POLiCOM
Dipartimento di Scienze Politiche e della Comunicazione

Lexical complexity and basic vocabulary of the Italian language

Nooj 2020 5 – 7 June '20
Università degli Studi di Salerno – Italy
Dipartimento di Scienze Politiche e della Comunicazione
Annibale Elia, Alessandro Maisto, Lorenza Melillo, Serena Pelosi.

Comprehensibility

- Comprehension's texts process starts from his decoding, and is built on basic language skills.

Comprehension's process is based on:
Enciclopedic knowledge
Inferential mechanism
Languages abilities

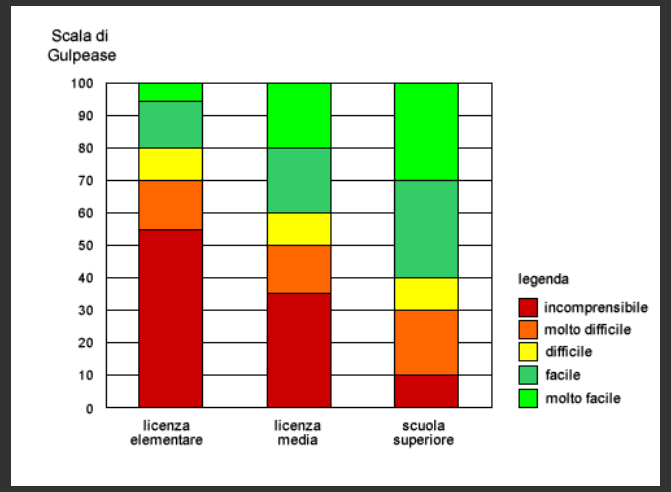
Readability

The Flesch formula

- Adaptation of the Flesh for English to Italian known as the Flesch-Vacca formula (Franchina and Vacca, 1986).

The GulpEase index

- Created from Lucisano and Piemontese, 1988.
- Based on: number of characters for word and the average number of words for sentence.



Comprehensibility and Readability

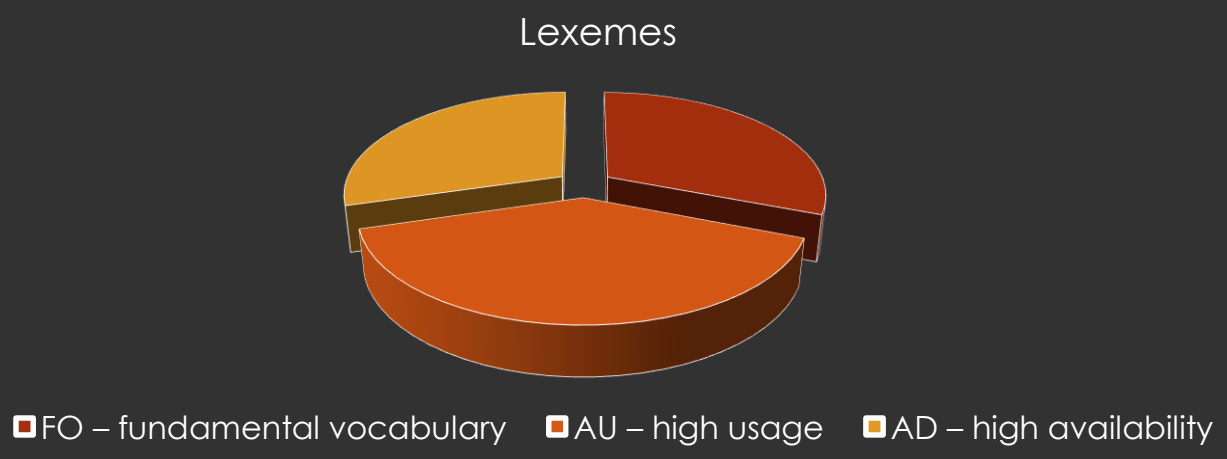
Comprehensibility	Readability
Qualitative analysis	Quantitative analysis
Logical and conceptual organization	Linguistic aspects Syntax and lexicon

Lexical complexity and the basic vocabulary

- **Basic Vocabulary of Italian** (VdB)
- First appeared as an annex to Guida all'uso delle parole Tullio De Mauro, 1980
- The VdB includes about 7000 words, those that have the higher statistical frequency in our language.

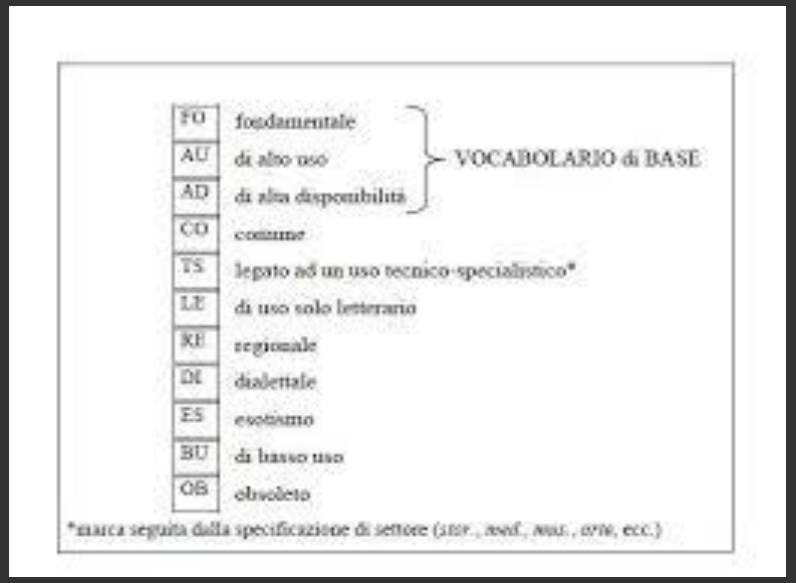
Table 1. VDB composition (Gradit 1999-2007)

