AcCompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian

Dominique Brunato[1], Cristiano Chesi[2], Felice Dell'Orletta[1], Simonetta Montemagni[1], Giulia Venturi[1], Roberto Zamparelli[3]
[1]ILC-CNR, Via G. Moruzzi 1, Pisa, Italy
[2]NETS-IUSS, P.zza Vittoria 15, Pavia, Italy
[3]CIMeC-UNITRENTO, Corso Bettini 31, Rovereto Italy
[name.surname]@ilc.cnr.it, cristiano.chesi@iusspavia.it, roberto.zamparelli@unitn.it

Abstract

The Acceptability and Complexity evaluation task for Italian (AcCompl-it) was aimed at developing and evaluating methods to classify Italian sentences according to Acceptability and Complexity. It consists of two independent tasks asking participants to predict either the acceptability or the complexity rate (or both) of a given set of sentences previously scored by native speakers on a 1-to-7 points Likert scale. In this paper, we introduce the datasets distributed to the participants, we describe the different approaches of the participating systems and provide a first analysis of the obtained results.

1 Motivation

The availability of annotated resources and systems aimed at predicting the level of grammatical acceptability or linguistic complexity of a sentence (see, among others, (Warstadt et al., 2018; Brunato et al., 2018)) is becoming increasingly relevant for different research communities that focus on the study of language. From the Natural Language Processing (NLP) perspective, the interest has been recently prompted by automatic generation systems (e.g. Machine Translation, Text Simplification, Summarization) mostly based on Deep Neural Networks algorithms (Gatt and Krahmer, 2018). In this scenario, resources and methods able to assess the quality of automatically generated sentences or devoted to investigate the ability of artificial neural networks to score linguistic phenomena on the acceptability and complexity scales are of pivotal importance. From the theoretical linguistics perspectives, controlled datasets containing acceptability judgments and analyzed with machine learning techniques can be useful to test the extent to which syntactic and semantic deviance can be induced from corpus data alone, especially for low frequency phenomena (Chowdhury and Zamparelli, 2018; Gulordava et al., 2018; Wilcox et al., 2018), while the same data, seen from a psycholinguistic angle, can shed light on the relation between complexity and acceptability (Chesi and Canal, 2019), and on the extent to which measures of on-line perplexity in artificial language models can track human parsing preferences (Demberg and Keller, 2008; Hale, 2001).

The Acceptability & Complexity evaluation task for Italian (AcCompl-it) at EVALITA 2020 (Basile et al., 2020) is in line with this emerging scenario. Specifically, it is aimed at developing and evaluating methods to classify Italian sentences according to Acceptability and Complexity, which can be viewed as two simple numeric measures associated with linguistic productions. Among the outcomes of the task, we also include the creation of a set of sentences annotated with acceptability and complexity human judgments that we are going to share with the linguistic community. While datasets annotated for acceptability exist for English, see in particular the COLA dataset (Warstadt et al., 2018), to our knowledge the present dataset is a first for Italian, and is also the first one to combine judgments of acceptability and complexity.

2 Definition of the task

We conceived AcCompl-it as a prediction task where participants were asked to estimate the average acceptability and complexity score of a set of sentences previously rated by native speakers on a 1-7 Likert scale and, if possible, to predict the actual standard error (SE) among the annotations. SE gives an estimation of the actual agreement between human annotators: the highest the SE, the

lowest the agreement. The task is articulated in three subtasks, as follows:

- the *Acceptability prediction* task (*ACCEPT*), where participants have to estimate the acceptability score of sentences (along with their standard error); in this case, 1 corresponds to the lowest degree of acceptability, while 7 corresponds to the highest level. The assignment of a score on a gradual scale is in line with the definition of perceived acceptability that we intend to empirically inspect. According to the literature, in fact, acceptability is a concept closely related to grammaticality but with some major differences (see, among others, (Sprouse, 2007; Sorace and Keller, 2005)). While the latter is a theoretical construction corresponding to syntactic wellformedness and it is typically interpreted as a binary property (i.e., a sentence is either grammatical or ungrammatical), acceptability can depend on many factors, such as syntactic, semantic, pragmatic, and non–linguistic factors;

