© Comunicación Social: Lingüística, Medios Masivos, Arte, Etnología, Folclor y otras ciencias afines. ISBN: 978-959-7174-36-3
Centro de Lingüística Aplicada, Santiago de Cuba, 2019

PAOLA CUTUGNO*
MELISSA FERRETTI**
LUCIA MARCONI*
Istituto di Linguistica Computazionale "Antonio Zampolli", Unità Organizzativa di Supporto di Genova, Consiglio Nazionale delle Ricerche*
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Sede secondaria di Genova, Consiglio Nazionale delle Ricerche**
Genova, Italy
{paola.cutugno | lucia.marconi}@ilc.cnr.it, melissa.ferretti@ieiit.cnr.it

Construcción, exploración y análisis en un corpus de producciones escolares

Bases del conocimiento para llevar a cabo estudios lingüísticos en el proyecto "Città Educante"

La metodología desarrollada para la extracción de conocimiento a partir de datos lingüísticos, requiere primero la existencia y/o la realización de una base de conocimiento, corpus, sobre la que elaborar el análisis estadístico y lingüístico.

La creación de corpus estructurados que contengan los trabajos producidos por estudiantes y la inserción de datos en estructuras informativas accesibles, que también contienen los formularios con el consentimiento informado de los padres para la recogida, análisis y publicación de los datos y las autorizaciones necesarias, es ciertamente un elemento clave del método.

El uso de nuevas tecnologías por parte de los jóvenes tanto en el ámbito extracurricular como en el escolar puede favorecer la creación de un *"Repositorio de producciones escolares de los estudiantes"*. Por un lado, el repositorio podría proporcionar una base de información completa sobre la cual llevar a cabo análisis psicológicos, sociológicos y lingüísticos por parte de todos aquellos involucrados en la educación y/o el bienestar de los estudiantes, que podría llegar a ser con la revisión y selección de los datos por parte de los docentes, una base de información a la que los propios niños podrían acceder para reutilizar los materiales existentes, actualizar y / o crear nuevos materiales y luego generar nuevos conocimientos. Los corpus además tienen una función fundamental para reconocer, analizar y clasificar un fenómeno lingüístico dado, para llevar a cabo estudios sobre la evolución del lenguaje, para contribuir a la construcción de herramientas de análisis lingüístico mediante la recopilación de información a partir de datos empíricos.

Actividades llevadas a cabo en la escuela

Las actividades se referían a la creación de una base de datos que consistía en las composiciones de los estudiantes destinados a la construcción de un corpus de temas históricos artísticos. El trabajo involucrado vio la participación de cinco clases de la Escuela Secundaria "Bernardo Strozzi" del Istituto Comprensivo de Genova Quarto. El objetivo fue recoger los temas desarrollados por los estudiantes para construir un corpus dedicado a las palabras utilizadas por los estudiantes y relacionadas con el patrimonio cultural; del corpus se extrajeron las palabras más utilizadas por los chicos. Las actividades se han distinguido en varias fases.

La primera fase, preliminar, previó algunas reuniones con los docentes participantes para definir la estructuración de las fases posteriores en sinergia. En esta fase también se eligió la imagen para los niños y la elección recayó en la pintura al óleo titulada *"Vista fantástica de los principales monumentos de Italia"*, realizada en 1858 por Petrus Henricus Theodor Tetar Van Elven y actualmente guardada en la Galería de Arte moderno de Génova Nervi (Fig. 1). Esta pintura ha sido elegida porque presenta numerosos monumentos italianos que se suponía, habrían sido parcialmente reconocidos por los estudiantes de forma independiente sin la intervención de profesores o investigadores de CNR. En esta etapa, se preparó un comunicado para que los padres de los niños firmaran al final de esta primera fase junto con la cláusula de exención de responsabilidad. También se produjo un módulo en el que se solicitó a las familias que proporcionaran información estadística útil y referida a: clase de pertenencia (Primera, Segunda, Tercera), nacionalidad, país de nacimiento, sexo, hijo único (sí o no), nacionalidad de los padres, grado de instrucción de los padres, profesión de los padres.