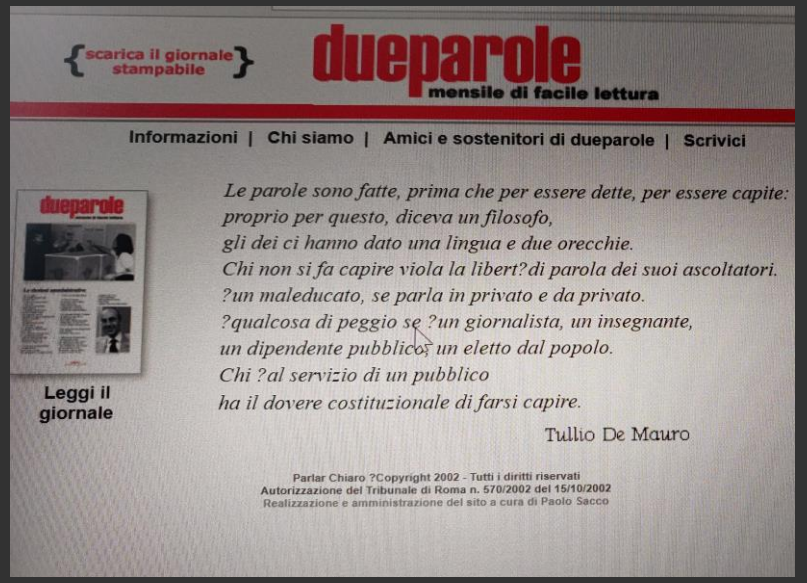
Range	Lexemes
FO – fundamental vocabulary	2,077
AU – high usage	2,663
AD – high availability	1,988



Lexical complexity and the basic vocabulary

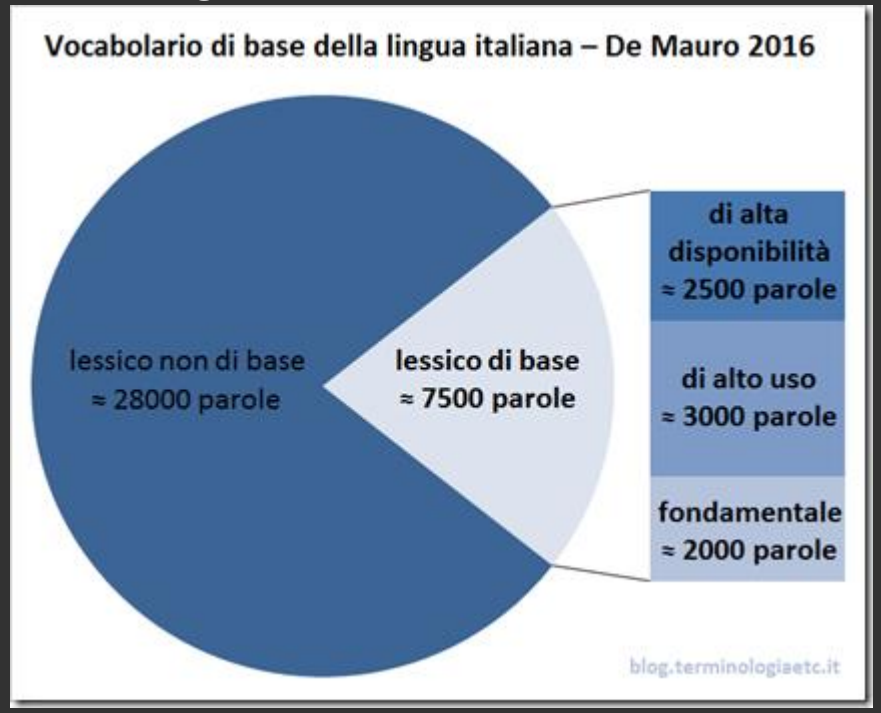
- Over the years the VdB has been used to mark words in fact the main dictionaries of use of the Italian language have used the vocabulary ranges, although with differences in labeling.
- a) vocabulary ranges of the GRADIT (Great dictionary of italian , dir. da T. De Mauro, 8 voll., Torino, UTET, 1999-2007)





