Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione

WORKING PAPER N. /2024

A Wordlist for Italian Grade 5 EFL Learners as a Reference Tool in the Assessment of Receptive Skills: A Preliminary Corpus-Based Investigation.

Francesca La Russa - INVALSI

https://orcid.org/0000-0002-8689-2454

Collana: Working Papers INVALSI

ISSN: 2611 - 5719

Abstract

Lexical competence is crucial for acquiring a second language and plays a significant role in developing receptive skills. In fact, several studies have shown that learners should know between 95% and 98% of the words in a text to adequately comprehend it (Laufer & Ravenhorst-Kalovski, 2010). However, determining the appropriate vocabulary for a given proficiency level is challenging, as lexical competence is influenced by an individual's daily and personal experiences with the target language. A valuable tool for English as a Foreign Language test developers is the English Vocabulary Profile (EVP), a Reference Level Description that categorizes English vocabulary according to the Common European Framework of Reference for Languages (CEFR) proficiency levels from A1 to C2. The EVP vocabulary lists are comprehensive and not specific to any age group or first language.

This study focuses on Italian grade 5 students expected to achieve A1 proficiency in English. Young learners are recognized as a distinct group, with linguistic needs and acquisition processes that may differ significantly from those of adults. Our hypothesis is that, compared to the A1 vocabulary lists in the EVP, some above-level words might nonetheless be part of the lexical repertoire of Italian young learners. To test this hypothesis a preliminary investigation was conducted on the IT_YL, a subcorpus from the Cambridge Learner Corpus containing the written productions of Italian learners aged 10–11 at the A1 proficiency level, with the aim of identifying the words that are listed above the A1 level in the EVP.

Keywords: vocabulary list, English as a Foreign Language, young learners, Italian grade 5 pupils, corpus-based research.

Introduction

Several research perspectives on language acquisition assign a fundamental role to the lexicon. According to Lewis's Lexical Approach (1993), learning a language primarily involves communicating meaning, which requires focusing on what conveys meaning the most, i.e. vocabulary. In this view, language is considered a "grammaticalized lexicon" rather than a "lexicalized grammar".

The close relationship between vocabulary knowledge and comprehension has been extensively studied and validated (Laufer, 1989, 1992; Liu & Nation, 1985; Laufer & Ravenhorst-Kalovski, 2010). Research indicates that achieving between 95% and 98% lexical coverage of a text is essential for adequate comprehension and for successfully inferring the meaning of unknown words from context. Therefore, it is essential for test developers to understand which words learners are expected to know at specific proficiency levels to accurately assess their receptive language skills.

The acquisition of lexical competence is largely shaped by an individual's personal experiences with the target language. As a result, creating a profile of the vocabulary that learners might know at different proficiency levels is particularly challenging.

Some useful guidelines on the sequence of objectives, skills, and lexical content to be developed at each level of competence are outlined in Chapter 5 of the *Common European Framework of Reference for Languages: Companion Volume* (Council of Europe, 2020). The *Vocabulary Range Scale* provides descriptors related to the range and variety of words and expressions that learners can comprehend or produce. The *Vocabulary Control Scale* focuses on the ability to select appropriate expressions.

Although the CEFR offers a comprehensive framework for describing language skills, its descriptors scales are language neutral. Further language-specific details are provided in the *Reference Level Descriptions*. For English, the *English Vocabulary Profile* (EVP)[2] fulfills this role, categorizing English vocabulary (i.e., words, fixed and semi-fixed phrases, phrasal verbs, and other multi-word units) according to CEFR proficiency levels ranging from A1 to C2. As stated by Capel (2010), levels are assigned not to the words themselves but to their individual meanings.

The EVP draws its data from various sources: the *Cambridge International Corpus* (CIC) containing a billion words of written and spoken English from diverse sources; the *Cambridge Learner Corpus* (CLC), which consists of over 40 million words written by students from more than 150 different nationalities; the *Cambridge English Lexicon* (Hindmarsh, 1980) and wordlists from course books and other teaching materials. The result of this huge research project is a list of 15696 entries distributed across 6 CEFR levels (from A1 to C2) and 22 topics (e.g., food and drink, arts and media, body and health, etc).

The EVP wordlists are comprehensive and not specific to any age group or first language (L1), however. both age and L1 can influence learners' lexical repertoire. As stated by Capel (2010:10): "the [EVP] Wordlists are recommended for anyone dealing with learners aged 11 and upwards; for young learners, a different grouping of words/senses seems inevitable".