Durante la segunda fase, en las clases involucradas en la actividad pudieron observar la imagen elegida en la fase previa por primera vez. Se pidió a los estudiantes que describieran esta imagen con solo la información que se podía inferir de su visión. No se proporcionó información adicional, como el título de la pintura o el nombre de su autor. Esta elección se hizo para dejar libertad de expresión a los niños que podrían haber descrito la imagen libremente, sin ser influenciados por las opiniones de profesores o investigadores. Esta fase también incluyó la recopilación por parte de los maestros de los temas desarrollados por los estudiantes que fueron tratados anónimamente. Se asignó un código alfanumérico a cada alumno, que fue adoptado en el otro tema realizado posteriormente, como se da a conocer más adelante. Los temas despojados de elementos de identificación, pero aún en formato papel, fueron retirados por el personal de ILC-CNR, leídos y archivados en carpetas especiales.

Fig. 1 Petrus Henricus Theodor Tetar van Elven, 1858
Fantástica vista de los principales monumentos de Italia

Fig. 2 *Fantástica vista de los principales monumentos de Italia* dividida en sectores rectangulares

Los investigadores de ILC-CNR han visitado la escuela varias veces para revisar la imagen junto con los estudiantes y describirla con ellos. Solo en esta fase los estudiantes recibieron información sobre el título de la pintura, el nombre de su autor, la fecha de realización, etc. Además, optamos por responder algunas preguntas que nos hubieran permitido profundizar en el contexto histórico en el que se pensó y realizó el trabajo: ¿Quién lo hizo? ¿Por qué? ¿Quién lo había comprado? Lo que sucedió en Italia en la segunda mitad del siglo XIX, etc. El trabajo fue descrito en detalle y se decidió describir la pintura al proceder con sectores verticales rectangulares que permitieron enfocar más la atención en porciones individuales del espacio, como se puede ver en la figura que se muestra (Fig. 2). Para cada sección de la pintura se han descrito los monumentos representados, y también se les ha dado información sobre la historia de los mismos. Por último, se señaló cómo la imagen parece dividirse en dos partes: una dedicada a la belleza construida por el hombre y la otra dedicada a la belleza natural, ambos declarados patrimonio nacional que deben protegerse, como lo dicta la segunda parte del Artículo 9 de la Constitución de la República Italiana que dice *"La República [...] protege el paisaje y el patrimonio histórico y artístico de la Nación"*. A continuación, se pidió a los estudiantes que volvieran a trabajar una segunda composición teniendo en cuenta la información recibida. Los temas desarrollados fueron recopilados por los profesores y entregados a los investigadores de ILC-CNR que verificaron, en particular, si todos los temas eran anónimos y si los códigos alfanuméricos atribuidos a los alumnos correspondían con los de los temas anteriores. Todos los temas de los estudiantes fueron digitalizados.

Estructuración, procesamiento de datos y objetivos

Se creó un corpus compuesto por las composiciones de los alumnos de la escuela y, una vez digitalizados los temas, se organizaron por clase y en experiencia previa y posterior. El corpus realizado consta de 222 composiciones, subdivididas en 114 referidas al período previo a la experiencia y 108 en referencia al período posterior. La participación fue constante tanto para la primera como para la segunda prueba en casi todas las clases.

Escuela "Bernardo Strozzi"	*n. temas Classe 1D*	*n. temas Classe 1E*	*n. temas Classe 2A*	*n. temas Classe 2E*	*n. temas Classe 3E*	*n. temas total*
1° composición	22	23	25	24	20	114
2°composición	22	22	25	19	20	108

Tab. 1 Número de composiciones realizadas por los alumnos en la 1a y 2a prueba

Las composiciones se sometieron al proceso de lematización y los documentos lematizados se organizaron como texto individual. El objetivo general del estudio fue investigar si las palabras usadas por los estudiantes en las composiciones habían cambiado después de la experiencia hecha; para luego analizar un conjunto de palabras extraídas de las composiciones más frecuentes que caracterizan el tema tratado. Los objetivos más específicos fueron identificar las posibles diferencias en las partes del discurso de los textos producidos por los estudiantes y al mismo tiempo tratar de establecer en qué rango de valores se ubica la relación entre sustantivos y verbos.