Lexical complexity and the new basic vocabulary

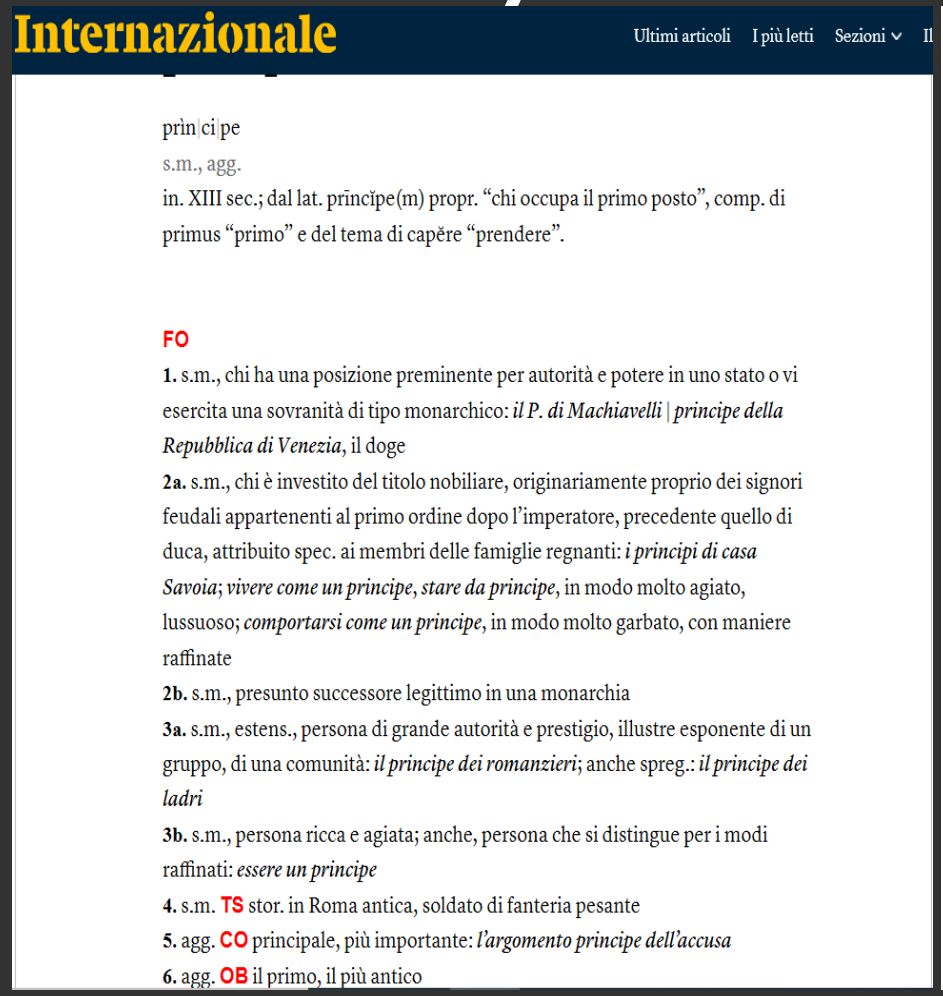
- The renovated version of VdB, **NVdB** (Chiari and De Mauro2016) also includes other vocabulary ranges, that we have included in dictionaries and grammars



Lexical complexity and the new basic vocabulary

- The VdB is based on a corpus of 500,000 occurrences of words collected in five categories,
- The NVdB is based on a corpus of 18,000,000 occurrences of words that are derived by the analysis of a specifically built corpus of contemporany Italian and includes multiword expressions .

Lexical complexity and the new basic vocabulary



Internazionale

Ultimi articoli I più letti Sezioni Il

prìn|ci|pe

s.m., agg.

in. XIII sec.; dal lat. prīncĭpe(m) propr. “chi occupa il primo posto”, comp. di primus “primo” e del tema di capĕre “prendere”.

FO

1. s.m., chi ha una posizione preminente per autorità e potere in uno stato o vi esercita una sovranità di tipo monarchico: *il P. di Machiavelli* | *principe della Repubblica di Venezia*, il doge

2a. s.m., chi è investito del titolo nobiliare, originariamente proprio dei signori feudali appartenenti al primo ordine dopo l’imperatore, precedente quello di duca, attribuito spec. ai membri delle famiglie regnanti: *i principi di casa Savoia*; *vivere come un principe*, *stare da principe*, in modo molto agiato, lussuoso; *comportarsi come un principe*, in modo molto garbato, con maniere raffinate

2b. s.m., presunto successore legittimo in una monarchia

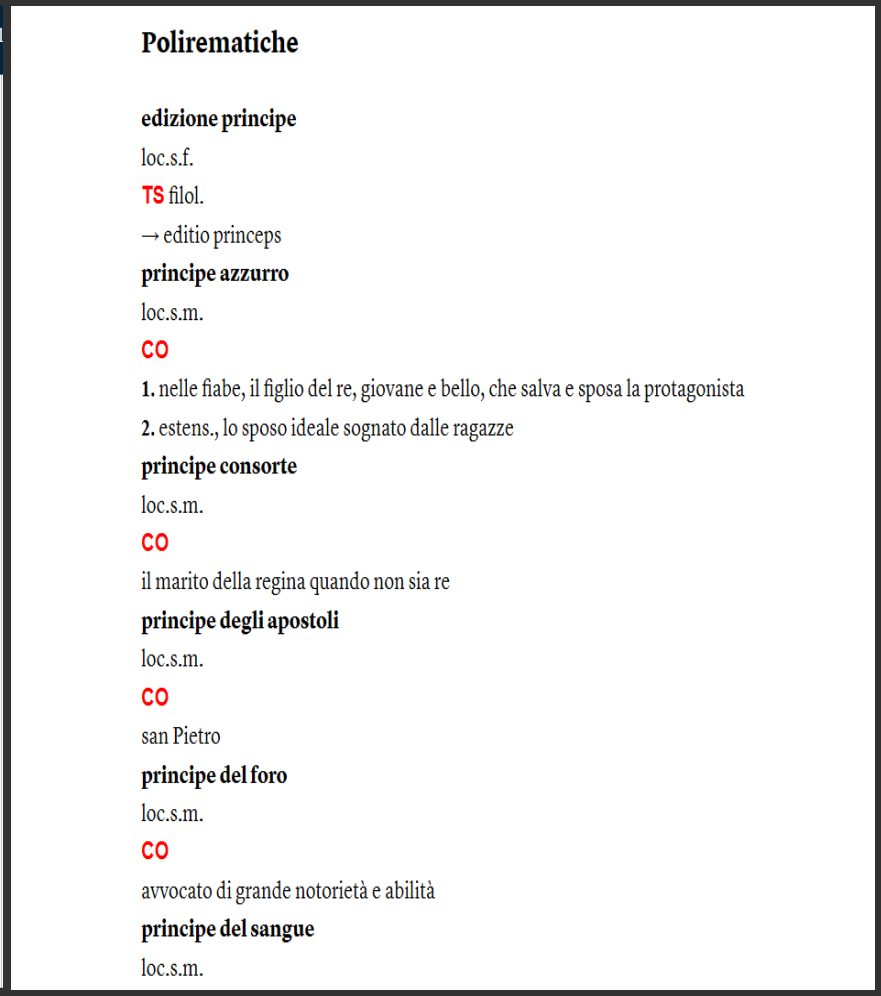
3a. s.m., estens., persona di grande autorità e prestigio, illustre esponente di un gruppo, di una comunità: *il principe dei romanzieri*; anche spreg.: *il principe dei ladri*

