ETIQUETACIÓN AUTOMÁTICA DE ROLES SEMÁNTICOS EN FRAMENET ESPAÑOL

MARC ORTEGA GIL
Universidad Autónoma de Barcelona

RESUMEN

El proyecto FrameNet Español ha desarrollado un lexicón basado en un corpus que contiene un parte significativa del vocabulario del español actual, basándose en la teoría de marcos semánticos (Fillmore 1982, 1985). El objetivo es crear un recurso on-line del español basado en la semántica de marcos y en la información textual del corpus y documentar las posibles combinaciones semánticas y sintácticas de cada palabra y sus posibles significados, mediante la anotación asistida de ejemplos oracionales. En este artículo se pretende aproximar el problema de la anotación parcial de roles semánticos en español a partir en los datos del proyecto FrameNet Español.

Palabras clave: semántica de marcos, etiquetación semántica automática, autómatas finitos, transductores, análisis sintáctico, análisis semántico

ABSTRACT

Spanish FrameNet Project has developed a corpus-based lexicon for a significant portion of the vocabulary of present-day Spanish in terms of Frame Semantics (Fillmore 1982, 1985). The Spanish FrameNet Projetct is creating an on-line lexical resource for English, based on frame semantics and supported by corpus evidence. The aim is to document the range of semantic and syntactic combinatory possibilities of each word in each of its senses, through computer-assisted annotation of example sentences. In this paper we approach the problem of partial automatic semantic role assignment in Spanish introducing a testing environment based on data from the Spanish FrameNet project.

Keywords: frame semantics, automatic semantic role assignment, finite-state machine, chunking, tagging

1. INTRODUCCIÓN

El proyecto FrameNet Español[1] (SFN) (Subirats 2009) ha desarrollado un lexicón basado en un corpus que contiene un parte significativa del vocabulario del español actual, basándose en la teoría de marcos semánticos (Petruck 1996). El objetivo de FrameNet Español es realizar un análisis semántico del léxico dentro de la semántica de marcos y la gramática de construcciones (Goldberg 1995). SFN pretende documentar las posibles combinaciones semánticas y sintácticas de cada palabra y sus posibles significados, mediante la anotación asistida de ejemplos oracionales y la generación automática de los resultados. Actualmente el conjunto de datos de SFN está formado por 10.270 oraciones anotadas, subcorporadas en 1.047 unidades léxicas que pertenecen a 308 marcos semánticos.

La etiquetación automática de roles semánticos tiene aplicaciones en otros campos del procesamiento del lenguaje natural como la extracción de información (Surdeanu et al. 2003), *question answering* (Narayanan y Harabagiu 2004) y traducción automática, y es un paso importante hacia el análisis de la información textual.

En este artículo se quiere mostrar una aproximación al problema de la etiquetación semántica automática a partir del trabajo que se está desarrollando dentro del proyecto FrameNet Español. Para ello se utiliza el conjunto de oraciones anotadas en SFN como corpus de entrenamiento de la herramienta Shalmaneser (Erk y Pado 2006). Esta herramienta es una aplicación estadística que permite la asignación automática de roles semánticos en oraciones anotadas sintácticamente. El resultado puede ser analizado y revisado gráficamente mediante la herramienta SALTO (Burchardt et al. 2006).

2. CORPUS DE ENTRENAMIENTO

2.1 Anotación semántica en FrameNet Español

El corpus de entrenamiento se construye a partir de la base de conocimiento del proyecto FrameNet Español. De ella se extraen las oraciones con anotación sintáctica y semántica las cuales, después de ser procesadas, se utilizan para entrenar a la herramienta Shalmaneser.

El proceso de construcción de la base de datos que integra el proyecto FrameNet Español implica tres tareas distintas:

(1) identificar los marcos semánticos a los que pertenecen las unidades léxicas estudiadas y determinar los argumentos semánticos que forman parte de dichos marcos;

(2) determinar, a partir de un corpus, cuáles son las construcciones sintácticas en las que se proyectan los argumentos semánticos de los marcos identificados en (1);

(3) anotar semánticamente las oraciones extraídas automáticamente del corpus.