Por ejemplo, para el corpus en las 5 clases (1D, 1E, 2A, 2E, 3E) para cada alumno se calcularon las palabras que pertenecen a las siguientes categorías gramaticales: adjetivos, adverbios, nombres y verbos para las dos composiciones y para cada escrito se calculó el número de palabras. Todos los datos fueron normalizados tanto en las dos composiciones de cada estudiante como dentro de la clase para obtener valores comparables dentro de cada clase. Por ejemplo, si analizamos la relación entre sustantivos y verbos para cada alumno y para cada clase, los resultados muestran que estos valores siempre son superiores a 1, por lo que el número de nombres

utilizados siempre es más alto que el de los verbos; de los gráficos podemos deducir cómo los valores de estas relaciones varían para cada alumno en las dos composiciones en las 5 clases (Fig. 3).

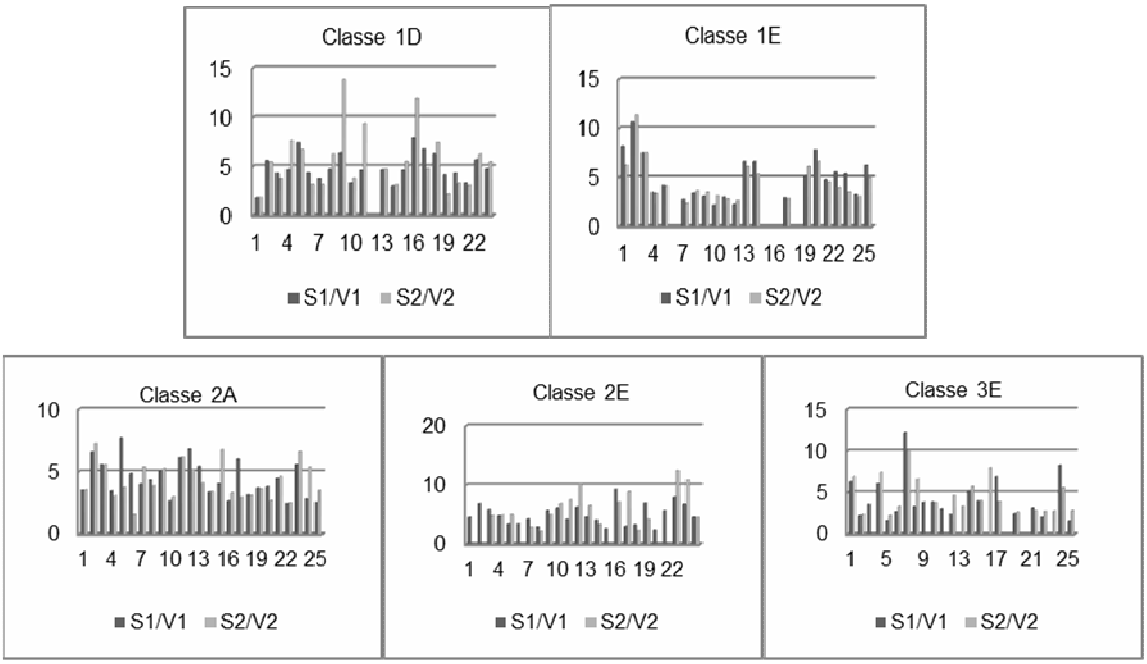

Fig. 3: Valores de las relaciones S/V para cada alumno en las dos composiciones

Para el análisis del corpus de la escuela Strozzi se utilizó, también, el software Sketch Engine[1]: un programa que permite la construcción, exploración y análisis de corpus. En particular, el programa permite, tomando como entrada un corpus, generar concordancias, bocetos de palabras, tesauros, listas de palabras, palabras múltiples, etc. Al insertar las primeras composiciones (corpus: S_1) y hacer la compilación con el software Sketch Engine, hemos obtenido, en primer lugar, información general como el número de tokens, palabras, frases, documentos e información léxica como palabras, etiquetas, etc. El corpus S_1 consta de 5 documentos, cada uno relacionado con las primeras composiciones de las cinco clases que participaron en el proyecto. S_1 tiene 931 oraciones, 23012 (token) entre palabras, puntuación, fechas, números y 20481 palabras, etc. Hemos extraído los verbos con una lista de palabras seleccionando la categoría de los verbos (VER): el verbo más frecuente es *ser* con 648 ocurrencias hasta los verbos: *saltare, capire, risaltare, scendere, rilassarsi, spargere, continuare, visitare, illuminare, spuntare, passare* cada uno con 5 ocurrencias; a continuación, la Tab.2 muestra los verbos hasta la frecuencia igual a 14.

