DARC-IT: a DAtaset for Reading Comprehension in ITalian

Dominique Brunato[◇], **Martina Valeriani**[•], **Felice Dell'Orletta**[◇]
[•] University of Pisa
marti.valeriani@gmail.com
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

English. In this paper, we present DARC-IT, a new reading comprehension dataset for the Italian language aimed at identifying 'question-worthy' sentences, i.e. sentences in a text which contain information that is worth asking a question about[1]. The purpose of the corpus is twofold: to investigate the linguistic profile of question-worthy sentences and to support the development of automatic question generation systems.

Italiano. *In questo contributo, viene presentato DARC-IT, un nuovo corpus di comprensione scritta per la lingua italiana per l'identificazione delle frasi che si prestano ad essere oggetto di una domanda*[2]*. Lo scopo di questo corpus è duplice: studiare il profilo linguistico delle frasi informative e fornire un corpus di addestramento a supporto di un sistema automatico di generazione di domande di comprensione.*

1 Introduction

Reading comprehension (RC) can be defined as "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (Snow, 2002). Such a definition emphasizes that RC is a complex human ability that can be decomposed into multiple operations, such as coreference resolution, understanding discourse relations, commonsense reasoning and reasoning across multiple sentences. In educational scenarios, student's comprehension and reasoning skills are typically assessed through a variety of tasks, going from prediction tasks (e.g. cloze test) to retellings generation and question answering, which are costly to produce and require domain expert knowledge. Given also the challenges posed by the broad diffusion of distance learning programs, such as MOOC (Massive Open Online Courses), the automatic assessment of RC is becoming a rapidly growing research field of Natural Language Processing (NLP). While much more work has been done on developing Automated Essay Scoring (AES) systems (Passonneau et al., 2017), recent studies have focused on the automatic generation of questions to be used for evaluating humans' reading and comprehension (Du and Cardie, 2017; Afzal and Mitkov, 2014). This is not a trivial task, since it assumes the ability to understand which concepts in a text are most relevant, where relevance can be here defined as the likelihood of a passage to be worth asking a question about. The availability of large and high-quality RC datasets containing questions posed by humans on a given text thus becomes a fundamental requirement to train data-driven systems able to automatically learn what makes a passage 'question-worthy'. In this regard, datasets collected for other NLP tasks, Question Answering above all, provide a valuable resource. One of the most widely used is the Stanford Question Answering Dataset (SQuAD), (Rajpurkar et al., 2016). It contains more than 100,000 questions posed by crowdworkers on a set of Wikipedia articles, in which the answer to each question is a segment of text from the corresponding reading passage. More recently, other large RC datasets have been released: it is the case of the 'TriviaQA' dataset (Joshi et al., 2017), which is intended to be more challenging than SQuaD since it contains a higher proportion of complex ques-

[1] The corpus will be made publicly available for research purposes at the following link: http://www.italianlp.it/resources/

[2] Il corpus sarà messo a disposizione liberamente per scopi di ricerca al seguente indirizzo: http://www.italianlp.it/resources/

tions, i.e. questions requiring inference over multiple sentences. The same holds for RACE (Lai et al., 2017), which is also the only one specifically designed for educational purposes. Indeed it covers multiple domains and written styles and contains questions generated by domain experts, i.e. English teachers, to assess reading and comprehension skills of L2 learners. While all these datasets are available for the English language, to our knowledge, no similar RC datasets exist for the Italian language. In this paper we introduce a new corpus for Italian specifically conceived to support research on the automatic identification of question-worthy passages. In what follows, we first describe the typology of texts it contains and the annotation process we performed on them. We then carry out a qualitative analysis based on linguistic features automatically extracted from texts with the aim of studying, on the one hand, which features mostly discriminate question-worthy sentences from other sentences and, on the other hand, whether the two classes of sentences have a different profile in terms of linguistic complexity.

