Building an Italian Written-Spoken Parallel Corpus: a Pilot Study

Elisa Dominutti, Lucia Pifferi
Università di Pisa
elisa.dominutti@gmail.com
luciapiff@gmail.com

Felice Dell'Orletta, Simonetta Montemagni
Valeria Quochi
ILC–CNR
name.surnamen@ilc.cnr.it

Abstract

This paper presents a pilot study towards the creation of a monolingual written–spoken parallel corpus in Italian, featuring two main novelties in the general landscape of spoken corpora: the alignment with the written counterpart of the same content and the spoken variety dealt with, represented by transcriptions of radio news broadcasting.

1 Introduction

Nowadays, the contrast between written and spoken language does no longer represent a clear-cut opposition. The emergence of modern communication technologies such as radio, television and new (digital) media led to important changes in the analysis of the diamesic variation. Under this view, the opposition spoken vs. written language is reformulated in terms of a continuum with prototypical written and spoken language at the extreme poles and within which a cline of intermediate linguistic varieties can be recognised, mixing, to a different extent, features of the two. Nencioni (1976) defined the extreme poles of this continuum as the *parlato-parlato* ('spoken-spoken') variety, i.e. casual, spontaneous conversation, and the *scritto-scritto* ('written-written') variety, i.e. planned, formal, written language. Besides the typical contexts envisaging the use of spoken language—which require all participants to be present in the same environment, that the conversation is held in turns and that speakers make sure their messages are getting across—different contexts can be imagined: among them, the radio and television language which, despite being spoken, present traces of textual organisation recalling the written language. Nencioni (1976) qualifies this variety of language use as *parlato-scritto* ('spoken-written'), a label that emphasises its hybrid nature characterised by the co-occurrence of traits typical of both written and spoken language. From a different perspective, Ong (1982) refers to this variety as 'secondary orality', i.e. "an orality not antecedent to writing and print, as primary orality is, but consequent and dependent upon writing and print".

In addition to this socio-linguistic interest, the issue also bears relevance for computational approaches as it has a substantial impact on the perceived naturalness of human-machine interaction. Indeed, one of the reasons why speech synthesis applications still produce unnatural speech, apart from bad prosody is that written language is generally not suitable, i.e. comprehensible, direct and effective, in spoken contexts (Kaji et al., 2004). With the rise and quick spread of Virtual Reality (VR) and Augmented-Reality (AR) applications, moreover, the mismatch between written and spoken language styles brings about serious technological limitations because unnaturalness of the virtual agents translates into bad human comprehension and/or distrust in those agents altogether. It is thus no longer sufficient to pass a written message to the speech synthesizer, but such a message needs to be transformed in a form suitable to be spoken in the specific context of use. In order to be able to do this, corpus data is needed such as a monolingual parallel aligned corpus of written and spoken texts about the same content. A corpus designed in this way is of fundamental importance for: a) investigating the features of the *parlato-scritto* language variety, its similarities and differences with respect to the written language; and b) for creating the prerequisites for the design and development of tools for monitoring the communicative effectiveness of texts with respect to their production mode and for support-

ing the semi-automatic generation or transformation of texts to be delivered orally. Such a corpus represents an important novel contribution in the area of language corpora; generally in fact corpora target either written or spoken language. Some corpora indeed also include sections with transcriptions of spoken language: see for instance the Brown corpus for English. On the front of spoken corpora, large corpora of spoken Italian were produced, some aiming at specific purposes, like CiT (*Corpus di Italiano Trasmesso*) (Spina, 2000) or LIR (Maraschio et al., 2004), while others aiming at representing Italian in a wider perspective like C-ORAL-ROM (Cresti and Moneglia, 2005). Some of them take into account only a few aspects of the linguistic variability, mainly the diaphasic and in some cases diamesic dimension.

Our *Corpus Italiano Parallelo Parlato Scritto* ('Spoken Written Italian Parallel Corpus', henceforth CIPPS) features two fundamental novelties in the general landscape of spoken corpora: the alignment with a written counterpart of the same content and the type of spoken variety dealt with.

2 Background and related works

Notwithstanding the differences between written and spoken language styles and the impact it bears on human-machine interaction, little computational work has been devoted to develop data and methods for "transforming" a written text in a text suitable for a specific spoken context.