Verbo	**Frecuencia**	**Verbo**	**Frecuencia**	**Verbo**	**Frecuencia**
essere	684	sedere	29	scorgere	17
vedere	181	infrangere	24	dipingere	17
trovare	128	guardare	24	circondare	15
rappresentare	93	raffigurare	23	pensare	15
sembrare	90	andare	22	appoggiare	14
potere	79	leggere	21	Volere	14
notare	52	osservare	20	Mettere	14
avere	47	affacciare	20	affiancare	14
stare	43	intravedere	19	ricoprire	14
fare	41	muovere	18		
piacere	35	dare	18		

Tab.2: Verbos en el corpus S_1 hasta la frecuencia 14

[1]http://www.sketchengine.co.uk/

Al seleccionar los adjetivos (ADJ) en la lista de palabras, pudimos extraer, del corpus objeto de estudio, 1973 diferentes adjetivos; a continuación, Tab. 3, muestra la lista de los adjetivos con una frecuencia mayor de 30 ocurrencias.

Adjetivo	Frecuencia	Adjetivo	Frecuencia	Adjetivo	Frecuencia	Adjetivo	Frecuencia
alto	114	marittimo	64	basso	45	nuvolo	39
piccolo	113	grande	58	scuro	41	italiano	39
primo	86	vicino	52	chiaro	41	secondo	37
bianco	76	bello	45	azzurro	41	importante	30

Tab. 3: Adjetivos en el corpus S_1 hasta la frecuencia 30

En S_1 hay 186 sustantivos, extraídos seleccionando la categoría sustantivo en la lista de palabras: el único sustantivo que supera las 250 ocurrencias es *mar* (276), hay 7 sustantivos con frecuencia entre 100 y 200 incluyendo: *quadro* (194), *monumento* (175), *montagna* (135), piano (118), *paesaggio* (115), *persona* (102) e *Italia* (100). Decidimos analizar el sustantivo *quadro (cuadro)* y en Fig. 4 hay unos ejemplos de su distribución.

