

Análisis mediante *Data Mining* de comentarios paulinos de Tomás de Aquino

Data Mining Analysis of Aquinas's Pauline Commentaries

Enrique Alarcón
Universidad de Navarra, España
ealarcon@unav.es
ORCID: 0000-0001-5527-1922

Resumen: *Data Mining* es una tecnología de estadística computarizada para la extracción de pautas reiteradas y significativas a partir de datos cuantiosos. En este artículo se exponen primero los fundamentos para su aplicación al análisis de los textos. Después, se detalla la aplicación del análisis de afinidad a buena parte de los escritos de Tomás de Aquino. Y, finalmente, se muestran los resultados obtenidos del análisis de varios de sus comentarios paulinos.

Abstract: *Data Mining* is a computerized statistical technology for the extraction of repeated and relevant patterns from large data. This article explains the fundamentals for its application to the analysis of texts; details the application of its affinity analysis to the Thomistic corpus; and shows results obtained from the analysis of some of Aquinas's Pauline commentaries.

Palabras clave: Tomás de Aquino, Epístolas de San Pablo, *Data Mining*, Análisis de afinidad

Keywords: Thomas Aquinas, Pauline Epistles, *Data Mining*, Affinity Analysis

Recibido: 15/05/21

Aprobado: 10/07/21

Introducción

Desde la Edad Moderna, la ciencia se caracteriza por el uso de las Matemáticas como el instrumento más básico y exacto para la demostración y expresión de los resultados. En efecto: desde el siglo XIV, la Física comienza a enfocarse de manera que pueda ser tratada matemáticamente, proceder que, en el siglo XVII, puede considerarse ya generalizado. Ya en el siglo XVIII, empieza a destacarse el tratamiento de la Química desde las premisas de esa nueva Física matemática. Y, entre el siglo XIX y el XX, sobre la base de esta Química nueva, se desarrolla la Biología contemporánea, gracias prin-

principalmente a la Bioquímica y a la Genética. La Psicología, a través de las neurociencias, parece la próxima candidata idónea para integrarse en este proceso de ordenación jerárquica de saberes. De este modo, la ciencia contemporánea usa implícitamente la teoría aristotélica de la subalternación, de manera que los saberes más particulares emplean como premisas propias las conclusiones de los saberes más universales.

Esta jerarquía moderna de las ciencias, subalternadas principalmente a la Matemática, ha sido una estrategia sólida y exitosa para incrementar el conocimiento y perfeccionar la tecnología. Con todo, solo es aplicable a objetos tratables matemáticamente: en principio, los que tienen cantidad y/o posición¹. Por ello, un amplio y nada desdeñable campo de estudio –las así llamadas *letras*– ha ido quedando fuera del ámbito de lo que la Modernidad ha llamado *ciencias*.

En rigor, ambos géneros de saber no son mutuamente excluyentes, como demuestra el caso de tantos saberes intermedios –la Economía o la Sociología, por ejemplo– susceptibles de cierto tratamiento matemático, pero limitado. A este respecto, el desarrollo de la Estadística y, más recientemente, de la Lógica formal y la Informática, ha permitido añadir a dichos casos fronterizos una parte de los saberes humanísticos: es aquí donde se sitúan las llamadas *Humanidades Digitales* (Burdick, Drucker, Lunenfeld, Presner, Jeffrey, 2016).

La complejidad de los asuntos humanísticos hace difícil abstraerlos a objetos matemáticos. Sin embargo, su cauce habitual de expresión, el texto escrito, es –hasta cierto punto– fácil de analizar matemáticamente, sobre todo en sus aspectos estadísticos. Tal análisis, para ser fiable, requiere de grandes números: por ello, el soporte digital de grandes cuerpos textuales y la computación automatizada han sido claves para su análisis matemático, la consiguiente expresión de los resultados y su almacenamiento. Tales han sido, a mi juicio, las razones para el reciente desarrollo de las Humanidades Digitales.