Previous works mostly deal with the transformation of spoken language into grammatically valid, correct written language that can be parsed by standard NLP tools—see for instance Marimuthu and Devi (2014) and Giuliani et al. (2014). However, the rise and spread of VR and AR applications seem to call for the need to appropriately tackle also the other direction, i.e. the transformation of written into (diamesically) appropriate spoken language, which presents different challenges[1].

Few studies have been devoted to the automatic transformation or generation of suitable spoken language, mostly on Japanese. Among these, Murata and Isahara (2001) describe an interesting model to perform different kinds of paraphrasing tasks, that is to transform sentences according to different predefined criteria. Interestingly, in their experiments both on sentence compression and on transformation from written language to spoken language they manage to apply the same algorithm applied to different data an dobtain good results. For the latter experiment, they used a monolingual parallel corpus of academic papers and transcripts of oral presentations and built a system that learns re-writing rules according to the defined criteria. In the former case re-writing rules were learnt from dictionaries.

Kaji and colleagues (2004; 2005) worked on the transformation of written language to spoken language style in Japanese, approaching the issue as a lexical paraphrasing problem, for which they constructed an ad-hoc written–spoken web corpora focused on the connotational differences related to the *suitability for orality* of expressions. Their method learns predicate paraphrases from a dictionary and then uses the corpus to statistically determine whether an expression is suitable to be spoken.

More recently, Matsubara and Hayashi (2012) report about an application for generating spontaneous news speech in a news speech delivery service. They approach the issue as a text generation task and develop a rule-based system for automatically generating news speech scripts—to be read via speech synthesis—starting from newspaper articles. Their approach however focuses on a specific stylistic difference peculiar to Japanese hardly portable to other languages and does not involve any kind of parallel aligned data.

3 Pilot corpus creation

In this work we describe our first attempts at building a parallel written–spoken corpus that might ultimately be useful to train a system for the transformation of written text into text suitable to be spoken. We focus on two different language varieties within the spoken-written language continuum, mentioned in section 1, namely radio spoken language and newspaper written language. This focus was dictated both by the need to neutralize the effects possibly deriving from considering different topics, textual genres and/or communication contexts, and by the practical need of finding readily available data to run the pilot. Thus the present data-set is built by aligning newspaper articles, taken as representatives of the written–written variety and news broadcasting via radio,

[1] VR/AR is currently a hot topic especially in both educational and industrial-training contexts (Akçayır and Akçayır, 2017; Żywicki et al., 2018; Gattullo et al., 2019; Heinz et al., 2019; Albayrak et al., 2019).

Day	Num of news	Average lenght
13/05/2003	150	479
15/05/2003	144	523
17/05/2003	148	480
23/05/1995	119	578
25/05/1995	125	547
27/05/1995	124	549
Tot	810	526

Table 1: Written corpus

Day	Num of news	Average lenght
13/05/2003	365	60
15/05/2003	321	57
17/05/2003	156	73
23/05/1995	1184	66
25/05/1995	1106	60
27/05/1995	598	83
Tot	3730	66.5

Table 2: Spoken corpus

taken as representatives of the spoken–written variety.

3.1 Data selection and preparation

Given the goals defined above, our first step was to collect the materials for building the pilot data-set.

For the spoken data-set we chose the *Lessico di italiano Radiofonico* corpus (LIR)(Maraschio et al., 2004)[2], which consists in transcriptions of various Italian radio broadcast channels sampled in 1995 and 2003 and contains various types of annotations among which: broadcaster, text genre, speaker, communication type, self-corrections, breaks, etc. In particular, we selected the transcriptions of radio news by Radio RAI1, Radio RAI2 and Radio RAI3[3] which amount to 6 days altogether: the 23rd, 25th, 27th May 1995, and the 13th, 15th and 17th 2003.

The written data-set was created by taking all news articles published in La Repubblica on the same dates[4]. Tables 1 and 2 report the figures of the data-sets.

In the case of the spoken corpus extensive extraction and cleaning work was required because the original transcriptions include many different genres (e.g. advertisements, interviews, entertainment,...) and several different annotation tags.

