



Pós-Graduação em Ciência da Computação

**Natália de Melo Franco**

**Vocabulary Selection and Organization for Augmentative and Alternative  
Communication of Children with Speech Impairment**



Universidade Federal de Pernambuco  
posgraduacao@cin.ufpe.br  
<http://cin.ufpe.br/~posgraduacao>

Recife  
2020

**Natália de Melo Franco**

**Vocabulary Selection and Organization for Augmentative and Alternative  
Communication of Children with Speech Impairment**

Tese de Doutorado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Doutor em Ciência da Computação.

**Área de Concentração:** Processamento de sinais e reconhecimento de padrões

**Orientador:** Robson do Nascimento Fidalgo

**Coorientador:** Rinaldo José de Lima

Recife

2020

Catálogo na fonte  
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

F825v Franco, Natália de Melo  
*Vocabulary selection and organization for augmentative and alternative communication of children with speech impairment* / Natália de Melo Franco. – 2020.  
127 f.: il., fig., tab.  
  
Orientador: Robson do Nascimento Fidalgo.  
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2020.  
Inclui referências e apêndice.  
  
1. Processamento de imagens. 2. Reconhecimento de padrões. I. Fidalgo, Robson do Nascimento (orientador). II. Título.  
  
006.4 CDD (23. ed.) UFPE - CCEN 2020 - 95

**Natália de Melo Franco**

**“Vocabulary Selection and Organization for Augmentative  
and Alternative Communication of Children with Speech Impairment”**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutora em Ciência da Computação.

Aprovado em: 04/03/2020.

---

**Orientador: Prof. Robson do Nascimento Fidalgo**

**BANCA EXAMINADORA**

---

Prof. Dr. Frederico Luiz Gonçalves de Freitas  
Centro de Informática/UFPE

---

Profa. Dra. Patrícia Cabral de Azevedo Restelli Tedesco  
Centro de Informática / UFPE

---

Prof. Dr. Alex Sandro Gomes  
Centro de Informática / UFPE

---

Prof. Dr. Munique Massaro  
Centro de Educação/ UFPB

---

Prof. Dr. Evandro de Barros Costa  
Instituto de Computação/UFAL

*I dedicate this work to my beloved parents, Dorgival and Petronila (in memoriam), for all the support and dedication throughout my academic life, from preschool to PhD. I also dedicate this work to Lili Passerino (in memoriam), for betting on me and encouraging me during my PhD, and also for adopting me as one of hers “Periquitas”.*

## ACKNOWLEDGEMENTS

To God for the angels put in my way and all the opportunities of learning and growth.

To my beloved family, especially my Mom (*in memoriam*) and Daddy, for being my guidance, support, and protection in this and in the other plane. To my brothers – Júnior, André –, my sister Renata, and my cousin Fabiola, for the true friendship, the sincere support, and the video calls, that always cheered me up.

To Robson, for being excellent in his functions of advisor, for extrapolating these functions and being my friend, godfather, and support in Recife. To Rinaldo, for all the guidance and learning during this PhD.

To my brothers in this journey – Edson, Thiago, Augusto, Jayr and Robério – for the really good moments of this PhD, sharing coffee (lots of coffee!) and unhealthy food, laughing at our misfortunes and, above all, taking care of each other (*“All for one and one for all”*). Our fake Daddy (Robson) made us a real family!

To my 702 girls – Carla, Rosalva and Helenise – for sharing the house, life, good and bad times with me. For Erica and Georgia, who also shared all of this with us of apartment 702. You girls made my life in Recife better.

To Filipe (my favorite brother-in-law), Dr. Gustavo and Jeferson, for helping me recover and maintain my mental health during the last year.

To the evaluation committee – Fred Freitas, Patrícia Tedesco, Lili Passerino (*in memoriam*), Alex Sandro, Munique Massaro, and Evandro Costa – for the valuable comments and corrections about this thesis.

To the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) for the financial support.

Finally, to all those who, in some way, contributed for me to achieve this goal, my most sincere thanks!

## ABSTRACT

Augmentative and Alternative Communication (AAC) systems are used to supplement or replace speech/writing of individuals with complex communication needs. These systems can be used for several applications beyond communication (e.g., inclusive education and serious games), and can be of two types: “low-tech” (i.e., non-technological systems based on pictures, objects, and communication books), or “high-tech” (i.e., technological systems such as speech-generating devices, and AAC software). To achieve good results, AAC systems must provide access to a large vocabulary, which is adequate to communication development and organized to facilitate the pictogram retrieval. Both, vocabulary selection and organization are essential tasks for successful AAC usage. To be easier, these tasks can be based on proposals of core word lists, and of vocabulary organization. However, there is no consensus among these proposals. In this thesis, we analyze the existing proposals of words and categories to propose a *Core Vocabulary* for children’s communication, which should also be organized to provide facilities for several high-tech AAC applications. In this context, we perform two Systematic Literature Review (SLR) to select, respectively, core word lists and category lists, and, based on these lists, we generate new core word lists using Natural Language Processing (NLP) techniques and new category lists using semantic clustering. Then, we conduct statistical analysis over corpora extracted from CHILDES Corpus Database to analyze the core word lists, and qualitative analysis with experts to analyze the category lists. Next, we organize the better results of words and categories into an ontology enriched with data from semantic databases (e.g., WordNet) and foundation ontologies (e.g., Ontolex). This thesis presents two main results: 1) a Core Vocabulary, formed by the best core word list organized in the categories that represent the essential concepts for children communication; and 2) an ontology (AACOnto), which organizes the core vocabulary and enriches it with syntactic and semantic data. Our Core Vocabulary supports vocabulary selection and organization tasks to implement a vocabulary on AAC systems that allows children to express themselves using different communicative functions. In addition, our ontology provides syntactic and semantic information about vocabulary instances, which allows the implementation of versatile high-tech AAC systems for several applications.

**Keywords:** Augmentative and Alternative Communication. Core Vocabulary. Children. Ontology.

## RESUMO

Os sistemas de Comunicação Aumentativa e Alternativa (CAA) são usados para complementar ou substituir a fala/escrita de indivíduos com necessidades complexas de comunicação. Esses sistemas podem ser usados para várias aplicações além da comunicação (e.g., educação especial e jogos sérios) e podem ser de dois tipos: de baixa tecnologia (i.e., sistemas não tecnológicos baseados em figuras, objetos e livros de comunicação) ou de alta tecnologia (i.e., sistemas tecnológicos como dispositivos de geração de fala e softwares de CAA). Para obter bons resultados, os sistemas de CAA devem fornecer acesso a um vocabulário grande, adequado ao desenvolvimento da comunicação e organizado para facilitar a recuperação de pictogramas. Tanto a seleção quanto a organização do vocabulário são tarefas essenciais para o uso bem-sucedido da CAA. Para ser mais fácil, essas tarefas podem ser baseadas em propostas de listas de palavras principais e de organização do vocabulário. No entanto, não há consenso entre essas propostas. Nesta tese, analisamos as propostas de palavras e categorias existentes para propor um *Vocabulário Base* para a comunicação das crianças, que também deve ser organizado para fornecer facilidades para as várias aplicações de CAA de alta tecnologia. Nesse contexto, executamos duas Revisões Sistemáticas da Literatura para selecionar, respectivamente, listas de palavras principais e listas de categorias e, com base nessas listas, geramos novas listas de palavras principais usando as técnicas de Processamento de Linguagem Natural e novas listas de categorias usando agrupamento semântico. Em seguida, realizamos: 1) análises estatísticas sobre corpora extraídos do CHILDES Corpus Database para analisar as listas de palavras; e 2) análises qualitativas com especialistas sobre as listas de categorias. Em seguida, organizamos os melhores resultados de palavras e categorias em uma ontologia enriquecida com dados de bancos de dados semânticos (e.g., WordNet) e ontologias de referência (e.g., Ontolex). Esta tese apresenta dois resultados principais: 1) um *Vocabulário Base*, formado pela melhor lista de palavras principais organizada nas categorias que representam os conceitos essenciais para a comunicação infantil; e 2) uma ontologia (AACOnto), que organiza o vocabulário principal e o enriquece com dados sintáticos e semânticos. Nosso *Vocabulário Base* dá suporte para as tarefas de seleção e organização de vocabulário em sistemas de CAA, para que as crianças se expressem usando diferentes funções comunicativas. Além disso, nossa ontologia fornece informações sintáticas e semânticas sobre as instâncias do vocabulário, o que permite a implementação de sistemas de CAA de alta tecnologia que sejam versáteis para várias aplicações.

**Palavras-chaves:** Comunicação Aumentativa e Alternativa. Vocabulário Base. Crianças. Ontologia.



## LIST OF FIGURES

Figure 1 – Sentence example using pictograms with captions . . . . .	17
Figure 2 – Reference Architecture for AAC Systems . . . . .	17
Figure 3 – <i>aBoard</i> app . . . . .	20
Figure 4 – The communication triad using AAC . . . . .	23
Figure 5 – Color-Coding Systems . . . . .	27
Figure 6 – Sentences examples with Colorful Semantics . . . . .	27
Figure 7 – Example of MLU calculation . . . . .	32
Figure 8 – WordNet full schema visualization . . . . .	34
Figure 9 – Ingestion frame relations of FrameNet . . . . .	35
Figure 10 – Ingestion frame of FrameNet with core frame elements . . . . .	35
Figure 11 – Ontology classification based on domain scope . . . . .	36
Figure 12 – Example of the main constructors of an ontology. . . . .	38
Figure 13 – Implementation of Figure 12 example on Manchester OWL Syntax . . . . .	39
Figure 14 – Components of Visual Vocabulary for Aphasia (ViVA) . . . . .	45
Figure 15 – Application of the User-centric Recommendation Model . . . . .	46
Figure 16 – User-centric Recommendation Model modules . . . . .	47
Figure 17 – Excerpt of the daily activities ontology . . . . .	48
Figure 18 – Excerpt of the data dictionary for ontology terms classification . . . . .	48
Figure 19 – Method to add personalized information to an existing ontology . . . . .	49
Figure 20 – Simple Upper Ontology (SUPO) hypothesis . . . . .	50
Figure 21 – Examples of properties of PictOntology . . . . .	51
Figure 22 – Examples of categories in PictOntology . . . . .	51
Figure 23 – Overview of materials and methods . . . . .	54
Figure 24 – Detailing of Material Selection Task . . . . .	54
Figure 25 – Studies selection flowchart of Word Lists . . . . .	55
Figure 26 – Studies selection flowchart of Category Lists . . . . .	59
Figure 27 – The Core OntoLex-Lemon Model . . . . .	63
Figure 28 – Morphosyntactic properties details of LexInfo . . . . .	63
Figure 29 – The class hierarchy of the WordNet schema . . . . .	64
Figure 30 – SKOS scheme . . . . .	64
Figure 31 – Detailing of Material Generation Task . . . . .	65
Figure 32 – MLU vs. Age distribution of Test Corpora samples . . . . .	66
Figure 33 – Test Corpora according to the AGE range of Brown’s Stages . . . . .	66
Figure 34 – Test Corpora according to the MLU range of Brown’s Stages . . . . .	66
Figure 35 – Detailing of Core Word Analyses Task . . . . .	72
Figure 36 – Detailing of Core Categories Analyses and Vocabulary Organization Task . . . . .	74

Figure 37 – CVR interpretation scale . . . . .	75
Figure 38 – Amount of synsets sense per word on WordNet . . . . .	77
Figure 39 – Detailing of Vocabulary Organization Task . . . . .	77
Figure 40 – Overview and detailed view of Major Lists and Super List recall over Test Corpora . . . . .	81
Figure 41 – Overview and detailed view of Corpus-based Lists and Composed Lists recall over Test Corpora . . . . .	82
Figure 42 – Recall of the Selected Core Word List over Age Corpora . . . . .	84
Figure 43 – Recall of the Selected Core Word List over MLU Corpora . . . . .	85
Figure 44 – Suggestion of a Category Hierarchy . . . . .	91
Figure 45 – Imported ontologies by AACOnto . . . . .	93
Figure 46 – AACOnto Schema . . . . .	95
Figure 47 – Example of instances for the concept “ <i>Cat</i> ” on AACOnto . . . . .	97
Figure 48 – Instance “ <i>Cat</i> ” on AACOnto . . . . .	98
Figure 49 – Instance “ <i>Where</i> ” on AACOnto . . . . .	98

## LIST OF TABLES

Table 1 – McShane’s Communicative Functions. . . . .	24
Table 2 – WordNet 2.0 database statistics . . . . .	33
Table 3 – Current FrameNet Status . . . . .	34
Table 4 – Start set of papers. . . . .	42
Table 5 – Selected papers. . . . .	43
Table 6 – Criteria met by ViVA. . . . .	46
Table 7 – Criteria met by User-centric Recommendation Model. . . . .	49
Table 8 – Criteria met by SUPO. . . . .	52
Table 9 – Criteria met by each Related Work. . . . .	53
Table 10 – The Sub-CHILDES Corpus details. . . . .	61
Table 11 – Characteristics of the Training and the Test corpora. . . . .	67
Table 12 – Summarization of AGE and MLU Corpora extracted from Test Corpora. . . . .	67
Table 13 – Super List Distribution over Commonalities . . . . .	68
Table 14 – High-commonality words of Super List . . . . .	68
Table 15 – Combined List Characterization considering corpus-based lists of 500 words . . . . .	69
Table 16 – High-commonality words that do not appear in the corpus-based lists . . . . .	70
Table 17 – Merged Category List. . . . .	70
Table 18 – Core word lists ordination. . . . .	73
Table 19 – Semantic group categories. . . . .	76
Table 20 – Maximum recall and number of lemmas of each lists over Test Corpora. . . . .	83
Table 21 – Maximum recall and number of words of the Selected Core Word List over AGE and MLU Corpora. . . . .	86
Table 22 – 1 <sup>st</sup> Delphi round summarization. . . . .	88
Table 23 – 2 <sup>nd</sup> Delphi round summarization. . . . .	89
Table 24 – AACOnto Ontology specification . . . . .	93
Table 25 – An excerpt of the Data Dictionary for terms classification and descriptions. . . . .	96
Table 26 – Examples of relations in ontology. . . . .	96
Table 27 – AACOnto metrics. . . . .	97
Table 28 – OntoQA schema metrics for AACOnto. . . . .	99
Table 29 – OntoQA knowledge metrics for AACOnto. . . . .	100
Table 30 – OntoQA Class Metrics. . . . .	101
Table 31 – Comparative Evaluation with Related Works. . . . .	103
Table 32 – Core Vocabulary. . . . .	122

## LIST OF ABBREVIATIONS AND ACRONYMS

<b>AAC</b>	Augmentative and Alternative Communication
<b>ARASAAC</b>	Aragonese Portal of Augmentative and Alternative Communication
<b>ASD</b>	Autism Spectrum Disorder
<b>ASHA</b>	American Speech-Language-Hearing Association
<b>BNC</b>	British National Corpus
<b>BPMN</b>	Business Process Model and Notation
<b>CHILDES</b>	Child Language Data Exchange System
<b>DF</b>	Document Frequency
<b>DL</b>	Description Logic
<b>EC</b>	Exclusion Criteria
<b>GUI</b>	Graphical User Interface
<b>IC</b>	Inclusion Criteria
<b>IDF</b>	Inverse Document Frequency
<b>MLU</b>	Mean Length of Utterance
<b>NLP</b>	Natural Language Processing
<b>NLTK</b>	Natural Language Toolkit
<b>NLU</b>	Natural Language Understanding
<b>OWL</b>	Web Ontology Language
<b>PLI</b>	Primary Language Impairment
<b>POS</b>	Part of Speech
<b>RA</b>	Reference Architecture
<b>RDF</b>	Resource Description Framework
<b>RDFS</b>	RDF Schema
<b>RQ</b>	Research Questions
<b>SKOS</b>	Simple Knowledge Organisation System
<b>SLR</b>	Systematic Literature Review
<b>SO</b>	Specific Objective
<b>SUpO</b>	Simple Upper Ontology
<b>TF</b>	Term Frequency
<b>TF-IDF</b>	Term Frequency – Inverse Document Frequency

<b>TTR</b>	Type-Token Ratio
<b>URI</b>	Unique Resource Identifier
<b>ViVA</b>	Visual Vocabulary for Aphasia
<b>W3C</b>	World Wide Web Consortium
<b>WD</b>	Word Density
<b>WN Schema</b>	WordNet Full Schema
<b>WSD</b>	Word Sense Disambiguation

## CONTENTS

<b>1</b>	<b>INTRODUCTION . . . . .</b>	<b>16</b>
1.1	CONTEXT AND MOTIVATION . . . . .	16
1.2	OBJECTIVES AND RESEARCH QUESTIONS . . . . .	19
1.3	SCOPE DELIMITATION . . . . .	19
1.4	TARGET AUDIENCE AND CONTRIBUTIONS . . . . .	20
1.5	DOCUMENT ORGANIZATION . . . . .	21
<b>2</b>	<b>THEORETICAL FOUNDATION . . . . .</b>	<b>22</b>
2.1	AUGMENTATIVE AND ALTERNATIVE COMMUNICATION . . . . .	22
<b>2.1.1</b>	<b>Customization Requirements . . . . .</b>	<b>24</b>
2.1.1.1	Vocabulary Selection . . . . .	24
2.1.1.2	Vocabulary Organization . . . . .	25
<b>2.1.2</b>	<b>Communication Requirements . . . . .</b>	<b>26</b>
2.1.2.1	Pictogram Selection . . . . .	26
2.1.2.2	Construction of Meaningful Sentences . . . . .	27
2.2	NATURAL LANGUAGE PROCESSING . . . . .	28
2.3	KNOWLEDGE REPRESENTATION . . . . .	31
<b>2.3.1</b>	<b>Corpus . . . . .</b>	<b>31</b>
2.3.1.1	CHILDES Corpus Database . . . . .	31
<b>2.3.2</b>	<b>Linguistic Knowledge Databases . . . . .</b>	<b>32</b>
2.3.2.1	WordNet . . . . .	33
2.3.2.2	FrameNet . . . . .	34
<b>2.3.3</b>	<b>Ontology . . . . .</b>	<b>35</b>
2.4	CHAPTER FINAL CONSIDERATIONS . . . . .	40
<b>3</b>	<b>RELATED WORKS . . . . .</b>	<b>41</b>
3.1	WORKS SELECTION . . . . .	41
3.2	WORK EVALUATION CRITERIA . . . . .	43
3.3	WORKS PRESENTATION . . . . .	44
<b>3.3.1</b>	<b>Visual Vocabulary for Aphasia (ViVA) . . . . .</b>	<b>44</b>
<b>3.3.2</b>	<b>User-centric Recommendation Model . . . . .</b>	<b>46</b>
<b>3.3.3</b>	<b>Simple Upper Ontology (SUPO) . . . . .</b>	<b>49</b>
3.4	COMPARATIVE ANALYSES . . . . .	52
3.5	CHAPTER FINAL CONSIDERATIONS . . . . .	53
<b>4</b>	<b>MATERIALS AND METHODS . . . . .</b>	<b>54</b>

4.1	SELECTION OF WORD LISTS . . . . .	54
4.1.1	<b>Major Lists</b> . . . . .	<b>56</b>
4.2	SELECTION OF CATEGORY LISTS . . . . .	58
4.2.1	<b>Category Lists</b> . . . . .	<b>58</b>
4.3	SELECTION OF KNOWLEDGE REPRESENTATIONS . . . . .	60
4.3.1	<b>Sub-CHILDES Corpus</b> . . . . .	<b>60</b>
4.3.2	<b>Selection of the Computational Representation</b> . . . . .	<b>61</b>
4.3.3	<b>Selection of Semantic Databases</b> . . . . .	<b>62</b>
4.3.4	<b>Reference Ontologies</b> . . . . .	<b>62</b>
4.4	CORPORA GENERATION . . . . .	65
4.4.1	<b>Training and Test Corpora</b> . . . . .	<b>67</b>
4.4.2	<b>Age and MLU Corpora</b> . . . . .	<b>67</b>
4.5	WORD LISTS GENERATION . . . . .	67
4.5.1	<b>Super List</b> . . . . .	<b>68</b>
4.5.2	<b>Corpus-based Lists</b> . . . . .	<b>68</b>
4.5.3	<b>Combined Lists</b> . . . . .	<b>69</b>
4.6	MERGED CATEGORY LIST GENERATION . . . . .	70
4.7	DATA PREPROCESSING . . . . .	71
4.8	CORE WORD ANALYSES . . . . .	72
4.9	CORE CATEGORY ANALYSES AND VOCABULARY CATEGORIZATION	74
4.9.1	<b>Categories Selection</b> . . . . .	<b>74</b>
4.9.2	<b>Word Categorization and Vocabulary Evaluation</b> . . . . .	<b>76</b>
4.10	VOCABULARY ORGANIZATION . . . . .	77
4.11	CHAPTER FINAL CONSIDERATIONS . . . . .	79
5	<b>RESULTS</b> . . . . .	<b>80</b>
5.1	CORE WORD ANALYSES . . . . .	80
5.1.1	<b>Recall Analyses over Test Corpora</b> . . . . .	<b>80</b>
5.1.2	<b>Recall Analyses over Age Corpora and MLU Corpora</b> . . . . .	<b>83</b>
5.1.3	<b>Summary of Results</b> . . . . .	<b>85</b>
5.2	CATEGORY ANALYSES AND VOCABULARY CATEGORIZATION . . . . .	87
5.2.1	<b>Category Analyses</b> . . . . .	<b>87</b>
5.2.2	<b>Vocabulary Categorization</b> . . . . .	<b>89</b>
5.2.3	<b>Summary of Results</b> . . . . .	<b>90</b>
5.3	CORE VOCABULARY ORGANIZATION . . . . .	92
5.3.1	<b>Ontology Specification</b> . . . . .	<b>92</b>
5.3.2	<b>Knowledge Acquisition</b> . . . . .	<b>92</b>
5.3.3	<b>Integration</b> . . . . .	<b>93</b>
5.3.4	<b>Conceptualization</b> . . . . .	<b>94</b>
5.3.5	<b>Implementation: The AACOnto Ontology</b> . . . . .	<b>94</b>

5.3.6	<b>Evaluation</b> . . . . .	<b>99</b>
5.3.7	<b>Summary of Results</b> . . . . .	<b>101</b>
5.4	COMPARATIVE EVALUATION . . . . .	102
5.5	USAGE GUIDELINES: HOW CAN OTHERS MAKE USE OF THIS WORK	104
5.6	CHAPTER FINAL CONSIDERATIONS . . . . .	105
<b>6</b>	<b>CONCLUSIONS</b> . . . . .	<b>106</b>
6.1	FINAL CONSIDERATIONS . . . . .	106
6.2	CONTRIBUTIONS . . . . .	107
6.3	LIMITATIONS AND FUTURE WORKS . . . . .	108
	 <b>REFERENCES</b> . . . . .	 <b>110</b>
	 <b>APPENDIX A – CORE VOCABULARY</b> . . . . .	 <b>121</b>



# 1 INTRODUCTION

This chapter presents this thesis proposal, highlighting its context, motivation, objectives and research questions, as well as the scope delimitation, the target audience and expected contributions. Finally, the structure of the other chapters is presented.