Su origen suele situarse en 1948, cuando el jesuita Roberto Busa inicia la base de datos *Index Thomisticus* (Busa et alii, 2005). Su inspiración original fue el clásico adagio *Thomas sui interpres* (Massoulié, 1692)²: a Tomás de Aquino se

¹ Ciertamente que cierta hibridación, imperfecta, de las Matemáticas contemporáneas con la Lógica formal vuelve algo difusa tal caracterización. En efecto: las propiedades y relaciones de abstracciones simbólicas que pueden representar tanto números como otros términos lingüísticos pueden pertenecer a cualquiera de dichas ciencias. Pero esto solo añade un cierto *campo de nadie* entre las *ciencias* y las *letras* como áreas distintas, campo intermedio donde también se sitúa el objeto de este estudio.

² La máxima, que Massoulié hizo famosa, parece provenir de un curso de V. Baron aún inédito.

le interpreta mejor desde sus propios textos. Por eso, el P. Busa trató de analizar en profundidad el lenguaje de santo Tomás, empezando por su nivel más básico, el morfológico, y tomando inicialmente como referencia las grandes recopilaciones clásicas del léxico latino, especialmente el monumental *Totius Latinitatis lexicon* de Facciolati y Forcellini (1771), que Busa usó como base para la clasificación del léxico en su *Index Thomisticus* (Busa, 1998). Con el fin de que tal estudio y clasificación fuesen exhaustivos, se propuso aprovechar la capacidad de los ordenadores que, justamente entonces, comenzaban a emplearse para tareas análogas. Como resultado, la base de datos *Index Thomisticus* contiene las obras completas de santo Tomás en formato digital, con diversas informaciones filológicas para cada palabra. Y todas sus formas de declinación o conjugación, o sus variantes ortográficas, forman un solo *lemma* o término (aunque se consideran tres lemas distintos sus respectivas derivaciones como sustantivo, verbo y adjetivo o adverbio).

Supuestos del siguiente análisis de contenidos textuales mediante *Data Mining*

Esta colosal fuente de información es muy adecuada para el análisis estadístico computarizado del corpus tomista, al menos en el nivel del léxico. Más específicamente, constituye una herramienta idónea para el denominado *Data Mining* o *minería de datos*, a saber: la extracción de pautas reiteradas y significativas a partir de datos cuantiosos (Feldman & Sanger, 2007).

El primer empleo de esta tecnología fue el llamado *análisis de afinidad*, concebido originariamente para resolver un problema de mercadotecnia. Estudiando las compras en los supermercados, se comprueba que ciertos productos tienden a adquirirse conjuntamente: por ejemplo, refrescos y aperitivos. Así pues, las empresas desarrollaron metodologías estadísticas e informáticas para detectar estas *afinidades* e implantar en consecuencia sus correspondientes estrategias de mercado, distribución y ventas: por ejemplo, situando los productos a la venta más afines en zonas contiguas.

Pienso que hay una cierta analogía entre el cliente que introduce productos afines en su bolsa de la compra y el escritor que incluye términos dentro de la misma oración. Ambas concurrencias pueden ser meramente fortuitas. Sin embargo, si se reiteran mucho más que otras, resulta probable –en igual proporción– que tales concomitancias no sean casuales, como resultaría improbable que una moneda tirada al aire cayese más frecuentemente del mismo lado por puro azar. La *afinidad* estadística entre términos escogidos por el escritor

indica que, para él, guardan una especial relación, expresada en su texto con relativa insistencia e indicativa, por tanto, de un tema con particular relevancia.

A este respecto, conviene hacer una distinción entre lo que llamaré *afinidades sintagmáticas* y *afinidades asertivas*. A veces, la concurrencia reiterada de términos obedece a la especificidad de cierta lengua o corriente doctrinal: por ejemplo, la terminología cristiana consagra sintagmas frecuentes, como “Espíritu Santo” o “apóstol Pedro”. Tales afinidades, precisamente por deberse al carácter social de las lenguas y de las corrientes doctrinales, resultan fáciles de detectar y valorar como sintagmáticas. Algo análogo ocurre respecto al habla de un autor concreto, o a su específico enfoque doctrinal: cualquier lector asiduo de santo Tomás conoce la relevancia en sus textos de sintagmas como “*actus essendi*” o “*esse ut actus*”; o del sintagma “*quodam modo*”, como grafía alternativa de un término tan característico de su estilo como “*quodammodo*”. A su vez, el paleógrafo conocedor de la escritura manuscrita de santo Tomás puede constatar que, para un copista francés de la época, su modo de escribir “*et ideo*” fácilmente podría confundirse con el término “*nam*”, de manera que sus frecuencias, en general, tienden a ser inversas. Estas afinidades *sintagmáticas* pueden ser relevantes según qué tipo de investigación se desarrolle: paleográfica, léxica, doctrinal... Sin embargo, corresponden a un nivel relativamente simple del lenguaje, el léxico, donde no se da la plenitud del significado. En cambio, la significación completa, junto con la verdad o falsedad, sólo se da en la oración asertiva. Por lo mismo, es allí, en el conjunto de la estructura *sujeto+predicado*, donde mejor queda reflejado el pensamiento del autor. En consecuencia, el análisis de contenidos de un texto debe centrarse en lo que podríamos llamar *afinidades asertivas*, más que solo sintagmáticas.