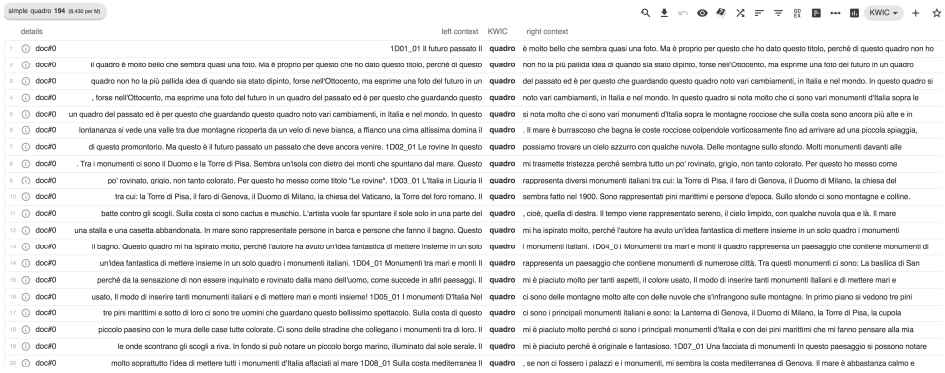

Fig. 4: distribución dentro del corpus S_1 de la palabra *"quadro"*

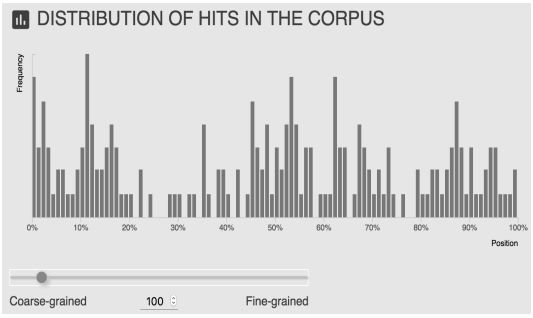

Fig. 5: Distribución de la palabra *"quadro"* en S_1

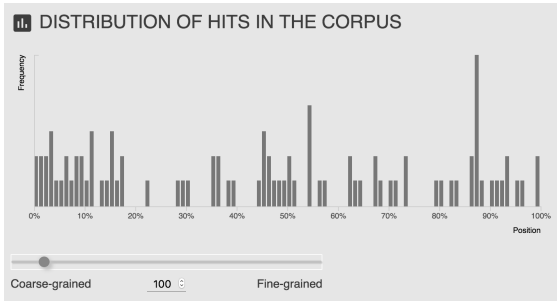

Fig.6: Distribución de "quadro" en S_1, filtro CQL, Tag: ver

Posteriormente, utilizando las diversas opciones para extraer las concordancias, hemos realizado una consulta para la palabra *"quadro"* respetando las reglas de CQL:

[lemma ="quadro"] [ ] {1,2} [tag = " VER.*"]

es decir, hemos especificado los criterios de búsqueda del nombre utilizando el Corpus Query Language descrito en el Corpus Query and Grammar Writing del software Sketch Engine. En este caso, extrajimos las concordancias del sustantivo *"quadro"* seguido por la categoría verbo en las posiciones 1 y 2 y luego también su distribución en el corpus. La Fig. 7 muestra un extracto de las 91 apariciones de las concordancias del sustantivo seguido en la posición 1 o 2 por un verbo y la Fig. 6 muestra su distribución con el mismo filtro; observe cómo las dos distribuciones de la misma palabra (Fig. 5 - Fig. 6) resultan ser diferentes debido a la inserción del filtro CQL.

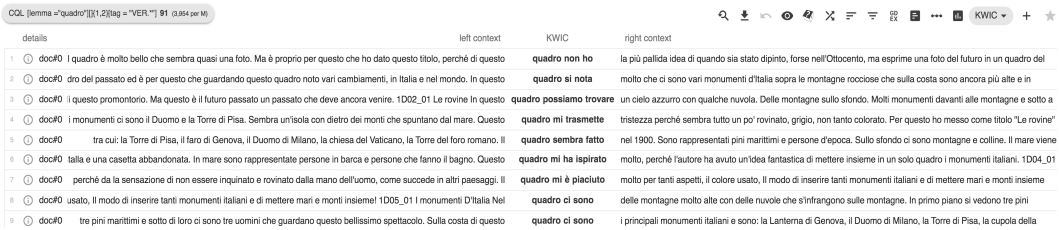

Fig. 7: Extracto de las concordancias de la palabra *"quadro"* seguido de un verbo en posición 1 o 2.

Por último, para el sustantivo *"quadro"* extrajimos el Word Sketch. El programa *Word Sketch* proporciona un análisis sobre las colocaciones gramaticales y léxicas de una palabra, es decir, usamos la función que nos permite observar la palabra seleccionada con las relaciones gramaticales en las que aparece la palabra; en la Fig. 8 se puede apreciar la búsqueda con Word Sketch de la palabra *"quadro"* caracterizada por similitud con un valor igual a 0.15. Como se muestra en la Fig. 8, los verbos utilizados en el corpus S_1 relacionados con la palabra *"quadro"* y caracterizados por similitud son: *guardare, sembrare, piacere, osservare, attraversare, splendere, considerare*.