- the *Complexity prediction* task (*COMPL*), where participants have to estimate the complexity score of sentences (along with their standard error); in this case, 1 corresponds to the lowest possible level of complexity, while 7 indicates the highest degree. Similarly to the Acceptability prediction task, the use of the Likert scale as a tool to collect perceived values is motivated by the assumption that sentence complexity is a gradient, rather than binary, concept;

- the *Open task*, where participants are requested to model linguistic phenomena correlated with the human ratings of sentence acceptability and/or complexity in the datasets provided.

The three subtasks were independent and participants could decide to participate in any one of them, though we encouraged participation in multiple subtasks, since the complexity metrics might be influenced by the grammatical status of an expression and vice versa. In line with this intuition, we distributed a subset of sentences annotated with both acceptability and complexity scores in order to investigate whether and to what extent there is a correlation between the two phenomena.

In all subtasks, participants were free to use external resources, and they were evaluated against a blind test set.

3 Dataset

3.1 Composition

Acceptability dataset: it contains 1,683 Italian sentences annotated with human judgments of acceptability on a 7–point Likert scale. The number of annotations per sentence ranges from 10 to 85, with an average of 16.38. The dataset is constructed by merging the data of four psycholinguistic studies on minimal variations of controlled linguistic oppositions with different levels of grammaticality with a subset of 672 sentences generated from templates.

The first subset (128 sentences), taken from (Chesi and Canal, 2019), focuses on person features oppositions in object clefts dependencies where Determiner Phrases (DPs) are either introduced by determiners or by pronouns used as determiner as in (1).

(1) {Sono | siete} {gli | voi} architetti che {gli | voi} ingegneri {hanno | avete} consultato.
{are$_{3Ppl}$ | are$_{2Ppl}$} {the | you} architects that {the | you} engineers {have$_{3Ppl}$ | have$_{2Ppl}$} consulted
'it is {the|you} architects that {the|you} engineers have consulted'

The second subset (515 sentences) is taken from the studies presented in (Greco et al., 2020) involving copular constructions (e.g. canonical (2a) vs. inverse (2b) (Moro, 1997).

(2) a. Le foto del muro sono la causa della rivolta.
the pictures of_the wall are the cause of_the riot

b. La causa della rivolta sono le foto del muro.
the cause of_the riot are the pictures of_the wall

This subset also contains declarative and interrogative (yes/no) sentences with a minimal verbal structure (contrasting preverbal vs postverbal subject position in unergatives (3a), unaccusatives (3b) and transitive predicates (3c))

(3) a. I cani hanno abbaiato | Hanno abbaiato i cani.
the dogs have barked | have barked the dogs

b. Gli autobus sono partiti | Sono partiti gli
The buses have left | have left the
autobus.
buses

c. Le bambine hanno mangiato il dolce |
the girls have eaten the dessert |
Hanno mangiato le bambine il dolce
have eaten the girls the dessert

The third set (320 sentences) is based on a study in which number and person subject-verb agreement and unagreement cases are tested (Mancini et al., 2018):

(4) Qualcuno ha detto che io_{1Psg} {$scrivo_{1Psg}$ |*
Somebody has said that I_{1Psg} {$write_{1Psg}$ |
$scriviamo_{1Ppl}$} una lettera.
*$write_{1Psg}$} a letter

The fourth one (48 sentences) contains experimental items from (Villata et al., 2015) involving different types of wh-islands violations.

(5) {Cosa | Quale edificio}$_i$ ti chiedi {chi |
{What | Which building}$_i$ do you wonder {who |
quale ingegnere} abbia costruito $\__i$?
which engineer} has built $\__i$?

The last set of 672 sentences was generated by creating all the possible content word combinations from various structural templates designed to test acceptability patterns due to: **(i)** extra or missing gaps in Wh-extractions (6a) vs. topic constructions (6b).