En la semántica de marcos una unidad léxica (LU) es un par formado por una palabra y su significado. Durante tarea (1) cada unidad léxica evoca un marco semántico y focaliza los elementos o aspectos del marco. Cada uno de los marcos definidos en FrameNet es una estructura conceptual que describe un tipo particular de situación, objeto, evento u otros roles. Cada uno de los roles definidos en un marco semántico se denomina elemento de marco, *frame element* (FE) (Johnson, et al. 2002).

En la segunda tarea (2) el lingüista determina, basándose en las evidencias que aparecen en el corpus, cómo las construcciones que acompañan al predicado (LU) se relacionan con los roles semánticos del marco. Para cada una de estas construcciones se define un transductor que la formaliza. El conjunto de transductores se aplica sobre todas las oraciones del corpus en las que aparece el predicado (en un proceso de transducción similar al que se describe en la sección 3.1). Este proceso consigue separar y clasificar, en diversos subcórpora, las oraciones del corpus que contienen el predicado, según la construcción o construcciones que le acompañan.

(3) De estos subcórpora se extraen de forma aleatoria un subconjunto de 30 oraciones que se cargan en la base de datos para ser anotadas semánticamente. La anotación de estas oraciones en FrameNet es un proceso semiautomático en el que el anotador etiqueta los constituyentes oraciones con uno de los roles semánticos (FE) definidos en el marco. El resultado de esta etiquetación es un conjunto de tripletas que muestran la realización de un rol semántico. Cada tripleta contiene un rol semántico del marco, como por ejemplo

Experiencer, una función gramatical, por ejemplo, objeto directo, y una función sintáctica, como por ejemplo grupo nominal. La etiquetación de las funciones gramaticales y sintácticas se realiza de forma supervisada validando, por parte del anotador, la etiquetación sugerida por el sistema.

2.4 Construcción del corpus de entrenamiento

El conjunto de todas las oraciones anotadas semánticamente en FrameNet Español, a partir del proceso descrito anteriormente, se utiliza para construir el corpus de entrenamiento para la herramienta Shalmaneser. Estas oraciones se extraen de la base de datos en un formato XML intermedio que contiene las tripletas de anotación de cada constituyente anotado y la información léxica de las formas que las componen. Actualmente este conjunto está compuesto por 10.270 oraciones. Cada una de estas contiene un marco semántico asociado a una de sus unidades léxicas y la proyección de los correspondientes roles semánticos sobre los constituyentes oracionales.

Una de las características del proceso de aprendizaje de Shalmaneser es que la información sintáctica debe estar jerarquizada, en forma de árbol oracional. En cambio, el análisis sintáctico de las oraciones en FrameNet no está jerarquizado. Este análisis es superficial y parcial, ya que no se etiquetan todos los constituyentes, únicamente aquellos sobre los que se proyecta un rol semántico. Otra de las características de los datos generados en FrameNet es que no se guardan los lemas que son necesarios en el proceso el aprendizaje. Esto requiere un preproceso previo de las oraciones que se extraen de la base de datos para que puedan ser utilizadas por Shalmaneser. En este proceso previo se lematizan cada una de las formas de la oración y se transforma el análisis sintáctico parcial no jerarquizado en un árbol oracional en el que no se añade más información sintáctica a la que ya tienen las oraciones anotadas en la base de datos.

A partir de estas oraciones Shalmaneser genera un entorno de experimentación en el que, de manera probabilística, aprende la asignación de marcos para cada predicado conocido y la asignación de los roles semánticos a sus constituyentes dentro del marco semántico asignado.

3 ETIQUETACIÓN AUTOMÁTICA DE ROLES SEMÁNTICOS

La construcción modular de Shalmaneser (Erk y Padó, 2006), apoyada sobre un formato XML de intercambio de información entre los diversos módulos que lo forman, permite su adaptación a diversos tipos de análisis y lenguas. Shalmaneser se desarrolló dentro del proyecto SALSA II[2] para realizar análisis semánticos basados en la teoría marcos semánticos y poder utilizarse en otros proyectos derivados de FrameNet[3].