En efecto: el autor expresa su pensamiento, básicamente, mediante afirmaciones o negaciones que integran sujetos con predicados, expresados mediante términos: principalmente, sustantivos, verbos, adjetivos y adverbios. Por eso, la concurrencia estadísticamente relevante de tales términos dentro de oraciones completas resulta, por lo general, significativa de los contenidos en los que el autor más ha insistido.

Hay otro tipo de términos, sin embargo, que, pese a ser muy frecuentes, no son significativos en el análisis estadístico de contenidos. Me refiero a los que cumplen una mera función gramatical, como los términos conectivos: “y”, “o”, “pero”, “aunque”, etc. Suelen emplearse casi siempre los mismos, independientemente del tema tratado. Por ello, son irrelevantes para el análisis de contenidos, pero muy frecuentes, de modo que destacan en cualquier análisis estadístico de un texto. Cuando el análisis de afinidad se ordena, no

a estudiar la lengua del autor, sino su pensamiento, conviene eliminarlos de los cómputos, para así evitar de raíz resultados irrelevantes.

Por razones análogas, conviene que el análisis de contenidos mediante afinidades asertivas considere los términos con independencia de variaciones meramente gramaticales: declinaciones, conjugaciones verbales, variantes gráficas, enclíticas, etc. Por eso, para el análisis estadístico informatizado de contenidos en las obras de santo Tomás, resulta de especial utilidad la agrupación en un solo *lema* o término de las diversas formas gramaticales de cada misma palabra que Roberto Busa realizó en su *Index Thomisticus*³. Y también su clasificación como lemas distintos de la mayor parte de los homógrafos, a saber: de las formas idénticas para palabras distintas, como es la forma *cano* del verbo *canere* y el adjetivo *canus*.

El análisis de contenidos del corpus tomista por afinidad de términos

Sobre dichos supuestos, y contando, gracias a la generosidad y apoyo pleno del P. Busa, con la base de datos *Index Thomisticus* tal como fue editada por él mismo, me propuse en el año 2012 realizar un análisis de afinidad de los escritos de santo Tomás⁴. Por diversas razones, que sería prolijo enumerar y omito aquí, no todos los textos contenidos en el *Index Thomisticus* se prestaban a tal análisis. Baste señalar, a manera de ejemplo, que algunos son tan breves que no obedecen a la ley estadística de los grandes números: carecen de una distribución normal estándar para los casos por azar y, por lo mismo, no cabe detectar diferencias ajenas a la hipótesis nula. Expresado de otro modo: son como una moneda tirada al aire solo dos veces, de manera que los resultados igualmente pueden deberse al azar o no. Seleccionando, pues, los textos de santo Tomás adecuados (y en algún caso sus porciones cuasi unitarias, como las partes de la *Summa Theologiae*) sumaban más de 70 escritos. A ellos añadí 40 textos de otros autores incluidos también por Busa en su base de datos, por su estrecha relación con el corpus tomista. Incluí, además, a efectos de sondeo

³ Inspirándose seguramente en la doctrina clásica sobre la determinación del sujeto y el predicado en las oraciones, Busa consideró como tres lemas distintos a los sustantivos, a los verbos y a los adjetivos o adverbios derivados de una misma palabra, como señalé más arriba. Criterio seguido también en mi análisis de afinidad del corpus tomista, por estar hecho sobre el *Index Thomisticus*.