2 Dataset Collection

The first step in the process of corpus construction was the selection of appropriate materials. As noted by Lai et al. (2017), a major drawback of many existing RC datasets is that they were either crowd-sourced or automatically-generated thus paying very little attention to the intended target user; this makes them less suitable to be used in real educational scenarios. To prevent these limitations, we relied on a corpus of reading comprehension tests designed by the National Institute for the Evaluation of the Education System (INVALSI), which is the Italian institution in charge of developing standardized tests for the assessment of numeracy and literacy skills of primary, middle and high school students.

To create the corpus, we focused only on tests designed to assess students' competences in the Italian language. We thus collected a total of 86 Italian tests administered between 2003 and 2013, of which 31 targeting primary school's pupils of the second, third and fifth grade, 29 targeting students of the first and third year of middle school and 26 targeting students of first, second and third grade of high school. To each text a number of questions is associated, which aim to deeply assess student's ability of reading and understanding. As documented by the last available technical report provided by the Institute[3], the INVALSI Italian test has been designed to cover seven main aspects underlying text comprehension, namely: understanding the meaning of words; identifying explicit information; inferring implicit information; detecting elements conveying cohesion and coherence in text; comprehending the meaning of a passage by integrating both implicit and explicit information; comprehending the meaning of the whole text; generating a meaningful interpretation (e.g. understanding the message, the purpose etc.). With respect to their form, questions can be either multiple-choice (typically with 3 or 4 options, see example (1)) or, more rarely, open-ended questions (example 2).

Example (1): *Dove abita il ragno del racconto?* (Where does the spider of the story live?)
A. In un albero del bosco. (On a forest tree)
B. Sopra un fiore del bosco. (Upon a forest flower)
C. In una siepe del bosco. (In a forest hedge)

Example (2): *Dopo aver letto il testo, qual è secondo te il messaggio che vuole dare l'autore?* (After reading the text, what do you think is the message the author wants to give?)

For the purpose of our study, we selected only the first type of questions, thus obtaining a total of 354 questions. Table 1 reports some statistics about the final corpus collected from the INVALSI tests.

SchoolGrade	Texts	Sentences	Questions
2nd Primary	10	195	75
4th Primary	9	205	36
5th Primary	12	427	50
1st Middle	19	513	72
3rd Middle	10	342	48
1st High	10	303	32
2nd High	7	211	18
3rd High	9	261	23
TOT	86	2457	354

Table 1: Total number of texts, total number of sentences and corresponding questions for each school grade in DARC-IT.

[3] http://www.invalsi.it/invalsi/doc_eventi/2017/Rapporto_tecnico_SNV_2017.pdf

2.1 Annotation Scheme

For each question of the corpus, the annotation process was meant to identify the sentence (or a sentence span) containing the corresponding answer. This information was marked on text by enclosing the relevant text span in opening and closing *xml* tags with a letter R in upper case.

The outcome of the annotation process was a tabular file with the following information reported in separate columns: i) the text segmented into sentences; ii) a binary value 1 vs 0 (1 if the sentence contains the answer to the question and 0 if not); iii) the corresponding question; iv) the answer provided by the annotator. Table 2 gives an example of the dataset structure.

A qualitative inspection of the corpus allowed identifying different typologies of 'question-worthy' sentences: sentences that were the target of one question only (this is the case of the second sentence reported in Table 2); sentences that were the target of multiple questions, such as (4), and sentences that only partially answered the question (i.e. the whole information required to give the answer is spread across multiple sentences), such as (5).

(4) Question-worthy sentence: *Leo decide di aiutare gli animali della giungla* (*Leo decided to help the jungle animals*)

Corresponding questions:

- Qual è la cosa più importante per Leo? (What is the most important think to Leo?)

  Multiple choice answers: A. Essere un bravo cacciatore. (To be a good hunter); B. Diventare il piú coraggioso di tutti. (To become the bravest of all); C. Rendersi utile agli altri. (To make himself useful to others); D. Fare nuove esperienze. (To make new experiences).

- Cosa sceglie di fare Leo nella giungla? (What does Leo choose to do in the jungle?)