## 1.1 CONTEXT AND MOTIVATION

Augmentative and Alternative Communication (AAC) is an Assistive Technology dedicated to expand functional communication skills of people with complex communication needs, including intellectual disability and neurological disorders (e.g., Cerebral Palsy<sup>1</sup>, Dysarthria<sup>2</sup> or Autism Spectrum Disorder (ASD)<sup>3</sup>) (World Health Organization, 2015). According to the American Speech-Language-Hearing Association (ASHA) (ASHA, 2019b), functional communication skills are forms of behavior that express needs, wants, feelings, and preferences that others can understand. The development of these skills promotes independence and social inclusion of people with complex communication needs (ASHA, 2019b). On the one hand, when communication problems affect literate people, they can type their messages using a conventional keyboard. On the other hand, in the case of children or non-literate people, the communication needs to be more intuitive. In this case, AAC based on a graphics system can be used, because it consists of a set of images/pictograms – such as the pictograms of the Aragonese Portal of Augmentative and Alternative Communication (ARASAAC) (PALAO, 2019) (cf. Figure 1) – with captions that represent objects, actions, feelings, etc. For this reason, AAC systems based on a graphics system are recommended for use in special education and in speech therapy interventions with children with complex communication needs (MIRENDA, 2003; WALKER et al., 2018; FRANCO et al., 2017; KAGOHARA et al., 2013). Notice that AAC can be used by people of any age, but the earlier it is used, the better the result (CRESS; MARVIN, 2003). In this sense, in this work, we will focus on the early childhood.

AAC systems can be of two types: “*low-tech*” (i.e., non-technological systems based on pictures, objects, and communication books), or “*high-tech*” (i.e., technological systems such as speech-generating devices, and AAC software). Concerning high-tech AAC systems, despite their importance and the number of systems available, we highlight that the knowledge produced is fragmented into several scientific proposals and technological

<sup>1</sup> Cerebral palsy is a group of disorders that affect a person’s ability to move and maintain balance and posture. These people also have related conditions such as intellectual disability and problems with hearing, or speech (CDCP, 2019).

<sup>2</sup> Dysarthria is a speech disorder caused by muscle weakness. It can make conversation difficult because interlocutors may have trouble understanding what people with dysarthria say. (ASHA, 2019).

<sup>3</sup> Autism Spectrum Disorder is a developmental disorder that affects communication and behavior (NIMH, 2019).

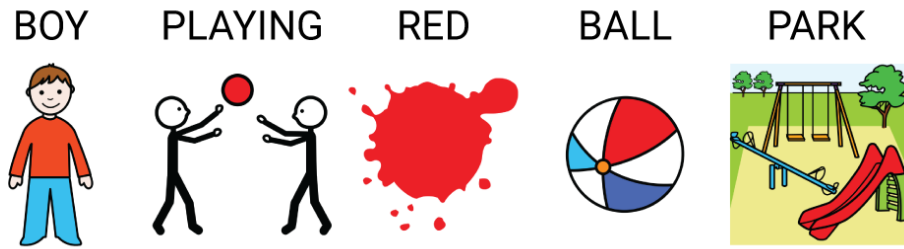


Figure 1 – Sentence example using pictograms with captions.

solutions, consequently, there is no consolidated baseline for developing robust AAC systems. To overcome these shortcomings, in Franco et al. (2018) we presented the first effort towards a Reference Architecture (RA) for developing and evaluating high-tech AAC systems robust enough to effective usage in communication and education of disabled people. In this context, a RA can be used as a conceptual guide that compiles key concepts, relationships, and features of a domain to define a template solution for a concrete software architecture (CLOUTIER et al., 2010). That is, a RA specifies an abstract and agnostic solution that guides and constrains the instantiations of concrete software architectures, ensuring standardization and interoperability, as well as, reducing costs/risks and increasing the quality of these instantiations. Figure 2 shows the RA proposed in Franco et al. (2018), which encompasses all the requirements of a robust AAC system shown in Section 2.1 and have components to deal with each requirement. In Franco et al. (2018) we detail all these requirements and components and how the RA connect them. In particular, this work will focus in the *Content Management* module, responsible for select and organize the AAC vocabulary.

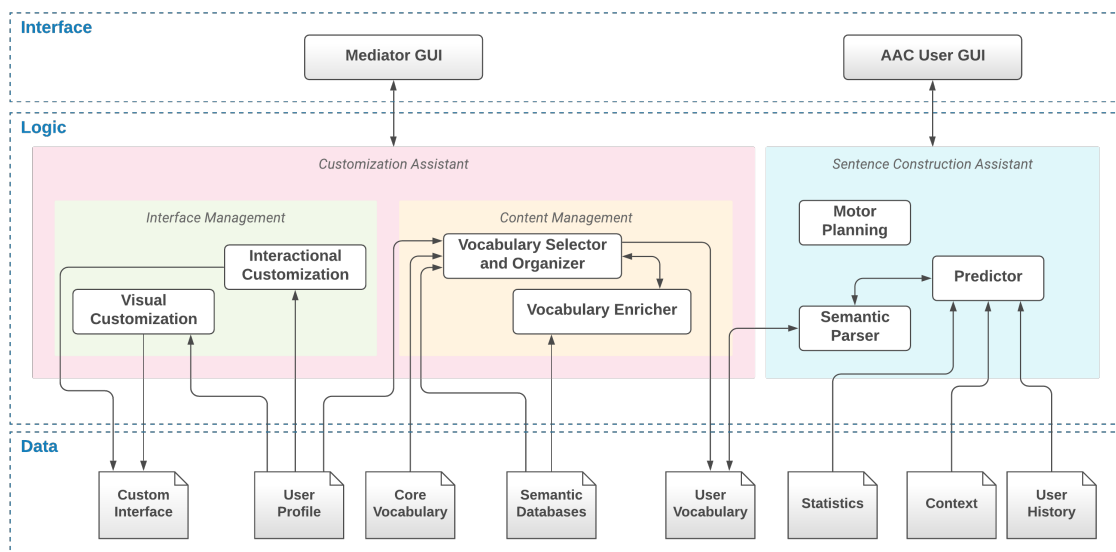


Figure 2 – Reference Architecture for AAC Systems.

By the age of 16 months, typical developing children can comprehend 195 words and produce 65 words (CASELLI et al., 1995), while by the age of 30 months they comprehend 900 words and produce 500 words (PAUL, 2007), approximately. These numbers show the increasing ratio of children vocabulary and, despite being for children of typical development, this increasing ratio may also be applied for children with complex communication needs, even if it takes longer. Ideally, AAC systems should be intuitive, offering maximum power of communication with minimal learning effort (LIGHT, 1997; FRANCO et al., 2018). That is, the AAC system must provide access to a large vocabulary that is adequate to the user communication development (BROWN, 1973) and organized to facilitate the pictogram retrieval (CAMPOS; GOMES, 2007). For this reason, both vocabulary selection and organization are essential tasks for successful use of AAC system, and are the first step for an effective and efficient communication development (FALLON; LIGHT; PAIGE, 2001). These tasks are usually done by mediators (e.g., health or education professionals or family members of children) who are not disabled and its outcome may have the bias of their conceptual model that does not necessarily correspond to children conceptual model (BEUKELMAN; JONES; ROWAN, 1989; DRAGER et al., 2003). Besides, mediators must also consider other aspects of language, such as syntax, semantic and pragmatic (cf. Section 2.2), to select words from different syntactic categories, also known as Part of Speech (POS), to provide a vocabulary that allows variability of communication and the production of syntactically and semantically correct sentences.

To achieve good results, the vocabulary selection task can be based on many scientific proposals that present *core word* lists (cf. Section 4.1.1), which covers the most common and most commonly used words by children. By the same token, the vocabulary organization task can be based on scientific proposals (cf. Section 4.2.1) and technological solutions (PORTER; CAFIERO, 2009; FRANCO et al., 2017; PALAO, 2019) which uses different approaches to organize AAC vocabularies. These proposals can be used by communication partners as a start point to facilitate vocabulary selection and organization tasks for low-tech and high-tech systems. However, despite having the same goal, there is no consensus among these proposals. Also, considering high-tech systems, there are only a few proposals that describe how to represent these vocabularies in a technological system to take advantage of available resources and provide more functionalities.

Considering the advantages of using high-tech devices over other AAC devices (STILL et al., 2014), in addition to the vocabulary selection and organization, we highlight the importance of the computational representation of this vocabulary. The ideal computational representation should be more than a simple representation of what is already done in low-tech AAC systems. This representation should allow the implementation of more complex functionalities through the use of existing computational resources (e.g., resources of input, processing, and output).

## 1.2 OBJECTIVES AND RESEARCH QUESTIONS

The main objective of this thesis is to select and organize the vocabulary used by children with complex communication needs on AAC systems. To reach this main objective, we have the following Specific Objective (SO):

- *SO-1* – To systematically generate new lists of core words and compare its coverage of children’s utterances in comparison with scientific proposals of core word lists;
- *SO-2* – To systematically generate a new category list and analyze if these categories are useful for children communication and sufficient to categorize all the core words of the resulting list of SO-1;
- *SO-3* – Select a computational approach to organizing the categorized vocabulary of SO-2 that allows new functionalities in a high-tech AAC system.

To accomplish these objectives, this work is guided by the following Research Questions (RQ):

- *RQ-1* – *Is it possible to generate a new core word list with better recall than the existing core word lists?*
- *RQ-2* – *Are the already proposed categories useful and sufficient to categorize all the core words of the best list of RQ-1?*
- *RQ-3* – *How can AAC vocabulary be computationally organized to allow new functionalities in a high-tech AAC system?*

## 1.3 SCOPE DELIMITATION

This work is delimited in 4 aspects:

1. *Language* – given the availability of lexical and semantic resources, this work is developed in English;
2. *Type of AAC System* – this work will focus on AAC systems based on a graphics system with a set of images/pictograms with captions that represent concepts (e.g., objects, actions, and feelings) Figure 3 shows the aBoard app (LIMA et al., 2017), an example of an AAC system based on a graphics system;
3. *Target Audience* – formed by children with complex communication needs who cannot write in a conventional manner nor use a conventional keyboard (e.g., QWERT) to communicate. Regarding the domain of written language, the public of this proposal can be literate or not. In the case of a literate child, cognitive deficits may

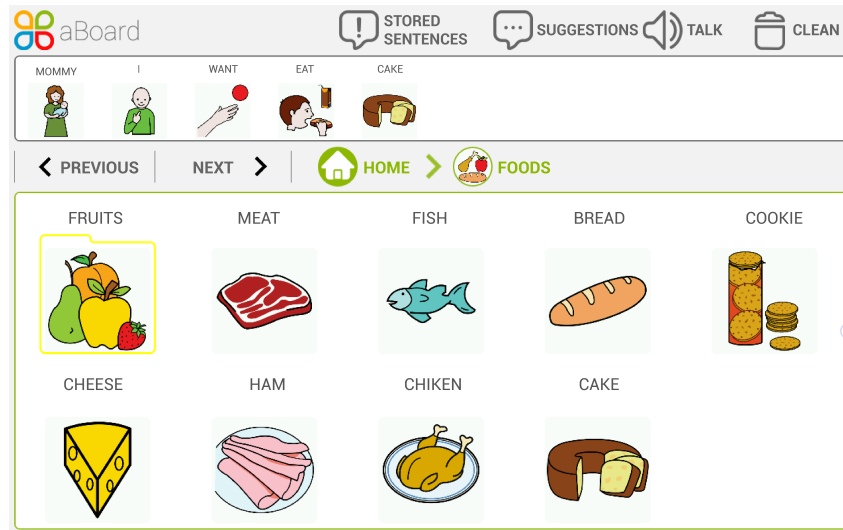


Figure 3 – *aBoard* app.

compromise the use of written language and, in this case, AAC is seen as a complementary resource (i.e., facilitator of communication). In the case of a non-literate child, AAC is considered an alternative resource for communication because it is based on a graphics system;

4. *RA Modules* – this work will focus on the *Content Management* module of the FRANCO et al. RA (FRANCO et al., 2018). The other modules, as well as the instantiation of the RA are outside the scope of this work.

#### 1.4 TARGET AUDIENCE AND CONTRIBUTIONS

The proposed solutions to solve the problems discussed in this chapter are aimed at three main audiences: mediators, AAC system developers, and AAC users. The first two are directly benefited, while the third is indirectly benefited, as follows:

- *Mediators* – this audience is responsible for selecting and organize the vocabulary that will be displayed in the AAC systems. For them, we deliver a *Core Vocabulary* formed by relevant words and categories for children that may help and guide the vocabulary selection and organization tasks;
- *AAC system developers* – this audience is responsible for develop high-tech AAC systems. For them, we deliver an ontology that computationally organizes the *Core Vocabulary*. This audience may use the semantic knowledge contained in the ontology to implement more complex functions in AAC systems, such as pictogram suggestion and construction of meaningful sentences;

- *AAC users* – this audience will be indirectly benefited by the proposed solutions of this work when using the systems, whether high or low technology prepared by the two previous audiences.

## 1.5 DOCUMENT ORGANIZATION

The remaining chapters of this thesis are structured as follows:

**Chapter 2 – Theoretical Foundation:** presents the theoretical foundations necessary to support this work.

**Chapter 3 – Related Works:** presents the scientific works related to this work.

**Chapter 4 – Materials and Methods:** presents and describes all the materials and methods used in the development of this work.

**Chapter 5 – Results:** presents and discusses the results of this work.

**Chapter 6 – Conclusions:** presents the final considerations on the main topics covered in this thesis, including the contributions achieved and the indications of future works.

Finally, appended to this document are 1 appendix:

**Appendix A – Core Vocabulary:** presents the complete Core Vocabulary produced in this work.

## 2 THEORETICAL FOUNDATION

This chapter presents the theoretical foundation that served as the basis for this research, with the necessary definitions for the understanding of this thesis. Section 2.1 conceptualize the AAC, and discusses the requirements to construct robust AAC systems. Section 2.2 presents the basic concepts of Natural Language Processing (NLP) necessities to understand the method we used in this work. Finally, Section 2.3 shows some methods to represent knowledge, focusing on corpus, linguistic knowledge databases, and ontologies.

### 2.1 AUGMENTATIVE AND ALTERNATIVE COMMUNICATION

According to the ASHA (ASHA, 2019a):

*“Augmentative and Alternative Communication (AAC) is an area of clinical practice that addresses the needs of individuals with significant and complex communication disorders characterized by impairments in speech-language production and/or comprehension, including spoken and written modes of communication.”*

Figure 4 shows the communication triad observed in AAC interventions with external support, which is composed of: 1) The AAC user – the person who uses AAC to communicate; 2) The AAC system – the external support, whether high or low technology, used in the intervention; and 3) The communication partner – the person who interacts with the AAC user, also known as mediator. There are several characteristics that must be considered in each point of the triad, for example: for the AAC user, it must be considered their communication competencies and quality of vision and hearing; for the AAC system, it must be considered the vocabulary selection and organization; and for the communication partner, it must be considered their engagement and the presumption of the AAC user competence. In this section, we will discuss only the aspects related to the external support of the AAC system.

Regarding the use of an external support, AAC systems can be *unaided* – which does not use any tools or devices outside the body (e.g., gestures, body language, facial expressions, and sign language) – or *aided* – which uses some kind of tool or device (e.g., a pen and a paper, pictures on a board, computer programs, and mobile applications). Aided systems can be of two types: “*low-tech*” or *non-technological* systems, which has low sophistication and is produced using low cost materials (e.g., a pen and a paper, and pictures on a board); or “*high-tech*” or *technological* systems, which is very sophisticated and is used with electronic devices (e.g., computer programs, and mobile applications) (BERSCH, 2013; COOK; POLGAR, 2014).



Figure 4 – The communication triad using AAC.

From a computer science perspective, a AAC system is a set of programs that collaborate with one another to provide an integrated hardware and software solution to help people overcome speech, reading, or writing disabilities. Given the specifics of AAC systems, developing this type of system is not a trivial task. AAC developers should prioritize some functional requirements for developing robust AAC systems (KAGOHARA et al., 2013; MCNAUGHTON; LIGHT, 2013; NEWELL; LANGER; HICKEY, 1998). These requirements include:

- Vocabulary customization – the system must allow the selection and organization of the vocabulary, as well as its language representation method;
- Interface customization – the system must allow different configurations for customizing the way the user see, listen and manipulate the interface;
- Communication flexibility – the system must allow freedom for the user to say whatever they want, shortcuts to increase the communication speed;
- Appropriate feedback – the system must provide an understanding of its status for the user through visual and sound notifications;
- Construction of meaningful sentences – the system must provide clues to help users to select pictograms and construct syntactic and semantically correct sentences.

We grouped these requirements into two groups: 1) Customization (i.e., vocabulary customization and interface customization); and 2) Communication (i.e., communication flexibility, appropriate feedback and construction of meaningful sentences). Franco et al. (2018) details these requirements and use them to propose a conceptual and technology-independent Reference Architecture for instantiate and evaluate concrete architectures for AAC systems. In this chapter, we detail only the requirements that are closely related to this work.



### 2.1.1 Customization Requirements

Here we deal with vocabulary customization, which comprises the vocabulary selection, and the vocabulary organization.

#### 2.1.1.1 Vocabulary Selection

Ideally, AAC systems should be intuitive, offering maximum power of communication with minimal learning effort (LIGHT, 1997; FRANCO et al., 2018). That is, the AAC system must provide access to a large vocabulary which is adequate to its communication development (BROWN, 1973). Moreover, the vocabulary selection should allow children to express themselves according to different communicative functions (LEONARD et al., 1982; MCSHANE, 1980). For instance, McShane (1980) proposes a system that consists of five major communicative functions (i.e., Regulation, Statement, Exchange, Personal, and Conversation) which cover the expected communication of a child (cf. Table 1).

Table 1 – McShane’s Communicative Functions.

Major Function [Specific Functions]	Objective
<b>Regulation</b> [Attention, Vocative, Request]	Attempts to regulate the behavior of another person by gaining it’s attention (Attention) or presence (Vocative), as well as requesting an object or assistance in some activity (Request).
<b>Statement</b> [Naming, Description, Information]	Make statements in the form of naming (Naming) or describing (Description) something, as well as giving information about a non-current event (Information).
<b>Exchange</b> [Giving, Receiving]	Indicates that the child is giving or trying to give an object to another person or is receiving an object from another person.
<b>Personal</b> [Doing, Determination, Refusal, Protest]	Express what the child is currently doing (Doing), is about to do (Determination), refuses to do (Refusal), or protests about when made to do (Protest).
<b>Conversation</b> [Question, Imitation, Answer, Follow-on]	Deals with child’s requests for information (Question) and responses to preceding utterances (Imitation, Answer, Follow-on).

The words that constitute the child vocabulary can be categorized according to three metrics that are not mutually exclusive (i.e., a content word can also be core and basic concept):

1. *Content words* or *Structure/function words* (BOENISCH; SOTO, 2015; TOMASELLO, 2009; MARVIN; BEUKELMAN; BILYEU, 1994) – *content words* are highly referenced and can be used alone for labeling and describing perceptible entities (e.g., objects, events, properties) that can provide the main idea of each utterance. This set of words includes POS as nouns, verbs, adjectives, and adverbs which are the first words learned by children. In turn, *structure words* provide linguistic connections among content words and need to be combined with other words within a sentence

to be meaningful. This set of words includes POS as pronouns, auxiliary verbs, conjunctions, and prepositions, which are later acquired by children because they are more abstract for them. Previous studies show that structure words are hardly included in the AAC systems for children (LUND; LIGHT, 2007; SUTTON et al., 2000), but even with less meaning than content words, the presence of structure words enable children to create a complete sentence and achieve a higher level of language (MARVIN; BEUKELMAN; BILYEU, 1994);

2. *Core words* or *Fringe words* (BANAJEE; DICARLO; STRICKLIN, 2003; YORKSTON et al., 1988; ROBILLARD et al., 2014) – *core words* have high degree of commonality (i.e., words that are shared among many users) and high frequency of use (i.e., words that are habitual to many users), while *fringe words* are highly individualized and have little commonality among users. In this context, a set of core words is called *core vocabulary* whereas a set of fringe words is called *fringe vocabulary*. According to the Center for Literacy and Disability Studies (2018), since the use of nouns is highly determined by its context, they are usually considered as fringe vocabulary, while verbs, pronouns, and articles are considered as core vocabulary. Brazas and Bourke (2016) also define the *personal core* as an individualized subset of the fringe vocabulary which includes words that a person uses all the time;
3. *Basic concept words* (BRACKEN; CRAWFORD, 2010; MCCARTHY et al., 2017) – comprehends primitive concepts in specific categorical areas (cf. Section 4.2) that represent a functional vocabulary needed for children to understand classroom conversations and teacher directions (BRACKEN; CRAWFORD, 2010; MCCARTHY et al., 2017), develop their vocabulary (RHYNER; BRACKEN, 1988) and readiness (PANTER; BRACKEN, 2009), to cite some.