This study focuses on Italian young learners completing the first cycle of education who, according to the national curriculum guidelines (*Indicazioni Nazionali*, 2012), are expected to achieve A1 proficiency in English as a Foreign Language (EFL). Our goal is to extend the EVP A1 wordlist by including words that, while classified as above level, may still be part of the vocabulary of young Italian learners.

In the following sections, we will first provide a brief overview of the relevant literature on lexical competence and wordlists for young learners. We will then outline the aims and hypothesis that guided our study, along with an explanation of the methods of analysis employed. Finally, the results will be presented and discussed, followed by reflections on the study's limitations and potential directions for future research.

1. Young learners' lexical competence

Young learners (YLs) can be defined as "those who are learning a foreign or second language and who are doing so during the first six or seven years of formal schooling" (McKay, 2006:1). This group typically includes children aged approximately five to twelve years, who are in primary or elementary school.

YLs are recognized as a distinct group, with linguistic needs and acquisition processes that may differ significantly from those of adults (Benigno & De Jong, 2016). Therefore, creating a functional and age-appropriate L2 vocabulary list tailored to this audience is essential.

Corpus-based research[1] has greatly expanded our understanding of vocabulary acquisition in both L1 and L2 contexts (Schmitt, 2010; Wray, 2009). However, studies on L2 acquisition primarily focus on older learners and research involving young L2 learners remains limited (see Hestetræet, 2018, for an overview).

Evidence suggests that YLs acquire vocabulary in chunks, associating words with concrete objects and often learning indirectly through exposure to spoken or written language (Cameron, 2001; Wray, 2002). Questions remain about which words they need to know and in what quantity.

As it concerns vocabulary size, research indicates that with knowledge of the 3000 most frequently occurring words, learners can achieve 95% comprehension of spoken English, making this a crucial learning goal for beginner EFL learners (Dale & Chall, 1948; Nation, 2001; Schmitt & Schmitt, 2014; Staehr, 2008).

[1] Corpus based research uses large collections of written and/or oral texts (corpora) produced by native speakers (L1 corpora) and/or L2 learners (L2 corpora) to investigate language patterns and usages systematically (Sinclair, 1991).

However, as Callies et al. (2009) points out, many words are polysemous, and different meanings of a single word may be acquired at varying stages of language development. Consequently, using the lemma[2] or the word family[3] as unit of count could be misleading. Instead, focusing on word meanings aligns more closely with the notion of vocabulary acquisition as a contextual and gradual process that goes from basic and useful units to more complex and specialized ones (Wolter, 2009).

As for the acquisition rate, while five-year-old children are expected to acquire approximately 1000 new words per year, EFL young learners can realistically learn 300–500 words annually (Nation, 1990; Cameron, 2001; Orosz, 2009). To progress from A1 to A2 level, they would need to add around 1500 words, and roughly 3000 words to reach B1 level (Milton, 2010)

Another crucial factor is selecting the age-appropriate vocabulary. As noted, high-frequency words are essential for children that start learning EFL vocabulary. However, frequency cannot be the only criterion for selecting target vocabulary. As noted by Callies et al. (2009:4), "learning is functional and theme-driven and some infrequent words are important for communicability". This is particularly true for YLs who use a great deal of vocabulary from specific domains, such as games, sports, school and so on. Therefore, the target vocabulary should be meaningfull, relevant and in line with children's communicative needs so as to promote their motivation and support learning (Hestetræet, 2018). Additionally, age-appropriate vocabulary should align with children's cognitive development. According to Cameron (1994, 2001), children learn basic-level concepts (e.g., chair, dog) before acquiring more specific lower-level concepts (e.g., rocking chair, spaniel) or more general higher-level concepts (e.g., furniture, animal).

Finally, evidence suggests that children learn vocabulary in chunks—prefabricated groups of words that commonly occur together and are likely stored and retrieved as whole units from memory (Lewis, 1993; Wray, 2000; Nation, 2013). These formulaic sequences help children develop fluency and provide them with ready-made phrases for expressing meaning in social interactions, such as satisfying their material needs, interacting at school, engaging with peers during play, and supporting their language learning.

To sum up, a wordlist designed for young learners should include high-frequency words and chunks that align with their cognitive development and communicative needs.