El entorno de experimentación de Shalmaneser se divide en tres pasos:

1) Análisis sintáctico y lematización
2) Asignación de marcos semánticos
3) Asignación de roles semánticos

3.1 Análisis sintáctico

Shalmaneser permite utilizar distintos analizadores sintácticos y, entre otros soporta el analizador *Collins* (Collins, 1997) y *Minipar* (Lin, 1993) para el inglés y el analizador *Sleepy* (Dubey, 2005) para el alemán. En el entorno de experimentación de SFN se ha optado por utilizar un análisis léxico y sintáctico propio basado en técnicas de estado-finito que permite aprovechar todas las herramientas de análisis léxico del español desarrolladas en SFN.

El proceso de análisis sintáctico de una oración se divide en dos partes: (1) análisis léxico de los elementos de la oración y (2) detección de los constituyentes oracionales y construcción del árbol oracional.

En la tarea (1) la etiquetación léxica permite lematizar y añadir información morfológica a los elementos oracionales. A cada unidad léxica se le asigna su correspondiente clase de palabra, lema y propiedades de flexión, en el caso de los nombres, adjetivos y verbos. En este análisis se reconocen tanto formas simples como locutivas. En el caso de las formas locutivas se reconocen tanto aquellas que se encuentran en el diccionario, p. ej., *bomba atómica*, como aquellas que no se pueden reconocer mediante el diccionario debido a que

requieren un análisis sintáctico, como p. ej., pasar revista, que se puede encontrar en la forma *pasándo[les siempre] revista*.

El proceso de etiquetación léxica se realiza en dos etapas. En la primera se lleva a cabo un análisis léxico a partir de un diccionario electrónico compuesto por 600.000 formas, simples y compuestas, como el adverbio *por la noche* que se etiqueta como la unidad léxica *por/la/noche* (Cf. Figura 1) . Este diccionario se representa como un transductor que se aplica sobre cada oración del corpus y como resultado genera el autómata finito determinista de la oración, cuyas transiciones recogen la etiquetación de cada unidad léxica y permiten representar de forma eficiente la ambigüedad generada durante el proceso de análisis.

Durante la segunda etapa se aplican sobre el autómata de la oración los 2.000 transductores léxicos que formalizan las propiedades sintácticas de las locuciones verbales y de los tiempos verbales compuestos en español. De este modo en el ejemplo anterior, *pasándo[les siempre] revista*, se consigue reconocer la unidad verbal *pasar/revista*, asignarle la información morfológica correspondiente y separar los elementos no locutivos para obtener *pasando revista [a ellos] siempre* (Subirats y Ortega 2000).

Una de las características de este proceso de transducción léxica es que los transductores no añaden elementos al autómata oracional. El proceso solo elimina y reorganiza elementos (estados y transiciones) cuando se transducen unidades léxicas locutivas; y elimina transiciones cuando la transducción de elementos locutivos permite eliminar ambigüedades.

(2) Una vez realizado el análisis léxico se aplica sobre el autómata de la oración el conjunto de gramáticas que permite detectar las construcciones sintáctica y generar el árbol oracional. Cada una de estas gramáticas se formaliza como un transductor sintáctico y el conjunto de transductores se organiza como una cascada de niveles de transducción sintáctica, que se aplica de forma secuencial. A diferencia de los transductores léxicos utilizados en el proceso anterior, los transductores sintácticos no modifican los elementos presentes (estados y transiciones con información léxica) en el autómata oracional. Estos solo añaden transiciones con información sintáctica entre los estados del autómata transducido.

Cada nivel de transducción puede contener uno o más transductores que permiten detectar construcciones sintácticas que se relaciona de forma recursiva. De este modo el nivel inferior utiliza únicamente la información léxica presente en el autómata de la oración y permite etiquetar las construcciones básicas: fechas, expresiones numerales, grupos nominales simples, grupos adverbiales, etc. Esta primera transducción sintáctica modifica el autómata de la oración añadiendo transiciones, etiquetadas con la información correspondiente a la construcción detectada, entre los estados del autómata que delimitan dichas construcciones. Cada una de estas nuevas transiciones oculta los elementos que contiene la construcción que representa. Por ejemplo el grupo nominal *Los inmigrantes subsaharianos* del ejemplo de la Figura 1, de modo que en el siguiente nivel de análisis solo es visible un elemento *NP*, que representa este grupo nominal.

