmente valiosos por su variedad de parámetros de búsqueda. Al ser el nivel gramatical altamente formalizable, la anotación automática de las características gramaticales, que suelen comprender la clase de palabra (dicho en otras palabras, parte de la oración) y sus características morfológicas, se vuelven muy fáciles, siendo la anotación sintáctica más trabajosa y más infrecuente. El español y las demás lenguas románicas, al igual que las eslavas, poseen un amplísimo sistema de flexiones, siendo las lenguas románicas más ricas en cuanto a flexiones verbales relativas a modos y tiempos, y las lenguas eslavas, en flexiones de nombres sustantivos debido a la categoría de casos, es decir, se caracterizan por un alto grado de sintetismo. Es precisamente en estos grupos de lenguas donde se da la necesidad de una detallada anotación morfológica. Las lenguas más analíticas, como el inglés, cuyos nombres sustantivos regulares disponen únicamente de una variación gramatical posible, que es el número singular/plural, marcado, en su mayoría, por "s/es" al final de la palabra, poseen un paradigma verbal reducido, en el caso del inglés, a cinco formas, que no presentan tanta dificultad de anotación gramatical. Por ejemplo, la consulta que sigue abarca todo el paradigma del verbo irregular 'take' en inglés:

take|takes|taking|took|token

Es precisamente de esta forma como funciona este tipo de consultas en *British National Corpus* (*BNC*). Dicho procedimiento, sin embargo, no es, ni mucho menos, válido para las lenguas flexivas. Mientras en la lengua inglesa moderna el verbo posee 5 formas, la RAE recoge 53 formas verbales en la descripción de los modelos de conjugación del verbo español (Modelos de conjugación), lógicamente, su enumeración exhaustiva en una consulta semejante sería engorrosa y quedaría absurda. Al mismo tiempo un trivial uso de asterisco, al parecer, efectuaría la extracción de todas las formas del verbo español (al tratarse de formas regulares, por supuesto), p. ej., el paradigma del verbo cantar es expresable mediante la expresión que sigue:

cant*

La aparente facilidad del uso de caracteres comodines se enfrenta con una serie de difi-

cultades: por un lado, los nombres sustantivos como "canto", "cante", "cantante", "cantor", "cantina" cuadran perfectamente en esta expresión, lo cual producirá resultados de búsqueda sobrantes; por otro lado, el principio utilizado en dicha expresión sigue sin cubrir los verbos irregulares, cuyas flexiones no se limitan al nivel de la terminación. Los corpus con anotación morfológica están pensados precisamente para subsanar esta carencia, al presentar otra serie de indudables ventajas: posibilidad de búsqueda por clases léxico-gramaticales, formas morfológicas específicas o combinaciones posibles de las características indicadas. Es decir, constituyen una herramienta de verdad "inteligente", capaz de manejar los conceptos gramaticales, que resulta de gran valía.

En lenguas eslavas la variedad de las flexiones verbales es también bastante rica: aunque el modo subjuntivo/condicional no posee desinencias específicas, en la mayoría de los tiempos presentes los verbos disponen de una desinencia particular para cada persona y número, con menor variedad en los tiempos pasados en bielorruso, ruso y ucraniano (que presenta formas idénticas para todas las personas de un mismo número), y en checo y eslovaco, donde son de naturaleza analítica; el futuro perfectivo en todas las lenguas indicadas usa las mismas desinencias que el presente.

Dados los volúmenes astronómicos de los corpus actuales, la anotación (i.e., indicación con marcas explícitas de la información gramatical de todos los componentes del texto, mayormente palabras) se lleva a cabo de forma automatizada mediante programas particulares, que necesitan pautas claramente comprensibles e inequívocas. Aun así, la anotación gramatical de algunas de las formas gramaticales en las lenguas mencionadas, al igual que de aquellas de otros grupos lingüísticos, se ve dificultada por dos razones: 1) una considerable homonimia entre las formas; 2) carácter multicomponencial de las formas gramaticales analíticas.

La homonimia se da con frecuencia en los tiempos pasados, p. ej. en ucraniano, ruso y bielorruso la forma del pasado, sea del aspecto perfectivo o imperfectivo, "(з)ходив/(с)ходил/(с)хадзіў vale para las tres personas del mismo género, en este caso, el masculino; en español, italiano y portugués la forma "lea/legga/leia" del verbo leer tanto vale para el Presente de Subjuntivo de la primera persona singular, tercera persona plural como para el Modo Imperativo de tercera persona singular. En francés, por otro lado, la forma 'cherche' (i.e., "busca"), puede corresponder a la primera o tercera persona singular del inficatif présent, de *subjonctif présent, mode imperatif* de segunda persona singular.