#### 2.1.1.2 Vocabulary Organization

Concerning vocabulary organization, children can organize their vocabulary in countless ways. However, the words in the AAC system must be retrieved using a preprogrammed organization that does not necessarily match the mental organization of the children (DRAGER et al., 2003). According to Light and Drager (2002), there are at least five different approaches to organizing vocabulary:

1. *Taxonomic* – uses hierarchical categories (e.g., people, places, food, actions);
2. *Schematic* – uses event schema (e.g., getting ready for bed, eating breakfast, circle time at preschool, snack time);
3. *Semantic-syntactic* – uses semantic groups organized from left to right to facilitate language development (e.g., agents, actions, descriptors, objects);

4. *Alphabetic* – uses the alphabetical order of words;
5. *Idiosyncratic* – uses unique organizational systems specific to an individual.

Further research by Drager et al. (2003) analyzes the effects of the three first approaches with young children using dynamic display AAC systems. In general, children had great difficulty learning to locate target vocabulary in the three first approaches, but the schematic one seems to be significantly better than the other two, maybe because this approach may be similar to the children conceptual model organization. Conversely, the taxonomic approach had better learning curve when compared to the other two approaches. Young children are able to understand taxonomic categories when the objects are familiar and the categories are labeled for them (KRACKOW; GORDON, 1998; LUCARIELLO; KYRATZIS; NELSON, 1992; MARKMAN; COX; MACHIDA, 1981), mainly when they enter school and are exposed to more adult mental models. Despite the initial difficulties, the use of taxonomies allows the establishment of standards for the classification of information through the use of inheritance mechanisms, and also allows users to learn from these conceptual structures (CAMPOS; GOMES, 2007).

## 2.1.2 Communication Requirements







Here we deal with pictogram selection and construction of meaningful sentences.

### 2.1.2.1 Pictogram Selection

The pictogram selection comprises (i) color-coding systems (FITZGERALD, 1949; BRYAN, 2003) and (ii) Motor Planning (HALLORAN; EMERSON, 2006). The color-coding systems are used to quickly find pictograms in a group by searching for colors that carry some meaning (cf. Figure 5). The most popular color coding system is the Fitzgerald Key (FITZGERALD, 1949) which groups concepts into six colors regarding its grammatical role and was used to teach deaf children to read by structuring sentences correctly. The Colorful Semantics (BRYAN, 2003) group concepts into five color-coded incremental levels based on the meaning of words (i.e., semantics) and it is used to develop children's grammar. Notice that only the purple color (i.e., Describe) can be moved to different places in the sentence to add a description to other words (cf. Figure 6).

The Motor Planning (HALLORAN; EMERSON, 2006) is related to the pictogram localization in the screen. The Motor Planning encompasses the planning and execution of a series of movements to select a specific pictogram, which brings consistency in pictogram search path and avoids variants in the same pictogram selection. These consistent motor patterns for pictogram selection allow the development of automaticity in communication. That is, the user learns the path (i.e., a sequence of clicks and swipes in the screen) to a specific pictogram and do not need to think about the pictogram path anymore.

## Fitzgerald Key

 Noun	 Verb	 Social
 Pronoun	 Adjective	 Miscellaneous

## Colorful Semantics



Figure 5 – Color-Coding Systems.

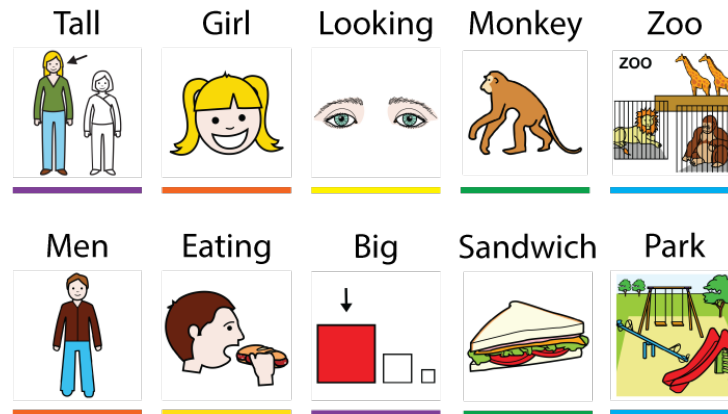


Figure 6 – Sentences examples with Colorful Semantics.

### 2.1.2.2 Construction of Meaningful Sentences

The construction of meaningful sentences comprises a variety of strategies that allow the construction of syntactically and semantically correct sentences. Here we detail the following strategies: (i) the theory of scripts (TODMAN et al., 2008) and the frame-semantic approach (LOWE, 1997), (ii) the prediction techniques, and (iii) the COMPANSION system (DEMASCO; MCCOY, 1992).

On the one hand, in the theory of scripts (TODMAN et al., 2008), a script denotes a structured communication dialog likely predictable. For example, a script can specify a sequence of steps required to complete a particular task (e.g., buy something in a grocery store). On the other hand, in the frame-semantic approach, a semantic frame is a data structure that represents familiar objects and common situations (LOWE, 1997). For example, the *Ingestion* frame model the verb *to eat* has two frame elements *Ingestor* and *Ingestible*. AAC developers can implement these strategies using questions and answers, where each answer interferes with the following question. This way, the communication is the result of the answers combination.

The prediction techniques guide the user’s communication by reducing the number of possible pictograms they have available at a given moment. More precisely, one can distinguish two specific kinds of prediction during sentence construction: when the user selects the very first pictogram or when the user has already chosen one or more pictograms. In the first case, the system can suggest pictograms based on (i) context knowledge, (ii) user’s log, or (iii) general statistical knowledge; in the second case, in addition to the previews points, the system can also suggest pictograms according to (iv) natural language aspects (i.e., syntax, semantics, and pragmatics) or (v) improved statistical knowledge considering previously selected pictograms, which will be detailed next. Contextual knowledge as time and geographical localization (e.g., School is a diary event that takes place between 8 am and noon when the user is near to the school) can be used to suggest contextual-related pictograms. User’s log with previously built sentences can also be used as a start point in prediction. Concerning the general statistical knowledge, one can use large text corpora (e.g., CHILDES Corpus Database – cf. Section 2.3.1) as a training data set to extract the most used single words and the most used collocations (i.e., the habitual juxtaposition of two words). To address natural language aspects, one can use simple grammatical rules based on color-code schemes (i.e., Fitzgerald Key and Colorful Semantics) and semantic-frames. The first one can be used as a clue for the user to identify missing pictograms in the sentence, while the second one can establish the correct complement for a given verb (e.g., the verb *eat* requires an *eatable thing* as a complement). Indeed, empirical evidence found in Bolderson et al. (2011) show that the adoption of Colorful Semantics yielded good results for children with severe communicative disorders since they often omit verbs and grammatical elements and fail to build complete sentences. One can also use semantic databases (e.g., WordNet and FrameNet – cf. Section 2.3.2) to add semantic and pragmatic restrictions on the simple grammar described above. Finally, concerning the improved statistical knowledge, it is based on the general statistical knowledge improved by restrictions established by previews selected pictograms. We highlight that the theory of scripts can also be regarded as a prediction technique since a given script can be employed to implement a conversational recommender (HERNÁNDEZ et al., 2014) (cf. Section 3.3.2) that makes suggestions and refinements based on the user’s feedback.

After the sentence construction, the COMPANSION system (DEMASCO; MCCOY, 1992) (COMPRESSED exPANSION) can be used to expand the telegraphic language (i.e., the compressed message made by sentences with missing words – “*Daddy here*”) into the full and grammatically corrected sentence (i.e., “*Daddy is here*”).

## 2.2 NATURAL LANGUAGE PROCESSING

According to Chowdhury (2003), Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech. The use of NLP in computer programs aims

to facilitate human-computer interaction through the use of natural language. NLP also provides theoretical and practical fundamentals for applications involving: text mining; word prediction; semantic parser; automatic translation; text summarizing; text generation; and ontology population. The NLP has several particularities that become challenges in this field, for example, the natural language has several ambiguities, and words can be combined into sentences in many ways, which makes it hard for computers to point the specific meaning of a word.

NLP applications require various kinds of knowledge and analysis of the language. According to Jurafsky (2000), there are 6 levels of knowledge of language: 1) Phonetics and Phonology – knowledge about linguistic sounds; 2) Morphology – knowledge of the meaningful components of words; 3) Syntax – knowledge of the structural relationships between words; 4) Semantics – knowledge of meaning; 5) Pragmatics – knowledge of the relationship of meaning to the goals and intentions of the speaker; and 6) Discourse – knowledge about linguistic units larger than a single utterance. In this section, we will present a simplified view of the levels 2 to 5 (i.e., from morphology to pragmatics):

- *Morphology* – The morphology deals with rules for inflection and word formation that provides a rich vocabulary for communication. It concerns the internal structure of words, which is made from one or more morphemes. A morpheme is the smallest unit in natural language that carries meaning. Words that are built from more than one morpheme can be split into a *stem* (the part of the word that never changes) and one or more *affix* (an additional element placed at the beginning or end of a stem to modify its meaning). The word “*produced*” is made from the stem “*produc*” and the affix “*ed*” representing the function morpheme for past tense. Also in this example, the word “*produced*” can be represented by its *lemma* “*produce*”, which is the form of the word that appears as an entry in a dictionary and is used to represent all the other possible *word forms* (e.g., “*producing*” and “*produces*”);
- *Syntax* – The syntax refers to the way words are arranged together, forming legal structures of a language. Words that present a similar behavior are categorized into syntactic categories, also known as Part of Speech (POS). Such categories include noun, verb, adjective, adverb, and pronouns, for example. A *grammar* is a set of valid syntactic combination of these categories into phrases. For example, the sentence “*The boy is going to the school*” is a syntactically correct sentence in English; whereas “*The boy going school to the is*” is not;
- *Semantics* – the semantics studies the meaning of a language unit: a word, a sentence, or the entire discourse. It focuses on how words can be combined into meaningful sentences. For example, although the sentence “*I’m eating an orange juice*” is syntactically correct, it is semantically incorrect since a beverage (i.e., “*orange juice*”) is not a proper complement to the verb “*eat*”; in contrast, the sentences

*“I’m drinking an orange juice”* and *“I’m eating an orange”* are syntactically and semantically correct;

- *Pragmatics* – the pragmatics studies the language use in terms of how sentences relate to one another and to the context it is used. For example, the sentence *“Pruning a tree is a long process”* can be interpreted in the literal sense as the process of cutting the physical tree, or in the context of a computer science algorithm (THANAKI, 2017).

The aforementioned knowledge provides the necessary basis for the execution of tasks involving NLP. Regarding semantic knowledge, the tasks that are based on it are directed to the Natural Language Understanding (NLU) (ALLEN, 1995). These tasks try to make computers able to understand texts written or spoken in natural language and then, take some action. For this, NLP applications may be based on statistical analyses of corpus (i.e., collections of text and speech – cf. Section 2.3.1), and on linguistic knowledge databases (i.e., databases that organize the lexicon according to its properties – cf. Section 2.3.2). To achieve good results, these applications should use well structured and annotated corpus, once the punctuation marks and the POS tagging annotation allows the better understanding of sentences and word meanings (JURAFSKY, 2000) to make the aforementioned analyses (i.e., the morphological, syntactic, semantic, and pragmatics analyses).

There are several tools and metrics to support the NLP with corpus. In this work we present examples using the Natural Language Toolkit (NLTK) (LOPER; BIRD, 2002) of Python language (PYTHON CORE TEAM, 2019), but there are other tools such as the Stanford CoreNLP (MANNING et al., 2014) for Java. Depends on the purpose, NLP analyses over corpus can be based on different sizes of *n-grams*, which is a contiguous sequence on *n* items from a given text. The counting of *n-grams* in a corpus are used to improve the word prediction, once it determines the most likely word to come next. Regardless of *n-grams*, the analyses can be based on the *word form* – as it appears in the corpus – that can produce long lists to analyze, or on the *lemma* – using the *lemmatization* function – that produces smaller lists to analyze. These analyses can also disregard a set of *stop words* (e.g., *“the”*, *“is”*, *“at”*, *“which”*, and *“on”*). There is no consensus on a stop word list, however, these words are generally the most common words in a language. Concerning the word count, analyses can be based on *types* – which means the number of distinct words in a corpus, that is, the size of the vocabulary – or *tokens* – which means the total number of running words. Notice that the word count can be based on word forms or lemmas. The Type-Token Ratio (TTR) is a metric that relates the total number of unique words (*types*) divided by the total number of words (*tokens*) in a given segment of language, that is,  $TTR = types/tokens$ . Other metrics include: the Term Frequency (*TF*), which points the most frequent words in a document; the Document Frequency (*DF*), which points the

number of documents analyzed that contain a specific word; and the Term Frequency – Inverse Document Frequency ( $TF * IDF$ ), which points the most relevant words in a set of documents. These and other metrics are well described in Section 4.5.2.

## 2.3 KNOWLEDGE REPRESENTATION

There are several ways to store and organize knowledge. In this section, we will focus on the representation of written texts and words. Concerning texts, in Section 2.3.1 we present the concept of *corpus* to organize these knowledge. Concerning words, it can be represented with and without associated semantics. Disregarding semantics, words can be represented in a bag-of-words (JURAFSKY, 2000), which is an unordered set of words. Regarding semantics, words can be represented, for example, in a: 1) Word list, which considers some kind of order among words; 2) Mental map, which is a mapping of concepts connected by subjective links; or 3) Taxonomy, which is a particular arrangement of words into a tree structure with more objective relations of *is-a* (e.g., “dog” *is-a* “animal”, and “animal” *is-a* “living being”). We present in Sections 2.3.2 and 2.3.3, two approaches for knowledge representation that are based on taxonomies: linguistic knowledge databases, and ontologies.

### 2.3.1 Corpus

Sardinha (2000) defines *Corpus* as a set of textual linguistic data collected for researching a language or linguistic variety and used to explore the language using empirical evidence. We highlight that *Corpus* also includes non-textual linguistic data sets (e.g., audible or visual data). As corpus example, we can cite the British National Corpus (BNC) (University of Oxford, 1995) – a collection of written and spoken words from various sources designed to represent a broad cross section of English British – and the Child Language Data Exchange System (CHILDES) Corpus Database (MACWHINNEY, 2000) – a collection of documents with speech transcripts and children’s texts that will be detailed next.

#### 2.3.1.1 CHILDES Corpus Database

The CHILDES Corpus Database was developed to serve as a central repository for the first language acquisition data, and has samples (transcripts, audio, and video) in 32 languages from 130 different corpora. For each sample, the following aspects are provided: 1) date and duration of the interaction; 2) the scenario in which the sample was collected (i.e., which people were involved and what was the context of the conversation); and 3) the child language development level, which can be measured by age or language proficiency. Regarding the last aspect, Brown (1973) defined five stages of child development based on the age range (in months) and the Mean Length of Utterance (MLU), which is a measure of language proficiency calculated as the ratio between the total number of morphemes



and the total number of sentences ( $MLU = \frac{N_{morphemes}}{N_{sentences}}$ ). Figure 7 shows two examples of how to calculate the MLU, the first example with 2 sentences containing 7 morphemes, and the second example with 1 sentence containing 12 morphemes. Notice that the higher the MLU, the more proficient in language the child is. The Brown’s Stages (BROWN, 1973) are still broadly used (KOVACS; HILL, 2017; DEMARIS; SMITH, 2017) and are defined as follows:

- **Stage I** (MLU range: 1.0-2.0; mean MLU: 1.75; age range: 12-26 months) – children who already have accumulated a vocabulary of 50 to 60 words;
- **Stage II** (MLU range: 2.0-2.5; mean MLU: 2.25; age range: 27-30 months) – children who learned how to use “-ing”, “in”, “on”, and “-s” plurals;
- **Stage III** (MLU range: 2.5-3.0; mean MLU: 2.75; age range: 31-34 months) – children who learned how to use irregular past tense, “’s” possessive, and the verb “to be” in its full form as the main verb in a sentence;
- **Stage IV** (MLU range: 3.0-3.75; mean MLU: 3.5; age range: 35-40 months) – children who learned how to use articles, regular paste tense, and third person regular in the present tense;
- **Stage V** (MLU range: 3.75-4.5; mean MLU: 4.0; age range: 41-46 months) – children who learned how to use third person irregular, the full form of the verb “to be” when it is an auxiliary verb in a sentence, and the shortened form of the verb “to be” when it is the only verb or an auxiliary verb in a sentence.

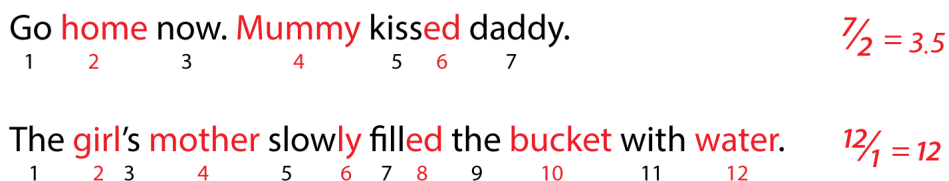


Figure 7 – Example of MLU calculation.

### 2.3.2 Linguistic Knowledge Databases

Linguistic knowledge bases organize the lexicon according to its properties and the structure of each word, being an important component in areas such as NLP, disambiguation of concepts and information retrieval. As linguistic knowledge bases example, we can cite: WordNet (MILLER, 1995), FrameNet (BAKER; FILLMORE; LOWE, 1998), VerbNet (SCHULER, 2005), and BabelNet (NAVIGLI ROBERTO E PONZETTO, 2012). In this work, we will present in detail, WordNet and FrameNet.

### 2.3.2.1 WordNet

The WordNet (MILLER, 1995) relates the various meanings of a term to other terms, according to human semantic organization. In this case, terms are grouped into a set of synonyms called *Synsets*, which connect by semantic (i.e., related to various senses of a word) and lexical (i.e., word-related) relationships, such as links of *hyperonymy* (super-class), *hyponymy* (subclass), *meronymy* (whole part), *holonymy* (everywhere), *antonymy* (opposite meanings) and *similarity* (similar meanings). For example, the word “car” is in both synsets: 1) {*car, auto, automobile, machine, motorcar*} and 2) {*car, railcar, railway car, railroad car*}. These are two different entities in WordNet because they have a different meaning. The first one is explained by the textual description (i.e., *gloss*) “*a motor vehicle with four wheels; usually propelled by an internal combustion engine*”, whereas the second one is explained by “*a wheeled vehicle adapted to the rails of railroad*”. In other words, a synset contains one or more *Word Senses* and each of them belongs to exactly one synset. In addition, each *Synset* is also defined by a set of phrases that illustrate usage and meaning of its *Word Sense* terms.

Boyd-Graber et al. (2006) notice that WordNet’s network is relatively sparse with a few connections among Synsets, specially cross-POS links, and no weighted arcs, once WordNet links are qualitative rather than quantitative. These shortcomings limit the results in tasks such as Word Sense Disambiguation (WSD). To move beyond this shortcoming, Boyd-Graber et al. (2006) proposed the use of the *evocation* technique – that is, how much the first concept brings to mind the second concept – to add quantified, oriented arcs between pairs of Synsets.

Table 2 shows the statistics of WordNet 2.0, and Figure 8 shows the WordNet Full Schema (World Wide Web Consortium, 2006) built in the WebVOWL Editor (WIENS; LOHMANN; AUER, 2018).

Table 2 – WordNet 2.0 database statistics

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	114648	79689	141690
Verb	11306	13508	24632
Adjective	21436	18563	31015
Adverb	4669	3664	5808
<b>Totals</b>	152059	115424	203145

Source: 2.0 wnstats(7WN), October 2019.

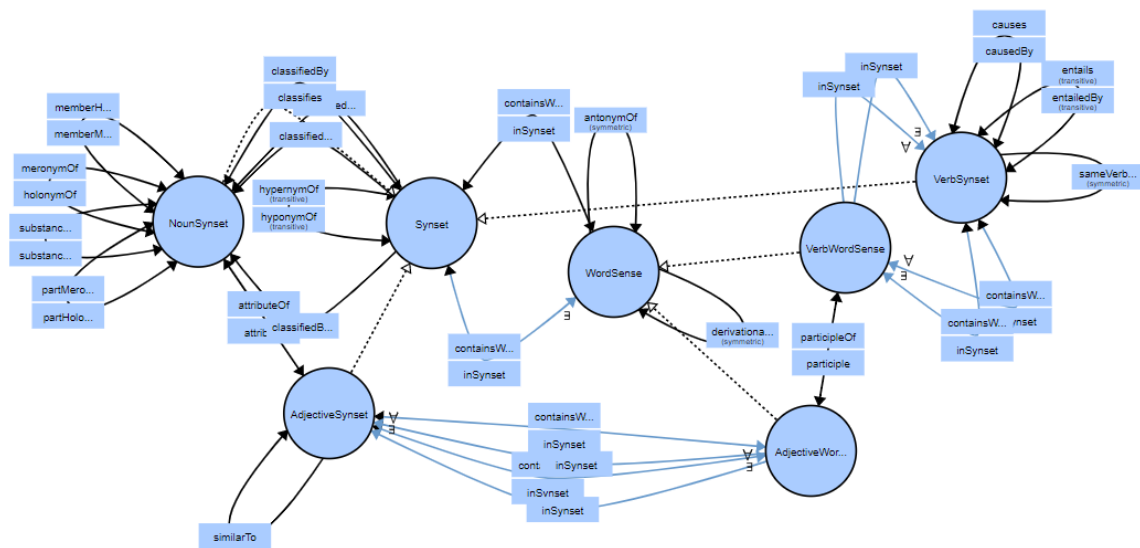


Figure 8 – WordNet full schema visualization.

#### 2.3.2.2 FrameNet

The FrameNet (BAKER COLLIN F. E FILLMORE, 1998) aims to build an English lexicon that is understandable to humans and machines using the semantic frames theory – which states that a semantic frame is a data structure that represents familiar objects and common situations (LOWE, 1997) – and supported by an annotated corpus of lexical items. Terms are organized into situations (known as *frames*) because the meaning of words is supposed to be better understood in the context of the situations and their participants (known as *frame elements*) and the properties involved. The *lexical unit* is a pairing of a word with a meaning. Figure 9 shows the *Ingestion* frame graph with the relations among the frames and frame elements, while Figure 10 shows the definition of this frame, with a sentence using the lexical unit “*devour.v*” in the form “*DEVoured*”, and the core frame elements *Ingestor* and *Ingestible*. Table 3 shows the current FrameNet project status, last updated October 2019.