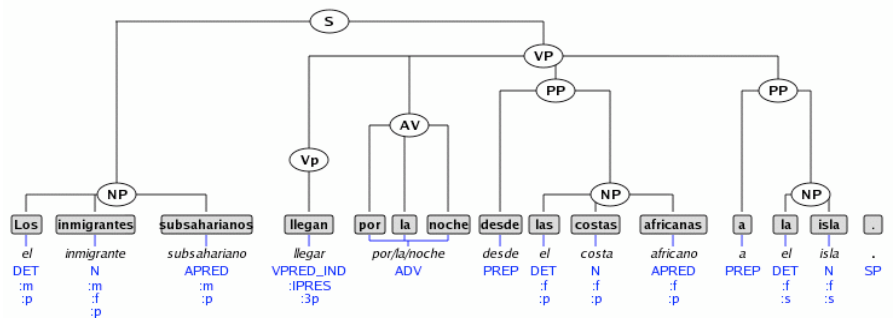

Figura 1. Análisis sintáctico de la oración *Los inmigrantes subsaharianos llegaron por la noche desde las costas africanas a la isla*

En los siguientes niveles la transducción utiliza la información léxica correspondiente a los elementos léxicos que quedan fuera de las construcciones detectadas en el/los nivel/es anterior/es y la información sintáctica añadida anteriormente, para detectar y etiquetar nuevas construcciones que se definen de forma recursiva a partir de las construcciones detectadas previa mente. En el ejemplo de la Figura 1 puede verse como para el grupo preposicional, *PP*, *a la isla* el transductor que lo detecta procesa los elementos *a NP*, de forma que no puede acceder al interior del grupo nominal.

Este proceso permite que en cada etapa disminuya la complejidad del análisis ya que el autómata oracional tiene menos elementos visibles. El nivel superior es el nivel oracional.

El analizador permite generar distintos niveles de análisis sintáctico produciendo como resultado un análisis superficial de la oración, en el que solo se detectan constituyentes sin utilizar la concordancia entre ellos , o un análisis parcial que permite generar el árbol oracional. Shalmaneser requiere este árbol oracional pero, como se comentará posteriormente, el hecho de que el corpus de entrenamiento solo tenga un análisis superficial puede ocasionar errores en la asignación de roles.

3.2 Asignación de marcos semánticos

FRED es el módulo de Shalmanser que se encarga de asignar el marco semántico correspondiente a los predicados aprendidos en el corpus de entrenamiento. Se trata de un sistema de desambiguación semántica supervisada que utiliza ventanas que pueden ser de una o más oraciones, y bigramas y trigramas centrados en el predicado.

En el entorno de experimentación de SFN esta desambiguación se limita al entorno de una sola oración. El índice de acierto que se consigue actualmente en la asignación de los marcos semánticos es del cien por cien ya que, en el estado actual del desarrollo de SFN, no hay demasiados casos de ambigüedad predicado-marco semántico.

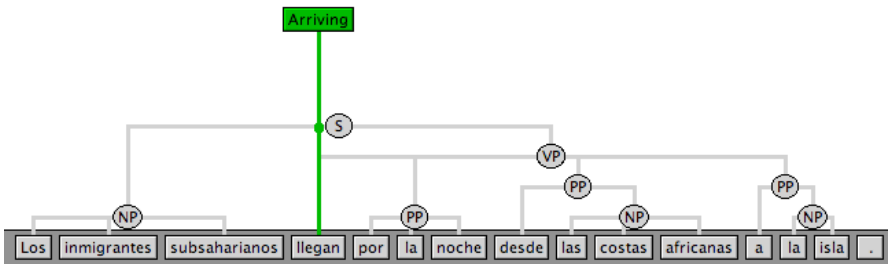

Figura. 2. Asignación del marco semántico del verbo *llegar* a partir del análisis de la oración de la Figura 1.

3.3 Asignación de roles semánticos

El módulo ROSY asigna roles semánticos a los constituyentes oracionales que forman parte del contexto de un predicado al que, previamente, se le ha asignado el marco semántico correspondiente.