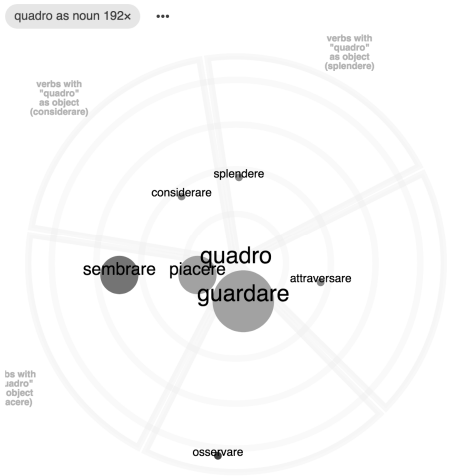

Fig. 8: verbos relacionados con "*quadro"* en S_1

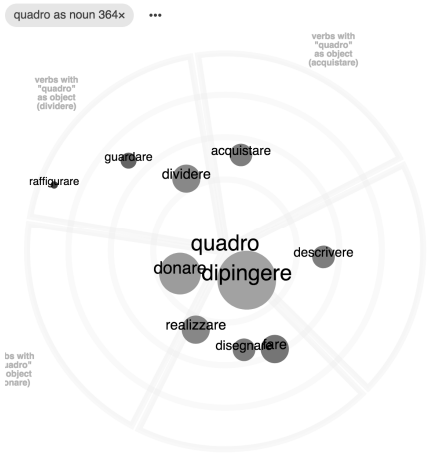

Fig. 9: verbos relacionados con "*quadro"* en S_2

Del mismo modo, todos estos estudios se llevaron a cabo para el corpus constituido por las segundas composiciones (S_2). El corpus S_2 consta de 5 documentos, cada uno relacionado con las cinco clases que participaron en el proyecto; S_2 tiene 30139 Tokens, 26141 palabras y 1038 oraciones. S_2 tiene 234 nombres diferentes y el más frecuente es el sustantivo *"quadro"* con 364 ocurrencias seguido de *"monumento"* (303). El sustantivo *"mar"*, que en el corpus S_1 fue el más frecuente, en el corpus S_2 resulta estar solo en la undécima posición, probablemente porque después de la explicación de la pintura por parte de los investigadores, los estudiantes pudieron caracterizar el objeto de estudio identificando los monumentos de las ciudades actuales.

La búsqueda con Word Sketch de la palabra *"quadro"* en el corpus S_2 caracterizada por similitud con valor igual a 0,15 (Fig. 9) muestra que los verbos son *dipingere*, *donare*, *dividere*, *realizzare*, *fare*, *acquistare*, *descrivere*, *guardare* e *raffigurare*. Comparando la Fig. 8 y la Fig 9 podemos ver la acción de los investigadores, que fueron varias veces a la escuela para revisar la imagen y describirla junto con los estudiantes.

De hecho, los verbos caracterizados por similitud de la palabra *"quadro"* en el corpus S_2 (Fig. 9) son:

donare: porque se les ha dicho que el dueño de la pintura Oddone di Savoia lo donó a la ciudad de Génova al morir;
dividere: porque para explicar mejor la imagen se ha dividido en 7 rectángulos de ancho diferente;
realizzare y *fare*: en la explicación de la gran pintura se le dijo que Petrus Henricus Tetar Theodor Van Elven quería crear una especie de pre-manifiesto de la Unidad de Italia. En este período se esperaba el logro de una identidad política unitaria y el artista Van Elven, con la realización de esta imagen, quería recordar que la belleza de las obras de arte italianas realmente unió a nuestra península durante muchos siglos.
acquistare: como Oddone de Savoia, el dueño de la pintura era un apasionado de la ciencia y el arte y su pasión lo llevó a comprar colecciones de pinturas, esculturas, conchas, algas y colibríes;
descrivere, *guardare*y *raffigurare*: fueron utilizados por los investigadores para explicar la pintura.

Conclusiones

Estudiar y analizar el idioma a partir de los textos realizados por los alumnos lleva, por un lado, a medir o detectar fenómenos lingüísticos, por otro, permite investigar tanto sus emociones como sus conocimientos.

En este trabajo ha sido posible verificar que, al observar las palabras utilizadas por los niños, analizar y comparar las composiciones realizadas antes y después de una experiencia, podemos ciertamente notar una variación lingüística. En particular, al comparar una sola búsqueda con Word Sketch, referida a la palabra *"quadro*", podemos afirmar que ha habido un cambio en el conocimiento por parte de los alumnos.

Bibliografia

Elisa Corino, Didattica delle lingue corpus-based, EL.LE, 3, 2, 2014, pp. 231-258, ISSN 2280-6792**,** art-10.14277-2280-6792-99p.pdf

Doaa Samy, Ana Fernández Pampillón Cesteros, Jorge Arús Hita,Taller sobre herramientas de análisis textual: La herramienta Sketch Engine, Junio 2011, Sketch Engine-manual_nov2011.pdf.