El carácter sintético es propio de varias categorías gramaticales de lenguas eslavas:

- el equivalente del modo condicional/subjuntivo, que necesita de una partícula ('б/би/бы/by'), la que, además, puede distanciarse del verbo y en polaco incluso pegarse a otras clases léxico-gramaticales como preposiciones ('*żebyś, żebym'*), que incluso puede cambiar su forma en función de la persona;
- formas analíticas de grados de comparación de adjetivos y adverbios (que coexisten con las sintéticas): mientras en ruso predomina la formación analítica de grados de comparación de adjetivos, el checo, eslovaco, polaco, ucraniano se caracterizan por un uso amplísimo o, quizás, predominante de formas sintéticas;
- tiempos pasados en checo y eslovaco, que se forman con el verbo 'být/byt' combinado con el participio pasado;
- tiempo futuro imperfectivo, que se forma con el equivalente del verbo "ser" combinado

con el infinitivo (o, alternativamente, de participio pasado de participio pasado activo en polaco).

Cada una de las categorías indicadas merece una particular atención de la lingüística del corpus. Nuestro propósito, por tanto, es el de realizar un recorrido por los corpus gramaticalmente anotados de las lenguas eslavas: bielorruso, checo, eslovaco, polaco, ruso y ucraniano explorando sus posibilidades técnicas en cuanto a la selección de formas analíticas del tiempo futuro, proponiendo soluciones a medida de cada uno de los corpus y generalizando las particularidades básicas de formulación de este tipo de consultas. Los corpus aprovechados para la presente exploración han sido el Corpus Nacional de la Lengua Checa (*Český národní korpus),* Corpus Nacional de la Lengua Eslovaca (*Slovenský národný korpus*), Corpus Nacional de la Lengua Polaca (*Narodowy Korpus Języka Polskiego*), Corpus Ruso Anotado de Helsinki (*HANCO),* Corpus General Regionalmente Anotado del Ucraniano (*Грак, Генеральний регіонально анотований корпус української мови*), Corpus Nacional de la Lengua Rusa (*Национальный корпус русского языка*), Corpus de la Lengua Ucraniana (*Корпус української мови)*, Corpus de la Lengua Bielorrusa (*Беларускі N-корпус*).

Problemas de la anotación gramatical

Las lenguas eslavas tratadas en este estudio poseen varias formas del tiempo futuro: las perfectivas e imperfectivas. El futuro imperfectivo se forma en estas lenguas mediante estructuras analíticas, disponiendo el ucraniano, además, de un futuro imperfectivo sintético (“ходитиму”, “ходитимеш”, “ходитиме”...) que lo destaca frente a la mayoría de otras lenguas eslavas que carecen de esta posibilidad. Es decir, las asimetrías entre las funciones y el uso del tiempo futuro han de ser notables no solo al contrastar lenguas de familias diferentes como es el caso del español y ruso cuyas formas futuras se contrastan en las obras de I.G. Miloslavskiy y V.S. Vinogradov (Милославский, 1987), H.H. Verba y R.Guzmán Tirado (Гусман Тирадо, 2005), R. Guzmán Tirado y E. Quero-Gervilla (Guzmán Tirado, 2007), O.Yu.Chuikova (Чуйкова, 2018).

Puesto que las formas del futuro sintético en bielorruso, checo, eslovaco, polaco, ruso y ucraniano vienen representadas por dos palabras, su anotación gramatical debería comprender grupos de palabras. Las formas que siguen ilustran en dichas lenguas eslavas las formas del futuro analítico de la primera persona singular:

Bielorruso: *буду рабіць*
Checo: *budu dělat*
Eslovaco: *budem robiť*
Polaco: *będę robił, będę robić*
Ruso: *буду делать*
Ucraniano: *буду робити*

No obstante, su anotación automática a la hora de compilar un corpus, al igual que su búsqueda, se ven obstaculizadas por el hecho de que los componentes de la forma analítica puedan distanciarse el uno del otro o invertirse bien por necesidad de énfasis, bien por razones estilísticas. En el ejemplo que sigue sacado del poema de Lesya Ukrainka “El canto del bosque” (“Лісова Пісня”) el infinitivo viene antepuesto al auxiliar “ser”; los dos

componentes están separados por dos palabras:

Будуть приходити люди,
вбогі й багаті, веселі й сумні,
радощі й тугу нестимуть мені,
їм промовляти душа моя буде
(Леся Українка)

Como se ve en el ejemplo, la forma del futuro, además de inversa, es discontinua. Es decir, la selección de ejemplos requiere una atención particular por parte del usuario, ya que la localización del infinitivo requiere una lectura consciente del fragmento respectivo. Como es de esperar, la anotación gramatical basada en palabras sueltas no rendiría resultados satisfactorios; aunque los programas de anotación automática actuales utilizan redes neuronales, que toman en consideración el contexto de la forma gramatical, su grado de precisión, muy alto, aunque no llega al 100%. Por ejemplo, el Instituto de Lingüística Formal y Aplicada de la Universidad Carolina (República Checa) cuenta con 282 modelos entrenados para anotar textos en lenguas diferentes, cuyo grado de precisión varía entre 61% y 97% (*Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty*).