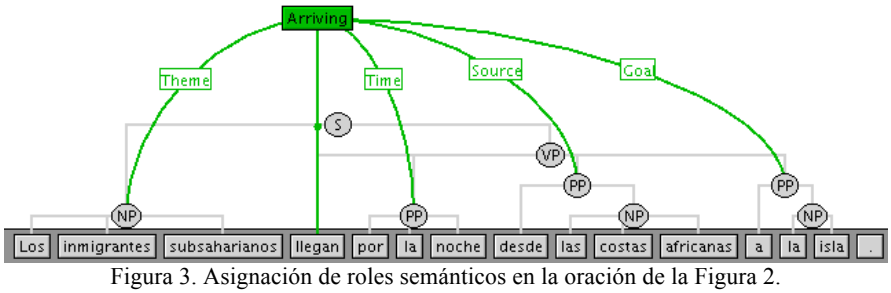

Figura 3. Asignación de roles semánticos en la oración de la Figura 2.

De los distintos constituyentes que forman parte del contexto del predicado solo unos pocos puede estar asociados a un rol semántico. Este hecho dificulta las asignación de roles. ROSY utiliza la estructura del árbol oracional para facilitar la clasificación y la asignación de roles.

3.4 Evaluación

Uno de los principales problemas de este experimento es el tamaño del corpus de entrenamiento que se extrae de SFN. 10.270 oraciones repartidas entre 1.047 unidades léxicas proporciona una media de 10 oraciones anotadas por unidad léxica. Por lo tanto la variedad de casos constituyente/rol semántico es baja para un sistema de aprendizaje automático. Este hecho afecta de forma significativa a los constituyentes preposicionales ya que Shalmaneser utiliza la preposición que los encabeza para su clasificación. En las figuras 4 y 5 pueden verse dos ejemplo de asignación de roles para el predicado *correr* que ilustran este problema. En la Figura 4 a la oración preposicional hasta subir al autobús se le asigna correctamente el rol *Result*, como resultado de la acción. En cambio en la oración de la Figura 5 el constituyente hasta que alcanzaron la meta se etiqueta como *Goal* (objetivo) cuando debería recibir el rol *Result*. En general

la precisión en la etiquetación de roles obtenida en los primeros experimentos es de 0.68, en parte debido al tamaño del corpus de entrenamiento.

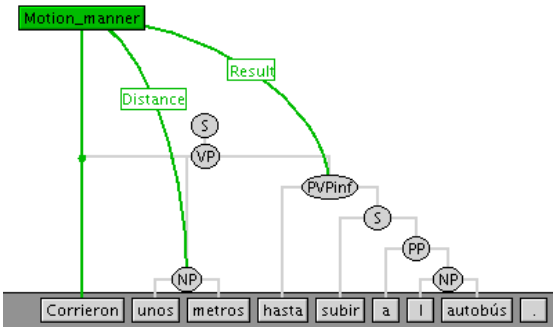

Figura 4. Asignación de roles semánticos en la oración: *Corrieron unos metros hasta subir al autobús.*

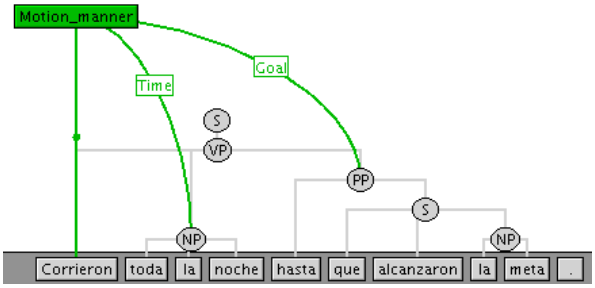

Figura 5. Asignación de roles semánticos en la oración: *Corrieron toda la noche hasta que alcanzaron la meta.*

El análisis sintáctico superficial que contienen las oraciones del corpus de entrenamiento también afecta a los resultados cuando la asignación se realiza sobre oraciones que contienen un análisis sintáctico completo. Actualmente se están desarrollando las herramientas que permitirán incorporar un análisis sintáctico completo a las oraciones del corpus de entrenamiento de modo que se conserve la información de los constituyentes etiquetados en la base de datos, conservando su información sobre las funciones sintácticas y la etiquetación de roles semánticos.