  Multiple choice answers: A. Giocare con tutti. (To play with everybody); B. Dormire e mangiare. (To sleep and eat); C. Aiutare chi è in difficoltà. (To help people in need); D. Nuotare nell'acqua del fiume (To swim in the river water)

(5) Question-worthy sentences: *"Io farò il postino!" Disse uno. "Io farò il maestro!" Disse un altro. "E io farò lo chef!". Urlò un terzo e salì sul vagone delle marmellate.* (I'm going to be a postman! One said. I'm going to be a teacher! Another said. And I'm going to be a chef! Shouted a third one and went up on the wagon of the jams).

Corresponding question: A che cosa pensano i bambini quando vedono gli oggetti sul treno? (What do children think when they see the items on the train?)

Multiple choice answers: A. Ai giochi che potranno fare. (To the plays they can do); B. A cose utili che si possono vendere. (To useful things that can be sold); C. Ai regali che vorrebbero ricevere. (To the presents they would like to receive); D. Ai lavori che faranno da grandi. (To the trades they will do as adults.)

3 Linguistic Analysis

As a result of the annotation process, we obtained 398 'question-worthy' sentences and 2059 'non-question' worthy sentences. Starting from this classification we carried out an in-depth linguistic analysis based on a wide set of features capturing properties of a sentence at lexical, morpho–syntactic and syntactic level. The aim of this analysis was to understand whether there are some linguistic features that mostly allow predicting the 'likelihood' of a sentence to be the target of a question. To allow the extraction of linguistic features, all sentences were automatically tagged by the part-of-speech tagger described in (Dell'Orletta, 2009) and dependency parsed by the DeSR parser described in (Attardi et al., 2009).

Table 3 shows an excerpt of the first 20 features (of 177 extracted ones) for which the average difference between their value in the 'question-worthy' and 'non question-worthy' class was highly statistically significant using the Wilcoxon rank sum test[4]. As it can be seen, sentences on which a comprehension question was asked are on average much more longer. This could be expected since the longer the sentence the higher the probability that it is more informative and thus containing concepts that are worth asking a question about. This is also suggested by the higher distribution of proper nouns [10], most likely referring to relevant semantic types (e.g. person, location) which typically occur in Narrative, i.e. the main textual genre of the Invalsi tests. The higher sentence length of 'question-worthy' sentences has effects also at morpho-syntactic and

[4] All significant features are shown in Appendix (A).

Sentence	Class	Tag	Question	Answer
La lucciola si preparò e, quando calò la sera, andò all'appuntamento.	0			
Entrò nel bosco scuro e raggiunse la siepe dove viveva il ragno.	1	Entrò <R>nel bosco scuro e raggiunse la siepe dove viveva il ragno.<\R>	Dove abita il ragno del racconto?	In una siepe del bosco.

Table 2: Sample output of the dataset structure.

syntactic level, as shown e.g. by the higher proportion of conjunctions introducing subordinate clauses ([7] *Subord. conj*: 1.63 vs 1.50) and by the presence of longer syntactic relations in which the linear distance between the 'head' and the 'dependent' is higher than 10 tokens ([20] *Max link*: 11.30 vs 8.30).

	Question		**NoQuestion**	
Features	**Avg**	**(StDev)**	**Avg**	**(StDev)**
Raw Text features				
[1] Sentence length*	29.00	(16.11)	20.00	(13.75)
Morpho–syntactic features				
[2] Punctuation*	4.74	(2.82)	7.70	(6.23)
[3] Negative adv*	1.23	(2.82)	1.19	(3.13)
[4] Coord. conj*	3.50	(3.40)	3.20	(3.81)
[5] Poss. adj*	0.96	(2.10)	0.89	(2.33)
[6] Relative pron*	1.14	(2.00)	1.12	(2.32)
[7] Subord. conj*	1.63	(2.80)	1.50	(2.90)
[8] Prepositions*	7.90	(5.01)	7.60	(6.20)
[9] Determiners*	9.13	(5.00)	9.00	(6.20)
[10] Proper nouns*	2.05	(3.90)	2.00	(4.30)
[11] Numbers	0.66	(1.87)	0.64	(2.25)
[12] Verbs	15.98	(6.32)	16.97	(8.18)
[13] Indicat. mood*	57.00	(30.70)	60.00	(33.82)
[14] Particip. mood	7.13	(14.22)	6.34	(14.88)
[15] 3^{rd}pers. verb*	55.15	(39.50)	45.20	(42.62)
[16] Conjunctions	5.1	(4.35)	4.34	(4.66)
Syntactic features				
[17] Clause length*	8.63	(4.34)	7.90	(4.24)
[18] Verbal heads*	4.00	(2.30)	3.00	(2.03)
[19] Postverb Subj*	13.60	(27.00)	15.70	(32.00)
[20] Max link*	11.30	(7.06)	8.30	(6.80)