El futuro en lenguas eslavas, pese a la aparente facilidad, presenta bastantes claroscuros. Una de las interrogantes es la coexistencia en ucraniano, bielorruso tarashkiano y rusino, de dos tiempos futuros imperfectivos: el sinético y el analítico (Кожанов, 2016). La diferencia formal, como es de esperar, trae consigo una variación funcional, la cual, a su vez, plantea otras interrogantes. L. Marchylo señala por su parte que las formas del futuro sintético a lo largo de su proceso evolutivo en ucraniano denotaban una acción más concreta (Марчило, 1999: 5). K.A. Kozhanov (que clasifica los tiempos futuros eslavos en 4 tipos) resume que el futuro analítico con el verbo SER predomina en lenguas eslavas orientales y occidentales, a diferencia de las meridionales... sin que quede claro el valor de esta forma, la cual, al parecer, denota intención de realizar una acción en futuro, frente al deseo, denotado por otros tipos (Кожанов, 2016). Es decir, tanto la diferenciación formal de varias formas coexistentes del futuro, al igual que sus funciones, siguen constituyendo un "campo sin explorar", lo cual se vuelve un desafío particular de cara al aprovechamiento de los recursos del corpus para la realización del muestreo. El problema de coexistencia de dos formas que se usan indistintamente es actual para algunas otras lenguas eslavas. En el eslavo antiguo el *Futurum II* también se forma con el verbo ser seguido de *l-participle* (Migdalski, 2006: 23). Además de la diferenciación del futuro sintético analítico, perfectivo e imperfectivo, en búlgaro, croata y macedonio existe un tiempo prefuturo (Penkova, 2018: 225).

Algunos de los valores funcionales del futuro, como, p. ej., el inmediato o progresivo, se ven plasmados gracias a marcadores contextuales (Чуйкова, 2018: 21-25), lo cual, a su vez plantea la perspectiva de uso de consultas formales, que además de las formas verbales, contengan otros marcadores complementarios.

Potencialidades del corpus para explorar las funciones del futuro

En contra de lo que pudiera creerse, el aspecto formal del tiempo futuro, que parece explorado de forma exhaustiva, pese a los evidentes avances de la lingüística computacional, sigue constituyendo, curiosamente, un nicho que sigue sin cubrir. El problema de la búsqueda

de formas compuestas o combinaciones de palabras está enfocado, en la mayoría de los casos, como un problema de selección de secuencia de morfemas o palabras. Por otro lado, los estudios centrados en la naturaleza formal y semántica de las formas gramaticales analíticas, mencionados en el apartado anterior, se van efectuando por separado, independientemente del problema de su procesamiento formal. En cuanto al aspecto formal, muchos se limitan a constatar la dificultad técnica de esta selección. T. Jelínek, B. Stindlová, A. Rosen y J. Hana comentan errores de etiquetado relacionados con formas verbales analíticas, verbos modales y predicados copulativos (Jelínek et al. 2012: 133). *A.* Rosen, J. Hana, B. Štindlová *et al.* indican, también que el etiquetado de las formas compuestas puede conducir a errores (Rosen et al., 2014: 12). M. Alexandrov, X. Blanco, O.Mitrofanova y M. Zakharov advierten que el planteamiento de las formas analíticas interrumpidas por otras palabras, muchas veces tratados como "blank characters" resulta demasiado simplista/simplificador y requiere una revisión (Alexandrov M., 2007: 14).

En el ejemplo que sigue podemos imaginarnos que una forma del futuro (ser + infinitivo), por más señas, puede encontrarse interrumpida por un infinitivo de otra estructura dando lugar a confusión:

Ucraniano: Вона буде свою здатність мислити розвивавти.

En este ejemplo, el infinitivo más cercano del auxiliar "бути" no forma parte del tiempo compuesto siendo complemento modificador del nombre al que sigue ("здатність"). En el ejemplo que sigue el verbo auxiliar y el infinitivo utilizado posteriormente no son componentes de una misma estructura gramatical sino una concurrencia casual de formas, como queda ilustrado en este otro ejemplo ruso:

Там будет много гостей — можно веселиться.

Está claro que una selección a rajatabla de todos los infinitivos precedidos por el equivalente del verbo "ser" en futuro cubrirá algunos ejemplos no válidos, como ocurre en casi todas las selecciones formales sin tomar en consideración el contenido. Es decir, la selección de las formas analíticas se ve dificultada por la necesidad de procesar características de dos palabras, no necesariamente seguidas, así como por su ambigüedad funcional. Totalmente conscientes de este problema, consideramos, sin embargo, que el procesar volúmenes grandes conlleva inevitablemente algún porcentaje de material sobrante que, a su vez, puede convertirse en un prolífico objeto de posteriores estudios. Sin pretender resolver el problema de las ambigüedades, que constituye un reto más bien en el marco de procesamiento de lenguaje natural por redes neuronales artificiales, nos centraremos aquí en la solución del problema de las formas inversas e interrumpidas que de por sí tiene bastantes aristas.

Los corpus actuales provistos de anotación gramatical disponen de una serie de soluciones a dicho problema, que, no obstante, varían de un corpus a otro. Por lo tanto, en vista de la creciente abundancia de recursos informáticos se vuelve cada vez más difícil el saber seleccionar la herramienta apropiada. Uno de los criterios más visibles es el volumen del corpus; sin embargo, existen otros tantos parámetros a tener en cuenta: el carácter sincrónico/diacrónico, equilibrado/no equilibrado del corpus, posibilidad o no de formar un subcorpus, anotación gramatical, anotación semántica, etc.