3.2 Spoken corpus cleaning

From the selected days of the LIR corpus we needed to extract only the transcriptions of news text. The original texts in fact contain several types of annotations, all in a proprietary tagging format, and news are easily recognisable. So, for each day mentioned, we created a data-set by collating the news of the different radio broadcasters, thus obtaining 6 spoken data-sets, one for each day. These were subsequently cleaned by using regular expressions that removed all annotation tags, which provided us with raw text data for the alignment experiment.

In Table 2 we can see the number of news extracted for each day and their average length in terms of tokens. Interestingly, but not surprisingly, we observe that newspaper articles on average are longer than radio news.

4 Alignment methodology

Once we gathered, cleaned and normalised the relevant data, we proceeded to align written and spoken texts on the basis of topic and semantic equivalence. Since the spoken transcriptions do not have an explicit marking of sentence boundaries, for the time being alignment is performed at text level; we leave sentence-level alignment for future work.

Given the six spoken data-sets and their corresponding written ones we experimented with two different methods to perform their alignment. One is based on the *Jaccard index* (Jaccard henceforth), the other method on *cosine similarity* (Cosine henceforth). Both algorithms followed one common preliminary step: for each data-set we took into consideration only nouns, verbs, adjectives and numerals, i.e. semantically heavy words.

The first method calculates similarity using the Jaccard index as a statistical index. In general, this coefficient measures the similarity of two samples through the ratio between the size of the intersection and the size of the union of the sample sets; so, in this case, the numerator is given by the overlap of words of the two documents, i.e. the number of relevant words present in both. The denominator instead is the sum of the relevant words of both documents. The computation can be represented

[2]Source: http://www.accademiadellacrusca.it/it/attivita/lessico-frequenza-dellitaliano-radiofonico-lir

[3]The news transcriptions of the other broadcasters were too short for our purposes.

[4]source: https://ricerca.repubblica.it/

as follows:

$$J(A,B) = \frac{|\text{overlapping words in A, B}|}{|\text{words A + words B}|} \quad (1)$$

The range of acceptable values stands between 0 (for the couples of documents that have no words in common) and 0,5 (for the couples of documents with the highest similarity, i.e. with all relevant words in common).

The second method computes the cosine similarity between a vector representing all the relevant words in a spoken text and a vector representing a written text. Each vector contains a number of components identical to the amount of relevant words contained in the texts, the value of each component being the *TFiDF* value of the corresponding word in the represented text. Once all vectors were built, we compared each spoken-vector with every written-vector and computed their cosine similarity. Finally, considering values of similarity in decreasing order we reorganised the pairs and completed document-alignment. The range of acceptable values for the Cosine method stands between 0 and 1, with values close to 1.0 indicating strong similarity.

4.1 Alignment evaluation

The two methods illustrated above produced twelve output files, six for each method, all ranked on the basis of their similarity score in decreasing order. For each of them we considered the first one hundred spoken-written text pairs and manually evaluated their alignments on a binary scale with respect to their information content. News about the same topics, events or facts were considered good alignments. We decided to stop the evaluation at the first one hundred pairs, because after this threshold the recognised alignments were no longer significant (i.e. algorithms aligned pairs of documents with different topics).

On the 1200 manually assessed pairs we than calculated the **accuracy** of the two methods. We considered accuracy as the ratio between the number of aligned pairs in particular range of distance values and the total number of couples in the same range.

The graphics in Figures 1 and 2 show method accuracy for each range of similarity values, using both the 1995 and 2003 data. For example, in the range of values between 0,1 and 0,2, the Cosine method has an accuracy of 6% with the 1995 data and 22% with the 2003 data. As we advance in the higher similarity bands, we notice a growing trend for both methods, but while for Cosine we observe a gradual growth, the Jaccard method shows a faster rise. Moreover, we notice that most of the alignments occur in the lowest similarity range of value, while in the higher similarity ranges we found very few alignments (see Table 3 and 4 for details).

Remembering that the range of admissible values are different for the two methods let us focus on the results.

Cosine alignment evaluation Cosine for both data-sets has an accuracy of 100% in the range of values 0,8-0,7 and 0,6-0,5, while for the range 0,2-0,3 it has an accuracy of 6% for 1995's data-sets and 22% for 2003's data. Figure 1 shows a gap between 0.7 and 0.6 for 2003's data. That is because, for this data-set, the cosine method did not assign values in this range. Overall, Cosine total accuracy is 61%, 53% on 1995 data and 69% on 2003 data.