⁴ En consecuencia, las ediciones de los textos analizados son aquellas mismas usadas por Busa en su *Index Thomisticus*. Concretamente, para los comentarios paulinos usa la edición de R. Cai publicada en 1953.

de validez, unos pocos textos tan breves que, por las razones antes expuestas, no eran idóneos para un análisis de afinidad; pero que, precisamente por eso, gozaban del interés de probar los resultados de aplicárselo.

Como beneficiario del *IBM Academic Program*, recibí gratuitamente del mismo su programa *DB2 Intelligent Miner*. Y, con la colaboración técnica del ingeniero Jaime García-Hoz, realizamos el análisis de afinidad de dichos textos en el año 2012.

Por la ley estadística de los grandes números, para evitar resultados azarosos, computamos solo los términos significativos más frecuentes en el corpus tomista. Concretamente, hay unos 260 lemas cuyas ocurrencias suman más de dos terceras partes de todas las palabras contenidas en el *Index Thomisticus*. De dicho elenco eliminamos aquellos cuyo empleo es principalmente gramatical, como antes expuse. Y analizamos la afinidad de los restantes lemas por su concurrencia frecuente dentro de las mismas oraciones (que, en santo Tomás, casi siempre son asertivas).

Precisamente porque en el análisis estadístico solo resultan significativos los grandes números, optamos por representar gráficamente los resultados y no mediante listados numéricos: en efecto, los gráficos elaborados por *DB2 Intelligent Miner* permiten una mejor visión de conjunto y una interpretación más intuitiva. Están publicados desde 2013 en el apartado *Chartae synopticae operum de Corpus Thomisticum* (Alarcón y García Hoz, 2013).

Los mapas de contenidos de los comentarios paulinos de Tomás de Aquino

En esos mismos años, por iniciativa de Piotr Roszak, la colección *Scholastica Thorunensia* de la Universidad Nicolás Copérnico de Toruń, comenzó una edición bilingüe, en latín y polaco, de los comentarios de santo Tomás a las epístolas paulinas (Tomasz z Akwinu, 2012). Su promotor me invitó amablemente a colaborar con él en este empeño, incluyendo los mapas de contenidos correspondientes obtenidos por análisis de afinidad, junto con una breve presentación de cada uno (Tomasz z Akwinu, 2016, 2017, 2021). Es por este motivo que aquí, como muestra de las diversas líneas actuales de investigación en la Teología Bíblica de santo Tomás, presento a los lectores de lengua española los fundamentos de dicho análisis y, a continuación, varios ejemplos de los resultados obtenidos, explicados brevemente.

Conforme a lo expuesto, estos resultados se muestran como mapas de contenidos estadísticamente significativos en los correspondientes comenta-

rios paulinos de santo Tomás sometidos a análisis de afinidad. Y son gráficos elaborados por el programa *IBM Intelligent Miner* a partir de los resultados numéricos obtenidos. Para interpretarlos adecuadamente es imprescindible conocer el sentido de cada recurso gráfico empleado por dicho programa.

Los lemas más *afines*, los más concurrentes con otros, vienen representados por círculos, etiquetados con la forma representativa de dicho lema en la clasificación del P. Busa⁵. El color de cada círculo indica su frecuencia: naranja, los más repetidos; en azul, los menos; y en blanco los de frecuencia intermedia.

La segunda información estadística presentada en cada gráfico es el llamado *soporte* de esas afinidades: hasta qué punto son frecuentes en el texto las oraciones donde concurre cada pareja de lemas afines. Se calcula midiendo la proporción de las oraciones donde coinciden los mismos lemas sobre el total de oraciones del texto. Por eso, a mayor extensión del texto, más significativo resulta el soporte de afinidad. En el gráfico, las flechas naranjas indican soportes de afinidad mayores; las azules, menores; y las blancas representan soportes con valores intermedios.

Finalmente, la anchura de las flechas indica la denominada *confianza* de la afinidad entre los términos: corresponde a la proporción matemática entre el número de veces en que un lema coincide con su afín y el número total de ocurrencias de dicho lema en el texto. De este modo, si un lema aparece casi siempre junto a su afín, guarda estrecha relación con él. Este valor de confianza es tanto más significativo cuanto más frecuente sea el lema en cuestión.