[2] A lemma is the base form of a word that appears as an entry in a dictionary and is used to represent all the other inflected forms. For example, "run" is the lemma for the words "runs," "ran," and "running."

[3] The concept of word family refers to a word and its main inflections and derivations (Hatch & Brown,1995). All the words in the same family share common meaning, e.g. "inform", "information", "informative", "informer", "informed", etc.

1.1 EFL wordlists for young learners

Although research on the L2 vocabulary of YLs is less extensive compared to research on adults, several EFL wordlists for young learners do exist.

One of these lists is the *Global Scale of English (GSE) Vocabulary for Young Learners*[4] (Benigno & De Jong, 2017), which is a structured lexical framework tailored for EFL learners aged 6 to 11, comprising approximately 3000 entries. Each entry corresponds to a word meaning and is accompanied by a definition, an indication about the topic, the grammatical category, the GSE value and CEFR level (from PreA1 to B1), and an example sentence. To account for both high frequency and low frequency words, which are nonetheless representative of YLs' language, entries were sourced from various corpora, including the *British National Corpus* (BNC), the *Child Language Exchange Data System* (CHILDES, MacWhinney, 2000), and *SUBTLEX-UK* (van Heuven et al., 2014), as well as from teaching materials designed for children, such as coursebooks, flashcards, the *Seward 4K Teaching List*, and the Dale-Chall list. Subsequently, a group of 18 primary teachers from diverse countries evaluated the communicative usefulness of each word meaning. A weighted measure that integrates both quantitative (frequency) and qualitative (teacher evaluations) criteria was utilized to rank the lexical items and assign them to the corresponding CEFR level.

Another list that is worth mentioning here is the CEFR-J wordlist (Negishi et al., 2013; Tono, 2013; Negishi & Tono, 2016). Initially developed from a frequency analysis of English textbooks used in primary and secondary schools across Japan and neighboring Asian countries (e.g., China, Korea, and Taiwan), the wordlist was later compared with the EVP to incorporate missing words.

Finally, several major ESOL certification bodies have published wordlists outlining the vocabulary that test takers are expected to know. While not all of these lists are derived from empirical research and should therefore be interpreted with caution, the words they contain appear annually in tests taken by large numbers of young learners, suggesting that these learners may be familiar with them. Examples of these lists include:

- Cambridge Starters (PreA1), Movers (A1) and Flyers (A2) wordlists[5] that present words in alphabetical order, along with indications of the corresponding parts of speech.
- Cambridge A2 Key for Schools Wordlist[6] which identifies receptive and productive vocabulary necessary for test takers. Initially based on the vocabulary from the Council of Europe's *Waystage 1990* specification and other high-frequency words supported by corpus evidence, the list is regularly updated to reflect current language use through analysis of the CLC and the EVP. It includes both

[4] https://www.english.com/gse/teacher-toolkit/user/vocabulary
[5] https://www.cambridgeenglish.org/images/149681-yle-flyers-word-list.pdf
[6] https://www.cambridgeenglish.org/images/506886-a2-key-2020-vocabulary-list.pdf

single words and multi-word verbs, with usage examples provided where necessary to clarify meaning. Additionally, a thematic list organizes words under categories like "Food and Drink," "House and Home," and so on.

- LanguageCert Young Learners Fox (Pre-A1) and Owl (A1) wordlists[7] which provide teachers and candidates with an alphabetic list of words for the ESOL tests. Each entry is accompanied by the part of speech and usage information to clarify meaning.
- the vocabulary List of Pearson English International Certificate for Young Learners[8] that lists in alphabetical order the vocabulary typically tested at GSE levels 1 to 4.

These vocabulary lists are all designed for young learners and may effectively reflect their lexical repertoire, making them a valuable reference tool for teachers, learners, researchers, and test developers. However, none of these lists specifically targets YLs from particular national backgrounds. YLs of English as a Foreign Language generally have minimal exposure to the target language outside the school environment (Benigno & De Jong, 2016). Consequently, their lexical repertoire is likely shaped by the vocabulary presented in class and, hence, by the specific national context in which English is learned. Additionally, typological differences between languages can impact ease of acquisition, making certain words easier or harder to learn depending on a learner's L1. For instance, an Italian learner may find the word *important* easy to learn due to its similarity to the Italian word *importante*, whereas the same word may be difficult for learners whose L1 is typologically more distant from English (e.g., Chinese or Russian). Thus, further research is needed on the lexical repertoires of YLs from diverse nationalities and L1 backgrounds, including Italian.