3b. s.m., persona ricca e agiata; anche, persona che si distingue per i modi raffinati: *essere un principe*

4. s.m. TS stor. in Roma antica, soldato di fanteria pesante

5. agg. CO principale, più importante: *l’argomento principe dell’accusa*

6. agg. OB il primo, il più antico



Polirematiche

edizione principe

loc.s.f.

TS filol.

→ editio princeps

principe azzurro

loc.s.m.

CO

1. nelle fiabe, il figlio del re, giovane e bello, che salva e sposa la protagonista

2. estens., lo sposo ideale sognato dalle ragazze

principe consorte

loc.s.m.

CO

il marito della regina quando non sia re

principe degli apostoli

loc.s.m.

CO

san Pietro

principe del foro

loc.s.m.

CO

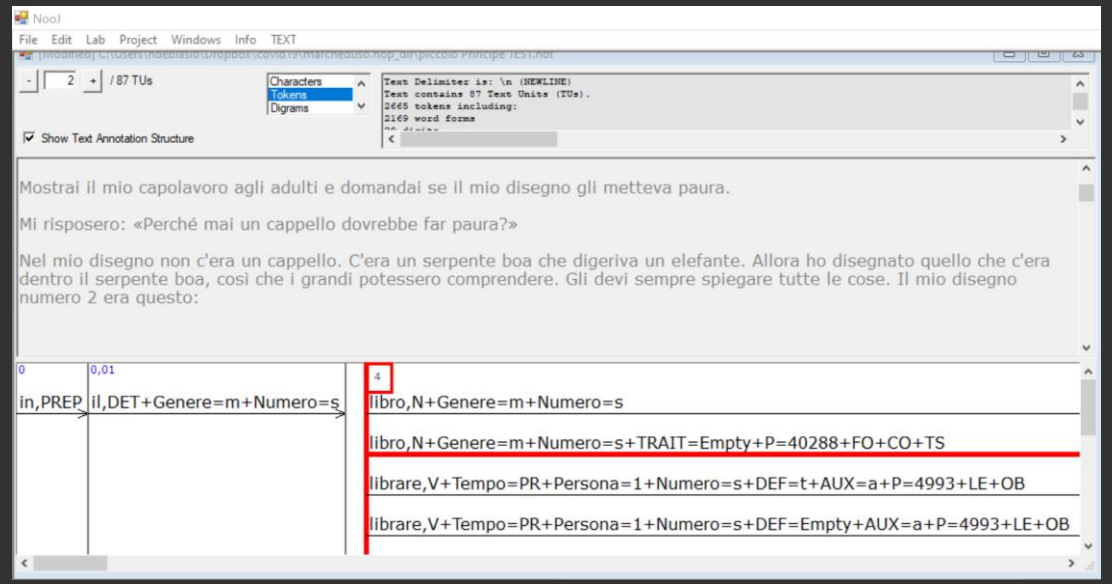
avvocato di grande notorietà e abilità

principe del sangue

loc.s.m.

Annotation structure

In our work we have computerized De Mauro's NVdB, including the Vocabulary ranges, and we have included it within the Italian for Nooj form.