Formulación de consultas para la selección automática de formas analíticas del futuro

Lo cierto es que en la actualidad muchos corpus anotados no disponen de etiquetado específico para formas gramaticales analíticas que abarcan secuencias de palabras, pues los programas actuales de anotación están entrenados para anotar textos con *tags* descriptivos palabra por palabra. Lo expuesto quiere decir que los usuarios que necesitan buscar ejemplos de uso de una forma gramatical compuesta se ven obligados bien a limitarse a la selección de uno de los componentes para seguir segregando los ejemplos "a mano", bien para componer una expresión regular o "máscara de búsqueda". Ya que esta última encarna precisamente el poder generalizador de la inteligencia en la extracción de datos, nos enfocaremos seguidamente en esta segunda opción.

Advertimos, no obstante, que algunas herramientas de corpus particularmente avanzadas desde el punto de vista gramatical ya poseen una anotación apropiada para tales formas analíticas como el modo condicional o el tiempo futuro, que es el *Corpus Ruso Anotado de Helsinki (HANCO),* aunque no exento de limitaciones, tratadas líneas antes. Aun así, la mayoría de los corpus modernos carecen todavía de esta valiosa propiedad. Además, aun suponiendo que en pocos años dispondremos de una excelente anotación gramatical, no nos servirá para la selección automática de otras incontables estructuras, abundantes en lenguas románicas, germánicas y bastante frecuentes en lenguas eslavas, como, p. ej., las compuestas del verbo "dar" (o sus equivalentes) con infinitivo. En lo que respecta al tiempo futuro, la perífrasis verbal que consta del verbo "стать" en lugar de "быть" con infinitivo es muy usual en ruso. O. Yu. Chuykova, por su parte, observa que su uso viene determinado por tales factores como el carácter agentivo del sujeto, persona y presencia/ausencia de la negación (Чуйкова, 2018: 15). Este tipo de estructuras, al igual que su análisis contrastivo se vuelve un suculento desafío para cualquiera que se dedique a explorar formas analíticas. Los ejemplos presentados demuestran que el reto del procesamiento de formas analíticas se extiende más allá de la gramática y comprende un sinnúmero de unidades léxico-gramaticales o incluso estructuras lexicalizadas.

Una de las herramientas más apropiadas para automatizar la selección de unidades multicomponenciales es la lengua *CQL* (*Context Query Language*). En breves palabras, esta lengua sirve para efectuar la búsqueda de una palabra o secuencias de palabras no solo por su forma literal sino por alguna de sus características o propiedades de manera generalizada, de ahí que una sola consulta abarque un conjunto considerable de ejemplos de uso (como las expresiones regulares en lenguas de programación). Por ejemplo, el atributo "lema" permite seleccionar del corpus todo el paradigma de un lexema particular; gracias al atributo "tag" la búsqueda es procesada en razón de sus características morfológicas (u otras, contenidas en los *tags*). Dichas consultas son válidas tanto para palabras sueltas, como para sus secuencias, lo cual resulta particularmente útil para la búsqueda de formas gramaticales analíticas que por su naturaleza se componen de dos o más palabras con unas propiedades morfológicas bien precisas.

Las estrategias de efectuar consultas descritas en adelante son igualmente válidas para otras lenguas y otras categorías con propiedades parecidas (la de componerse de varias formas) y resultan fácilmente adaptables a sus parámetros.

Según queda apuntado, el futuro imperfectivo analítico en las lenguas indicadas se compone mayoritariamente del equivalente del verbo "ser" en forma personal del tiempo futuro con el infinitivo, el cual representa la acción realizada en el futuro. En polaco el tiempo futuro analítico se usa con regularidad con el participio pasado activo en lugar del infinitivo. En todo caso, el denominador común para todas las formas del futuro analítico es el verbo "ser" en futuro como un marcador bastante exclusivo y casi inconfundible. Dicha propiedad demuestra que el verbo auxiliar puede servir de anzuelo para "pescar" todas las formas del futuro de un corpus de textos. En consecuencia, en el Corpus General Regionalmente Anotado del Ucraniano (*ГРАК, Генеральний Регіонально Анотований Корпус*), al efectuar la consulta

[tag= «.*fut.*» & lemma= «бути»]

se seleccionarán automáticamente todos los casos del verbo "ser" usado en futuro. Al tratarse de una forma analítica, para que la selección sea completa, el usuario necesita localizar, además, el infinitivo correspondiente, lo cual no parece tan fácil, pues en las formas del tiempo futuro analítico el auxiliar en ocasiones no está pegado al infinitivo; es más, aunque en la mayoría de los casos el auxiliar es seguido del infinitivo, dicha estructura puede invertirse. Además, el verbo "ser" usado en futuro, aun siendo marcador inconfundible del tiempo futuro analítico, no forma parte necesariamente del futuro sintético, pues una vez usado sin infinitivo que lo acompañe indica existencia en futuro; entonces, ha de considerarse sintético:

Bielorruso: *І будзе новая тут хата* (Якуб Колас, *Беларускі N-корпус*).
Ucraniano: *А буде син, і буде мати*, *І будуть люде на землі* (Тарас Шевченко).
Polaco: *Nie pasuję ja do tych państwa, co tam będą* (Tadeusz Dołęga Mostowicz, NKJP).