2. Aims

The underlying hypothesis of this study is that, in addition to the 786 entries on the EVP A1 vocabulary list, there may be above level words and phrases that could nonetheless form part of the vocabulary of Italian young learners.

Accordingly, this study aims to enhance the EVP lists through a preliminary corpus-based investigation by incorporating words that, although categorized above the A1 level, appear in the language production of Italian young learners and may therefore be part of their lexical repertoire.

3. Methods

[7] https://www.languagecert.org/en/language-exams/english/languagecert-young-learners-esol/pre-a1-fox-281234

[8] https://qualifications.pearson.com/content/dam/pdf/pearson-test-of-english/pte-young-learners/pte-yl-vocabulary-list.pdf

Given that learner corpora that collect large samples of written and/or spoken learner productions are a valuable resource for investigating learner language features (cf. §1), this methodology was adopted to gain insights into the language—particularly the vocabulary—of Italian EFL young learners at A1 level. Therefore, the first step was to identify a learner corpus containing language samples from the target learners.

Numerous corpora[9] collecting language samples from learners of English as a Foreign or Second Language having diverse L1 backgrounds have been developed. However, due to the difficulties in collecting language samples from younger learners, only a few corpora include YLs' productions. With no claim of being exhaustive, we can mention: the *Trinity Lancaster Corpus* (TLC, Gablasova et al. 2019), the *International Corpus of Crosslinguistic Interlanguage* (ICCI, Tono, 2012), the *Cambridge Learner Corpus* (CLC), the *Corpus of Young Learner Interlanguage* (CYLIL, Housen, 2002), the CHILDES English-L2 Paradis Corpus (Paradis, 2005), the *Corpus of Young Learner Language* (CORYL, Hasselgreen, & Sundet, 2017).

Among these corpora, only the CYLIL and the CLC contain language samples from our target audience. Due to difficulties in obtaining online access to the CYLIL, and given that the CLC has been used in the development of the English Reference Level Description (i.e., the EVP), and may therefore be considered a reputable reference source, we selected the CLC as the primary data source for our study.

The CLC is a collection of thousands of examination scripts from Cambridge ESOL exams written by learners worldwide. In *Sketch Engine*[10], the CLC is divided into two main parts: the error coded learner corpus (CLC coded) that contains exam scripts that have been coded to indicate learner error, and the larger uncoded learner corpus (CLC uncoded) that contains uncoded scripts from a wider range of nationalities and first languages. To extract scripts exclusively from Italian young learners at the CEFR A1 level, a subcorpus of the uncoded CLC, referred to as IT_YL, was created. A frequency list of all lemmas in this subcorpus was then generated using *Sketch Engine*. The uncoded version of the CLC was chosen because, in the coded CLC, corrected words are counted as tokens. This approach ensured that the frequency list reflected only the learners' productions, excluding any corrections.

Finally, to identify lemmas above the A1 level, the resulting list was compared with the EVP lists using *Text Inspector*[11], a software that, among other functions, analyzes texts according to the EVP and assigns each word a CEFR level.

4. Results

The IT_YL subcorpus was created by applying the following filtering criteria to the uncoded CLC 2021:

[9] For an overview, visit https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[10] *Sketch Engine* (https://www.sketchengine.eu/) is a software for corpus analysis.