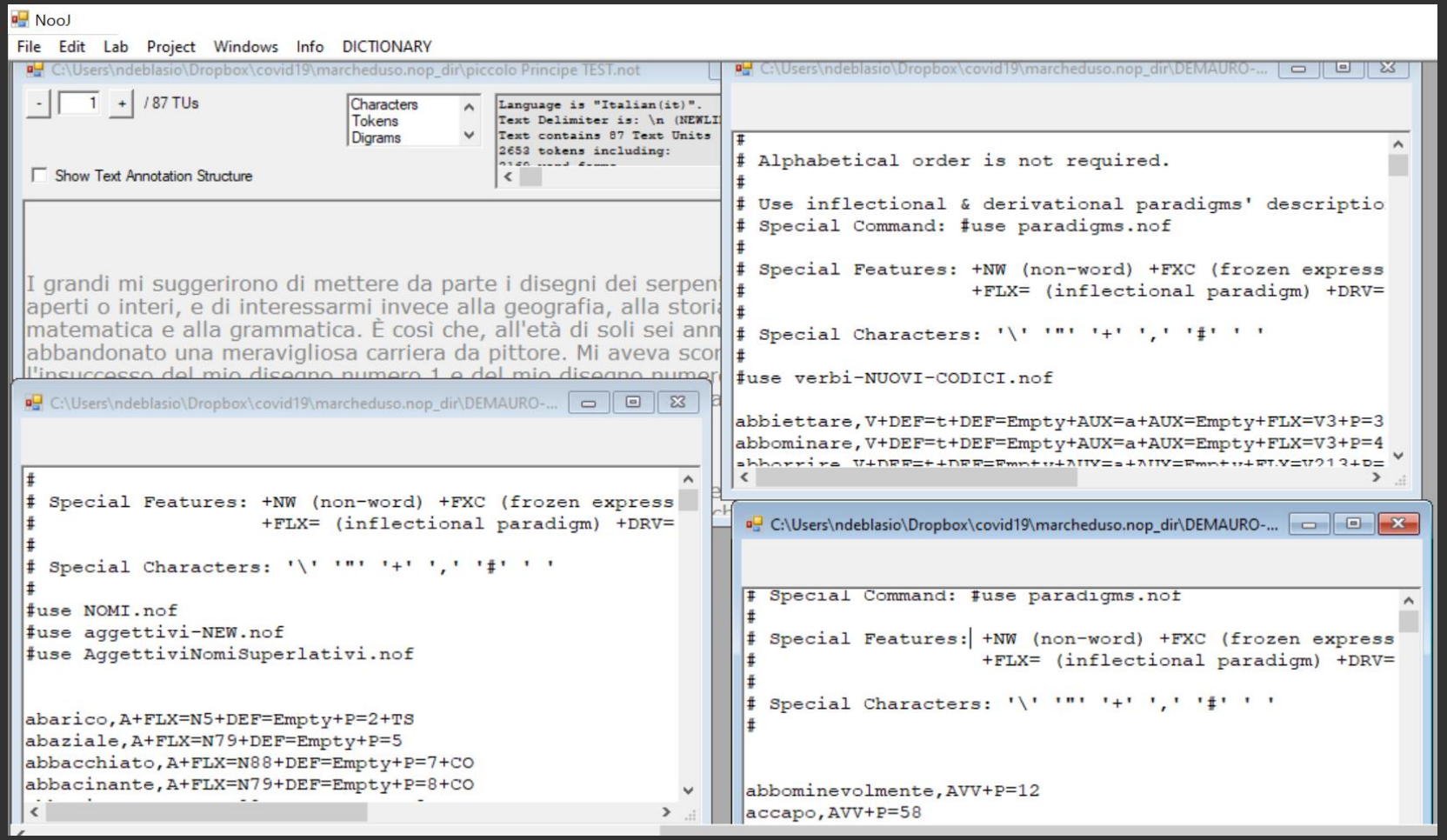


The texts analyzed are:
The Little Prince of Antoine de Saint-Exupéry,
Marcovaldo of Italo Calvino
and five research papers on the subject Coronavirus -19.