Table 3 – Current FrameNet Status

<b>Total Frames</b>	1224
Lexical Frames	1087 (89%)
Non-Lexical Frames	137 (11%)
Lexical Units	13675
Frame Elements in Lexical Frames	10542
Frame Elements/Lexical Frame	9.7
Frame Relations	1878
Frame Element Relations	10749

Source: FrameNet Current Project Status, October 2019.

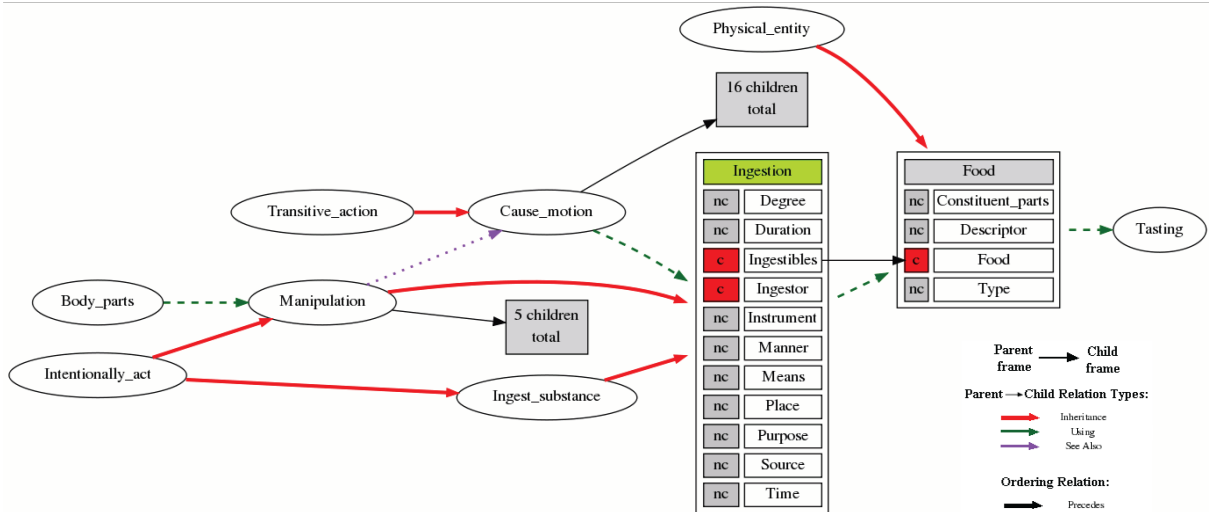


Figure 9 – Ingestion frame relations of FrameNet.

## Ingestion

### Definition:

An **Ingestor** consumes food or drink (**Ingestibles**), which entails putting the **Ingestibles** in the mouth for delivery to the digestive system. This may include the use of an **Instrument**. Sentences that describe the provision of food to others are NOT included in this frame.

**The wolves DEVoured the carcass completely.**

### FEs:

#### Core:

**Ingestibles** [Ingible]

The **Ingestibles** are the entities that are being consumed by the **Ingestor**.

**Ingestor** [Ing]

The **Ingestor** is the person eating or drinking.

Semantic Type: Sentient

Figure 10 – Ingestion frame of FrameNet with core frame elements.

### 2.3.3 Ontology

The term “*Ontology*” comes from philosophy where it is used to describe the nature of beings (CORAZZON, 2019). In computer science, the term was introduced by Artificial Intelligence and began to be treated as part of knowledge-based systems (GRUBER, 1993). There are different definitions of ontology in literature (GRUBER, 1995; STUDER; BENJAMINS; FENSEL, 1998; BORST, 1999), however, the most accepted definition was described by Gruber (1995), who states that *an ontology is an explicit specification of a conceptualization*. In this context, “*conceptualization*” is related to a model of some domain of knowledge by means of definitions of concepts, properties, and the relationships between them; and “*explicit specification*” means that the model must be specified in

unambiguous language, accessible to machines, and humans.

According to Roussey et al. (2011) and considering the scope of the objects described by the ontology, ontologies can be classified in six groups (cf. Figure 11):

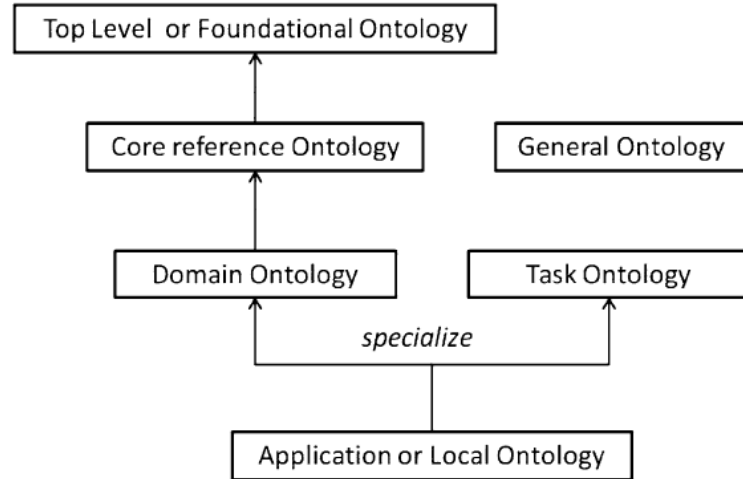


Figure 11 – Ontology classification based on domain scope.

- *Foundational ontology* – can be viewed as meta ontology that describe the top level concepts or primitives used to define others ontologies;
- *Core reference ontology* – contains the fundamental concept of a domain being a standard used by different group of users;
- *Task ontology* – contains the knowledge to accomplish a specific task;
- *Domain ontology* – only applicable to a domain with a specific view point of a group of users, and also describes the knowledge to which the task of the *task ontology* is applied;
- *Local ontology* – has a narrower scope than the *domain ontology*, which may mean that there is no consensus or knowledge sharing, representing a domain-specific model from the point of view of a single user or developer;
- *General ontology* – are not dedicated to a specific domain or fields, containing general knowledge of a huge area. Its concepts can be as general as those of *core reference ontologies*.

Ontologies are used in the formalization of domains (GRUBER, 1995), defining a set of primitives that represent a particular area of knowledge (MIKA PETER E AKKERMANS, 2005). Thus, the ontology provides a common vocabulary to represent the domain knowledge, which can be shared and reused by people who develop applications in that particular domain. Several advantages have been presented in the literature for the adoption of ontologies (AZEVEDO et al., 2008), including:

- Reuse, adaptation, and expansion of ontologies – since the development of knowledge databases is the most expensive, complex and time consuming task in systems development;
- Availability of well-defined ontologies – allowing the reuse and adaptation with specific-domain concepts;
- Computational semantics – since ontologies provide formal definitions that prevent misinterpretations through formally defined restrictions;
- Translation among languages and formalisms – which can be made using ontology editors such as the *Protégé* (STANFORD CENTER FOR BIOMEDICAL INFORMATICS RESEARCH, 2019);
- Online servers for ontologies – which allows the storage of a huge data quantity and serves as a tool to keep the data updated.

Figure 12 shows a graphical representation of an ontology containing its main building blocks (BECHHOFFER et al., 2004), as follows:

- *Instances* – denotes the elements or individuals of a given ontology;
- *Classes* – denotes an abstraction mechanism for grouping instances with similar characteristics. Classes are usually organized in taxonomies, that is, an hierarchy with generalization/specialization relationships. For example, the classes *Professor* and *Person* are related through a taxonomic relationship “*Is a*”, stating that professor is a person;
- *Properties* – represents either an association (relationship) between individuals or an association between individuals and values (data type values). These properties can be of two types: (i) *Object properties* – link individuals to individuals; and (ii) *Data properties* – link individuals to data values that can be represented by a string or number, for instance.

Ontologies can be specified in a wide variety of languages and notations. They are generally represented using Resource Description Framework (RDF) and Web Ontology Language (OWL) (MCGUINNESS; HARMELEN et al., 2004), which are part of the technologies stack defined by the World Wide Web Consortium (W3C) for Semantic Web. These two representation languages are presented next.

RDF has been developed to annotate web resources that are uniquely identified by a Unique Resource Identifier (URI). The basic statement in RDF is a triple of the form (*subject*, *property*, *property value*) that express a binary relation between *subject* (i.e., a resource) and *property value*. The *property* is also denoted by a URI, and the *property value* might be another resource or a datatype value. When the property value is a resource,

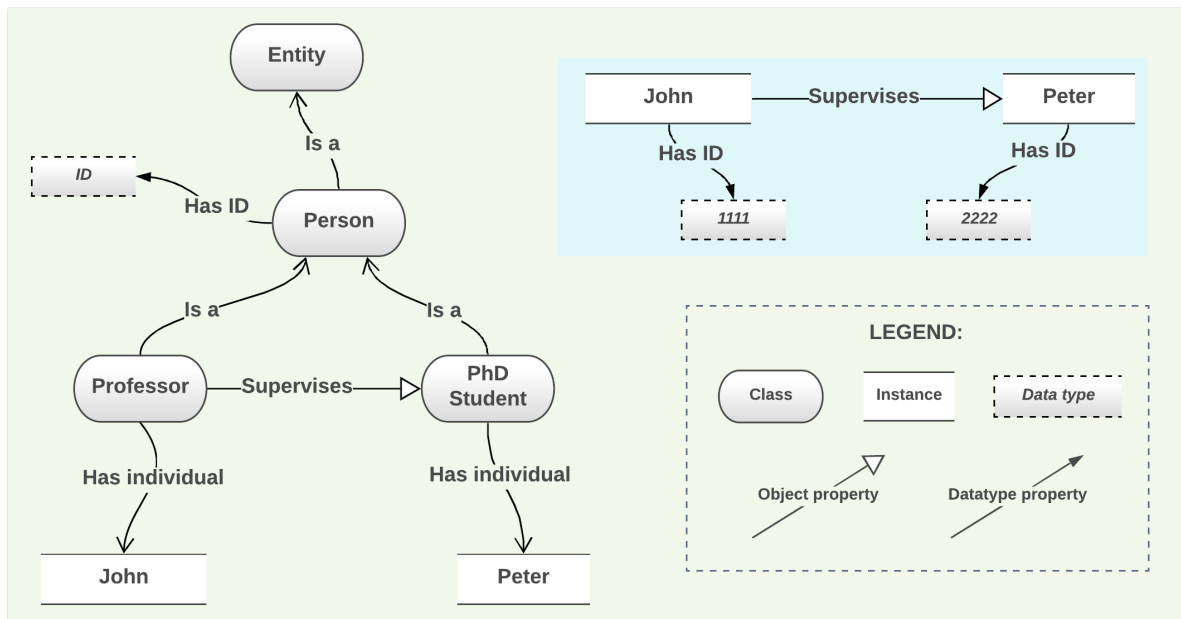


Figure 12 – Example of the main constructors of an ontology.

the property is called *object property*, otherwise, it is called *datatype property*. The RDF Schema (RDFS) provides a data-modelling vocabulary for RDF data, and is an extension of the basic RDF vocabulary (W3C, 2014). Each resource may be associated with one or several concepts (e.g., *rdfs:Class*) via the *rdf:type* property, which is an instance of the *rdf:Property*. RDFS also allows the definition of restrictions on properties and concepts, for example, the definition of domain and range properties (e.g., the *rdfs:domain* of *rdf:type* is *rdfs:Resource*, while the *rdfs:range* of *rdf:type* is *rdfs:Class*).

The second language for representing ontologies is the OWL, which extends RDFS formulating more expressive schemas and hierarchies of subclass with additional logical constraints (W3C, 2012). OWL is based on the Description Logic (DL) (BAADER; HORROCKS; SATTler, 2008) that defines knowledge representation formalisms and reasoning techniques with *axioms* to specify constraints on the ontology, verify its consistency and infer new knowledge. Figure 13 shows the implementation of the Figure 12 example on Manchester OWL Syntax. In this example, we show the definition of: (i) one object property (*Supervises*); (ii) one data property (*Has\_ID*); (iii) three classes (*Person*, *PhD Student*, and *Professor*); and (iv) two individuals (*John* and *Peter*).

The construction of ontologies can be performed in one of two approaches: it can start from scratch (CRISTANI; CUEL, 2005) or it can be built on an existing ontology (GÓMEZ-PÉREZ; ROJAS-AMAYA, 1999). In both approaches, development methodologies and techniques for evaluating the resulting ontology must be used. Concerning development methodologies, different strategies were proposed, for example: 1) *METHONTOL-OGY* (FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ; JURISTO, 1997) that suggests the ontology construction based on knowledge levels; 2) *On-to-knowledge methodology* (SURE; STAAB;

Figure 13 – Implementation of Figure 12 example on Manchester OWL Syntax.

```

Prefix: onto: <http://www.semanticweb.org/nmf/ontologies/2020/3/onto#>
Prefix: owl: <http://www.w3.org/2002/07/owl#>
Prefix: rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Prefix: rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Prefix: xml: <http://www.w3.org/XML/1998/namespace>
Prefix: xsd: <http://www.w3.org/2001/XMLSchema#>

Ontology: <http://www.semanticweb.org/nmf/ontologies/2020/3/onto>

Datatype: xsd:integer

ObjectProperty: onto:Supervises
  Domain:
    onto:Professor
  Range:
    onto:PhD_Student

DataProperty: onto:Has_ID
  Domain:
    onto:Person
  Range:
    xsd:integer

Class: onto:Person

Class: onto:PhD_Student
  SubClassOf:
    onto:Person

Class: onto:Professor
  SubClassOf:
    onto:Person

Individual: onto:John
  Types:
    onto:Professor
  Facts:
    onto:Supervises onto:Peter,
    onto:Has_ID 1111

Individual: onto:Peter
  Types:
    onto:PhD_Student
  Facts:
    onto:Has_ID 2222

```



STUDER, 2004) that proposes to base the development of ontologies on use cases; and 3) *NeOn methodology* (SUÁREZ-FIGUEROA; GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ, 2012) that proposes different developing scenarios by re-using or not ontological resources.

Regarding ontology evaluation, Brank, Grobelnik and Mladenic (2005) and, Raad and Cruz (2015) made two surveys on ontology evaluations approaches to find an efficient one by presenting the existing techniques, and discussing their advantages and drawbacks. The available techniques fall into four categories: 1) Gold standard approach – attempts to compare the new ontology with a previously created reference ontology known as the gold standard; 2) Corpus-based approach – attempts to compare the learned ontology with the content of a text corpus that covers significantly a given domain; 3) Task-based approach – considers that a given ontology is intended for a specific task and is evaluated only according to its performance in that task, regardless of all structural characteristics; and 4) Criteria-based approach – measures how far an ontology or taxonomy adheres to certain desirable criteria. In this context, *OntoQA* (TARTIR; ARPINAR, 2007) is an criteria-based approach for ontology evaluation that uses a set of metrics measuring different aspects of the ontology schema and knowledge base to provide an overview of the general characteristics of ontology.

## 2.4 CHAPTER FINAL CONSIDERATIONS

This chapter presented a brief theoretical foundation that approaches the main concepts necessary to understand the remainder of this thesis. This chapter consists of 3 subjects: 1) The Augmentative and Alternative Communication (AAC), where we presented an overview of the AAC area, including the requirements of a robust AAC system; 2) The Natural Language Processing, where we presented few definitions necessary to understand the method of this thesis; and 3) The knowledge representations method, where we detailed the three methods used in this thesis (i.e., corpus, linguistic knowledge database, and ontology). Next chapter presents the related works of this thesis.

### 3 RELATED WORKS

Since this work has different contributions, there are several categories of related work. Thus, in this chapter we will present in detail the most related works of our work, while in Chapter 4 we present works related to specific contributions of our work, that is, vocabulary selection (cf. Section 4.1) and vocabulary organization (cf. Section 4.2).

Section 3.1 shows the method we use to select the related works; Section 3.2 lists of criteria to be analyzed in each work to perform a fair comparison among them; Section 3.3 presents in details each related work; Section 3.4 shows the comparative analyses among related works; and Section 3.5 presents the final considerations of this chapter.

#### 3.1 WORKS SELECTION

To perform a systematic selection of related works, we applied a forward snowballing approach (WOHLIN, 2014), which is based on identifying new papers based on the papers that cite the paper being examined. We consider the result of the two SLR to select the related works of our specific contributions (cf. Sections 4.1 and 4.2) as the start set of papers to use in the forward snowballing. Table 4 shows the start set of papers. Notice that one paper may cited several papers of the start set of papers.

The citations to these papers were analyzed in October 2019 using Google Scholar<sup>1</sup>. Likewise the SLR protocol (KITCHENHAM; CHARTERS, 2007), the inclusion or exclusion of a paper is done by a sequence of screenings, starting from the information provided in Google Scholar, the abstract, and the full text. In this approach, we do as many iterations as necessary until there are no more studies to be examined. In addition to these documents, we also consult with experts working in this field to request more related work.

Table 5 shows the selected related works, which will be presented in details in Section 3.3. Notice that there is more than one paper from the same research group, so we will present them together.

---

<sup>1</sup> <https://scholar.google.com/>

Table 4 – Start set of papers.

Source	Paper	Reference	Citations
SLR Core Words	Vocabulary requirements for writing activities for the academically mainstreamed student with disabilities	McGinnis and Beukelman (1989)	35
SLR Core Words	Frequency of word usage by non disabled peers in integrated preschool classrooms	Beukelman, Jones and Rowan (1989)	114
SLR Core Words	An initial vocabulary for non speaking preschool children based on developmental and environmental language sources	Fried-Oken and More (1992)	97
SLR Core Words	Vocabulary-use patterns in preschool children: Effects of context and time sampling	Marvin, Beukelman and Bilyeu (1994)	95
SLR Core Words	Core Vocabulary Determination for Toddlers	Banajee, Dicarlo and Stricklin (2003)	111
SLR Core Words	Vocabulary selection for Australian children who use augmentative and alternative communication	Trembath, Balandin and Togher (2007)	55
SLR Core Words	Words needed for sharing a story: implications for vocabulary selection in augmentative and alternative communication	Crestani, Clendon and Hemsley (2010)	21
SLR Core Words	Vocabulary Use Across Genres: Implications for Students with Complex Communication Needs	Clendon, Sturm and Cali (2013)	17
SLR Core Words	Monolingual and bilingual children with and without primary language impairment: core vocabulary comparison	Robillard et al. (2014)	20
SLR Core Words	The Oral Core Vocabulary of Typically Developing English-Speaking School-Aged Children: Implications for AAC Practice	Boenisch and Soto (2015)	29
SLR Core Words	Core Vocabulary in Written Personal Narratives of School-Age Children	Wood, Appleget and Hart (2016)	2
SLR Core Words	Core vocabulary of young children with Down syndrome	Deckers et al. (2017)	10
SLR Core Words	Core vocabulary in the narratives of bilingual children with and without language impairment	Shivabasappa, Peña and Bedore (2017)	0
SLR Categories	Enhancing vocabulary selection for preschoolers who require augmentative and alternative communication (AAC)	Fallon, Light and Paige (2001)	70
SLR Categories	Basic Concepts in Early Childhood Educational Standards: A 50-State Review	Bracken and Crawford (2010)	57
SLR Categories	<i>Estudo de vocábulos para avaliação de crianças com deficiência sem linguagem oral</i>	Paura and Deliberato (2014)	9

Table 5 – Selected papers.

Source	Paper	Reference
1 <sup>st</sup> iteration	A semantic grammar for beginning communicators	Martínez-Santiago et al. (2015)
1 <sup>st</sup> iteration	Pictogram Tablet: A Speech Generating Device Focused on Language Learning	Martínez-Santiago, Montejo-Ráez and García-Cumbreras (2018)
Expert suggestion	Building Semantic Networks to Improve Word Finding in Assistive Communication Tools	Nikolova and Cook (2010)
Expert suggestion	Conceptualizing a Daily Activities Ontology for an Augmentative and Alternative Communication System	Mancilla et al. (2013)
Expert suggestion	User-centric Recommendation Model for AAC based on Multi-criteria Planning	Hernández et al. (2014)
2 <sup>nd</sup> iteration	Towards ontology personalization to enrich social conversations on AAC systems	Mancilla, Sastoque and Iregui (2015)
2 <sup>nd</sup> iteration	Pictogrammar: an AAC device based on a semantic grammar	Martínez-Santiago et al. (2016)
3 <sup>rd</sup> iteration	Computational model based on language development theories for languages learning and training: Vocabulary module	Moreno et al. (2016)
3 <sup>rd</sup> iteration	SMART-ASD, model and ontology definition: a technology recommendation system for people with autism and/or intellectual disabilities	Sevilla et al. (2018)
3 <sup>rd</sup> iteration	An ontology-based recommendation system for people with autism and technology apps: Ontology application for helping persons with autism	Sevilla, Zapater and Herrera (2018)

### 3.2 WORK EVALUATION CRITERIA

To perform a fair comparison among related works, we define a set of criteria to be analyzed in each work. The criteria are:

- **Controlled vocabulary** – analyses if the work proposes and/or uses any controlled vocabulary for AAC, which indicates a previous study or justification for the chosen vocabulary;
- **Vocabulary categorization** – analyzes if the work proposes and/or uses any categorization for the vocabulary, which indicates a concern about concepts (i.e., word or pictogram) organization and retrieval by the AAC user;
- **Vocabulary organization** – analyzes if the work uses some method (e.g., taxonomy, ontology) to organize the vocabulary in an AAC system;
- **Semantic enrichment** – analyzes if the work uses some semantic enrichment (e.g., semantic databases) that allow the system to infer relations among words.

Our analysis considers whether the article meets each criterion and whether there is sufficient evidence to justify the author’s choices. The criteria may be: fully satisfied (●),

partially satisfied (●) or dissatisfied (○).

### 3.3 WORKS PRESENTATION

In this section, we present the selected papers (cf. Table 5) grouped by research group in three. To better understand related works, we also cite other works on the same subject and research group.

#### 3.3.1 Visual Vocabulary for Aphasia (ViVA)

The work of Nikolova and Cook (2010) aims to solve the problem of symbols retrieval in AAC systems. According to them, some word collections are organized in hierarchies, which often leads to deep and non-intuitive research; whereas others are simply a list of arbitrary categories that cause excessive scrolling and a sense of confusion (NIKOLOVA et al., 2009). The Inefficiency in organizing and navigating vocabulary undermines usability principles and makes difficult the adoption of AAC systems. In face of that, Nikolova et al. (2009) proposes a more effective approach for people with aphasia to look for symbols in AAC systems. This approach is based in both, WordNet connections and the evocation technique, which is based on theories that explain how the human brain organizes concepts to provide better access and retrieval of information. This approach was developed to be both, adaptable – able to be customized by the user – and adaptive – able to dynamically change to better suit the user’s past actions and future needs.