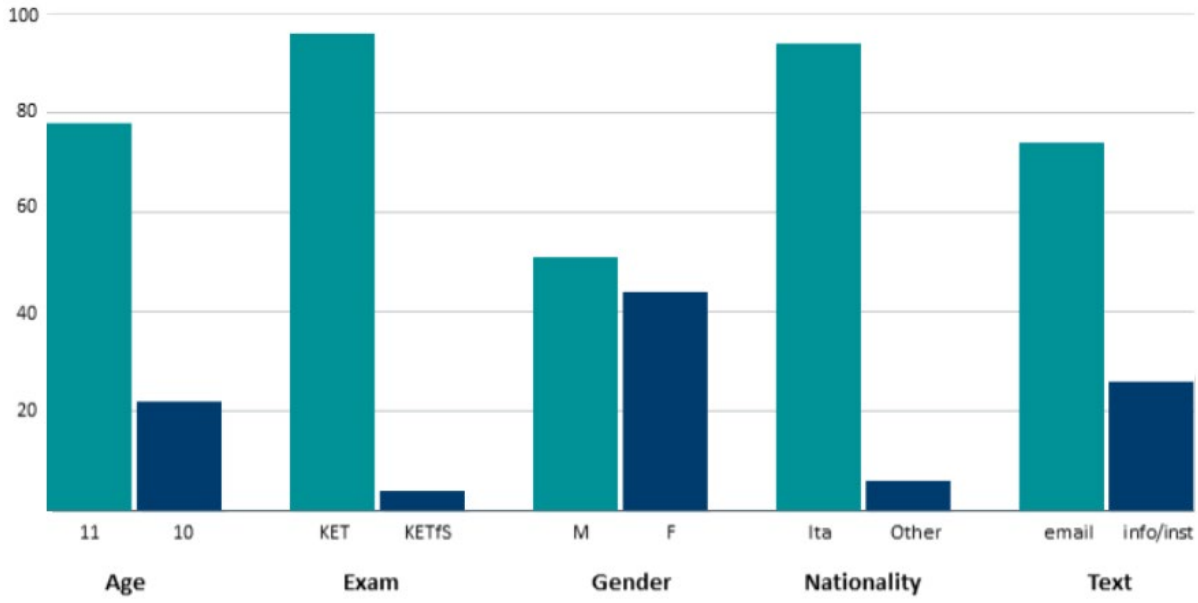
[11] https://textinspector.com/

- First language: Italian
- Age: 10 and 11 years old
- CEFR level performance: A1

The resulting subcorpus consists of 112 texts (4671 tokens and approximately 4091 words) written by 10- and 11-year-old students in response to the A2 KET and KET for Schools exams administered between 2002 and 2012. These texts are emails, notes, memos (74% of tokens); or informative/instructional texts (26% of tokens). Texts length ranges from 22 to 62 words. The majority of the texts were produced by Italian students, with a few productions made by Swiss, Spanish, Argentinian, British, Iranian and Brazilian test-takers. All written performances were classified at the A1 level of competence.

Graphic 1 shows the percentage of the IT_YL tokens covered across various parameters: age, type of exam, gender, nationality and type of text produced.

Graphic 1. IT_YL token coverage in percentage.



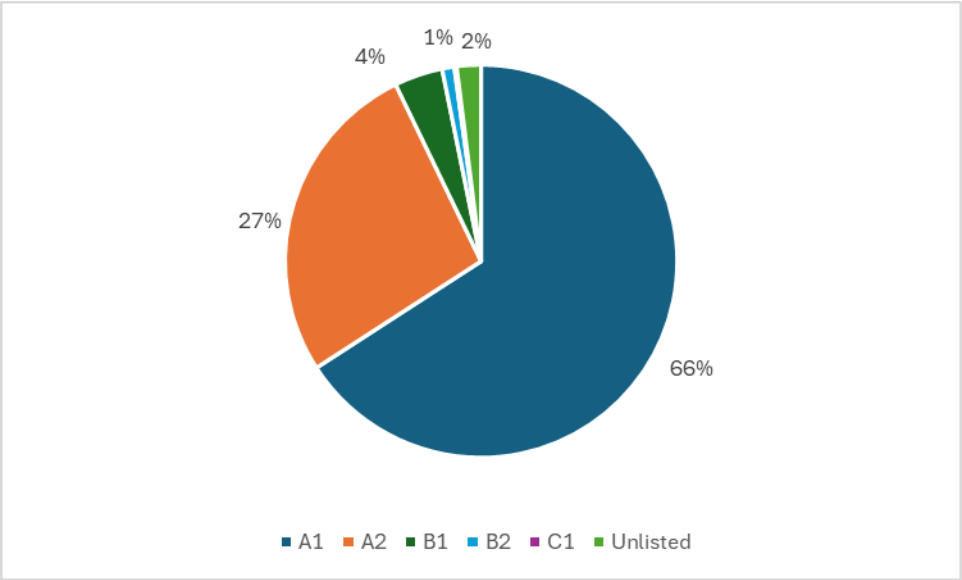
Using Sketch Engine, a frequency list of the 611 lemmas in the subcorpus was generated. Due to the limited size of the IT_YL no frequency cutoff was applied. This list was then further refined by manually excluding lemmas that are usually not present in wordlists and that are thus irrelevant to the study's objectives. Excluded words are determiners, personal pronouns, possessives, proper nouns, toponyms, and misspelled words already listed in their corrected forms. The refined IT_YL wordlist contains 433 lemmas.