Seguido de adjetivo, participio o adverbio, el tiempo futuro denota calidad o estado:

Checo: *Buď budou tyto obce platit, nebo je mateřské školy přijmou, jen pokud bude volné místo*. (Týdeník Sokolovska, *Český národní korpus*).
Ucraniano*: Не поправлять сльози щастя, серцю легше буде* (Іван Котляревський).
Polaco: *Doły, do których ich złożą, będą bezimienne* (Zdzisław Smektała, NKJP).

Algunos de los corpus como el *HANCO*, el Corpus de la Lengua Bielorrusa *Беларускі N-корпус* no incluyen la posibilidad de consultas *CQL*, aunque disponen de un riquísimo abanico de opciones de búsqueda con una parametrización gramatical detallada, aunque en el Corpus de la Lengua Bielorrusa prevé unicamente la posibilidad de seleccionar todas las formas del futuro sin diferenciación de sus formas sintéticas y analíticas, al igual que el Corpus de la Lengua Ucraniana (*Корпус української мови*). No obstante, por muy minuciosa que sea dicha parametrización, puede fallar en ocasiones sin que el usuario pueda corregir el sesgo. Así pues, si realizamos la consulta del verbo "быть" en futuro seguido de participio pasivo, en el corpus *Hanco* obtendremos 20 resultados correspondientes a la voz pasiva en futuro que incluye la consulta de todos los tiempos futuros, entre los cuales, e.g.:

Глава парижского клуба отметил, что даже если выплат не будет или они будут осуществлены не вовремя, россия вряд ли будет объявлена банкротом (Ирина Андреева, *корпус HANCO*).

> *Елена Мизулина обещает, что новый кодекс будет принят уже в этом году* (Маша Гессен, *корпус HANCO*).

Aun así, al realizar una consulta compleja que comprende el tiempo futuro analítico en voz pasiva proporciona únicamente 13 resultados del verbo "быть" seguido de una forma pronominal, no necesariamente pasiva sin incluir, curiosamente, las construcciones pasivas de la consulta anterior, p. ej.:

> *Никто не берется сказать, сумеет ли новый президент выработать свой неповторимый стиль, который будет отличаться от того, что здесь уже видели* (Аркадий Орлов, корпус HANCO).

Aun suponiendo que los resultados produjesen una selección exhaustiva e inequivocable, varios propósitos de la investigación pueden requerir consultas más específicas. En la lengua polaca, p. ej., hay dos posibilidades de formar el tiempo futuro analítico imperfectivo: mediante el verbo 'być' seguido de participio o de infinitivo, excepcionando los verbos modales 'móc', 'chceć', 'musieć', que se usan en estas construcciones únicamente en forma de participio. No obstante, constituye un indudable interés investigador la variación de uso de cada una de estas formas desde el punto de vista diacrónico, estilístico, sociolingüístico, etc. Mediante las consultas *CQL* las dos estructuras son fácilmente segregables. Lo dicho significa que la consulta orientada al verbo "ser" usado en alguna de las formas del futuro produce en todo caso un resultado más amplio del que se necesita. Las opciones visualizables mediante la interfaz son muy cómodas para el usuario, pero están ancladas en patrones preestablecidos careciendo de flexibilidad frente a las consultas *CQL*, que sí la poseen.

Estas y otras manifestaciones del grado de complejidad del problema demuestran la importancia de llevar a cabo estudios basados en un vasto material empírico que, a su vez, supone un voluminoso muestreo. Es imprescindible, además, prever la posible inversión del infinitivo frente al auxiliar en forma personal y las posibles distancias entre estos dos componentes. Al disponer de las valiosísimas herramientas de la lingüística del corpus y, al carecer al mismo tiempo de anotación particular para dichas categorías, la automatización de la selección del material empírico requiere el uso de consultas con una sintaxis particular. En consecuencia, exponemos aquí adelante las particularidades de las susodichas consultas para el checo, eslovaco, polaco y ucraniano (los corpus rusos y bielorruso analizados no prevén la posibilidad de consultas *CQL*) e ilustrar su uso en corpus provistos de anotación morfológica.

Ilustraremos paso a paso la construcción de la consulta necesaria en base a la lengua checa. La consulta:

```
[lemma=»být»&tag=».*F.*»]
```

cubre las formas del verbo 'být' ("ser") en futuro, mientras la secuencia que sigue selecciona infinitivos:

```
[tag=».*V.*»&tag=».*f.*»]
```

Los *tags* "F.", "f.", "V." y otros varían de un corpus a otro y se pueden consultar sobre la página del corpus correspondiente. En el Corpus Nacional de la Lengua Checa (*Český*

národní korpus), "f." designa "infinitivo", mientras en el corpus eslovaco el infinitivo viene etiquetado con la letra "I.".