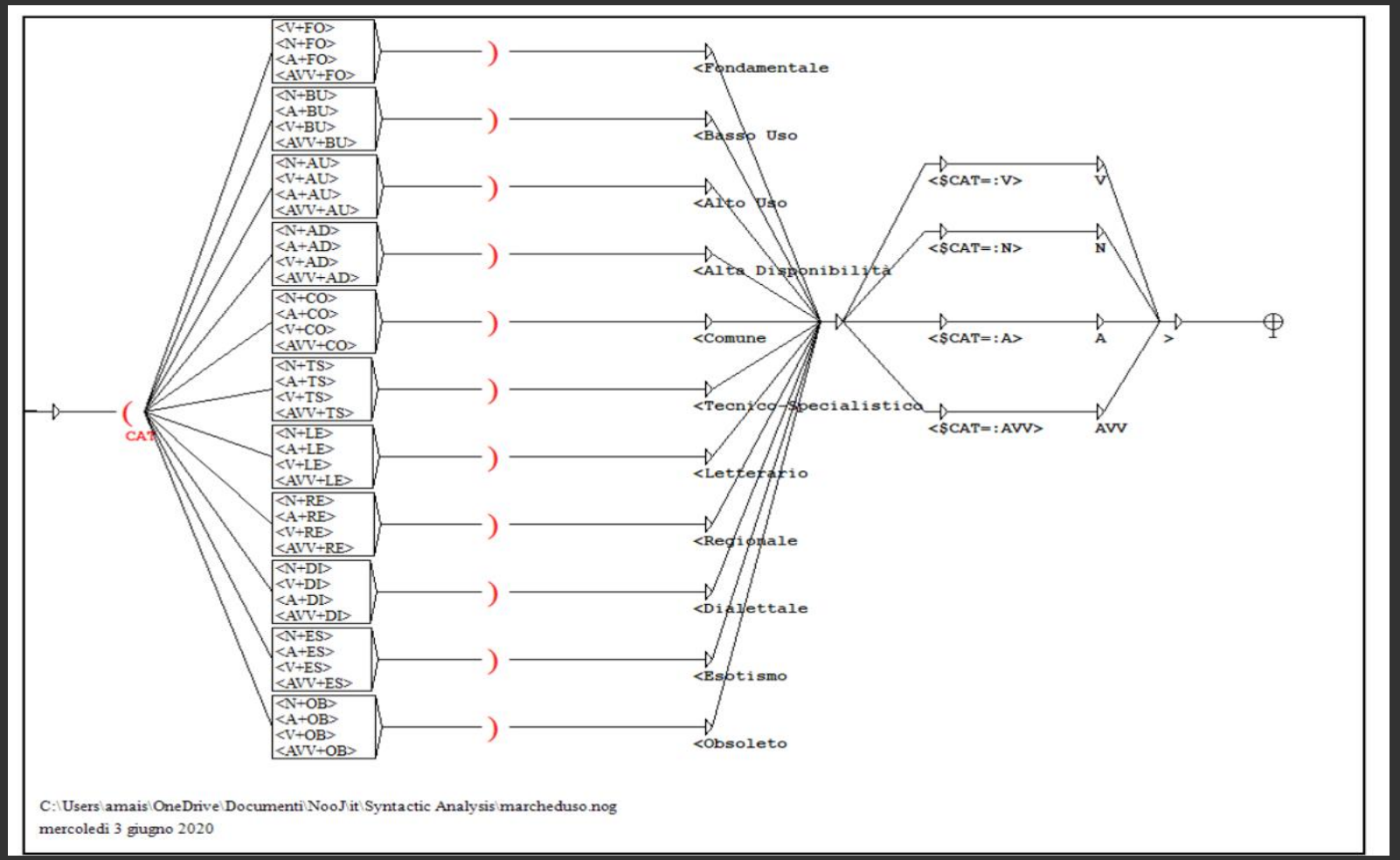
Nooj

- As you can see, electronic dictionaries and grammars have been created and enriched with the Vocabulary ranges of the New basic vocabulary.
- In this analysis, the Vocabulary ranges taken into consideration are:
  - FO – fundamental vocabulary
  - AU – high usage
  - AD – high availability
- The application to the texts shows us that the use of words within a text allows to classify the text as high comprehensibility or not.
- The Vocabulary ranges the have been added to nouns, adjectives, verbs and adverbs.

Dictionary of “vocabulary ranges “



Sintactic grammar of “vocabulary ranges”of the New basic vocabulary



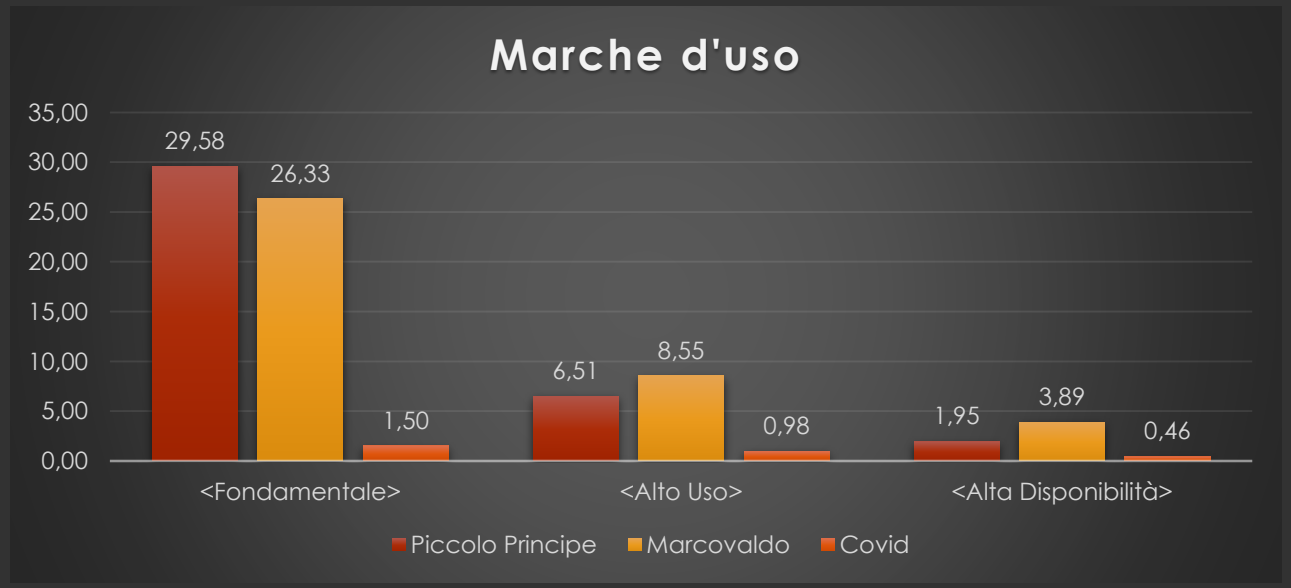
C:\Users\amais\OneDrive\Documenti\NooJ it\Syntactic Analysis\marcheduso.nog
mercoledì 3 giugno 2020

Nooj

- We can see a comparison of Vocabulary ranges between literary text for children and scientific articles with grammar category and a comparison of Vocabulary ranges without grammar category .
- The data resulting from this analysis show us how the three corpus give us feedback on the frequency of the words of the Fundamental vocabulary (FO) completely opposite. The Little Prince have a frequency of 5300 (five thousand three hundred) tokens on about 20,000 twenty thousand tokens , from the Fundamental vocabulary (FO), Marcovaldo have a frequency of about 11,000 eleven thousand tokens on 40,000 forty thousand and research papers 330 (three hundred thirty) token on about 20,000 twenty thousand tokens.