Table 3: Linguistic features whose average difference between the two classes was statistically significant. For each feature it is reported the average value (avg) and the standard deviation (StDev). All differences are statistically significant at p<.005; those with * also at p<.001. (Note: Question=question-worthy sent.; NoQuestion=Non question-worthy sent.)

A further analysis was meant to investigate the profile of question-worthy sentences with respect to linguistic complexity. To this end, we exploit READ-IT (Dell'Orletta et al., 2011), a general-purpose readability assessment tool for Italian, which combines traditional raw text features with lexical, morpho-syntactic and syntactic information to operationalize multiple phenomena of text complexity. READ–IT assigns different readability scores using the following four models: 1) Base Model, relying on raw text features only (e.g. average sentence and word length); 2) Lexical Model, relying on a combination of raw text and lexical features; 3) Syntax Model, relying on morpho-syntactic and syntactic features; 4) Global Model, combining all feature types (raw text, lexical, morpho-syntactic and syntactic features).

Results are reported in Table 4. As it can be noted, question-worthy sentences have a higher complexity with respect to all models. Especially at syntactic level, this could be expected given the higher values obtained by features related to syntactic complexity which turned out to be significantly involved in discriminating these sentences.

	Question	NoQuestion
READ-IT Base	59,9%	21,1%
READ-IT Lexical	98,9 %	66,4%
READ-IT Syntactic	69,3%	37,5%
READ-IT Global	100%	95%

Table 4: Readability score obtained by different READ-IT models.

4 Conclusion

We presented DARC-IT, a new reading comprehension dataset for Italian collected from a sample of standardized evaluation tests used to assess students' reading and comprehension at different grade levels. For each text, we annotated 'question-worthy' sentences, i.e. sentences which contained the answer to a given question. A qualitative analysis of these sentences showed that the likelihood of a sentence to be 'question-worthy' can be modeled using a set of linguistic features, which are especially linked to syntactic complexity. We believe that this corpus can support research on the development of automatic question generation systems as well as question answering systems. Current developments go into several directions: we are carrying out a first classification experiment to automatically predict 'question-worthy' sentences and evaluate the impact of linguistic features on the classifier performance. We are also planning to enlarge the corpus and to investigate more in-depth the typology of questions and answers it contains, in order to study what characterizes sentences answering, for instance, to factual vs non-factual questions.

5 Acknowledgments

The work presented in this paper was partially supported by the 2–year project (2018-2020) SchoolChain – Soluzioni innovative per la creazione, la certificazione, il riuso e la condivisione di unità didattiche digitali all'interno del sistema Scuola, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

Naveed Afzal and Ruslan Mitkov. 2014. Automatic generation of multiple choice questions using dependency-based semantic relations *Soft Computing*, 18 (7), 1269–1281.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, Edimburgo, UK: 73–83.

Xinya Du and Claire Cardie. 2017. Identifying Where to Focus in Reading Comprehension for Neural Question Generation. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark.

Mandar Joshi, Eunsol Choi, Daniel Weld and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 1601–1611.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, Eduard H. Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark.

Rebecca J. Passonneau, Ananya Poddar, Gaurav Gite, Alisa Krivokapic, Qian Yang and Dolores Perin. 2016. Wise Crowd Content Assessment and Educational Rubrics. *International Journal of Artificial Intelligence in Education*, 28, 29–55.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, pages 2383–2392.

Catherine Snow. 2002. *Reading for understanding: Toward an RD program in reading comprehension*. Rand Corporation.

Appendix (A).