(6) a. {Cosa | Quale problema}$_i$ lo studente
{what | which problem}$_i$ the student
dovrebbe descriver(e) {$\__i$| -lo$_i$ | questo
should describe {$\__i$| it | this
problema}?
problem}

b. Questo problema$_i$, lo studente dovrebbe
this problem, the student should
descriver(e) {$\__i$| -lo$_i$ | questo problema}
describe {$\__i$| it$_i$ | this problem}

(ii) Wh- and relative clauses with gaps inside VP conjunctions (in all conjuncts, i.e. "Across the Board", in only one conjunct, or not at all, see e.g. (7)).

(7) Chi$_i$ ... Maria vuole chiamar(e) {$\__i$| -lo} e
who$_i$ ... Mary wants $call_{inf}$ {$\__i$| him} and
il dottore medicar(e) {$\__i$| -lo}?
the doctor cure {$\__i$| him}?

(iii) embedded Wh-clauses and the possibility of subextractions from them (similar to (5)).

(8) Quale provvedimento Maria ha saputo {che |
Which measure M. has heard {that |
dove | perché | quando} il ministro prenderà?
where | why | when} il ministro prenderà?

(iv) extractions from VPs in subject vs. object positions (9) (cf. (2)).

(9) Carlo conosceva bene il compagno$_i$ di classe
Carlo knew well the classmate$_i$
che {incontrare $\__i$ divertiva sempre Anna | Anna
that {$meet_{inf}$ $\__i$ amused always Anna | Anna
voleva sempre incontrare $\__i$}
wanted always $meet_{inf}$ $\__i$}

(v) NEGPOLs (*nessuno, alcunché, mai* 'any, anything, ever') that are licensed by a higher negation, by a question, or not licensed, in simple or (deeply) embedded sentences (e.g. (10)).

(10) {Maria | Nessuno} si aspetta che qualcuno
{M. | No-one} self expects that someone
possa aver {già | mai} finito questo
could have {already | never} completed this
esercizio (?)
exercise (?)

The use of expanded templates was designed to minimize the potential effect of collocations or specific lexical choices.

Whenever possible each sentence was also manually annotated according to the linguistic-theoretic expectations for "grammaticality", on a 4-points scale: * (ungrammatical, coded as 0), ?? (very marginal, coded as 0.66), ? (marginal, coded as 0.33) and OK (grammatical, coded as 1).

Complexity dataset: it comprises 2,530 Italian sentences annotated with human judgments of perceived complexity on a 7–point Likert scale as for the acceptability dataset. The number of annotations per sentence ranged from 11 to 20, with an average of 16.753. The corpus was internally subdivided into two subsets representative of two different typologies of data, i.e. 1,858 naturalistic sentences extracted from corpora and 672 artificially-generated sentences drawn from the *Acceptability* dataset, and chosen to cover the range of linguistic phenomena represented in its templates. The first subset contains sentences taken from the Universal Dependency (UD) treebanks (Nivre et al., 2016) available for Italian, representative of different text genres and domains. In this regard, the largest portion contains 1,128 sentences taken from the newswire section of the Italian Stanford Dependency Tree-

bank (ISDT) (Bosco et al.,), annotated with complexity judgments by Brunato et al. (2018). Beside these, we chose to include in this corpus smaller subsets of sentences representative of a non-standard language variety and of specific constructions, i.e. Wh-questions and direct speech. Non-standard sentences (for a total of 323) are in the form of generic tweets and tweets labelled for irony taken from two representative treebanks, i.e. PoSTWITA and TWITTIRÒ (Sanguinetti et al., 2018; Cignarella et al., 2019). Wh-questions (164 sentences) were extracted from a dedicated section (prefixed by the string 'quest') included in ISDT. Direct speech sentences (243) mainly include transcripts of European parliamentary debates (taken from the 'europarl' section of ISDT) and extracts from literary texts (mostly contained in the UD Italian VIT (Delmonte et al., 2007)). The choice of annotating a shared portion of data with both acceptability and complexity scores was explicitly motivated by the attempt to empirically investigate whether there is a correlation between the two sentence properties, and whether complexity is judged differently in the case of ill-formed constructions.