Lógicamente, para seleccionar los verbos "ser" en futuro seguidos de infinitivos, necesitamos combinar los dos bloques, el uno seguido del otro:

```
[lemma=»být»&tag=».*F.*»] [tag=».*V.*»&tag=».*f.*»]
```

Con el fin de incluir los casos de inversiones, recurrimos al operador de disyunción ("o" lógico), expresado mediante una barra vertical:

```
[lemma=»být»&tag=».*F.*»]
[tag=».*V.*»&tag=».*f.*»]|[tag=».*V.*»&tag=».*f.*»][lemma=»být»&tag=».*F.*»]
```

Somos consientes, además, de que las formas del futuro analítico pueden verse interrumpidas por otras palabras o signos de puntuación, técnicamente llamados *tokens*. Para evitar material sobrante, prescindiremos de signos de puntuación, etiquetados en el corpus checo mediante "Z.". Ya que el tiempo futuro en cuestión se compone del verbo "ser" en futuro e infinitivo o participio pasado, en las palabras que interrumpen no han de aparecer dichos fenómenos gramaticales, lo cual permitirá que la consulta se enfoque en las estructuras mínimas (según el así llamado "cuantificador perezoso"). Dicho en otras palabras, al explorar el verbo "ser" en futuro seguido de dos infinitivos, la consulta se limitará al primero.

La exclusión en consultas formales viene expresada con el signo de desigualdad "!=". El cuantificador {0,7} significa que el grupo anterior comprendido entre corchetes puede aparecer entre 0 y 7 veces (*i.e.*, la forma del futuro puede interrumpirse por 7 *tokens* como mucho). Vistas todas estas consideraciones, la consulta final será:

```
[lemma="být"&tag=".*F.*"][tag!=".*Z.*"&tag!=".*F.*"&tag!=".*f.*"]{0,7}[pos=".*V
.*"&tag=".*f.*"]|[tag=".*V.*"&tag=".*f.*"][tag!=".*Z.*"&tag!=".*F.*"&tag!=".*f.*"]{0,7}
[lemma="být"&tag=".*F.*"]
```

Finalmente, ya que el tiempo futuro analítico puede comprender participio pasado activo en lugar del infinitivo (en el caso de la forma pasiva del verbo), incluiremos también esta posibilidad al extender la subexpresión [pos=".*V.*"&tag=".*f.*"], que describe el infinitivo, a [pos=".*V.*"&tag=".*f.*"]|[tag=".*s.*"], pues ".*s.*" es el descriptor respectivo del participio pasado.

Necesitamos agragar al final de la consulta "within <s/>" (o "within s" para Poliqarp) para que ambas partes de la forma analítica se sitúen dentro de una misma oración.

Hemos optado por el número 7 como una constante de distancia máxima, suponiendo que tal es el número de unidades que un individuo es capaz de retener en la memoria operativa, el famoso "número mágico" establecido por George A. Miller. Dichas unidades, o porciones (*bits* o *chunks* en términos de G.Miller), se corresponden lógica-, aunque no exclusivamente, con palabras (Miller, 1956: 94). No descartamos, ni mucho menos, que el número 7 pueda ser insuficiente quedando abierto a debate y consiguiente exploración en una próxima publicación.

Resultado

Resumiremos, por tanto, los resultados obtenidos en la tabla 1, que comprende la indicación de las fuentes (corpus textuales), consulta *CQL* válida para seleccionar formas del tiempo futuro analítico (en el caso del polaco se utiliza su alternativa, *Poliqarp*), así como referencias a las guías con *tags* convencionales correspondientes a características morfológicas.

Lengua/Corpus	Página web	Consulta CQL para seleccionar las formas del futuro analítico	Lista de tags
Checo *Český národní korpus*	https://kontext.korpus.cz/first_rm?corpname=syn2010	[lemma="být"&tag=".*F.*"][tag!=".*Z.*"&tag!=".*F.*"&tag!=".*f.*"]{0,7}([pos=".*V.*"&tag=".*f.*"]|[tag=".*s.*"])|([pos=".*V.*"&tag=".*f.*"]|[tag=".*s.*"])[tag!=".*Z.*"&tag!=".*F.*"&tag!=".*f.*"]{0,7}[lemma="být"&tag=".*F.*"] within <s/>	https://wiki.korpus.cz/doku.php/seznamy:tagy
Eslovaco Slovenský národný korpus	http://korpus.juls.savba.sk:8080/manatee.ks/index	[lemma="byť"&tag=".*B.*"][tag!=".*Z.*"&tag!=".*I.*"&tag!=".*B.*"]{0,7}([tag=".*V.*"&tag=".*I.*"]|[tag=".*G*"&tag=".*t.*"])|([tag=".*V.*"&tag=".*I.*"]|[tag=".*G*"&tag=".*t.*"])[tag!=".*Z.*"&tag!=".*I.*"&tag!=".*B.*"]{0,7}[lemma="byť"&tag=".*B.*"] within <s/>	https://korpus.sk/morpho_en.html
Polaco NKJP (Narodowy Korpus Języka Polskiego). Balanced NKJP subcorpus (300M segments)	http://nkjp.pl/poliqarp/nkjp300/query/	[base=".*być.*"&tag=".*bedzie.*"][tag!=".*bedzie.*"&tag!=".*inf.*"&tag!=".*interp.*"]{0,7}([tag=".*inf.*"]|[tag=".*praet.*"]|[tag=".*ppas.*"])|([tag=".*inf.*"]|[tag=".*praet.*"]|[tag=".*ppas.*"])[tag!=".*bedzie.*"&tag!=".*inf.*"&tag!=".*interp.*"]{0,7}[base=".*być.*"&tag=".*bedzie.*"] within s	http://nkjp.pl/poliqarp/help/ense2.html#x3-20002
Ruso Национальный корпус русского языка	http://ruscorpora.ru/new/search-main.html	Mediante la interfaz del usuario, sin distinción de formas sintéticas y analíticas	
Ruso Helsinki Annotated Russian Corpus HANCO (ХАНКО - Хельсинкский аннотированный корпус русского языка)	http://h248.it.helsinki.fi/hanco/	Mediante la interfaz del usuario (formas analíticas y sintéticas por separado)	