	Question-worthy sentences		**Non Question-worthy Sentences**	
Features	**Average**	**(StDev)**	**Average**	**(StDev)**
Raw Text features				
Sentence length***	29.00	(16.11)	20.00	(13.75)
Lexical features				
% Basic Italian Vocabulary (BIV)*	88.54	(8.53)	88.99	(10.66)
% Fundamental BIV**	78.26	(10.83)	79.59	(13.23)
% 'High Usage' BIV*	12.31	(8.12)	12.50	(10.28)
Lexical density*	0.56	(0.08)	0.58	(0.11)
Morpho–syntactic features				
% Adjectives*	5.20	(4.71)	4.35	(5.55)
% Articles***	9.13	(5.00)	9.00	(6.20)
% Conjunctions**	5.1	(4.35)	4.34	(4.66)
% Coordinat. conj***	3.50	(3.40)	3.20	(3.81)
% Demonstrative determiners***	0.61	(1.61)	0.55	(1.90)
% Indefinite pronouns	0.87	(2.26)	0.66	(2.24)
% Interrogative determiners*	00.5	(0.52)	0.06	(0.67)
% Interjections*	0.03	(0.31)	0.09	(0.72)
% Numbers**	0.66	(1.87)	0.64	(2.25)
% Negative adverbs***	1.23	(2.82)	1.19	(3.13)
% Ordinal numbers*	0.27	(1.04)	0.14	(0.83)
% Possessive adjectives***	0.96	(2.10)	0.89	(2.33)
% Prepositions***	7.90	(5.01)	7.60	(6.20)
% Proper nouns**	2.05	(3.90)	2.00	(4.30)
% Punctuation***	4.74	(2.82)	7.70	(6.23)
% Relative pronouns***	1.14	(2.00)	1.12	(2.32)
% Subordin. conj***	1.63	(2.80)	1.50	(2.90)
% Verbs**	15.98	(6.32)	16.97	(8.18)
% Verb_Participial mood**	7.13	(14.22)	6.34	(14.88)
% Verb_Indicative mood***	57.00	(30.70)	60.00	(33.82)
% Verb_Conditional mood**	1.37	(6.13)	2.35	(9.58)
% Verb_Past tense**	22.19	(34.80)	23.88	(37.73)
% Verb_Imperfect tense**	29.08	(39.35)	29.04	(41.13)
% Verb_Present tense*	45.04	(43.50)	38.40	(44.91)
% 3rdpers. verb***	55.15	(39.50)	45.20	(42.62)
% 2ndpers. verb*	1.37	(7.34)	1.84	(10.25)
TTR ratio (first 100 lemmas)**	0.84	(0.10)	0.89	(0.10)
Syntactic features				
Clause length (in tokens)***	8.63	(4.34)	7.90	(4.24)
Avg verbal heads/sentence***	4.00	(2.30)	3.00	(2.03)
Avg prep. links length*	1.11	(0.45)	0.93	(0.58)
Max link length***	11.30	(7.06)	8.30	(6.80)
Verb arity	34.93	(29.74)	33.37	(32.70)
% Postverbal subject***	13.60	(27.00)	15.70	(32.00)
% Preverbal objects*	10.17	(25.17)	9.22	(25.55)
% DEP Root**	5.52	(3.31)	8.20	(6.30)
% DEP Mod_rel***	1.50	(2.21)	1.30	(2.50)
% DEP Copulative Conj**	5.34	(4.92)	4.65	(5.26)
% DEP Determiner***	9.10	(5.00)	8.80	(6.20)
% DEP Disjuntive Conj	0.14	(0.76)	0.20	(0.99)
% DEP Locative Compl*	0.73	(2.03)	0.53	(1.81)
% DEP_neg***	1.20	(2.80)	1.13	(2.84)
% DEP conj**	4.58	(4.12)	3.91	(4.62)
% DEP concatenation*	0.06	(0.52)	0.08	(0.8)

Table 5: Linguistic features whose average difference between the two classes was statistically significant. For each feature it is reported the average value and the standard deviation (StDev). *** indicates a highly significant difference ($p<.001$); ** a very significant difference ($p<.01$); * a significant difference ($p<.05$).