Figure 14 shows the components of ViVA, as well as an usage example. The vocabulary organization is modeled based on user input and symbol associations. In this case, the symbol network, which is centered on the “*doctor*” symbol, encompasses related symbols (e.g., “*medication*”, “*pain*”, and “*hospital*”), as well as in previously composed sentences (i.e., “*Call my doctor*”, and “*I cannot find my medication*”).

Nikolova and Cook (2010) uses as a controlled vocabulary the vocabulary provided by *Lingraphica*<sup>2</sup>, a commercial AAC system designed specifically for people with aphasia. This vocabulary has been enriched by the connections provided by WordNet and the evocation technique (i.e., how much a concept resembles another concept) (BOYD-GRABER et al., 2006), which aims to add connections between different grammar classes. This way, ViVA is able to suggest symbols according to the connections provided by the vocabulary network, WordNet, the evocation technique, as well as the suggestions based on the system use. ViVA was evaluated by experiments involving users with and without aphasia, and by a case study with a single patient. Two vocabularies (the original *Lingraphica* and ViVA) were implemented in an AAC system, in which the number of clicks and the construction time of each message would be evaluated. As a result, the use of ViVA decreased the response time and the path (i.e., number of clicks) among associated words.

<sup>2</sup> <https://www.aphasia.com/>

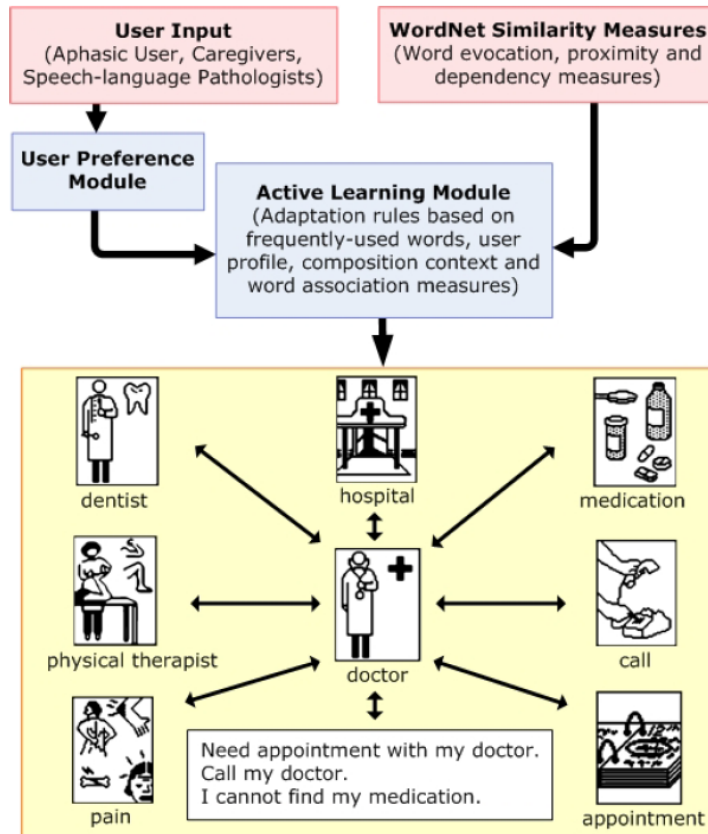


Figure 14 – Components of ViVA.

Next, we present our evaluation of ViVA concerning the listed evaluation criteria (cf. Section 3.2), and summarize this evaluation on Table 6:

- *Controlled vocabulary* – ViVA uses a controlled vocabulary originated from two resources: the *Lingraphica* vocabulary, and the “core” WordNet. This vocabulary consists in all the verbs of *Lingraphica*, and all nouns and adjectives in both, *Lingraphica*’s vocabulary and the core 5,000 synsets of WordNet. The authors do not mention the final vocabulary size, nor if there have been previous studies on this vocabulary or any justification for your choice. This criterion was partially satisfied;
- *Vocabulary categorization* – ViVA organizes the initial collection of words in the *Lingraphica*’s hierarchy and the authors do not mention any previous study that justifies this choice. This criterion was partially satisfied;
- *Vocabulary organization* – the authors of ViVA uses the expression “semantic network”, which uses the *Lingraphica* hierarchy, the semantic relations of WordNet, and the evocation technique to organize the vocabulary. However, the authors do not explain how this semantic network is implemented, as an ontology or a database, for example. This criterion was fully satisfied;

- *Semantic enrichment* – ViVA uses the semantic enrichment of WordNet and the Evocation data. This criterion was satisfied.

Table 6 – Criteria met by ViVA.

Controlled Vocabulary	Vocabulary Categorization	Vocabulary Organization	Semantic Enrichment
●	●	●	●

### 3.3.2 User-centric Recommendation Model

The work of Hernández et al. (2014) proposes a recommendation model that shortens sentence construction time through questions and answers in order to improve the experience of people with aphasia in building consistent and semantically correct sentences. Figure 15 shows an application of this model.

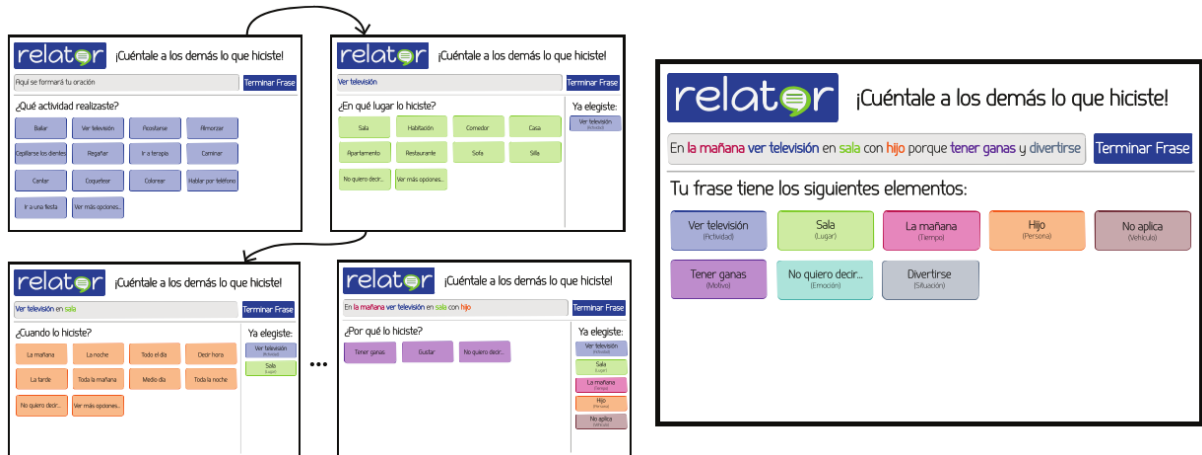


Figure 15 – Application of the User-centric Recommendation Model.

The user-centric recommendation model links domain-specific knowledge to recommendation techniques, allowing the composition of coherent sentences, regardless of word order and message syntactic structure. This model is based on five modules (cf. Figure 16):

- Knowledge Representation – in this case, an ontology was used to represent the concepts;
- User Archetype Extraction – represents user behavior patterns, goals and needs;
- Knowledge-based Recommender – uses information and rules from a specific domain to list different items that will be recommended logically and coherently according to real scenarios;

- Conversational Recommender – uses a user dialogue flow to guide the question-and-answer item selection process;
- Memory-based Recommender – uses data extracted from users over time to suggest items that best fit user preferences.

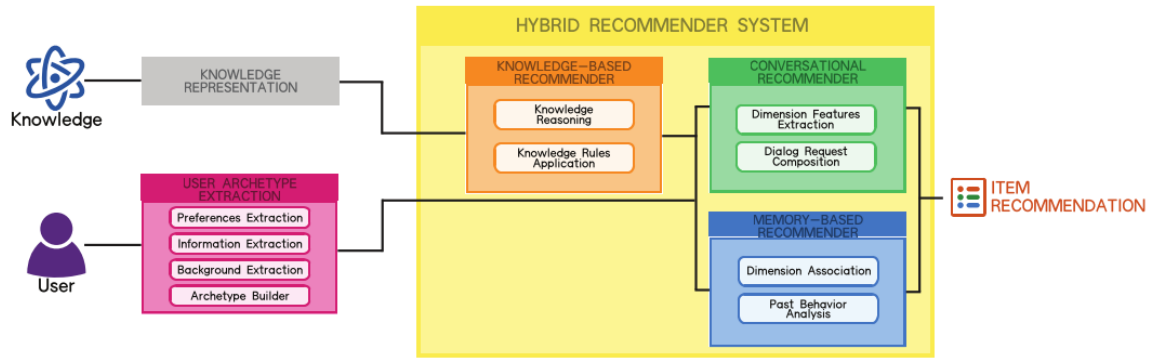


Figure 16 – User-centric Recommendation Model modules.

Mancilla et al. (2013) present the ontology (i.e., the knowledge representation model) that covers the domain of daily routines for people with aphasia to be used in an AAC system. For the ontology knowledge acquisition, the authors used a set of sources were used (e.g., documents describing a normal day for different people, non-structured interviews with speech therapists, and brainstorming). The result is the ontology vocabulary. However, the author highlight that this vocabulary is limited to the main concepts and it does not go into details, for example, the activity “to eat” is in the vocabulary and, if we want to be more specific, than ontology of types of food should be included. Figure 17 show an excerpt of the ontology, detailing the 8 upper concepts (i.e., classes) and some of their relations; whereas Figure 18 show an excerpt of the data dictionary built to classify and describe ontology terms. From the former figure, we can see evidences of a hierarchy of concepts – Class (“Activity”) → Subclass (“Recovery Activity”) → Instance (“Go to speech therapy”); however, the authors do not present more evidences of this hierarchy.

Mancilla, Sastoque and Iregui (2015) focus on the ontology personalization to enrich social conversations on AAC systems. According to the authors, the previous ontology allows users to build the sentence “I told my friend that I want to go to the mall.”. However it is not possible to specify the friend’s name or the name of the mall. Figure 19 shows the method used for ontology personalization, which has four stages:

- Knowledge retrieval – collects the ontology classes that are customizable, that is, classes that have data properties (attributes);
- Knowledge personalization – obtains users information to give values to classes data properties;



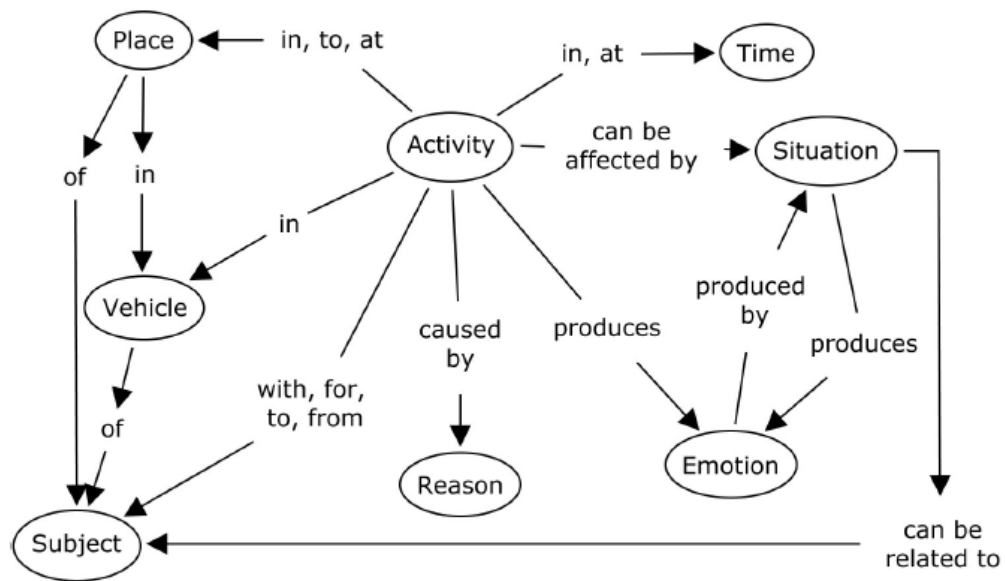


Figure 17 – Excerpt of the daily activities ontology.

Term	Synonyms	Description	Type
Activity	-	Actions made by a person through the day.	Class
Recovery Activity	-	Activities related to recovery and medical activities	Subclass
Go to speech therapy	Go to the speech therapist	Speech recovery activity	Instance

Figure 18 – Excerpt of the data dictionary for ontology terms classification.

- Formalization – links the personalized individuals to the previous ontology. In this stage there are two ontology representations, the *master ontology* (i.e., the previous ontology) and the *user-specific ontology* (i.e., an extension of the master ontology);
- Integration – involves the persistent storage of the personalized ontology that now contains users information.

Next, we present our evaluation of the user-centric recommendation model concerning the listed evaluation criteria (cf. Section 3.2), and summarize this evaluation on Table 7:

- *Controlled vocabulary* – This work uses as controlled vocabulary the result of the analysis of a set of sources. The authors says that this vocabulary is limited to the main concepts, but do not show which concepts are these. This criterion was partially satisfied;

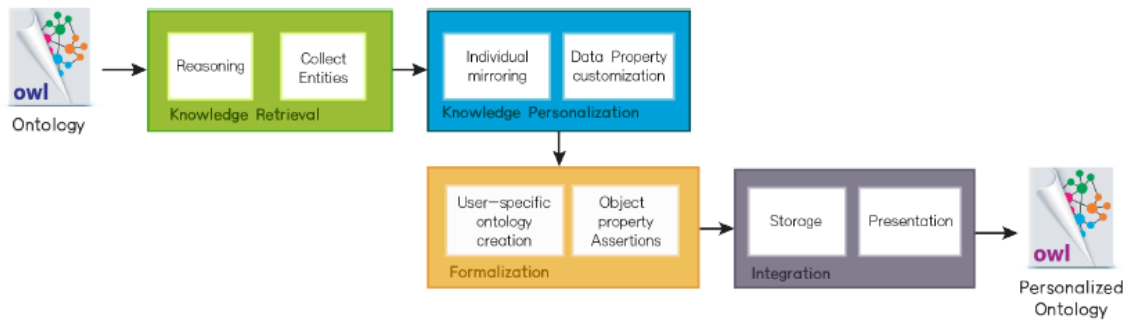


Figure 19 – Method to add personalized information to an existing ontology.

- *Vocabulary categorization* – the authors present a few evidences that may have a hierarchy to organize the concepts. However, it is not possible to state this. This criterion was partially satisfied;
- *Vocabulary organization* – this work uses an ontology to organize the controlled vocabulary. This criterion was fully satisfied;
- *Semantic enrichment* – the authors state that the ontology is semantically enriched with relations among classes. The authors do not mention the use of any semantic database for this purpose, neither present the full set of relations used. This criterion was fully satisfied.

Table 7 – Criteria met by User-centric Recommendation Model.

Controlled Vocabulary	Vocabulary Categorization	Vocabulary Organization	Semantic Enrichment
●	●	●	●

### 3.3.3 Simple Upper Ontology (SUPO)

The work of Martínez-Santiago et al. (2015) hypothesizes that a framework based on an upper ontology which includes a formal representation of a controlled natural language is an adequate framework for developing tools for beginning communicators with speech difficulties. Figure 20 shows this hypothesis. For the authors, controlled natural language is any part of the language that is specified with a set of formal rules (e.g., the symbolic communication used on AAC systems); Simple Upper Ontology (SUPO) is a semantic model to represent the controlled natural language; and the authoring tool is the AAC systems. The authors emphasize that the upper ontology is language-independent; describes the controlled natural language with no, or only a few, syntactic details; and is feasible because the size of the vocabulary of the controlled language is relatively small and it is used to make straightforward assertions.

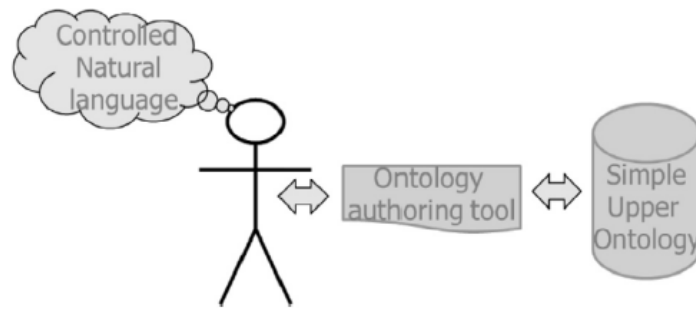


Figure 20 – SUpO hypothesis.

The goal of SUpO is to develop a semantic grammar whose coverage is sufficient to model the language of a beginner communicator (i.e., preschool age). For this, the authors deal with both, a controlled natural language and a semantic model. For the controlled natural language, the authors selected the vocabulary commonly used by preschoolers and children with complex communication needs. The purpose of this vocabulary is to express spontaneous and simple language, without concern for morphology or syntactic complexity. The selection resulted in 621 basic words that include concrete concepts and their properties (e.g., shape, color, size, temperature, sensations and emotions), events, prepositions, interjections and pronouns. This vocabulary can be expanded by applying morphosyntactic rules for gender and number, where applicable. The semantic model of SUpO is an adaptation of FrameNet while the syntactic model is implemented with the *Grammatical Framework* (RANTA, 2011). Thus, SUpO was not built from scratch, but from the integration of various ontologies. From FrameNet, SUpO obtains the concepts of *frame*, *lexical unit*, *frame element*, and *pattern*; and the taxonomy, which is based on the *inherits-from* relations between frames, to classify the selected vocabulary.

SUpO was evaluated by analyzing the sentences it is capable to produce. As a result, the authors ensure a multilingual semantic grammar with the potential to improve communication for beginning communicators. Despite the presented conclusions, this work does not detail how the evaluation was made, nor does it use SUpO in a real AAC scenario. In addition, there is no evidence to prove the multilingual aspect of grammar or if there is any ambiguity treatment.

In another work (MARTÍNEZ-SANTIAGO et al., 2016), the authors present two other concepts: the *Pictogrammar* and the *PictOntology*. The *Pictogrammar* is the authoring tool (i.e., a pictogram-based AAC system that contains the *PictOntology*, which can be defined as a specialization of the SUpO ontology (in a 1:1 relation) for media resources, establishing additional properties and a pictogram taxonomy (cf. Figure 21). *PictOntology* organizes pictograms into 35 categories that evoke various ideas about the same concept and share the same syntactic class (cf. Figure 22). In the newest work, Martínez-Santiago, Montejo-Ráez and García-Cumbreras (2018) focuses on the authoring tool called

*Pictogram Tablet* – which is designed for developing new communication skills rather than improving the act of communication –, and in the evaluation of this tool during supervised sessions with users.

Name	Data Type	Description
ma:identifier	identifier:URI, type:String	The file which contains the pictogram
ma:title	title:String, type:String	The English name of the pictogram. This name is usually equivalent to its expression in English.
ma:language	String	Usually, pictograms are language-independent but sometimes are localized for a given language or culture, e.g. calendars.
ma:creator	String	The creator of the resource
ma:contributor	identifier:URI—String, role:String	The ID of the person who added the pictogram to the ontology
ma:collection	URI—String	Name of the collection it belongs to
ma:relation	identifier:URI, relation:String	“is-a” relation between a pictogram and its category
pt:expressions	List of lang:String, expression:String	The textual translation of the pictogram in the group of supported languages
pt:level	{“transparent” “learned” “abstract”}	<ul style="list-style-type: none"> <li>• “transparent” symbols are very obvious depictions of the concepts that they illustrate.</li> <li>• “learned”: the meaning needs to be learned. The consistent nature of learned symbols means that the concepts they represent become obvious when they are shown together.</li> <li>• “abstracts”: symbols that have no obvious meaning when viewed on their own, and typically represent determiners or adpositions.</li> </ul>
pt:learned_group	String	If a pictogram is learned, then it is consistent with other pictograms relative to the same concept. This label is shared for the learned pictograms relative to the same concept. For example, <i>in</i> , <i>on</i> , <i>under</i> and <i>behind</i> are all about “relative position”
pt:SUpO_concepts	{ identifier:URI, type:List of Strings }	The ID of SUpO concepts depicted by the pictogram

Figure 21 – Examples of properties of PictOntology.

Category	Description	Examples
Flavours	Usual adjectives regarding food	tasty, spicy, salty, sweet
Moving	verbs related to movement	come, climb, dance, drive, drop, fly, follow, kick, quit, run
Professions	Names of well-known jobs	teacher, bus driver, doctor, nurse

Figure 22 – Examples of categories in PictOntology.

Next, we present our evaluation of SUpO concerning the listed evaluation criteria (cf. Section 3.2), and summarize this evaluation on Table 8:

- *Controlled vocabulary* – SUpO uses as a controlled vocabulary a set of 621 basic words commonly used by preschoolers and children with complex communication

needs. Despite using different resources for select the vocabulary, the authors do not mention the criteria used in this selection nor present the selected set. This criterion was partially satisfied;

- *Vocabulary categorization* – the *PictOntology* organizes pictograms into 35 categories. However, the articles do not present these 35 categories and only show 3 categories. This criterion was partially satisfied;
- *Vocabulary organization* – the authors use an upper ontology (i.e., SUpO) that uses the taxonomy, among others characteristics of FrameNet, the FrameNet taxonomy to organize the concepts. This criterion was fully satisfied;
- *Semantic enrichment* – the use of FrameNet is also a proof of SUpO semantic enrichment. This criterion was fully satisfied.

Table 8 – Criteria met by SUpO.