Comparison of Vocabulary ranges between literary text for children and scientific articles

Covid			22020	Piccolo Principe			17916	Marcovaldo			41345
Rank	Term	Frequency		Rank	Term	Frequency		Rank	Term	Frequency	
1	<Fondamentale>	330	1,50	1	<Fondamentale>	5300	29,58	1	<Fondamentale>	10888	26,33
2	<Alto Uso>	216	0,98	2	<Alto Uso>	1166	6,51	2	<Alto Uso>	3533	8,55
3	<Alta Disponibilità>	102	0,46	3	<Alta Disponibilità>	350	1,95	3	<Alta Disponibilità>	1609	3,89



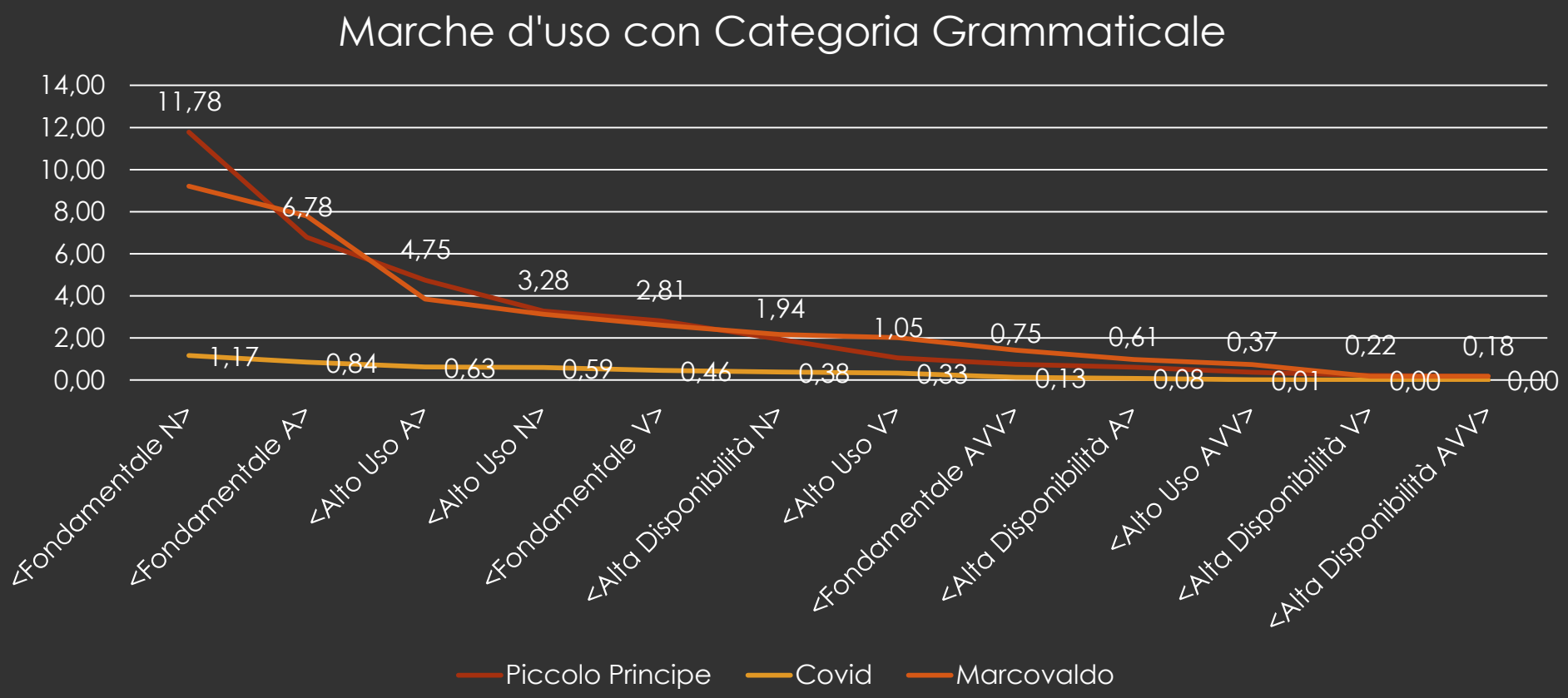
Comparison of Vocabulary ranges between literary text for children and scientific articles with grammar category

	Covid		22020
Rank	Term	Frequency	
1	<Fondamentale N>	257	1,17
2	<Fondamentale A>	186	0,84
3	<Alto Uso A>	138	0,63
4	<Alto Uso N>	130	0,59
5	<Fondamentale V>	101	0,46
6	<Alta Disponibilità N>	83	0,38
7	<Alto Uso V>	72	0,33
8	<Fondamentale AVV>	28	0,13
9	<Alta Disponibilità A>	18	0,08
10	<Alto Uso AVV>	3	0,01
11	<Alta Disponibilità V>	1	0,00

	Piccolo Principe		17916
Rank	Term	Frequency	
1	<Fondamentale V>	2110	11,78
2	<Fondamentale N>	1214	6,78
3	<Fondamentale A>	851	4,75
4	<Fondamentale AVV>	587	3,28
5	<Alto Uso N>	504	2,81
6	<Alto Uso V>	348	1,94
7	<Alto Uso A>	189	1,05
8	<Alta Disponibilità N>	135	0,75
9	<Alta Disponibilità V>	109	0,61
10	<Alta Disponibilità A>	67	0,37
11	<Alta Disponibilità AVV>	39	0,22
12	<Alto Uso AVV>	32	0,18