For the purpose of the task, both datasets were split into training and validation samples with a proportion of 80% to 20%, respectively.

3.2 Annotation with Human Judgments

For the collection of judgments of sentence acceptability and complexity by Italian native speakers we relied on crowdsourcing techniques using different platforms. More specifically, for *Acceptability*, the set of sentences drawn from the psycholinguistic studies described in Section 3.1 was annotated using an on-line platform based on jsPsych scripts (De Leeuw, 2015). For the *Complexity* dataset, the annotation of the subcorpus of sentences taken from (Brunato et al., 2018) was performed through the CrowdFlower platform[1] (more details are reported in the reference paper), while the remaining sentences in this dataset were annotated using Prolific[2]. To make the annotation process comparable to the one followed by (Brunato et al., 2018), the whole process was split into different tasks, each one consisting in the annotation of about 200 sentences randomly mixed for the various typologies. For all tasks, workers were asked to read each sentence and answer the following question:

> *"Quanto è complessa questa frase da 1 (semplicissima) a 7 (molto difficile)?"*
> *'How difficult is this sentence from 1 (very easy) to 7 (very difficult)?'*

Beyond complexity, the 672 artificially-generated sentences were also labelled for perceived acceptability according to the following question:

> *"Quanto è accettabile questa frase da 1 (completamente agrammaticale) a 7 (perfettamente grammaticale)?" 'How acceptable is this sentence from 1 (completely ungrammatical) to 7 (completely grammatical)?'*

After collecting all annotations, we excluded workers who performed the assigned task in less than 10 minutes, which we set as the minimum threshold to accurately complete the survey.

3.3 Analysis of Judgments across Corpora

Table 1 shows the average value, standard deviation and minimum and maximum score of complexity and acceptability labels for the whole dataset. As it can be noticed, complexity values are on average lower and less scattered than the acceptability ones. For this corpus, the lowest value on the Likert scale (1) – which should have been used to label sentences perceived as very easy, in line to the task question – was given only twice, specifically to the following sentences:

(11) Dimmi il nome di una città finlandese.
tell me the name of a town Finnish
'Tell me the name of Finnish town'

(12) Quali uve si usano per produrre vino?
Which grapes PRT they_use to make wine?

Conversely, for the acceptability corpus, the highest value on the Likert scale (i.e. 7, meaning in this case *completely acceptable*) was attributed to 26 sentences. For space reasons, we report here only two examples:

(13) Le sorelle sono sopravvissute.
The sisters are survived.
'the sisters have survived'

(14) I lupi hanno ululato.
The wolves have howled.

[1] Now known as Figure Eight, https://appen.com/
[2] www.prolific.co

With respect to the 'worst' values, two sample sentences judged respectively as the most complex (i.e. 6.46 on the Likert scale) and (among) the least acceptable (1.55) in each dataset are the following ones, respectively:

(15) Chi è che lui ha affermato che il professore
who is that he has claimed that the professor
aveva detto che lo studente avrebbe dovuto
had said that the student had$_{subj}$ must
considerare questo candidato?
consider this candidate?

(16) Il falegname è arrivato mentre noi
The carpenter has arrived while we
montavo la mensola.
were_assembling$_{1Psg}$ the shelf.

	COMPL		ACCEPT	
	SCORE	SE	SCORE	SE
μ	3.12	0.332	4.45	0.36
σ	1.04	0.08	1.7	0.14
min	1	0	1.13	0
max	6.46	0.63	7	0.74

Table 1: Statistics collected for the two corpora of the AcCompl-it dataset.

If we consider the internal composition of the two datasets we can see a more articulated picture depending on its various subparts (see Table 2 and 3). For complexity, average scores are higher for sentences created to display specific acceptability patterns, thus proving that acceptability does affect the perception of complexity. Note that the most complex sentence (reported in (15)) is contained in this set, and is ungrammatical (no gap).