Ucraniano General Regionally Annotated Corpus of Ukrainian, GRAC (Генеральний регіонально анотований корпус української мови, ГРАК)	http://www.parasolcorpus.org/bonito/run.cgi/	[lemma="бути"&tag=".*futr.*"][tag!=".*futr.*"&tag!=".*inf.*"&tag!=".*punct.*"]{0,7}([tag=".*inf.*"]|[tag=".*adjp.*"&tag=".*pasv.*"])|([tag=".*inf.*"]|[tag=".*adjp.*"&tag=".*pasv.*"])[tag!=".*futr.*"&tag!=".*inf.*"&tag!=".*punct.*"]{0,7}[lemma="бути"&tag=".*futr.*"] within <s/>	https://github.com/brown-uk/dict_uk/blob/master/doc/tags.txt?fbclid=IwAR1-Pp8e2ZSLQzzY1W-IwLa4tLW29OIfMo3OEZsA0mcPhBBnGEeKm-9Vw
Ucraniano Корпус української мови	http://www.mova.info/corpus.aspx	Mediante la interfaz del usuario, sin distinción de formas sintéticas y analíticas	
Bielorruso Беларускі N-корпус	https://bnkorpus.info/index.html	Mediante la interfaz del usuario, sin distinción de formas sintéticas y analíticas	

Tabla 1. Consultas *CQL* para selección de las formas del tiempo analítico futuro.

Conclusiones

Las herramientas de los corpus actuales permiten automatizar el muestreo gracias a la anotación gramatical detallada. Las dificultades de anotación y selección automática de las formas gramaticales analíticas se deben a su estructura compleja, por lo cual, en la mayoría de los corpus actuales su anotación morfológica a nivel de palabras sueltas no es posible o, de serlo, produce resultados incompletos o, en ocasiones, erróneos. No descartamos que en un futuro muy cercano la susodicha carencia sea subsanada. No obstante, los propósitos particulares de la investigación pueden requerir que las consultas sean más específicas: selección de una sola estructura de las dos variantes posibles (como es el caso del polaco, futuro con infinitivo o con el participio pasado), presencia/ausencia de verbos modales, filtración de las formas pasivas o activas en la búsqueda.

Las herramientas de la lingüística del corpus, pese a una serie de limitaciones respecto a la selección automática de formas gramaticales analíticas, se vuelven mucho más manejables gracias a las consultas *CQL* o sus alternativas, como el lenguaje *Poliqarp*, mucho más flexibles frente a las opciones de la interfaz del usuario, válidas tan solo para una característica específica. Dichas consultas han de comprender la descripción formal de dos (o más) componentes de la forma analítica en cuestión. Los componentes variables de la forma analítica han de acompañarse del atributo "lemma"/ "base" en la consulta, mientras sus características gramaticales han de referirse con (*omitir por estar duplicado*) el atributo *"tag"*. En las consultas es imprescindible tener en cuenta que los componentes de una forma gramatical analítica pueden invertirse, interrumpirse con otras palabras, por lo cual las consultas han de comprender respectivas disyuntivas de subconsultas con el orden directo y el inverso, al indicar las distancias mínimas y máximas entre los componentes. Puesto que la distancia máxima es difícil de establecer *a priori*, recomendamos partir del número 7 como

una posible referencia, no sin descartar la necesidad de explorar la posibilidad de extensión de dichas distancias mediante la estrategia de pruebas y errores, lo cual es planteable como una interesante perspectiva para estudios posteriores tanto en el plano contrastivo entre las lenguas indicadas, como para investigaciones monolingües.

REFERENCES

Гусман Тирадо, Р., Верба, Г.Г. (2005). Употребление личных глагольных форм в официально-деловых и юридических текстах испанского и русского языков. *Вестник Удмуртского университета*. 5(2), сс.153-160.