Jaccard alignment evaluation In the range 0,3-0,2 the Jaccard method has an accuracy of 100% on both datasets; while for the 1995 data it drops to 53% in the range 0,2-0,1 and to 47% in the range 0,1-0,6. For the 2003 data in the range 0-2,01 the accuracy is 86%, which decreases to 44,8% in the range 0,1-0,07. Also in this case, as reported in Table 4, we have few alignments in higher distances despite the number of lower ones.

Overall, Jaccard total accuracy is 50%, 50% on 1995 data and 51% on 2003 data.

According to this evaluation, Cosine using TFiDF values is the best method for aligning our data.

Here is an example of text pairs with high cosine similarity values (0,7-0,8):

[Spoken]: `[...] il diario di Paul Mccartney [...] rottura con i Beatles è stato riconsegnato [...] al cantante il giorno dopo il concerto dei fori imperiali [...] Mccartney ha avuto modo di rileggere quel preziosissimo diario stracolmo di ricordi e ha confermato l'autenticità [...] alcune frasi portano il segno della storia "Arriva John per discutere lo scioglimento della partnership" giugno millenovecentosettanta la fine dei Beatles`

[Written]: `[...] il diario di Paul Mccartney [...] rottura con i Beatles è stato riconsegnato [...] al cantante, il giorno dopo il concerto dei fori imperiali. [...] sir Paul ha avuto modo di rileggere quel preziosissimo diario stracolmo di ricordi, e ha confermato l'autenticità dell'agenda. [...] alcune frasi portano il segno della storia: ''arriva John per discutere lo scioglimento della partnership''. giugno 1970, la fine dei Beatles. [...]`

What follows instead is an example of a good alignment with lower cosine similarity values (0,3-0,2)[5]:

[Spoken]: `se non mi attaccassero non mi difenderei [...] spiega Berlusconi [...] "Io sono un moderato" ripete il premier "Mi difendo da teoremi folli che non attaccano me ma il presidente del consiglio" [...]`

[Written]: `Berlusconi al contrattacco "Denuncerò chi mi offende". [...] E aggiunge che le accuse contro di lui si basano su "Teoremi folli". Teoremi ai quali [...] "Ho dato la risposta più moderata, contenuta e misurata che si potesse dare". [...]`

The first example is also an example of high Jaccard similarity values (0,3-0,2).

In general, with both methods, the pairs of documents correctly aligned in the lower ranges of similarity show considerable differences in terms of lexical items and possibly linguistic structures, and thus represent a very interesting set of pairs for future investigation. Regarding higher ranges, we find a greater lexical overlap and a lower variation in linguistic structure. Comparing the pairs correctly aligned by the two methods we counted 77 identical ones, while the number of different pairs derived from Jaccard is 220, and from Cosine 260. In total we obtained 557 different correctly aligned pairs.

5 Pilot corpus profiling

The final pilot CIPPS corpus consists of 557 text pairs corresponding to the correctly aligned and manually validated pairs of spoken and written

[5]For reasons of space the example texts have been arbitrarily shortened.