	Marcovaldo		41345
Rank	Term	Frequency	
1	<Fondamentale V>	3806	9,21
2	<Fondamentale N>	3222	7,79
3	<Alto Uso N>	1591	3,85
4	<Fondamentale A>	1290	3,12
5	<Fondamentale AVV>	1076	2,60
6	<Alto Uso V>	893	2,16
7	<Alta Disponibilità N>	831	2,01
8	<Alto Uso A>	586	1,42
9	<Alta Disponibilità V>	402	0,97
10	<Alta Disponibilità A>	303	0,73
11	<Alto Uso AVV>	75	0,18
12	<Alta Disponibilità AVV>	73	0,18

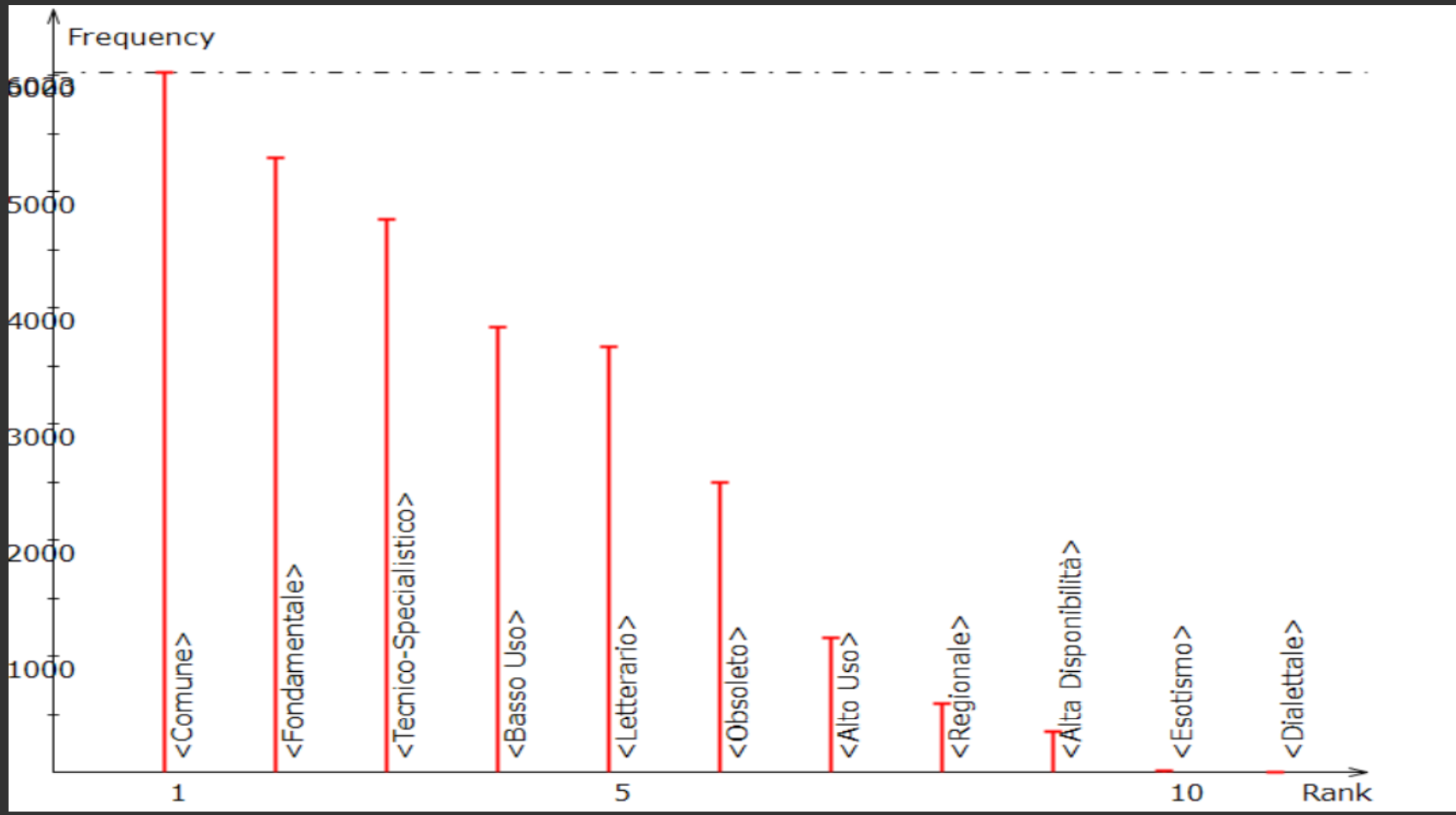
Comparison of Vocabulary ranges between literary text for children and scientific articles with grammar category



Nooj

- The Little Prince and Marcovaldo are texts of classic fiction for children aged between 8 and 11, the other corpus is a scientific text.
- The results are consistent with the hypothesis the high frequency in the examined text of words of the **fundamental vocabulary (FO)** can be considered a high index of comprehensibility.
- In fact, children's literature has a high index of comprehensibility compared to scientific articles with a techno-scientific lexicon

In this statistics analysis you can see the application to the text “Little Prince” of all the Vocabulary ranges include in Nooj's



Conclusion

- This integrated work typology between the calculation of the readability and the level of comprehensibility of the lexicon of the texts is important for the linguistic education of primary school children.

Future work

- the lexical-grammatical
- analysis of the syntactic complexity

Thank you !

Merci pour votre attention !