Among the treebank sentences, those extracted from journalistic texts (ISDT_news) were judged on average as the most complex, questions as the easiest ones. Twitter and direct speech sentences obtained scores in between the highest and the lowest value and very close to each other. This is in line with stylistic and linguistic analysis showing that the language of social media inherits many features from spoken language.

For the whole acceptability dataset, the Spearman's rank correlation coefficient between theoretically-driven grammaticality and mean acceptability labels is very strong (r(656)=.83, p<.001). While this could be somehow expected, when we focus only on the 672 sentences annotated for both complexity and acceptability, we still observe a significant but lower correlation

	SCORE	SE	MIN	MAX
ISDT_news	3.28	0.33	1.25	5.7
Twitter	2.59	0.31	1.13	4.69
DirectSpeech	2.68	0.31	1.14	6
Wh-Quest	1.61	0.22	1	2.94
ArtifSent	3.63	0.37	1.42	6.47

Table 2: Average complexity score, standard error and minimum and maximum value across the different subsets of the **Complexity** dataset.

between expected grammaticality and mean complexity (r(656)=.34, p<.001). Still considering this subset, an additional outcome is the moderate (and negative) correlation between the two metrics (r(672)=.49, p<.001), further suggesting that the more a sentence is perceived as complex, the less acceptable it is.

4 Evaluation measures

For both the ACCEPT and COMPL Task, the evaluation metric was based on Spearman's rank correlation coefficient between the participants' scores and the test set scores. For each task, two different ranks were produced according to the prediction of the relative scores and to standard errors. In each task a different baseline was defined:

- in the ACCEPT task, it corresponds to the score assigned by a SVM linear regression using unigram and bigram of words as features;
- in the COMPL task, it corresponds to the score assigned by a SVM linear regression using sentence length as its sole feature.

5 Participation and results

The AcCompl-it task received three submissions for each subtask from two different participants, for a total of 6 runs. Unfortunately, neither participant took part in the Open Task. Results for the other ones are reported in Tables 4 and 5.

The systems from the two participants in the task follow very different approaches: one is based on deep learning and trained on raw texts (Sarti, 2020), the other relies on (heuristic) rules applied to semantic and syntactic features automatically extracted from sentences (Delmonte, 2020). In spite of their very different nature, the two approaches also present some commonalities, such

	SCORE	SE	MIN	MAX
clefts (1)	4.27	0.39	1.36	6
copular	5.01	0.48	2.90	6.5
canonical (2a)	5.47	0.44	3.58	6.5
inverse (2b)	4.56	0.51	2.90	5.9
unerg V (3a)	5.91	0.30	3.70	7
SV	6.64	0.21	5.94	7
VS	5.18	0.40	3.70	6.2
unacc V (3b)	6.28	0.27	4.86	7
SV	6.61	0.20	5.82	7
VS	5.96	0.33	4.86	6.72
trans V (3c)	4.91	0.34	2	7
SV	6.47	0.24	5.06	7
VS	3.34	0.43	2	4.52
S V agree (4)	3.81	0.31	1.25	6.93
match	5.87	0.30	3.14	6.92
mismatch	1.74	0.32	1.25	2.72
wh-island (arg) (5)	3.85	0.17	1.68	5.63
filler-gap dep.	3.56	0.51	1.5	6.69
doubly filled (6)	3.28	0.42	1.5	6.69
coord (7)	4.25	0.47	2.5	6
wh-island (adj) (8)	3.02	0.41	1.38	5.6
no extraction	6.26	0.29	5.26	7
subj/obj (9)	3.70	0.51	2.66	5.15
NPIs (10)	4.75	0.45	2.27	6.6
bad fillers	1.13	0.06	1.13	1.13
good fillers	6.76	0.07	6.76	6.76
medium fillers	4.07	0.18	4.07	4.07

Table 3: Average acceptability score, standard error and minimum and maximum value across the different linguistic phenomena of the **Acceptability**. Numbers in (·) refer to examples in the text.