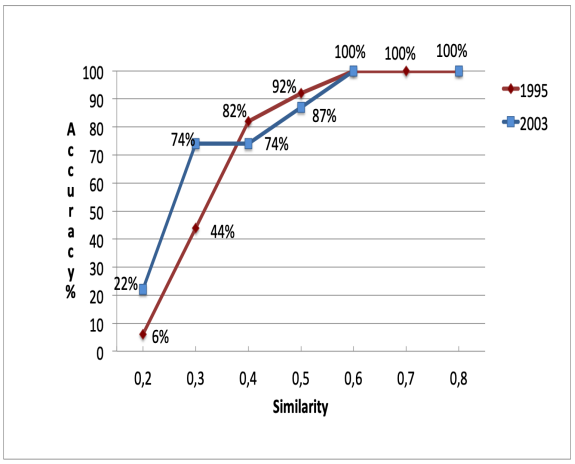

Figure 1: Cosine accuracy

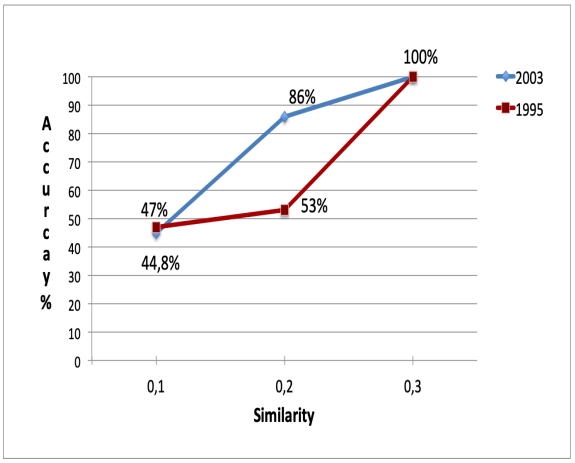

Figure 2: Jaccard accuracy

Distance	1995		2003	
	Correct	Tot	Correct	Tot
0,8-0,7	1	1	2	2
0,7-0,6	3	3	0	0
0,6-0,5	6	6	4	4
0,5-0,4	12	13	26	30
0,4-0,3	45	55	45	61
0,3-0,2	90	206	123	167
0,2-0,1	1	16	8	36
TOT	158	300	208	300

Table 3: Cosine Accuracy (1995-2003)

Distance	1995		2003	
	Correct	Tot	Correct	Tot
0,3-0,2	5	5	3	3
0,2-0,1	41	77	37	43
0,1-0,065	103	218	114	254
TOT	149	300	154	300

Table 4: Jaccard accuracy (1995-2003)

documents resulting from both alignment methods. It can thus be taken as a gold-standard corpus of content aligned text pairs of news for the dates and years mentioned in section 3.1.

This section reports on our preliminary contrastive analysis of CIPPS using Monitor-IT (Montemagni, 2013), so as to establish basic linguistic profiling of the two language varieties represented in the corpus. This analysis was done with a specific view to investigating similarities and differences in the distribution of multi-level linguistic cues (we focus here on lexical and morpho-syntactic features) both within the corpus and against prototypical written and spoken language (in the future, we plan to extend this analysis to the underlying syntactic structure).

Let us first compare the two sections of the CIPPS corpus. On the one hand, highly correlated features between the CIPPS written and spoken sections concern the distribution of nouns (both common and proper) and adjectives as well as verbal forms used in the third person singular; the correlation was calculated with the Spearman's Correlation Coefficient (p-value ≤ 0.05). On the other hand, statistically significant different features across the spoken and written corpus sections detected with the Wilcoxon test (p-value $\leq$ 0,05) include specific verbal forms, deictic elements and determiners, prepositions and acronyms, as well as lexical richness (measured in terms of Token/Type Ratio). In particular, if verbal moods such as gerundive, subjunctive, infinitive and conditional are typically associated with written articles, the 1st and 2nd person of verbs in both singular and plural forms are typical of the spoken news reports. Demonstrative determiners and pronouns represent significant features of the spoken variety, whereas acronyms and lexical richness measured in terms of Token-Type Ratio characterise the written CIPPS section.

For what concerns the comparison of the linguistic profiling results sketched above with what we know from the literature about features of spoken vs. written language, we observe that the widely acknowledged fact that spoken language is less complex than written language is declinated here in quite a peculiar way. Differently from the 'spoken-spoken' variety characterised by a reduced number of nouns and consequently by a lower noun/verb ratio (ranging between 0,80 and 1, (Montemagni, 2013)), the 'spoken-written' variety shares with prototypical written language a twice higher noun/verb ratio, which, according to Biber (1988), is typical of informative texts. On the other hand, it shares with prototypical spoken language the more frequent use of deictic elements, of 1st/2nd person reference in verbal forms, lexical repetition.

These findings, which need to be further elaborated and explored, confirm the hybrid nature of the spoken language variety represented in the CIPPS corpus, which is in line with the trend reported in the literature that the language of the radio shares features with both spontaneous oral and written language varieties.