Controlled Vocabulary	Vocabulary Categorization	Vocabulary Organization	Semantic Enrichment
●	●	●	●

### 3.4 COMPARATIVE ANALYSES

Table 9 summarizes the criteria met by the three groups of related works. None of the works fully satisfied all the criteria, likewise, none of them dissatisfied any criteria, as follows:

- *Controlled vocabulary* – all three works use sources to select the AAC vocabulary, however, none of them provide evidence to justify the choice of vocabulary, nor any proof that these vocabularies are suitable for their users. The three works partially satisfied this criterion;
- *Vocabulary categorization* – Nikolova and Cook (2010) use the *Lingraphica*’s hierarchy, Hernández et al. (2014) show a few evidences of a hierarchy to organize the vocabulary, and Martínez-Santiago et al. (2015) propose a set of 35 categories to categorize pictograms. None of them justify the categories choice neither the show the full set of categories or the hierarchy used. The three works partially satisfied this criterion;
- *Vocabulary organization* – Hernández et al. (2014) and Martínez-Santiago et al. (2015) use an ontology to organize the vocabulary; whereas Nikolova and Cook (2010) use a “semantic network” for this purpose, however, the authors do not mention how this was implemented. The three works fully satisfied this criterion;

- *Semantic enrichment* – all of the works satisfied this criteria. Nikolova and Cook (2010) use WordNet and the evocation technique, and Martínez-Santiago et al. (2015) use FrameNet. Hernández et al. (2014) state that they use a set of relations among ontology classes, however, they do not mention the use of any semantic database for this purpose. The three works partially satisfied this criterion.

Table 9 – Criteria met by each Related Work.

Related Work	Controlled Vocabulary	Vocabulary Categorization	Vocabulary Organization	Semantic Enrichment
Visual Vocabulary for Aphasia (ViVA)	●	●	●	●
User-centric Recommendation Model	●	●	●	●
Simple Upper Ontology (SUPO)	●	●	●	●

From Table 9 we can see that the results present the same pattern, they do not bring evidence to fully satisfy the first two criteria (i.e., controlled vocabulary and vocabulary categorization). This means that these works do not give the necessary importance to the basis of AAC communication, that is, vocabulary selection and organization.

### 3.5 CHAPTER FINAL CONSIDERATIONS

This chapter presented the three works that are related to this thesis. These works were selected during a forward snowballing over the results of two SLR (cf. Sections 25 and 26). For each of them, we made a detailed presentation focusing on four evaluation criteria: 1) Controlled vocabulary; 2) Vocabulary categorization; 3) Vocabulary organization; and 4) Semantic enrichment. Next chapter presents the materials and methods of this thesis.

## 4 MATERIALS AND METHODS

Figure 23 shows a Business Process Model and Notation (BPMN) diagram with the main flow of the methods employed in this work, whereas Figures 24, 31, 35, 36, and 39 present, throughout this chapter, the five subprocess of Figure 23, showing all the materials used.

Figure 23 – Overview of materials and methods.

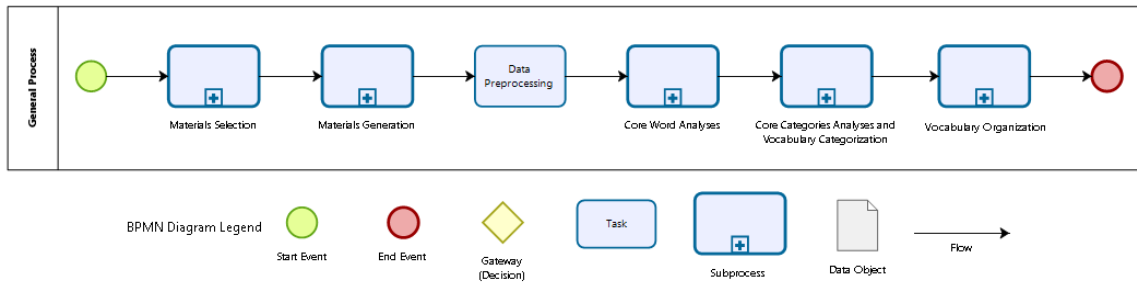
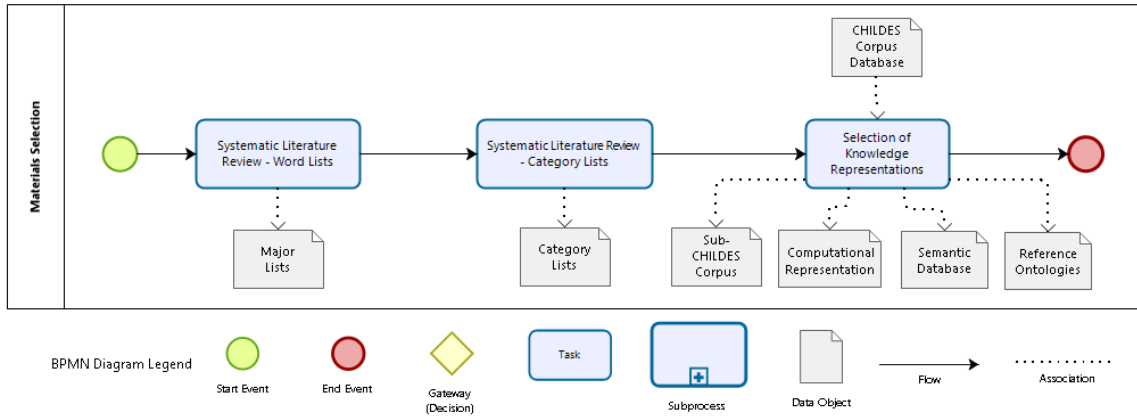


Figure 24 shows the *Materials Selection* subprocess of Figure 23 that will be detailed next, in Sections 4.1, 4.2, and 4.3.

Figure 24 – Detailing of Material Selection Task.



### 4.1 SELECTION OF WORD LISTS

To perform a systematic selection of core word lists, we conducted a Systematic Literature Review (SLR) based on the Kitchenham and Charters (2007) protocol and guided by the Covidence Tool<sup>1</sup>. The objective of our SLR is to select and analyze the available core word lists for children who uses AAC devices. In order to do this, we applied the search string “*Core Vocabulary*” AND (*Child OR Children OR Toddler*) AND “*Alternative Communication*” in four electronic databases (i.e., PubMed<sup>2</sup>, Taylor & Francis Online<sup>3</sup>,

<sup>1</sup> <https://www.covidence.org>

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/pubmed/>

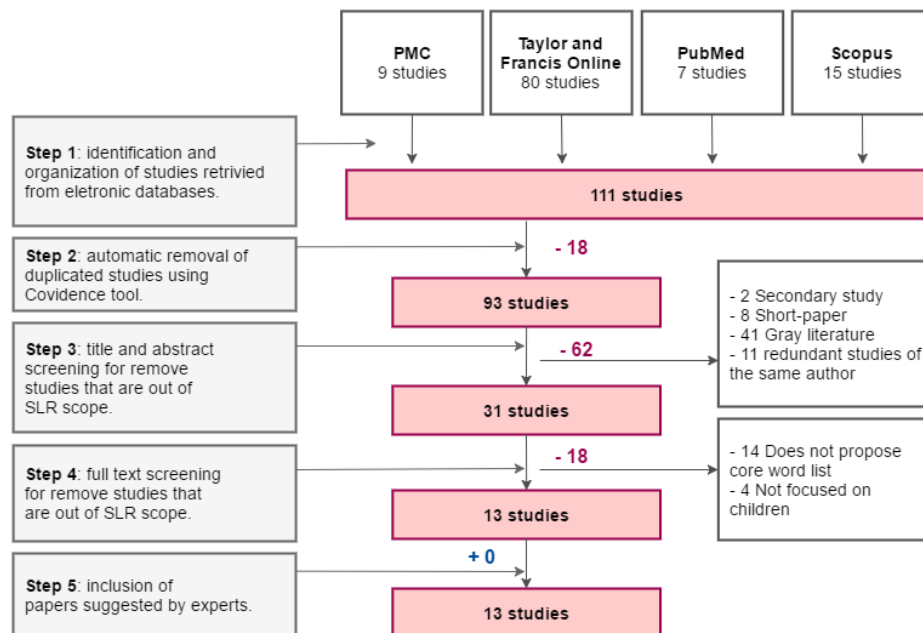
<sup>3</sup> <http://tandfonline.com/>

Scopus<sup>4</sup>, and PMC<sup>5</sup>) in December 2018 to select manuscripts which propose English-based core word for children. To filter the results, we used two Inclusion Criteria (IC) and four Exclusion Criteria (EC):

- IC-1) Studies that propose a new core word list;
- IC-2) Studies focused on children (i.e., people under 14 years);
- EC-1) Secondary study (e.g., a SLR);
- EC-2) Short-paper (e.g., abstracts or extended abstracts);
- EC-3) Gray literature (e.g., books and thesis);
- EC-4) Redundant studies of the same author.

Figure 25 displays the flowchart of our SLR. At the end of the process we selected 13 studies that proposes core word lists based on data of English-speaking children from preschool to 6<sup>th</sup> grade ages collected from different sources (i.e., conversation and written samples, and suggestions of parents and clinicians).

Figure 25 – Studies selection flowchart of Word Lists.



<sup>4</sup> <https://www.scopus.com/>

<sup>5</sup> <http://www.ncbi.nlm.nih.gov/pmc/>



#### 4.1.1 Major Lists

Composed by the 13 lists collected from the SLR and identified by the letters A-M:

- **Major List A** (MCGINNIS; BEUKELMAN, 1989) – composed by the 329 most frequent and distinct words appearing on written samples of different communication levels (letter writing, language and arts, and science) collected from 374 students. The aim of this study was to assist the AAC team as they developed an augmented writing system for a 12-year-old student with severe dysarthria due to cerebral palsy;
- **Major List B** (BEUKELMAN; JONES; ROWAN, 1989) – encompasses 250 words accounting for approximately 85% of the spontaneous conversation samples taken from six non-disabled and verbally active preschool children aged between 44 to 57 months. The aim of this study was to identify the vocabulary used by non-impaired peers who have been integrated into preschool classrooms for disabled children;
- **Major List C** (FRIED-OKEN; MORE, 1992) – comprises 206 distinct words produced from the analysis of a set of 90 sources which includes: word lists generated by parents and clinicians of 15 nonspeaking children diagnosed with palsy aged between 36 to 71 months; and language samples of 30 peers which match typical developing children for gender and age to the children with cerebral palsy.
- **Major List D** (MARVIN; BEUKELMAN; BILYEU, 1994) – composed by 329 words considering both the most frequent structure and content words of spontaneous conversation samples from 60 non-disabled children aged between 48 to 62 months taken in two contexts (home and preschool). The analyses discussed in this study suggest that children have similar vocabulary-use patterns at home and at preschool and that one third of vocabulary were used across both, home and preschool contexts;
- **Major List E** (BANAJEE; DICARLO; STRICKLIN, 2003) – formed by the 25 most common and frequent words from spontaneous conversation samples of 50 toddlers aged between 24 to 36 months. The samples were collected during two different activities (play within interest centers and snack time) aiming to identify if the common words used by toddlers change across different activities.
- **Major List F** (TREMBATH; BALANDIN; TOGHER, 2007) – containing 262 distinct words which account for 79.8% of the spontaneous conversation samples of six children engaged in preschool activities and aged between 36 to 60 months. The aim of this study was to identify the words most frequently and commonly used by typically developing preschool-aged children, in order to assist the vocabulary selection for their classmates who use AAC;
- **Major List G** (CRESTANI; CLENDON; HEMSLEY, 2010) – composed of 173 words which recalls 80% of the total words used by 18 typical development children aged

between 60 to 68 months in three modalities of oral story sharing activities: story-retelling, personal narrative and script narrative (based on a sequence of pictures). The objective of this study is to guide vocabulary selection for children with complex communication needs. We highlight that Crestani paper only presents the top 50 words used;

- **Major List H** (CLENDON; STURM; CALI, 2013) – comprises 140 words which represent 70% of written samples of self-selected topics produced by 65 non-disabled students from kindergarten and 59 non-disabled students from 1<sup>st</sup> grade. The authors suggest that these 140 words are very important to beginning writers and should serve as a starting point to vocabulary selection for educational teams;
- **Major List I** (ROBILLARD et al., 2014) – composed of 48 distinct words from spontaneous conversation samples of 57 monolingual and bilingual (English-French, French-English) children aged between 53 to 77 months. These children were segmented into three groups: French (6 children), bilingual children (22 French predominant children and 19 English predominant children), and French-speaking children with Primary Language Impairment (PLI) (10 children). The analyses of this study showed that there were no significant differences among the core words from the three groups;
- **Major List J** (BOENISCH; SOTO, 2015) – formed by the 395 most frequent words used by 30 developing school-aged English-speaking children (20 from elementary school and 10 from middle school) aged between 84 to 168 months. These children include English native speakers and English as second language speakers. The main objective of this study was to collect core words from a variety of common activities in academic settings;
- **Major List K** (WOOD; APPLEGET; HART, 2016) – comprises 191 distinct words which represent 70% of written samples of 211 children belonging to 1<sup>st</sup> and 4<sup>th</sup> grades. The two grades were selected to allow the analysis of potential differences in word choice by grade. Moreover, the authors were based on previous researchers hypothesis (CLENDON; ERICKSON, 2008) that suggests that the core words of beginning writers may be most relevant to children with complex communication needs;
- **Major List L** (DECKERS et al., 2017) – composed of the 58 most frequently words of spontaneous language samples of 30 Dutch children with Down syndrome aged between 28 to 84 months and developmental age below 48 months. These samples were collected across three settings: free play with parents, snack-time at home or at school, and speech therapy sessions. The final list accounted for 67.2% of the total sample;

- **Major List M** (SHIVABASAPPA; PEÑA; BEDORE, 2017) – formed by the 30 most frequently words in story telling in Spanish and English of children belonging to two groups: the normative group (15 children with PLI and 15 typically developing children) and the bilingual group (65 Spanish-dominant children and 37 English-dominant children). Analyzing core vocabulary in a structured narrative language task rather than spontaneous speech is more informative when the goal is to compare lexical access and use of words among children with and without language impairment.

## 4.2 SELECTION OF CATEGORY LISTS

Likewise in Section 4.1, to perform a systematic selection of core categories lists, we conducted another SLR aiming to select the available core categories lists for children. In order to do this, we applied the search string (*“Word Category” OR “Word Categories”*) *AND (Child OR Children OR Toddler) AND (Semantic OR Pragmatic)* in December 2018 in the same four electronic databases (cf. Section 4.1) to select manuscripts which propose core categories for children. The terms *“Semantic”* and *“Pragmatic”* where used as a restriction to avoid studies that deal with only syntactic categories (e.g., POS and phrasal categories). To filter the search results, we used two IC and five EC:

- IC-1) Studies that propose a new semantic or pragmatic categories;
- IC-2) Studies focused on children (i.e., people under 14 years);
- EC-1) Secondary study (e.g., a SLR);
- EC-2) Short-paper (e.g., abstracts or extended abstracts);
- EC-3) Gray literature (e.g., books and thesis);
- EC-4) Redundant studies of the same author;
- EC-5) Studies that discuss only syntactic categories.

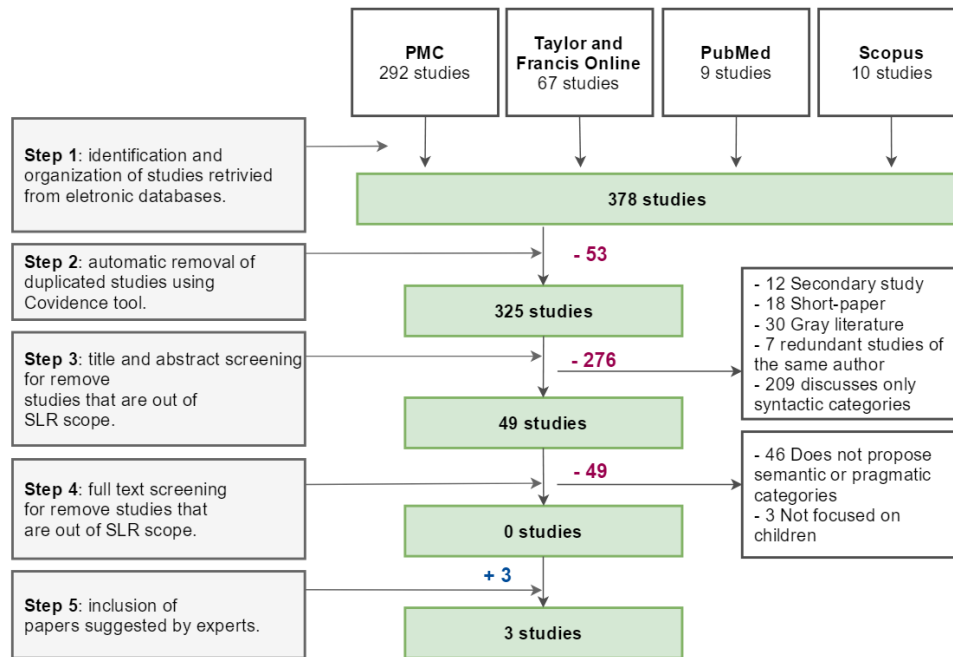
Figure 26 presents the studies selection flowchart of this SLR. At the end of the process, no article was selected from electronic databases, but 3 studies were included by suggestions from AAC experts.

### 4.2.1 Category Lists

Composed by the 3 resulting lists from the SLR and identified by the letters A-C:

- **Category List A** – Semantic-syntactic Categories (FALLON; LIGHT; PAIGE, 2001)  
– composed by 30 categories as part of a categorical framework to elicit individual

Figure 26 – Studies selection flowchart of Category Lists.



child vocabulary. These categories are identified by the codes A1-A30: [A1] Activities, [A2] Adverbs, [A3] Animals, [A4] Articles, [A5] Body Parts, [A6] Clothing, [A7] Conjunctions, [A8] Contractions, [A9] Demonstratives, [A10] Descriptors, [A11] Emotion/Feelings, [A12] Foods and Drinks, [A13] Furniture, [A14] Household Items, [A15] Interjections, [A16] Locations/Places, [A17] Nature, [A18] People/Relationships, [A19] People's Name, [A20] Positions and Equipment, [A21] Prepositions, [A22] Pronouns, [A23] Quantity/Time, [A24] Question Words, [A25] School Materials, [A26] Social/Greetings, [A27] Toys, [A28] Verbs, [A29] Weapons, and [A30] Yes/No Responses;

- **Category List B** – Basic Concepts Categories (BRACKEN; CRAWFORD, 2010) – composed by 10 concept categories that help children in the vocabulary development since it provides words needed to understand classroom conversation and teacher directions. These categories are identified by the codes B1-B10: [B1] Colors, [B2] Direction/position, [B3] Letters, [B4] Numbers/counting, [B5] Quantity, [B6] Self-/social-awareness, [B7] Shapes, [B8] Size/comparison, [B9] Texture/material, and [B10] Time/sequence;
- **Category List C** – Semantic-syntactic Categories (PAURA; DELIBERATO, 2014) – composed by 18 themes of a semantic and syntactic word classification. These themes are identified by the codes C1-C18: [C1] Animals, [C2] Behavior, [C3] Body parts, [C4] Clothing, [C5] Descriptors, [C6] Foods, [C7] Furniture and rooms, [C8] Home routines and activities, [C9] Nature, [C10] People, professions and personal pronouns, [C11] Places, [C12] Pronouns and prepositions, [C13] Questions, answers

and social expressions, [C14] School routines and activities, [C15] Toys, entertainment, sports, instruments and music, [C16] Transport, [C17] Utensils and objects, [C18] Verbs or actions.

### 4.3 SELECTION OF KNOWLEDGE REPRESENTATIONS

In this section, we show the criteria we used to select corpus (cf. Section 4.3.1), computational representation (cf. Section 4.3.2), semantic databases (cf. Section 4.3.3), and reference ontologies (cf. Section 4.3.4).

#### 4.3.1 Sub-CHILDES Corpus

From CHILDES, we selected a subset of five corpora described below for composing the Sub-CHILDES Corpus. These five corpora includes longitudinal and cross-sectional studies that allow the analysis of 1,271 transcriptions with 331,444 sentences and 807,515 words. Table 10 details the Sub-CHILDES Corpus according to each component corpus, which are detailed next:

- **Corpus A** – Manchester corpus (THEAKSTON et al., 2001) – composed of 804 samples of a longitudinal study with 12 children aged between 20 to 36 months and MLU average of 2.97;
- **Corpus B** – Belfast corpus (HENRY, 1995) – composed of 89 samples of a longitudinal study with 8 children aged between 28 to 53 months and MLU average of 5.41;
- **Corpus C** – Cruttenden corpus (CRUTTENDEN, 1978) – composed of 42 samples of a longitudinal study with two children aged 17 and 43 months respectively and MLU average of 4.52;
- **Corpus D** – Tommerdahl corpus (TOMMERDAHL; KILPATRICK, 2014) – composed of 43 samples of a cross-sectional study with 23 children aged between 30 to 42 months and MLU average of 4.95;
- **Corpus E** – Wells corpus (WELLS, 1981) – composed of 297 samples of a longitudinal study with 32 children aged between 18 to 61 months and MLU average of 2.85.

Table 10 – The Sub-CHILDES Corpus details.

Corpus	Name	N_Children	N_Docs	N_Words	N_Sentences	MLU Range	AGE Range
<b>A</b>	Manchester	12	804	564,284	249,317	1.19 - 5.38 (avg = 2.97)	21 - 36 (avg = 28.9)
<b>B</b>	Belfast	8	89	88,485	23,002	1.75 - 8.73 (avg = 5.41)	24 - 54 (avg = 41.6)
<b>C</b>	Cruttenden	2	42	10,308	3,061	1.12 - 9.75 (avg = 4.52)	18 - 44 (avg = 30.2)
<b>D</b>	Tommerdahl	23	40	37,892	11,367	3.64 - 6.98 (avg = 4.95)	29 - 45 (avg = 35.7)
<b>E</b>	Wells	32	296	106,546	44,697	1.07 - 5.39 (avg = 2.85)	18 - 61 (avg = 31.5)
<b>TOTAL</b>		77	1,271	807,515	331,444	1.07 - 9.75 (avg = 3.22)	18 - 61 (avg = 30.6)

### 4.3.2 Selection of the Computational Representation

There are several approaches to computational represent the vocabulary (cf. Section 2.3). In this work, we will use an ontology (cf. Section 2.3.3), because it is used to model a domain of knowledge, it can be shared and reused by people who develop applications that uses this particular domain. Moreover, the backbone of an ontology consists of a taxonomy (i.e., an hierarchy with generalization/specialization of concepts), which was considered by Drager et al. (2003) as the better approach for vocabulary organization for children because:

- The taxonomic approach had better learning curve when compared to the other approaches (cf. Section 2.1.1.2);
- Young children are able to understand taxonomic categories when the objects are familiar and the categories are labeled for them (KRACKOW; GORDON, 1998; LUCARIELLO; KYRATZIS; NELSON, 1992; MARKMAN; COX; MACHIDA, 1981), mainly when they enter school and are exposed to more adult mental models;
- Taxonomies allows the establishment of standards for the classification of information through the use of inheritance mechanisms, and also allows users to learn from these conceptual structures (CAMPOS; GOMES, 2007).