PARTICIPANT	SCORE	SE
UmBERTO-MTSA (Sarti)	**0.88****	**0.52****
ItVenses-run1 (Delmonte)	0.44**	0.25**
ItVenses-run2 (Delmonte)	0.49**	0.41**
Baseline	0.30**	0.35**

Table 4: ACCEPT task results.**p value<0.001; *p value <0.05

as the reliance on external resources. In particular, both make use of additional sentences taken from existing Italian treebanks, either to enrich the original training sets with additional annotated examples (Sarti's case) or to check the frequency of a given construction and use this info among the features of the proposed system (ItVenses).

Sarti's systems obtained the best performance on both tasks using a similar multi-task learning (MTL) approach, which consists in leveraging the predictions of a state-of-the-art neural language model for Italian (i.e. UmBERTO[3]) fine-tuned on the two downstream tasks to augment the original development sets with a large set of unlabeled ex-

[3]https://github.com/musixmatchresearch/umberto

PARTICIPANT	SCORE	SE
UmBERTO-MTSA (Sarti)	**0.83****	**0.51****
ItVenses-run1 (Delmonte)	0.31**	0.09*
ItVenses-run2 (Delmonte)	0.31**	0.07
Baseline	0.50**	0.33**

Table 5: COMPL task results. **p value<0.001; *p value <0.05

amples extracted from available Italian treebanks. The bigger dataset was then split into different portions to train an ensemble of classifiers. The resulting MTL model was finally used to predict the complexity/acceptability labels on the original test sets.

Delmonte's *ItVenses* system parses the sentences to obtain a sequence of constituents and a set of sentence-level semantic features (presence of agreement, negation markers, speech act and factivity). These features, along with constituent triples and their frequency in the training set and in the Venice Italian Treebank are weighed with various heuristics and used to derive a prediction. Agreement mismatches were checked using morphological analysis of verb and subject, while the argumental structure is inferred using a deep parser. The two versions of the system (*run1* and *run2*) differ only in their use of features (*run2* dispenses with proposional negation and certain verb agreement features).

As it can be seen, ItVenses's performance were considerably lower than Sarti's system (lower, in fact, than the baseline based on sentence length, in the COMPL prediction task). However, as better explained in the following section, in the artificial data subset, which has complex but far less diverse structures, the gap with the winning system is reduced in the COMPL task (cfr. Table 7) and, even more robustly, in the ACCEPT task (Table 6).

6 Discussion

The extremely good performance of the winning system in both tasks is not wholly unexpected in light of the impressive results obtained by current neural networks models across a variety of NLP tasks. In this regard, it is worth noticing that, in his report, the author compared the performance of the best system based on multi-task learning to the one obtained by a simpler version of the UmBERTO-based model with standard fine-tuning on the two downstream tasks, achieving al-

ready very good results (.90 and .84 for acceptability and complexity predictions on the training corpus, respectively). Similarly, and especially for the automatic assessment of sentence acceptability, the scores obtained by the winning system (.88) are in line with those reported in (Linzen et al., 2016), who train a classifier to detect subject-verb agreement mismatches from the hidden states of an LSTM, achieving a .83 score. Most other systems at work on the ability of neural models to detect acceptability or grammaticality in a broader range of cases report much lower scores, but they try to read (minimal pair) judgments from metrics associated to the performance of systems that have not been expressly trained on giving judgments, reasoning that 'judgment giving' is not a task humans have a life-long training for, but which is nonetheless feasible.

To have a better understanding of the potential impact of different types of data on the predictive capabilities of the two systems, we further inspected the final results by testing each system on sentences representative of diverse linguistic phenomena and textual genres. To this end, we split the whole test set into the distinct subsets defined in the corpus collection process (cfr. Section 3.1) and we assessed the correlation score between predicted and real labels for each type: note that, for the ACCEPT predictions, this analysis was performed considering only two 'macro–classes', i.e. artificial vs psycholinguistics-related data, in order to have a significant number of examples in the test set. Similarly, for COMPL, we distinguished the artificially–generated sentences from sentences drawn from all treebanks. Results of this fine-grained analysis are shown in Tables 6 and 7.