Кожанов, К.А. (2016). Аналитическое будущее время в языке русских цыган как калька с восточно-славянских языков. *International Journal of Slavic Studies,*1, 249-262. https://cyberleninka.ru/article/n/analiticheskoe-buduschee-vremya-v-yazyke-russkih-tsygan-kak-kalka-s-vostochno-slavyanskih-yazykov/viewer

Марчило, Л.М. (1999). *Історія форм майбутнього часу дієслова в українській мові.* Автореф. Дис. Канд. Філол. Наук, Київ. http://enpuir.npu.edu.ua/bitstream/123456789/6156/1/Marchylo.pdf

Милославский, И. Г., Виноградов, В.С. (1987). *Сопоставительная морфология русского и испанского языков.* Москва: Русский язык.

Пенькова, Я.А. (2018): Славянское второе будущее: семантика, структурные особенности и эволюция. *Славянское языкознание*. *XVI Международный съезд славистов*. Белград, 20–27 августа 2018 г. Доклады российской делегации. В С. М. Толстая (Отв. Редактор). Москва: ИСл РАН, сс. 225-243. DOI: 10.31168/0417-6.2.6

Чуйкова, О.Ю. (2018). *Семантика будущего времени в русском, английском и испанском языках: взаимодействие темпоральности, аспектуальности и модальности.* Автореф. Дисс. Канд. Филол. Наук, Санкт-Петербург. https://www.dissercat.com/content/semantika-budushchego-vremeni-v-russkom-angliiskom-i-ispanskom-yazykakh-vzaimodeistvie-tempo/read

Alexandrov, M., Blanco, X., Mitrofanova, O. M., Zakharov, V. (2007). Nooj Applications for Document Clustering and Corpus Linguistics. In X. Blanco, Silberztein M. (Eds.) *Proceedings of the 2007 International NooJ Conference*, Cambridge Scholars Publishing: Newscastle, pp. 6-19. https://www.cambridgescholars.com/download/sample/60082

Guzmán Tirado, R. & Quero-Gervilla, E. (2007). Estudio comparado de las construcciones subordinadas generativas que expresan relaciones concesivas en ruso y en español. *Eslavística complutense,* 7, 115-133. https://www.researchgate.net/publication/27593243_Estudio_comparado_de_las_construcciones_subordinadas_generativas_que_expresan_relaciones_concesivas_en_ruso_y_en_espanol

Jelínek, T., Stindlová, B., Rosen, A., & Hana, J. (2012). Combining Manual and Automatic Annotation of a Learner Corpus. In: Sojka P, Horák A, Kopeˇcek I, Pala K (eds.) *Text, Speech and Dialogue – Proceedings of the 15th International Conference*. *TSD 2012*, Brno, Check Republic, pp 127-134. https://core.ac.uk/download/pdf/208679553.pdf

Migdalski, K. (2006). *The Syntax of Compound Tenses in Slavic*. Utrecht: Lot.

Miller, G.A. (1956). The Magical Number Seven, Plus or Minus Two. *The Psychological*

Review, vol. 63, pp. 81-97.
Modelos de conjugación verbal. *Real Academia Española* https://www.rae.es/dpd/ayuda/modelos-de-conjugacion-verbal
Rosen, A., Hana, J., Štindlová, B. et al. (2014). Evaluating and automating the annotation of a learner corpus. *Lang Resources & Evaluation*, 48, pp. 65-92. https://doi.org/10.1007/s10579-013-9226-3
Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulty. Univerzita Karlova. https://ufal.mff.cuni.cz/udpipe/models

Recursos del corpus

Беларускі N-корпус https://bnkorpus.info/index.html
Грак, Генеральний Регіонально Анотований Корпус Української Мови http://www.parasolcorpus.org/bonito/run.cgi/first_form
Грак, Генеральний Регіонально Анотований Корпус Української Мови. Набір тегів. https://github.com/brown-uk/dict_uk/blob/master/doc/tags.txt?fbclid=IwAR1-YmPp8e2ZSLQzzYlW-IwLa4tLW29OIfMo3OEZsA0mcPhBBnGEeKm-9Vw
Корпус української мови КНУ імені Тараса Шевченка http://www.mova.info/corpus.aspx?l1=209
Adam Przepiórkowski, Aleksander Buczyński Ściągawka do Narodowego Korpusu Języka Polskiego http://nkjp.pl/poliqarp/help/ense2.html#x3-20002
BNC, British National Corpus https://www.english-corpora.org/bnc/
Český národní korpus (Manual) https://wiki.korpus.cz/doku.php/seznamy:tagy
Český národní korpus. Křen, M. – Bartoň, T. – Cvrček, V. – Hnátková, M. – Jelínek, T. – Kocek, J. – Novotná, R. – Petkevič, V. – Procházka, P. – Schmiedtová, V. –
Petkevič, V. – Procházka, P. – Schmiedtová, V. – Skoumalová, H.: SYN2010: žánrově vyvážený.
Helsinki Annotated Russian Corpus HANCO (ХАНКО - Хельсинкский аннотированный корпус русского языка) http://h248.it.helsinki.fi/hanco/
NKJP (Narodowy Korpus Języka Polskiego) http://nkjp.pl/poliqarp/nkjp300/query/
Slovenský národný korpus – prim-7.0-public-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2015. Available at WWW http://korpus.juls.savba.sk:8080/manatee.ks/index
Slovenský národný korpus. Morfologická anotácia textov Slovenského národného korpusu https://korpus.sk/morpho_en.html