The ontology backbone can also be enriched with properties of semantic databases (cf. Section 4.3.3) that allow the inference and automatic reasoning about ontological elements, as well as knowledge from well-known and consolidated ontologies (cf. Section 4.3.4), that are well accepted by the community.

AAC system developers can take advantage of the use of an ontology to develop systems with more complex functionalities for AAC users. These functionalities may be seen in two perspectives: data input (when the AAC user writes the message using symbols) and output (when the system reads what the user wrote). In the first situation, ontologies may be used as a predictive semantic grammar (GRUZITIS; PAIKENS; BARZDINS, 2012), applying filters to display the symbols according to the linguistic model (syntactic and semantics), as well as the context of the communication. In the second situation, ontologies

may help the tasks of expand the telegraphic message to the full and grammatically corrected sentence (DEMASCO; MCCOY, 1992) or translate the message to another language.

### 4.3.3 Selection of Semantic Databases

We selected WordNet (BAKER; FELLBAUM, 2009) as the semantic database for this work over FrameNet (BAKER; FILLMORE; LOWE, 1998) for the following reasons:

- Coverage is one of the main weaknesses of the current FrameNet lexical database (JOHANSSON; NUGUES, 2007) – FrameNet lists only 13,675 lexical units, compared to 203,145 word–sense pairs in WordNet 2.0 (cf. Section 2.3.2);
- Organization of the concepts – WordNet implements a more complete taxonomy among synsets, while FrameNet organizes the concepts using the theory of semantic frames. On the one hand, WordNet has many *is-a* relationships, allowing comparisons of word senses in the same POS. On the other hand, FrameNet’s semantic frames have only a few *is-a* relationships and prioritize other relationships. We highlight the use of taxonomies allow a better understand of concepts in the same POS, bringing facilities for WSD tasks;
- Reference database – WordNet has become the lexical database of choice for NLP and has been incorporated into other language tools, including VerbNet (SCHULER, 2005). Furthermore, several online dictionaries, including Google’s “define” function, rely significantly on WordNet (BAKER; FELLBAUM, 2009).

Given the above information, WordNet is the first selected semantic database because it allows the implementation of a better base for our ontology.

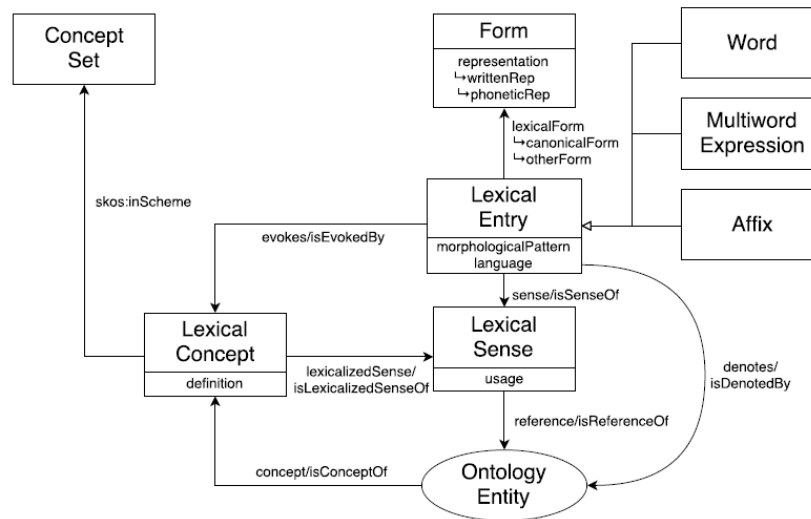
### 4.3.4 Reference Ontologies

Since ontologies are developed to be reused, we selected some well-known and consolidate ontologies, which are well accepted by the community, to integrate into our ontology. The 4 selected reference ontologies are a reference for specific areas important to our propose (i.e., lexical description and representation, word sense disambiguation, and knowledge representation):

- **OntoLex-Lemon Model** – the primary mechanism for the representation of lexical data on the Semantic Web. This model aims to provide a linguistic grounding that includes the representation of morphological and syntactic properties of lexical entries as well as the meaning of these entries with respect to an ontology or vocabulary (MCCRAE et al., 2017; CIMIANO; MCCRAE; BUITELAAR, 2016). Figure 27 shows the core of OntoLex Lemon Model, which is centered in the *lexical entry*, representing a single word, that is connected to *forms* that groups all morphological

expressions of a word, and *lexical senses* that represents all possible concepts in the ontology it can refer to. Moreover the *lexical concept* allows the definition of a word meaning independently of an external ontology. In addition to the core, the OntoLex-Lemon Model provides four more modules for deal with: 1) syntax and semantics; 2) decomposition of words; 3) variation & translation; and 4) metadata.

Figure 27 – The Core OntoLex-Lemon Model.



- **LexInfo** – it imports the OntoLex-Lemon Model and allows the association of linguistic information regarding any level of linguistic description and expressiveness to elements of an ontology (CIMIANO et al., 2011). With LexInfo is possible to maintain the linguistic descriptions associated with ontology elements and independent of the specific ontology application. By doing so, language information can be reused on different systems. Figure 28 shows the detailing of the morphosyntactic properties of LexInfo.

Figure 28 – Morphosyntactic properties details of LexInfo.

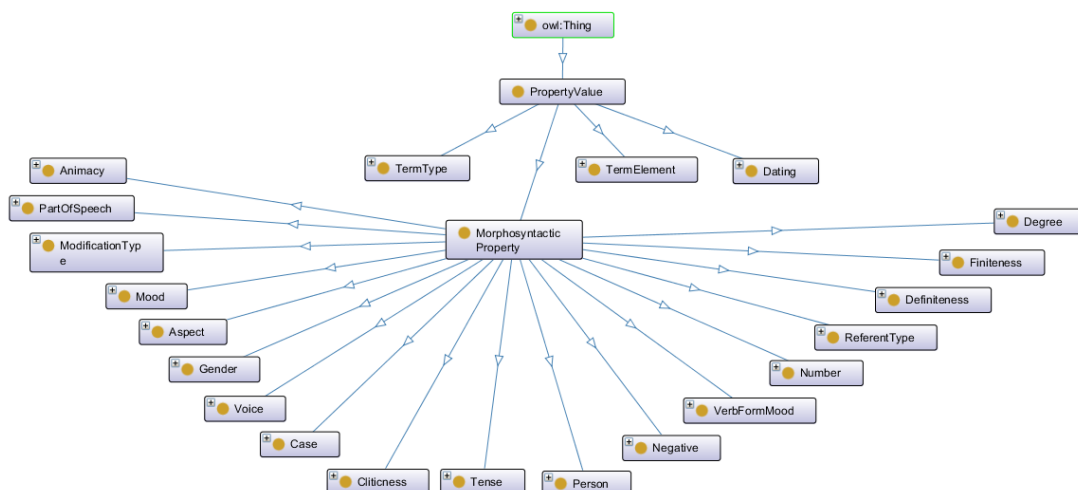
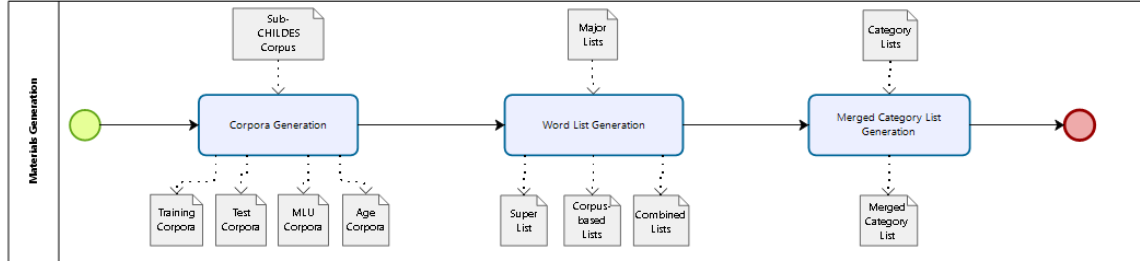






Figure 31 shows the *Materials Generation* subprocess of Figure 23 that will be detailed next, in Sections 4.4, 4.5, and 4.6.

Figure 31 – Detailing of Material Generation Task.



#### 4.4 CORPORA GENERATION

To eliminate the bias of analyzes results and ensure the disjunction of the corpus samples that will generate new word lists (cf. Section 4.5) of those that will compose the test set for recall analyses, we split the Sub-CHILDES Corpus into Training and Test Corpora (4.4.1). In order to do this, we randomly took 60% of each five corpora to compose the *Training Corpora* and 40% to compose the *Test Corpora*. Since there is no baseline for this ratio, we selected the 60:40 ratio to have representative data for each purpose (i.e., training and testing).

Figures 32, 33, and 34 show the distribution of the samples of Test Corpora according to children’s Age and MLU. Figure 32 presents the distribution of MLU Score *vs.* Age in months of Test Corpora samples. Figure 33 presents the distribution of the Age in months of the Test Corpora samples over Age ranges according to the Brown’s Stages. In turn, Figure 34 presents the distribution of the MLU Score of the Test Corpora samples over MLU ranges according to the Brown’s Stages. Notice that, besides the fifth stage of Brown’s model, Figures 33 and 34 also shows two groups: 1) Upper Age – for age greater than 46 months; and 2) Upper MLU – for MLU greater than 4.5. Notice that these ranges are not covered by Brown’s Stages, but these are important since children belonging to these groups also uses AAC systems. The dense vertical distribution in Figure 32 indicates that the age in months is not a proper parameter to analyze child communication development, once there are many children with the same age (x-axis) in different Brown’s Stages (y-axis). Also, in Figure 33, excepting Stage I, the more significant part of the age distribution is below that of the Brown’s Stages. However, in Figure 34, the distribution of samples properly follow the MLU behavior (i.e., min, max and mean MLU) of Brown’s Stages. Therefore, MLU is a better metric than age to analyze the child communication development, and the concomitant use of these two metrics mix children of different development stages in the same statistical analysis.

Figure 32 – MLU vs. Age distribution of Test Corpora samples.

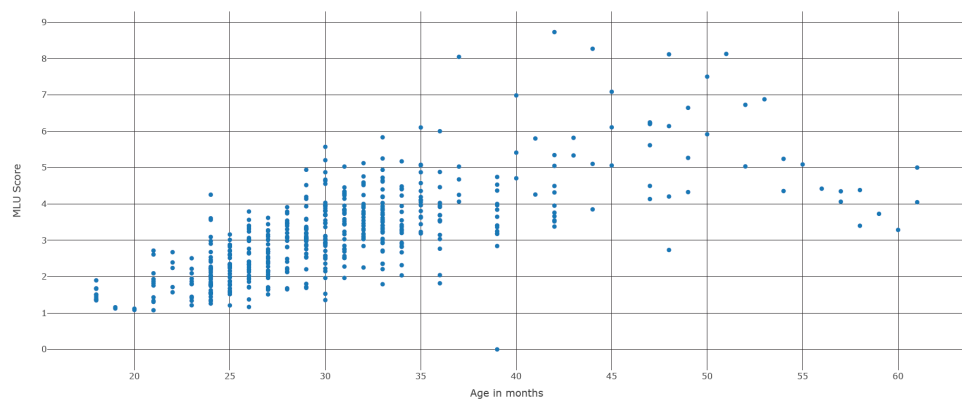


Figure 33 – Test Corpora according to the AGE range of Brown's Stages.

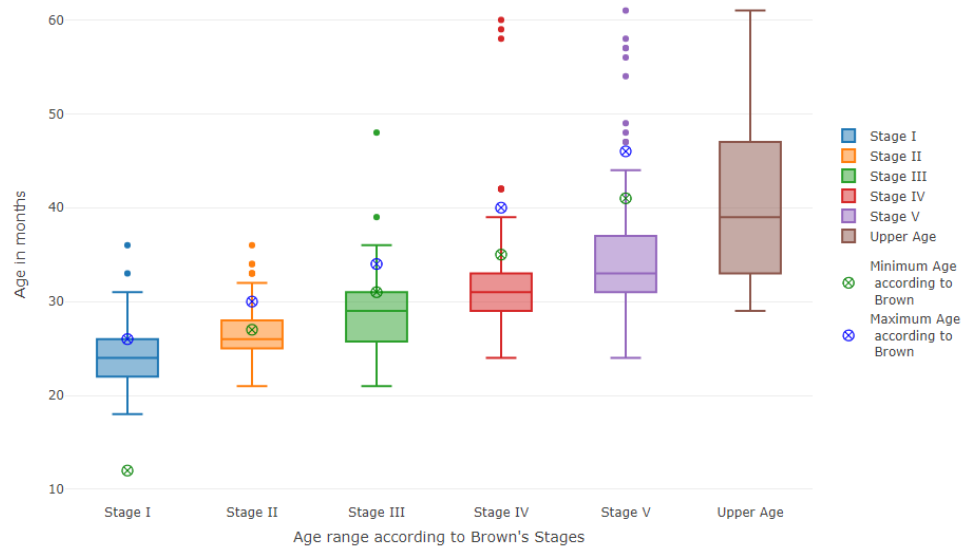
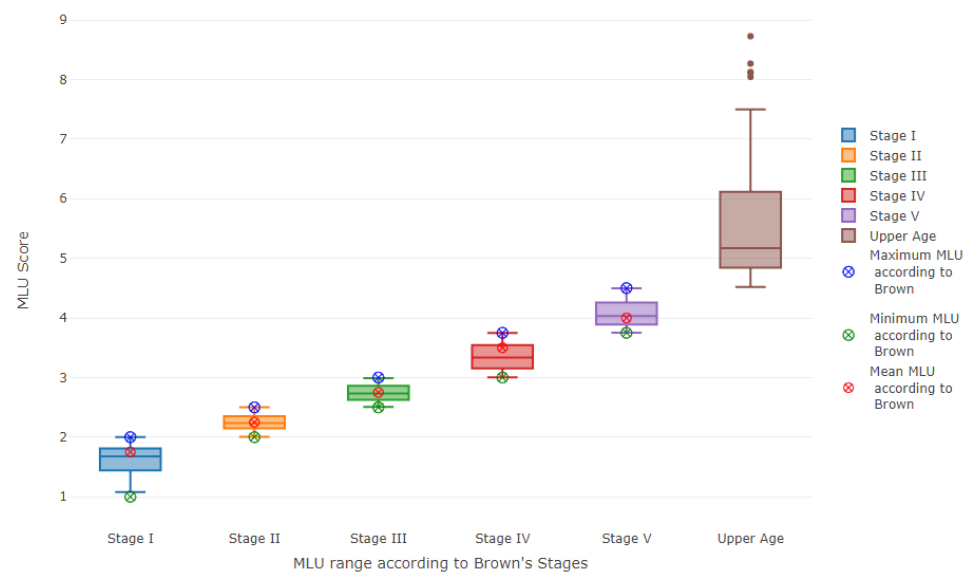


Figure 34 – Test Corpora according to the MLU range of Brown's Stages.



To investigate the Test Corpora samples, regardless the context where they were collected and considering only one metric at a time, we created another two corpora – the *Age Corpora* and the *MLU Corpora* (cf. Section 4.4.2) – by dividing the Test Corpora according to each metric.

#### 4.4.1 Training and Test Corpora

Table 11 summarizes the Training and Test Corpora characteristics.

Table 11 – Characteristics of the Training and the Test corpora.

	Training Corpora (60%)			Test Corpora (40%)		
	N_Docs	N_Words	N_Sentences	N_Docs	N_Words	N_Sentences
<b>Corpus A</b>	483	336,717	149,002	321	227,567	100,315
<b>Corpus B</b>	54	49,222	13,055	35	39,263	9,947
<b>Corpus C</b>	26	7,803	2,216	16	2,505	845
<b>Corpus D</b>	24	22,631	6,781	17	15,261	4,586
<b>Corpus E</b>	178	60,185	25,679	118	46,361	19,018
<b>TOTAL</b>	765	476,558	196,733	506	330,957	134,711

#### 4.4.2 Age and MLU Corpora

Table 12 shows the Age and the MLU Corpora. The Age Corpora divide the Test Corpora according to the age range from Brown’s Stages; whereas the MLU Corpora divide the Test Corpora according to the MLU range from Brown’s Stages.

We highlight that one transcription sample of Tommerdahl corpus was excluded from all corpora (i.e., Test Corpora, Age Corpora, and MLU Corpora) because its MLU was 0.

Table 12 – Summarization of AGE and MLU Corpora extracted from Test Corpora.

	N_Docs	N_Words	N_Sentences		N_Docs	N_Words	N_Sentences
<b>AGE I</b>	143	58,971	33,333	<b>MLU I</b>	86	24,467	17,515
<b>AGE II</b>	136	88,107	38,265	<b>MLU II</b>	67	32,493	18,048
<b>AGE III</b>	120	95,588	36,733	<b>MLU III</b>	70	40,192	19,184
<b>AGE IV</b>	54	34,452	11,832	<b>MLU IV</b>	134	100,002	40,289
<b>AGE V</b>	21	13,939	3,933	<b>MLU V</b>	84	71,809	24,000
<b>Upper AGE</b>	32	39,900	10,615	<b>Upper MLU</b>	65	61,994	15,675
<b>TOTAL</b>	506	330,957	134,711	<b>TOTAL</b>	506	330,957	134,711

## 4.5 WORD LISTS GENERATION

To generate new lists, we began from three assumptions concerning the coverage of children’s utterances:

1. The union of Major Lists may have better coverage than each Major List (cf. *Super List* on Section 4.5.1);
2. The use of NLP metrics over Training Corpora may generate a list with better coverage than each Major List (cf. *Corpus-based Lists* on Section 4.5.2);
3. The merge of the previous assumptions, that is, complement each new NLP list with the union of Major Lists, may generate a list with better coverage than each Major List (cf. *Combined Lists* on Section 4.5.3).

#### 4.5.1 Super List

The *Super List* is composed of 753 words and was created by the union of the 13 *Major Lists* and assigning a commonality score – number of *Major Lists* that the word appears, denoting its relevance – for each word. For example, if a word has commonality score 13, it means that this word is relevant because it appears in all lists. Table 13 shows the distribution of the Super List over commonalities, and Table 14 shows the words in the highest commonalities.

Table 13 – Super List Distribution over Commonalities

Commonalities	13	12	11	10	9	8	7	6	5	4	3	2	1	TOTAL
N_Words	1	6	8	17	16	30	25	36	35	59	75	114	331	<b>753</b>

Table 14 – High-commonality words of Super List

Commonality	N_Words	Words
13	1	<i>“the”</i>
12	6	<i>“I”, “and”, “go”, “in”, “my”, and “on”</i>
11	8	<i>“a”, “for”, “have”, “it”, “out”, “that”, “there”, and “you”</i>
10	17	<i>“all”, “at”, “be”, “do”, “get”, “he”, “his”, “like”, “no”, “of”, “one”, “she”, “then”, “they”, “this”, “to”, and “want”</i>
9	16	<i>“big”, “but”, “her”, “here”, “home”, “know”, “little”, “more”, “not”, “see”, “so”, “some”, “up”, “we”, “what”, and “with”</i>
8	30	<i>“yes”, “too”, “two”, “put”, “dog”, “time”, “said”, “was”, “good”, “back”, “me”, “eat”, “mom”, “off”, “away”, “dad”, “him”, “first”, “can”, “play”, “our”, “look”, “mine”, “when”, “over”, “went”, “house”, “them”, “because”, and “is”</i>
7	25	<i>“help”, “going”, “why”, “are”, “come”, “right”, “down”, “water”, “these”, “your”, “just”, “new”, “something”, “had”, “today”, “if”, “ball”, “did”, “three”, “how”, “open”, “who”, “oh”, “other”, and “where”</i>

#### 4.5.2 Corpus-based Lists

The *Corpus-based Lists* were created by employing four NLP metrics (i.e., Term Frequency, Document Frequency, Term Frequency-Inverse Document Frequency, and Word

Density) over words in the *Training Corpora*. For this, we rank by descending the words in *Training Corpora* using each of the metrics and then applying a threshold cut to position 500, which encompasses a large number of distinct words and cover more than 70% of children communication (cf. Figure 40). Moreover, from 500 words there was no significant variance in the recall of the corpus. These metrics and the lists are detailed next:

- **Corpus-based TF List** – contain the first 500 most relevant words based on the Term Frequency (TF) metric (ROBERTSON, 2004) that measures the frequency of the word in the corpora;
- **Corpus-based DF List** – contain the first 500 most relevant words based on the Document Frequency (DF) (ROBERTSON, 2004) which counts the number of documents in the analyzed corpora which contain that word;
- **Corpus-based TF-IDF List** – contain the first 500 most relevant words based on the  $TF * IDF$  metric (ROBERTSON, 2004) which combines the TF metric multiplied by the Inverse Document Frequency (IDF) metric. This metric is guided by the intuition that a word which occurs many times in a document or in many documents in a corpora is not a good discriminator and needs to be balanced (i.e.,  $TF * (\frac{1}{DF})$ );
- **Corpus-based WD List** – contain the first 500 most relevant words based on Word Density (WD) metric which is calculated by the TF of a word divided by the sum of the TF of all words (i.e.,  $WD_i = \frac{TF_i}{\sum TF}$ ) (CHEN; WU, 2004).

### 4.5.3 Combined Lists

The *Combined Lists* were created by the union of each Corpus-based List (cf. Section 4.5.2) with its relative complement (i.e., the set composed by all the elements of A that are not in B:  $A \setminus B = \{\forall x | x \in A \wedge x \notin B\}$ ) of the Super List (cf. Section 4.5.1). Table 15 summarizes the composition of each *Combined List* by showing the total quantity of words (i.e., N\_Words) in the list, as well as the size (in words) of the relative complement and its distribution over commonalities. Notice that all lists have the same missing words in higher commonalities (i.e., from 13 to 9 – cf. Table 16).