This list was subsequently compared to the lexical lists of the EVP using Text Inspector to determine the proficiency level of each lemma and identify words above the A1 level. Since the EVP assigns proficiency levels to individual word meanings, an additional manual check was conducted to confirm that the EVP levels align with the intended meanings of words in the IT_YL subcorpus.

The analysis identified 288 A1 words, 10 unlisted words and 135 words above the A1 level (see Appendix 1): 115 words at the A2 level, 16 at the B1 level, 3 at the B2 level, and 1 at the C1 level.

Graphic 2 displays the percentage distribution of lemmas in the IT_YL wordlist across different CEFR levels.

Graphic 2. CEFR level of the lemmas in the IT_YL wordlist.



5. Discussion and conclusions

This preliminary corpus-based study reveals that, although the vast majority of words produced by Italian YLs fall within the A1 proficiency level (66%), some words from higher levels do appear in their productions. This supports the hypothesis that Italian YLs may know a handful of words beyond the A1 level.

Most of the higher-level words are at the beginner A2 level (see Graphic 2), with only a small number at intermediate (B1 and B2) or advanced levels (C1, with no C2 words observed). Notably, the higher-level words in the IT_YL wordlist (e.g., *laser* at B2; *compliment* at C1) are identical or closely resemble their Italian counterparts (*laser*, *complimenti)*. This suggests that typological similarities between languages aid vocabulary acquisition.

However, this study is only a preliminary investigation and has limitations that affect the generalizability of its findings. The primary limitation is the small size of the IT_YL subcorpus. While we might reasonably

assume that some of the more frequent A2 words (e.g., *cost* or *bring*, which appear 34 times in the IT_YL) are likely part of the lexical repertoire of Italian YLs, this assumption does not extend to words at higher CEFR levels, which occur far less frequently in our subcorpus (B1, B2, and C1 words appear only once or twice). Consequently, further research using larger corpora is needed to determine if these higher-level words truly belong to the lexical repertoire of our target population.

Moreover, while evidence suggests that YLs acquire vocabulary in chunks (cf. §1), our current wordlist include only single words. Future research may be conducted to incorporate the phraseological dimension.

Furthermore, frequency data can be biased by the nature of the corpus and therefore be misleading. For instance, the topics of exam tasks assigned to learners can influence the lexical items they produce. As a result, if certain topics are missing, learners are unlikely to have used related vocabulary, which would then be absent from our wordlist. Furthermore, since the CLC only includes learners' written productions, words more commonly used in spoken language may also be missing from the corpus and the resulting wordlist.

Additionally, the texts in the IT_YL subcorpus date back several years (2002 to 2012), raising concerns about their relevance to current language use. Language evolves rapidly, as does the world around us; just consider the vocabulary related to new technologies that have become part of our daily lives. To accurately reflect the lexical repertoire of today's Italian YLs, an investigation based on more recent texts is needed.

As discussed in § 3, corpora that collect the language production of Italian YLs are limited. Some of these limitations could be addressed by developing a larger and up-to-date corpus of Italian young learners of English in the future.

Finally, our preliminary investigation is based on a corpus of learners' written productions. The question of the potential disparity between receptive and productive vocabulary has been a topic of research for many years (e.g., Melka, 1997). It is commonly assumed that receptive vocabulary is larger than productive vocabulary and that reception precedes production. To gain a more comprehensive understanding of the lexical repertoire of our target audience, future research could involve comparing our wordlist with the wordlists from widely used EFL primary school textbooks. This would provide insight into the vocabulary learners are exposed to in class that potentially reflects their receptive vocabulary. In future research, the wordlist might be expanded to incorporate this additional dimension.

Further validation of the resulting vocabulary list could include administering a survey to Italian primary school EFL teachers regarding the communicative usefulness of the words included, as well as conducting a vocabulary test with Italian EFL learners in primary schools to assess their understanding of the target words.

In summary, this study represents the very first step in a longer and more complex research project, and there is still much work ahead to achieve a comprehensive representation of the lexical repertoire of Italian young learners.