Supuesto estos criterios de interpretación, presentaré seguidamente algunos ejemplos de los gráficos obtenidos al aplicar el análisis de afinidad a algunos comentarios paulinos de santo Tomás. El lector puede encontrar todos los gráficos obtenidos de sus comentarios bíblicos, con sus colores originales, en la dirección: <https://www.corpusthomicum.org/ichartae.html#CB>.

La Lectura super I Epistolam B. Pauli ad Thessalonicenses

El primer gráfico que presentaré aquí corresponde a la *Lectura super I Epistolam B. Pauli ad Thessalonicenses*:

⁵ El P. Busa usó un procedimiento automatizado para seleccionar dicha forma paradigmática de cada lema, lo que, ocasionalmente, dio lugar a resultados inesperados: así, en el *Index Thomisticus*, la forma representativa del término *causa*, *causae* es *caussa*. Los gráficos usan estas mismas formas paradigmáticas.

en torno a *Cristo*, considerado principalmente como *causa* de la *resurrección* de los *cuerpos*. A su vez, en lo que respecta a dicha *resurrección*, el análisis de afinidad hace evidente que el otro tema destacado es su *orden*. En efecto, el comentario incide repetidamente en el orden de la *resurrección* de los *cuerpos* y su relación con el orden establecido por la *sabiduría* de *Dios*.

Además, en esta primera *galaxia* conceptual de *Super I Thes.*, centrada en *Cristo*, hay otros elementos destacados: primero su *divinidad* y *poder divino*, después su *humanidad* y finalmente la *salvación*.

La *galaxia* de *Cristo* aparece vinculada, a través de la *palabra*, con otra cuyo lema central es *Dios*. El *hombre* aparece como un tercer elemento principal. En efecto, santo Tomás reitera en su comentario que la *palabra* de *Cristo* es *palabra* de *Dios* y no de *hombre*. Tal es la razón de su *poder divino*. Aunque con menor relevancia, también destaca en esta *galaxia* temática el *don* de *Dios*.

La tercera agrupación principal de temas en el texto detectada por afinidad estadística se centra en el *bien*. Se contrapone al *mal*, y aparece vinculada a la *galaxia* temática centrada en *Cristo* mediante la *caridad* y la *fe*. Se observa además su vinculación a *virtus*, referente aquí a la firmeza en la *fe*.

Otros *sistemas gravitatorios* muestran por su color claro menor relevancia: más baja frecuencia en el texto de los lemas implicados, más escaso número de afinidades, y carencia de vinculación directa con las *galaxias* temáticas principales.

La Lectura super I Epistolam B. Pauli ad Timotheum

El segundo ejemplo que presentaré aquí es simple y fácilmente interpretable. Corresponde a la *Lectura super I Epistolam B. Pauli ad Timotheum*:

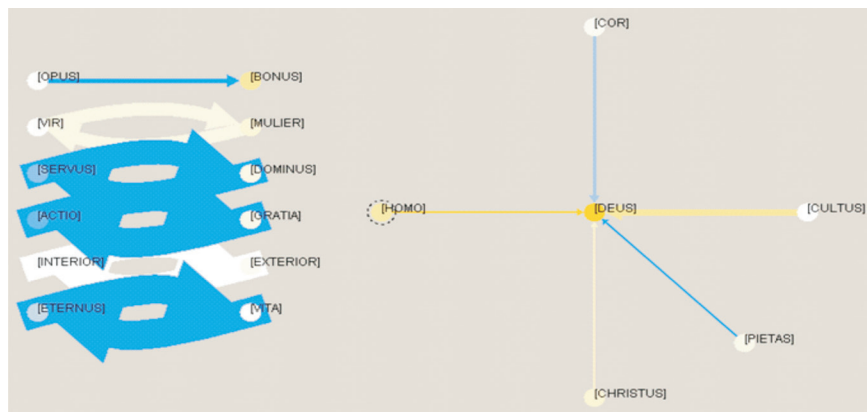


Gráfico 2: *Super I Tim.*

con *beneficio* (referente aquí al recibido por los efesios del apóstol Pablo). En cuanto a los valores de confianza en las distintas afinidades encontradas, resultan más significativos (por la mayor frecuencia de los lemas implicados) los correspondientes a los lemas *ley/antigua, vida/eterna, alma/cuerpo, voluntad/divina, y Padre, Hijo y don*.