6 Conclusions and Future work

In this paper we have presented our first experiments towards the creation of the CIPPS, a monolingual written-spoken parallel aligned corpus. The data for this pilot was drawn from existing corpora and archives, it was automatically aligned on the basis of two statistical methods and finally manually validated. To the best of our knowledge, this is the first attempt to build such a corpus and more research is needed to improve its potentials and increase its magnitude.

Among the open issues to be approached first is the lack of punctuation in the spoken part of the corpus, which makes automatic alignment with the written counterpart too coarse. As mentioned in the introduction, a corpus like ours might also be precious as a training set for the development of a system for transforming written into suitable spoken texts. Although little work has been done in this direction, the time is now ripe to tackle the challenge and we plan to start experimenting with both paraphrasing methods—as mentioned in section 1— and with monolingual machine translation, taking inspiration from Quirk et al. (2004) and Wubben et al. (2012). In this perspective, however, the first necessary step is to increase corpus size and improve alignment.

Acknowledgments

This work was partially supported by the 2-year project ADA, Automatic Data and documents Analysis to enhance human-based processes, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

Murat Akçayır and Gökçe Akçayır. 2017. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, 20:1 – 11.

M. S. Albayrak, A. Öner, I. M. Atakli, and H. K. Ekenel. 2019. Personalized training in fast-food restaurants using augmented reality glasses. In *2019 International Symposium on Educational Technology (ISET)*, pages 129–133, July.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins.

M. Gattullo, V. Dalena, A. Evangelista, A. E. Uva, M. Fiorentino, A. Boccaccio, M. Ruta, and J. L. Gabbard. 2019. A context-aware technical information manager for presentation in augmented reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 939–940, March.

Manuel Giuliani, Thomas Marschall, and Amy Isard. 2014. Using ellipsis detection and word similarity for transformation of spoken language into grammatically valid sentences. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 243–250.

Mario Heinz, Sebastian Büttner, and Carsten Röcker. 2019. Exploring training modes for industrial augmented reality learning. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2019, Island of Rhodes, Greece, June 5-7, 2019*, pages 398–401.

Nobuhiro Kaji and Sadao Kurohashi. 2005. Lexical choice via topic adaptation for paraphrasing written language to spoken language. In *Natural Language Processing - IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005, Proceedings*, pages 981–992.

Nobuhiro Kaji, Masashi Okamoto, and Sadao Kurohashi. 2004. Paraphrasing predicates from written language to spoken language using the web. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 241–248.

Nicoletta Maraschio, Stefania Stefanelli, Stefania Buccioni, and Marco Biffi. 2004. Dal corpus lir: prove e confronti lessicali. In Federico Albano Leoni, Francesco Cutugno, Massimo Pettorino, and Renata Savy, editors, *Atti del Convegno Nazionale "Il Parlato Italiano"*, page 36.

K Marimuthu and Sobha Lalitha Devi. 2014. Automatic conversion of dialectal tamil text to standard written tamil text using fsts. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland, USA, June 27, 2014*, pages 37–45.

Shigeki Matsubara and Yukiko Hayashi. 2012. Personalization of news speech delivery service based on transformation from written language to spoken language. In Toyohide Watanabe, Junzo Watada, Naohisa Takahashi, Robert J. Howlett, and Lakhmi C. Jain, editors, *Intelligent Interactive Multimedia: Systems and Services*, pages 449–457, Berlin, Heidelberg. Springer Berlin Heidelberg.

Simonetta Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi italiani di linguistica teorica ed applicata*, (XLII(1)):145–172.

Masaki Murata and Hitoshi Isahara. 2001. Universal model for paraphrasing - using transformation based on a defined criteria. *CoRR*, cs.CL/0112005.

Giovanni Nencioni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti critici*, (29).

Walter J. Ong. 1982. *Orality and Literacy: The Technologizing of the Word.* Methuen.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 142–149.

Stefania Spina. 2000. Il corpus di italiano televisivo (cit): struttura e annotazione. In *Atti del Convegno SILFI*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 1015–1024, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krzysztof Żywicki, Przemysław Zawadzki, and Filip Górski. 2018. Virtual reality production training system in the scope of intelligent factory. In Anna Burduk and Dariusz Mazurkiewicz, editors, *Intelligent Systems in Production Engineering and Maintenance – ISPEM 2017*, pages 450–458, Cham. Springer International Publishing.