Acknowledgement: we would like to express our gratitude to Cambridge University Press & Assessment for granting us access to the Cambridge Learner Corpus, without which this research would not have been possible.

References

Benigno, V., & De Jong, J. (2016). The "global scale of English learning objectives for young learners": a CEFR-based inventory of descriptors. In: Nikolov M. (ed.) *Assessing young learners of English: Global and local perspectives* 43-64. New York, NY: Springer.

Benigno, V., & De Jong, N. (2017). *Global Scale of English (GSE) vocabulary for young learners*. Pearson.

Callies, M., Benigno, V., & Hober, N. (2019). Developing a CEFR-based vocabulary inventory for young learners of English: Comparing native-speaker and learner corpus data. In *Widening the Scope of Learner Corpus Research.* Selected papers from the fourth Learner Corpus Research Conference. Corpora and Language in Use–Proceedings (Vol. 5).

Cameron, L. (1994). Organizing the world: children's concepts and categories, and implications for the teaching of English. *ELT Journa*l, 48, 28-39.

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge University Press.

Capel, A. (2010). A1–B2 Vocabulary: Insights and Issues Arising from the English Profile Wordlists Project. *English Profile Journal*, 1, 1-11.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Council of Europe Publishing.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 27*(1), 11–28.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research*, *5*(2), 126-158.

Hasselgreen, A., & Sundet, K. T. (2017). Introducing the CORYL Corpus: What it is and how we can use it to shed light on learner language. *Bergen language and linguistics studies*, 7,197- 215.

Hatch, E., & Brown. S. (1995). *Vocabulary, Semantics, and Language Education*. New York: Cambridge University Press.

Hestetræet, T. I. (2018). Vocabulary teaching for young learners. In: Garton, S. & Copland, F. (eds.) *The Routledge handbook of teaching English to young learners* 220-233. Routledge.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In: Granger, S., Petch-Tyson, S., & Hung, J. (eds.) *Computer learner corpora, second language acquisition and foreign language teaching* 77-116. Amsterdam: John Benjamins.

Hindmarsh, R. (1980). *Cambridge English Lexicon*. Cambridge University Press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In: Lauren C. & Nordman M. (eds.), *Special Language: From Humans Thinking to Thinking Machines* 316-323. Multilingual Matters.

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability. *Journal of Research in Reading*, *15*(2), 95-103.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward*. Language Teaching Publications.

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal, 16*(1), 33-42.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates

McKay, P. (2006). *Assessing young language learners*. Cambridge University Press.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagog*y. Cambridge: Cambridge University Press, 84-102.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In: Barning, I. Martin, M. Vedder, I. (eds.). *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 211-232. Eurosla Monographs Series.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2013). *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In: Galaczi, E. D., & Cyril, J. W. (eds.), Exploring language frameworks, Proceedings of the ALTE Krakow Conference, 135-163.

Negishi, M., & Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. *Studies in Language Testing*, 44, 113-133.

Orosz, A. (2009). The growth of young learners' English vocabulary size. In: Nikolov, M. (ed.) *Early learning of modern foreign languages: Processes and outcomes*, 181-194. Multilingual Matters.

Paradis, J. (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with Specific Language Impairment. *Language, Speech and Hearing Services in the Schools,* 36, 172-187.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N., & Schmitt, D. (2014). A reassessment of vocabulary learning: Implications for instruction. *Language Teaching Research, 18*(1), 32-54.

Sinclair, J. (1991). *Corpus, Concordance, Collocation. Oxford.* UK: Oxford University Press.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Tono, Y. (2012). International Corpus of Crosslinguistic Interlanguage: Project overview and a case study on the acquisition of new verb co-occurrence patterns. In: Tono, Y., Kawaguchi, Y., & Minegishi, M. (eds.). *Developmental and crosslinguistic perspectives in learner corpus research* 27-46. John Benjamins Publishing Company.

Tono, Y. (2013). *The CEFR-J handbook: A resource book for using CAN-DO descriptors for English language teaching*. Tokyo: Taishukan.

Van Ek, J. A., & Trim, J. L. M. (1991). *Waystage 1990*. Council of Europe.