La riqueza de afinidades en este texto contrasta con la relativa simplicidad del precedente, e invita a ir las considerando pormenorizadamente a lo largo de la lectura del texto. Se hace patente cómo el análisis de afinidades representado en el texto sirve como sondeo inicial, sinopsis de contenidos y guía para la lectura detallada del texto.

Algunas consideraciones finales

Pienso que el propio Tomás de Aquino simpatizaría con el desarrollo de instrumentos de investigación como los arriba descritos, pues él mismo dirigió la confección de uno muy notable, la *Tabula fratris Thomae*, que conservó hasta el final de su vida. Y, de hecho, santo Tomás es –si no yerro– el pensador al que se han dedicado más instrumentos y herramientas de investigación: incluso –para confusión de quienes desdeñan el tomismo como ajeno a lo contemporáneo– las *Humanidades digitales* comenzaron con el *Index Thomisticus*, que –a día de hoy– sigue siendo banco de trabajo privilegiado para el desarrollo más avanzado de las bases de datos textuales, como ilustran el *Index Thomisticus Treebank*, el proyecto LiLa o la anotación semántica actualmente en marcha en *Corpus Thomisticum*.

Asumido esto, conviene recordar que *Thomas sui interpres*: ningún instrumento sustituye la lectura de los textos tomistas, pues dicha lectura –y no otra cosa– fue lo presupuesto por el propio Tomás al elaborar su obra escrita.

Los mapas de contenidos aquí expuestos pueden servirnos para atender a aspectos determinados que, desde cierto punto de vista, destacan sobre los demás. Con todo, esto no deja de ser una abstracción de lo restante, que es mayoritario en el conjunto y cuya desconsideración implica un cierto riesgo: los grandes autores suelen ser sutiles, y no siempre basta para entenderlos atender a pasajes sueltos, por importantes que puedan ser.

Instrumentos de investigación como el análisis textual por *Data Mining* deben subordinarse a su fin: han de ser escalas de acceso a la mente del autor manifiesta en sus escritos. Y solo sin negar lo omitido se cumplirá aquella observación de Aristóteles que el tiempo ha consagrado como condensado de sabiduría: *abstrahentium non est mendacium*.

Referencias

- Alarcón, E. y García Hoz, J. (2013). *Chartae synopticae operum S. Thomae de Aquino*. <https://www.corpusthomisticum.org/ichartae.html>
- Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., Jeffrey, S. (2016). *Digital Humanities*. Cambridge [Mass.]: MIT Press.
- Busa, R. y Forcellini, Ae. (1998). *Totius Latinitatis lemmata quae ex Aeg. Forcellini Patavina editione 1940: a fronte, a tergo atque morphologicae*. Milano: Istituto Lombardo. Accademia di Scienze e Lettere.
- Busa, R. et al. (2005). *Index Thomisticus*. Ed. E Bernot, E. Alarcón. <https://www.corpusthomisticum.org/it/>
- Facciolati, J. et Forcellini, A. (1771). *Totius Latinitatis lexicon*. Patavii: Typis Seminarii.
- Feldman, R. & Sanger, J. (2007). *The Text Mining Handbook*. Cambridge: Cambridge University Press.
- Massoulié, A. (1692). *Divus Thomas sui interpres de divina motione et libertate creata*. Romae: Typis et sumptibus Iosephi Vannaccii.
- Thomas de Aquino (1953). *Super Epistolas S. Pauli lectura*. Ed. R. Cai. Taurini: Marietti.
- Tomasz z Akwinu (2012). *Wykład Listu do Kolosan. Super epistolam B. Pauli ad Colossenses lectura*. Ed. P. P. Roszak. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2012.
- (2016). *Wykład Pierwszego listu do Tymoteusza. Super primam epistolam B. Pauli ad Timotheum lectura*. Ed. P. P. Roszak y E. Alarcón. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
- (2017). *Wykład Listu do Efezjan. Super primam epistolam B. Pauli ad Ephesios lectura*. Ed. P. P. Roszak y E. Alarcón. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
- (2021). *Wykład Listu do Galatów. Super primam epistolam B. Pauli ad Galatas lectura*. Ed. P. P. Roszak, E. Alarcón y M. J. Janecki. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.