4 CONCLUSIONES

En este artículo se ha mostrado el proceso de construcción de un entorno de experimentación basado en la herramienta Shalmaneser que permite la asignación automática de roles semánticos a partir de los datos generados en el proyecto FrameNet Español.

El entorno de experimentación que se está desarrollando tiene como objetivo crear un sistema de asignación automática de roles semánticos en español que, a su vez, permita realimentar los datos de SFN permitiendo incorporar más oraciones al corpus de entrenamiento mediante un sistema supervisado que permita corregir, utilizando la herramienta SALTO, la anotación automática de los roles semánticos.

Este entorno de experimentación también debe servir como paso previo al desarrollo de un sistema de anotación semántica automática de textos, en el que la asignación de roles no se restrinja a una oración. Este aspecto es fundamental en lenguas como el español en las que la elisión del sujeto es muy frecuente.

NOTAS

[1] FrameNet Español (SFN) se está desarrollando en la Universidad Autónoma de Barcelona con la cooperación del International Computer Science Institute (ICSI). SFN está financiado por el Ministerio
de Ciencia e Innovación (MICINN, FFI2008-0875) y la Fundación Comillas
[2] http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index
[3] http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index

REFERENCIAS BIBLIOGRÁFICAS

Burchardt, A., Erk, K.; Frank, S., Kowalski, A., Pado, S., and Pinkal; M. 2006. SALTO – a versatile multi-level annotation tool. En *Proceedings of LREC 2006*, Genoa, Italy.

Collins, M. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of ACL/EACL 1997*, pág, 16–23, Madrid, España.

Dubey, A. 2005. What to do when lexicalization fails parsing German with suffix analysis and smoothing. En *Proceedings of ACL 2005*, Ann Arbor, Michigan.

Erk, K. 2005. Frame assignment as word sense disambiguation. En *Proceedings of IWCS 2005*, Tilburg, Holanda.

Erk, K., Pado, S. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. En *Proceedings of LREC-06*, Genoa.

Fillmore, C.J. 1982. Frame Semantics. En The Linguistic Society of Korea, ed. Linguistics in the Morning Calm. Seoul: Hanshin.

Fillmore, C.J. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica*, Vol 6, No. 2, pp. 222-254.

Fillmore, Charles J.; Kay, Paul; O'Connor. Catherine. 1988. Regularity and idiomaticity in grammatical constructions: The case of *Let alone*. *Language* 64/3: 501-538.

Goldberg, Adele. 1995. *Constructions. A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Johnson, C.R., Fillmore, C.J., Petruck, M.R.L., Baker, C.F, Ellsworth, M., Ruppenhofer, J., and Wood, E.J. 2002. FrameNet:

Theory and Practice. International Computer Science Institute, Technical Report-02009. Berkeley, CA.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Lin, D. 1993. Principle-based parsing without overgeneration. En *Proceedings of ACL-93*, Columbus, OH, USA.

Narayanan, S., y Harabagiu, S. 2004. Question answering based on semantic structures. En Proceedings of COLING 2004, 693–701.

Petruck, Miriam R. L. 1996. Frame Semantics. In J. Verschueren, J.-O. Östman, J. Blommaert y C. Bulcaen, eds. *Handbook of Pragmatics*. Ámsterdam / Philadelphia: John Benjamins. http://framenet.icsi.berkeley.edu/papers/miriamp.FS2.pdf

Subirats, C., Ortega, M. 2000. Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores. *Estudios de Lingüistica del Español* 10: http://elies.rediris.es/elies10/

Subirats, C. 2004. FrameNet Español. Una red semántica de marcos conceptuales. VI International Congress of Hispanic Linguistics, Leipzig University (Alemania).

Subirats, C.; Petruck, Miriam R.L. 2003. Surprise: Spanish FrameNet! En Proceedings of the International Congress of Linguists, Praga

Subirats, C. 2009. Spanish FrameNet: A Frame Semantic analysis of the Spanish lexicon. En Hans Boas, ed. *Multilingual FrameNets in Computational Lexicography*. New York/Berlin: Mouton de Gruyter.

Surdeanu, M., Harabagiu, S., Williams, J. y Aarseth, P. 2003. Using predicate-argument structures for information extraction. En Proceedings of ACL 2003, 8–15.