Van Heuven, W., J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British English. Quarterly *Journal of Experimental Psychology*, 67, 1176-1190

Wolter, B. (2009). Meaning-last vocabulary acquisition and collocational productivity. In Fitzpatrick, T. & Barfield, A. (eds.) Lexical Processing in Second Language Learners: Papers and Perspectives in Honour of Paul Meara 128–140. Bristol: Multilingual Matters.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. A*pplied linguistics, 21*(4), 463-489.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. 2009. Future directions in formulaic language research. *Journal of Foreign languages, 32*(6),2-17.

Appendix 1: Above A1 level words in the IT_YL subcorpus.

Lemma	Level	Part of speech	IT_YL Frequency
along	A2	preposition	1
amazing	A2	adjective	1
amazing	A2	adjective	1
another	A2	determiner	1
anymore	A2	adverb	2
arrive	A2	verb	3
art	A2	noun	5
autumn	A2	noun	1
bicycle	A2	noun	1
borrow	A2	verb	1
bottle	A2	noun	2
break	A2	noun	3
bring	A2	verb	34
brush	A2	noun	1
by	A2	preposition	12
call	A2	noun	2
card	A2	noun	1
care	A2	noun	1
centre	A2	noun	4
cloudy	A2	adjective	1
club	A2	noun	3
cola	A2	noun	1
competition	A2	noun	5
concert	A2	noun	42
cool	A2	adjective	2
cost	A2	noun	34
cupboard	A2	noun	1
decide	A2	verb	1
disco	A2	noun	1
drum	A2	noun	1
enough	A2	adverb	1

everything	A2	pronoun	2
exercise	A2	noun	1
exit	A2	noun	1
fall	A2	verb	2
fantastic	A2	adjective	1
finally	A2	adverb	1
footballer	A2	noun	1
free	A2	adjective	2
friendly	A2	adjective	1
glad	A2	adjective	1
high	A2	adjective	1
hills	A2	noun	1
hope	A2	verb	1
idea	A2	noun	2
if	A2	conjunction	16
in front of	A2	preposition	4
information	A2	noun	2
instrument	A2	noun	1
jazz	A2	noun	1
jumper	A2	noun	1
just	A2	adverb	6
kiss	A2	noun	5
large	A2	adjective	2
let	A2	verb	1
library	A2	noun	1
line	A2	noun	1
lose	A2	verb	1
lovely	A2	adjective	2
match	A2	noun	2
most	A2	adverb	1
mountain	A2	noun	1
mushroom	A2	noun	1
must	A2	modal verb	8

notebook	A2	noun	2
nothing	A2	pronoun	2
opposite	A2	preposition	5
out	A2	adverb	1
over	A2	adverb	1
per	A2	preposition	2
perfect	A2	adjective	2
plastic	A2	adjective	1
police	A2	noun	1
pool	A2	noun	4
pop	A2	noun	2
pound	A2	noun	3
practice	A2	noun	1
prepare	A2	verb	1
price	A2	noun	2
rap	A2	noun	2
receive	A2	verb	1
return	A2	verb	1
right	A2	adverb	2
rock	A2	noun	3
rubber	A2	noun	7
ruler	A2	noun	1
scarf	A2	noun	2
sell	A2	verb	1
shall	A2	modal verb	1
sheet	A2	noun	1
should	A2	modal verb	6
simple	A2	adjective	1
size	A2	noun	1
snack	A2	noun	2
so	A2	adverb	3
sock	A2	noun	2
sofa	A2	noun	1

soft	A2	adjective	1
song	A2	noun	1
straight	A2	adjective	1
sweater	A2	noun	3
telephone	A2	noun	1
thank	A2	verb	2
theatre	A2	noun	1
toy	A2	noun	1
trainer	A2	noun	1
trip	A2	noun	1
turn	A2	verb	3
type	A2	noun	1
underground	A2	noun	2
upstairs	A2	adverb	1
volleyball	A2	noun	1
way	A2	noun	1
win	A2	verb	1
wish	A2	noun	2
altogether	B1	adverb	1
central	B1	adjective	1
collection	B1	noun	1
competitor	B1	noun	1
direction	B1	noun	1
disappear	B1	verb	1
gun	B1	noun	1
horror	B1	noun	1
organize	B1	verb	2
regard	B1	noun	1
reply	B1	noun	1
ski	B1	noun	2
statue	B1	noun	1
stripe	B1	noun	1
total	B1	adjective	1

tournament	B1	noun	1
client	B2	noun	1
laser	B2	noun	1
theme	B2	noun	2
compliment	C1	verb	1