Interestingly, although the gap between the two systems is still evident, we observed that artificial data have an opposite effect on their performance. In particular, as anticipated in the previous section, *ItVenses* is more accurate in predicting both the complexity and, especially, the acceptability level of this group of sentences. The opposite holds for Sarti's system, which although still very good in both tasks, achieves lower correlation scores when tested against artificial data.

Running an exploratory analysis based on expected grammaticality, we observed that Sarti's system performs much better in predicting the acceptability score on expected grammatical sentences (r=.80, p<.001) than on expected ungrammatical ones (r=.76, p<.001). Similarly, but less robustly, the same numerical asymmetry is observed in both Delmonte's runs: for grammatical predictions, RUN1 r=.33, RUN2 r=.35; for ungrammatical ones RUN1 r=.32, RUN2 r=.34, all correlations being equally significant (p<.001).

PARTICIPANT	SCORE	SE
Psycholinguistics related		
UmBERTO-MTSA (Sarti)	0.90**	0.55**
ItVenses-run1 (Delmonte)	0.42**	0.24**
ItVenses-run2 (Delmonte)	0.50**	0.48**
Artificial data		
UmBERTO-MTSA (Sarti)	0.74**	0.33**
ItVenses-run1 (Delmonte)	0.50**	0.20*
ItVenses-run2 (Delmonte)	0.46**	0.25*

Table 6: ACCEPT task results on different subsets of the official test set. **p value<0.001; *p value <0.05

PARTICIPANT	SCORE	SE
Treebank sentences		
UmBERTO-MTSA (Sarti)	0.86**	0.61**
ItVenses-run1 (Delmonte)	0.25**	0.13*
ItVenses-run2 (Delmonte)	0.24**	0.10*
Artificial data		
UmBERTO-MTSA (Sarti)	0.70**	0.06
ItVenses-run1 (Delmonte)	0.44**	-0.07
ItVenses-run2 (Delmonte)	0.51**	-0.11

Table 7: COMPL task results on different subsets of the official test set. **p value<0.001; *p value <0.05

References

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

C. Bosco, S. Montemagni, and M. Simi. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*.

Dominique Brunato, Lorenzo De Mattei, Felice

Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Cristiano Chesi and Paolo Canal. 2019. Person features and lexical restrictions in Italian clefts. *Frontiers in Psychology*, 10:2105.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*.

Joshua R De Leeuw. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT - Venice Italian Treebank: Syntactic and quantitative features. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*.

Rodolfo Delmonte. 2020. Venses@AcCompl-it: Computing complexity vs acceptability with a constituent trigram model and semantics. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Matteo Greco, Paolo Lorusso, Cristiano Chesi, and Andrea Moro. 2020. Asymmetries in nominal copular sentences: Psycholinguistic evidence in favor of the raising analysis. *Lingua*, 245:102926.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

T. Linzen, E. Dupoux, and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 521–535.

Simona Mancini, Paolo Canal, and Cristiano Chesi. 2018. The acceptability of person and number agreement/disagreement in Italian: an experimental study. *Lingbuzz preprint: https://ling.auf.net/lingbuzz/005514*.

Andrea Moro. 1997. *The raising of predicates: Predicative noun phrases and the theory of clause structure*, volume 80. Cambridge University Press.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018)*.

Gabriele Sarti. 2020. UmBERTo-MTSA @ AcCompl-it: Improving complexity and acceptability prediction with multi-task learning on self-supervised annotations. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115:1497–1524.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental synta. *Biolinguistics*, pages 1123–134.

Sandra Villata, Paolo Canal, Julie Franck, Andrea Carlo Moro, and Cristiano Chesi. 2015. Intervention effects in wh-islands: An eye-tracking study. In *Architectures and Mechanisms for Language Processing (AMLaP 2015)*, pages 195–195.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.