Table 15 – Combined List Characterization considering corpus-based lists of 500 words

List Name	N_Words	N_Relative Complement	Distribution over Commonalities												
			13	12	11	10	9	8	7	6	5	4	3	2	1
Combined TF List	699	199	1	4	5	7	7	11	13	11	7	19	9	19	91
Combined DF List	711	211	1	4	5	7	7	12	13	11	7	19	10	20	100
Combined TF-IDF List	704	204	1	4	5	7	7	11	13	12	7	19	9	20	95
Combined WD List	704	204	1	4	5	7	7	11	13	11	7	19	9	19	96

Table 16 – High-commonality words that do not appear in the corpus-based lists

Commonality	N_words	Words
13	1	<i>“the”</i>
12	4	<i>“my”, “and”, “I”, and “on”</i>
11	5	<i>“it”, “a”, “out”, “for”, and “that”</i>
10	7	<i>“at”, “all”, “of”, “this”, “like”, “his”, and “no”</i>
9	7	<i>“what”, “with”, “see”, “but”, “little”, “her”, and “some”</i>

#### 4.6 MERGED CATEGORY LIST GENERATION

In order to decrease the amount of data to be analyzed, we have grouped similar categories and given them a common name and description. We generate another category list by grouping the Category Lists A, B and C (cf. Section 4.2) in the Category List M (i.e., the *Merged Category List*). Aiming to avoid ambiguity of meanings for this study, Table 17 shows the 34 resulting categories with their names, relations with the previous proposed categories, and descriptions.

Table 17 – Merged Category List.

Category Name	Related to	Description
[M1] Actions / Verbs	A28, C18	Content words that denote an action, occurrence, or state of existence.
[M2] Activities	A1, C8, C14	Any specific behavior.
[M3] Animals	A3, C1	A living organism characterized by voluntary movement.
[M4] Body Parts	A5, C3	Any part of an organism such as an organ or extremity.
[M5] Clothes	A6, C4	A covering designed to be worn on a person’s body.
[M6] Colors	A10, B1	A visual attribute of things that results from the light they emit or transmit or reflect.
[M7] Comparisons	B8	Relations based on similarities and differences.
[M8] Demonstratives	A9	Referring words.
[M9] Descriptors	A2, A10, B9, C5	Words that describe nouns, adjectives, verbs or adverbs.
[M10] Directions / Positions	A20, B2	A course along which a point moves. The particular portion of space occupied by something.
[M11] Emotion / Feelings	A11, B6, C2	Affective and health/physical feelings.
[M12] Entertainment	C15	An activity that is diverting and that holds the attention.
[M13] Expressions / Contractions	A8, A15, A26, A30, C13	A word or phrase that particular people use in particular situations, and contractions of words.

Table 17 continued from previous page

Category Name	Related to	Description
[M14] Foods and Drinks	A12, C6	Any substance that is used as a source of nourishment.
[M15] Furnitures	A13, C7	Furnishings that make a room or other area ready for occupancy or use.
[M16] Household Items	A14	They are the tangible and movable personal property placed in the house.
[M17] Letters and Numbers	B3, B4	Any of the symbols of an alphabet. A concept of quantity involving zero and units.
[M18] Location / Places	A16, C7, C11	A point or extent in space.
[M19] Materials	B9	Things needed for doing or making something.
[M20] Nature	A17, C9	The natural physical world including plants and animals and landscapes etc.
[M21] Objects	C17	A tangible and visible entity that can cast a shadow.
[M22] Particle and Connectors	A4, A7, A21, C12	Words used as affix to nouns, or that joins two groups of words together.
[M23] People / Professions	A18, A19, C10	A human being. Types of people and specific people.
[M24] Positions and Equipments	A20	Personal positions (sit on the lap, horsey ride) and position in equipment (sit in the wheelchair, turn sideways on the bed).
[M25] Pronouns	A22, C10, C12	A function word that is used in place of a noun or noun phrase.
[M26] Quantities	A23	Words referring to amounts.
[M27] Question Words	A24, C13	Interrogative words.
[M28] School Materials	A25	Objects that would be used in a school setting.
[M29] Sequence	B10	Words describing orders.
[M30] Shapes	A10, B7	Any spatial attributes (especially as defined by outline).
[M31] Size	A10, B8	The physical magnitude of something (how big it is).
[M32] Time	A23, B10	A unit for measuring time periods.
[M33] Toys	A27, A29	An artifact designed to be played with.
[M34] Transport	C16	Something that serves as a means of transportation.

## 4.7 DATA PREPROCESSING

Words (mainly nouns and verbs) can be flexed in gender, number, person, and tenses, which can produce lists that are long and laborious to compare. Moreover, for functional communication, AAC systems often use the telegraphic language (i.e., a compressed mes-

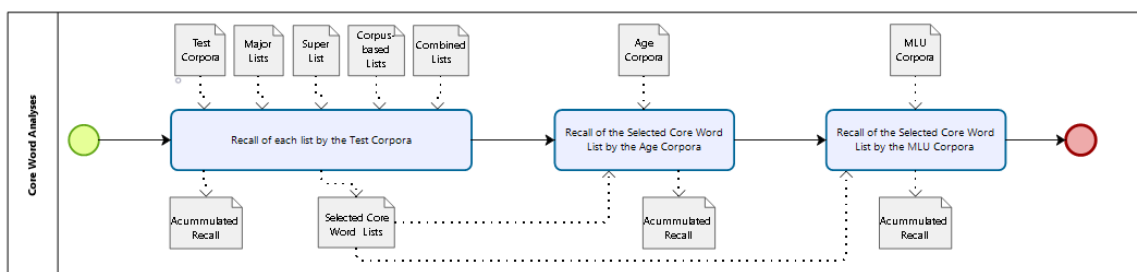


sage made by sentences with no inflections and with some missing words – “*Daddy eat chocolate now*”) rather than the full and grammatically corrected sentence (i.e., “*Daddy is eating chocolate now*”). Therefore, to compare words independent of its inflections, we preprocess our materials of word lists and corpora samples to make a new version of these materials. For this, we use the Python language (PYTHON CORE TEAM, 2019), the NLTK (LOPER; BIRD, 2002), and the WordNet (MILLER, 1995) data to perform processing considering three aspects:

1. Contractions – we expand all the contractions to their full forms (e.g., “*can’t*” is replaced by “*can*” and “*not*”) and consider only the constituents without ambiguity. For example, the contraction “*I’ll*” can be expanded to “*I will*” or “*I shall*” generating an ambiguity. In this case we consider only the word “*I*” and disregard the words “*will*” and “*shall*”;
2. Word form – considering the telegraphic language, we use the lemma form, that is, the canonical form of a set of words. For example, the lemma “*eat*” is used to represent the words “*eat*”, “*eats*”, “*ate*”, and “*eating*”;
3. Part of Speech (POS) – we use the NLTK Perceptron Tagger (LOPER; BIRD, 2002) for tagging all words in corpora regarding its context of use, that is, the sentence in which the word appears. The core word lists have the particularity of do not bring information about POS, neither its context of use. For each of these words, we consider all the possible POS. For example, “*right*” can be a noun, a verb, an adjective or an adverb, as well as, “*fish*” can be a noun or a verb. This POS overlap only occurs with content words, thus we use WordNet to make this POS tagging, by searching all possible POS of each word on WordNet.

Figure 35 shows the *Core Word Analyses* subprocess of Figure 23 that will be detailed next, in Section 4.8.

Figure 35 – Detailing of Core Word Analyses Task.



## 4.8 CORE WORD ANALYSES

This analysis aims to answer the first RQ (“*Is it possible to generate a new core word list with better recall than the existing core word lists?*”), and it is based on discovering the

coverage of each word list over a corpus. We subdivide the core word analyses into three analyses:

1. *Recall of each list by the Test Corpora* – to analyze the recall behavior of each 22 list (i.e., 13 Major Lists, 1 Super List, 4 Corpus-based Lists and 4 Combined Lists) considering all the samples at once, discarding the influence of the metrics age and MLU. The list with better recall in this analysis will be the *Selected Core Word List* for the next two analyses;
2. *Recall of the Selected Core Word List by the Age Corpora* – to analyze the recall behavior of the *Selected Core Word List* (i.e., the list with better recall in the first analysis) considering only the age range according to the Brown’s Stage;
3. *Recall of the Selected Core Word List by the MLU Corpora* – to analyze the recall behavior of the *Selected Core Word List* considering only the MLU range according to the Brown’s Stage.

For each analysis, we sort descending all the core word lists by its relevance according to the used measure (cf. Table 18). Then the *Accumulated Recall* (cf. Equation 4.1) is calculated starting from the most relevant word, where  $Recall_i$  is the percent of the corpus that the  $i$ -th word represents, and  $n$  is the word list size.

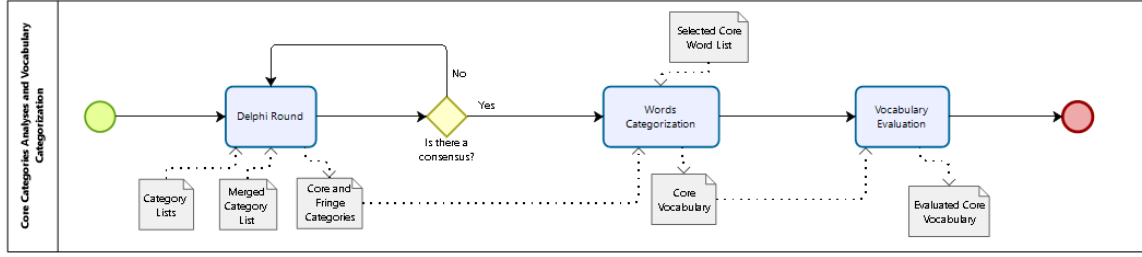
$$Accumulated Recall_n = \sum_{i=1}^n Recall_i \quad (4.1)$$

Table 18 – Core word lists ordination.

List Type	Name	Measure	Most relevant words
<b>Major Lists</b>	All Major Lists	Frequency in Test Corpora	High frequency words
<b>Super List</b>	List S	Frequency in Test Corpora	High frequency words
<b>Corpus-based Lists</b>	TF List	TF in the Training Corpora	High TF words
<b>Corpus-based Lists</b>	DF List	DF in the Training Corpora	High DF words
<b>Corpus-based Lists</b>	TF-IDF List	TF-IDF in the Training Corpora	High TF-IDF words
<b>Corpus-based Lists</b>	WD List	WD in the Training Corpora	High WD words
<b>Combined Lists</b>	All Derived List	Frequency in Test Corpora	High frequency words

Figure 36 shows the *Core Category Analyses and Vocabulary Categorization* subprocess of Figure 23 that will be detailed next, in Section 4.9.

Figure 36 – Detailing of Core Categories Analyses and Vocabulary Organization Task.



## 4.9 CORE CATEGORY ANALYSES AND VOCABULARY CATEGORIZATION

Aims to answer the second RQ (“Are the already proposed categories useful and sufficient to categorize all the core words of the best list of RQ-1?”) and is based in two phases, detailed in the next subsections.

### 4.9.1 Categories Selection

This phase is based on empirical and semantic evidences over core categories and expert evaluation combining two methods:

1. Delphi Method (ZARTHA et al., 2018) – based on successive rounds of anonymous experts consultation about a particular topic to achieve a consensus;
2. Lawshe’s Method (LAWSHE, 1975) – a quantitative approach to content validity that involves experts rating items into one of three categories: “*Essential*”, “*Useful, but not essential*”, or “*Not necessary*”.

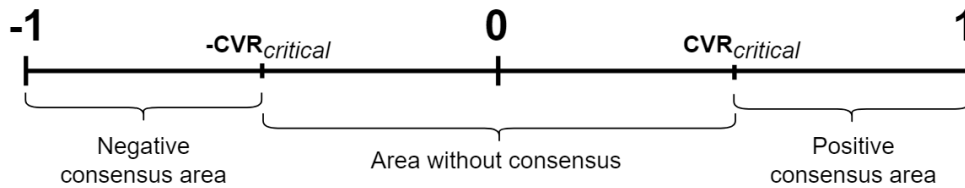
We make a 2-round-Delphi method for consulting experts (i.e., teachers, specialists and post-grad students working in the Assistive Technology, Special Education, or Health field, and AAC users relatives and friends) with a 3-point Likert scale questionnaires, considering the three Lawshe’s categories, to establish the importance of each topic searched. To determine if there is a consensus among experts in each questionnaire response, we use the Content Validity Ratio ( $CVR$ ) proposed by Lawshe (1975) (cf. Equation 4.2), where  $n_e$  is the amount of experts indicating that the item is “*Essential*”, and  $N$  is the total number of experts in the Delphi round. The same author also suggested that a level of 50% agreement gives some assurance of content validity and reported a table of critical  $CVR$  and  $n_e$  values, where the  $CVR_{critical}$  is the lowest level of  $CVR$  for expert consensus and  $n_{e\ critical}$  is the lowest number of experts to have a consensus. The  $CVR_{critical}$  were improved in recent papers (WILSON; PAN; SCHUMSKY, 2012; AYRE; SCALLY, 2014) and we use the newest one.

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}} \quad (4.2)$$

Figure 37 shows the  $CVR$  interpretation ranging from -1 (perfect agreement that the item is *not Essential*) to 1 (perfect agreement that the item is *Essential*). For values above

$CVR_{critical}$ , there is a positive consensus area (i.e., the item is important); for values below  $-CVR_{critical}$  there is a negative consensus area (i.e., the item is not important); and for values between  $-CVR_{critical}$  and  $CVR_{critical}$  there is an area without consensus. We can also use the formula above to calculate the  $CVR$  considering the remain 2 Lawshe's categories: "Useful, but not essential", and "Not necessary".

Figure 37 – CVR interpretation scale.



In the first Delphi round we analyze if the already proposed categories are essential to provide functional communication for children who use AAC. In order to do this, we cluster all the 60 different categories showed in Sections 4.2 and 4.6 (i.e., the Category Lists A, B, and C, and the Merged Category List) into 9 groups according to its meaning (cf. Table 19). Notice that some categories are organized in more than one group (e.g., *Quantity* appears in *Words of Description*, and *Measure and Order*), and some concepts are covered by more than one category (e.g., *People/Relationship*, *People's Name*, and *People/Professions*). This redundancy aims to make the respondent to think in the same category over different contexts and to decide the better nomenclature for each concepts set. We ask the experts to analyze the 9 groups and assign an option in a 3-point Likert scale ("Essential", "Useful, but not essential", or "Not necessary") concerning the category importance for an AAC vocabulary. We also ask the experts to list other categories that were not considered but are important for each group. Categories taken into consensus as: 1) "Essential" or "Useful, but not essential" are added to our Useful Categories; and 2) "Not necessary" are discarded. Categories without consensus and the suggested categories are considered to the second round.

In the second Delphi round we take the categories without consensus (i.e.,  $-CVR_{critical} < CVR_{category} < CVR_{critical}$ ) as well as the different categories suggested in the first round and try to achieve a consensus. For this, we discard the redundancies of the first round, trying to mitigate the respondent interpretation bias, and make more straightforward questions. That is, we ask questions such as "What do you think about that the inclusion of a Transport category which includes words as car, bus, train, and bike in an AAC system for children?" and the respondents answer in the same 3-point Likert scale. Categories taken into consensus as: 1) "Essential" or "Useful, but not essential" are added to our Useful Categories; and 2) "Not necessary" are discarded. In this round, categories without consensus are added to our Useful Categories because these categories were already considered as important for other authors and experts.

Table 19 – Semantic group categories.

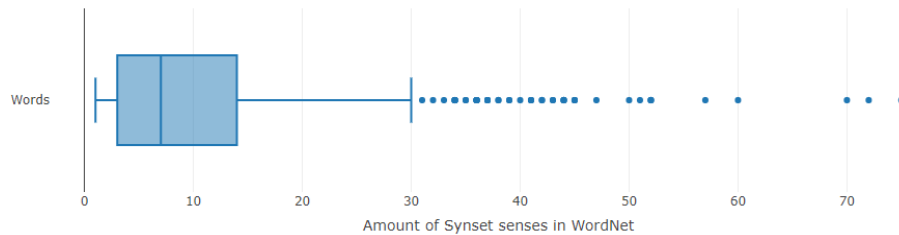
Semantic group	Definition	Categories
Action words	Any word that denote, for example, actions, occurrences, modes of being, and routine activities.	M1, M2, M12
Closed class words	Any word used to provide syntactic connections among nouns, verbs, adjectives and adverbs.	M8, M22, M25, M27
Living things	Any word that denote a living (or once living) thing or a part of it.	M3, M4, M20, M23
Inanimate things	Any word that denote tangible and visible instances.	M5, M14, M15, M16, M19, M21, M28, M33, M34
Words of description	Any word or expression that represents/ describes something in words.	M6, M7, M9, M11, M19, M26, M29, M30, M31, M32
Personal feelings	Any word that denote a physical or psychological sensation that one can experience.	M11
Measure and order	Any word used to determine the quantity and size of something as well as its order in a sequence.	M7, M8, M26, M29, M31, M32
Places and orientation	Any word that denote a point in space or the position of something relative to points of reference.	M10, M18, M20, M24, M32
Communication	Any word or expression used to communicate something to somebody.	M11, M13, M17, M27

#### 4.9.2 Word Categorization and Vocabulary Evaluation

For the word categorization we use WordNet (MILLER, 1995) for word sense disambiguation. With this database we annotate each category with the synset name (i.e., sets of cognitive synonyms used in WordNet) that better denotes the category sense, and each word from the *Selected Core Word List* with:

1. Its position – considering the frequency order in corpus;
2. Its POS and classification of content or structure word – for helping the word classification;
3. A set containing at most seven senses (i.e., synsets) of this word in WordNet – for helping the identification of the different meanings of a word (e.g., *Fish* as an animal, as a food or as an activity). We select seven synsets because it is the median of the distribution of the amount of senses of each word in WordNet (cf. Figure 38);
4. A set of exact categories and possible categories – we analyze the full hierarchy of the word. For example, the word *Foot* belongs to the *animal\_foot.n.01* synset (defined by “the pedal extremity of vertebrates other than human beings”) can be exactly

Figure 38 – Amount of synsets sense per word on WordNet.



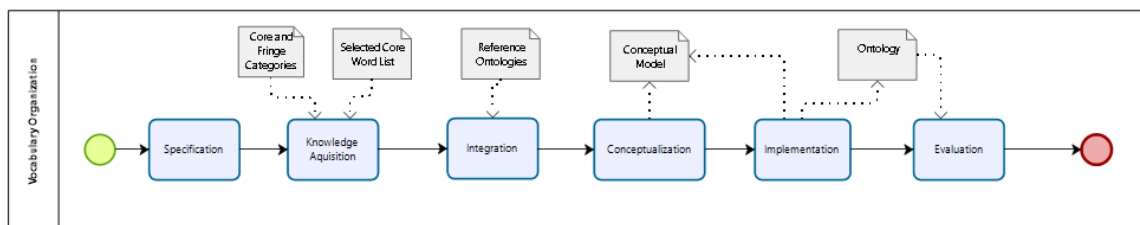
categorized into *Body Parts* because the synset *body\_part.n.01* is in its WordNet hierarchy and could also be categorized into the category *Animals*, since the word *animal* appears in the string of its WordNet hierarchy.

After annotating each category and each word from the *Selected Core Word List*, we manually analyze all the senses of each word and select those that are adequate to children. For example, we discard those senses that are pejorative or sexual (e.g., “cow” denoting a woman in a pejorative way, and “dog” denoting a sexual position) or are out of the childish context (e.g., “will” denoting a testament). These analyses allow us to choose the better category for each word.

For the evaluation and improvement of the vocabulary categorization, we make a 1-section Focus Group (SHULL; SINGER; SJØBERG, 2007) with experts on linguistics, AAC, and ASD. The Focus Group is a research technique that collects data through group interaction on a topic determined by the researcher (MORGAN, 1996) and it was chosen because it allows researchers to quickly get multiple points of view from a group of experts in a specific area. Moreover, with this technique the researcher can intervene during the session, allowing discussion and clarification of doubts.

Figure 39 shows the *Vocabulary Organization* subprocess of Figure 23 that will be detailed next, in Section 4.10.

Figure 39 – Detailing of Vocabulary Organization Task.



#### 4.10 VOCABULARY ORGANIZATION

Aims to answer the third RQ (“How can AAC vocabulary be computationally organized to allow new functionalities in a high-tech AAC system?”) and it is based on the computational representation of the vocabulary that allows new functionalities for a high-tech AAC system.

For the computational representation we chose the ontology (cf. Section 4.3.2) that will be developed using the METHONTOLOGY (FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ; JURISTO, 1997), a framework composed by a list of seven tasks to perform. We highlight that this framework does not indicate an order in which to perform such tasks. Next, we detail the seven tasks identified by METHONTOLOGY:

1. *Specification* – aims to answer questions about why the ontology is being built and what are its intended uses and end-users. This way, in this phase we establish the purpose of the ontology, its scope, domain, level of formality (USCHOLD; GRUNINGER, 1996), and sources of knowledge;
2. *Knowledge Acquisition* – uses different techniques to understand the context and to extract information for knowledge construction;
3. *Integration* – considers the reuse of definitions (concepts and relations) that have been built in other ontologies in order to speeding up the construction of the new ontology, and to take advantage of available reference ontologies;
4. *Conceptualization* – structures the domain knowledge in a conceptual model that describe both, the problem and the solution. Here the knowledge acquired is structured through different techniques such as data dictionary and graphical representations (GÓMEZ-PÉREZ; FERNÁNDEZ; VICENTE, 1996);
5. *Implementation* – code the ontology in a formal language to make it computable;
6. *Evaluation* – makes a technical judgment of the ontology with respect to a frame of reference (e.g., requirements specification document). In the evaluation process there is a difference between verification and validation. *Verification* refers to the process that guarantees the correctness of an ontology, that is, if the ontology is able to answer domain competency questions. *Validation* determines if the definitions of the ontology satisfy the necessities of the domain, through the use of real contexts and end users;
7. *Documentation* – produces documents in each of the other six tasks throughout the ontology development process.

Concerning the evaluation task, we will evaluate the ontology with the metrics of OntoQA (TARTIR; ARPINAR, 2007) by using the author’s implementation<sup>6</sup>. We have chosen OntoQA because its model considers how classes are organized in the schema and on how instances are distributed across the schema, enabling the determination of the quality of an ontology. OntoQA divide the metrics in two categories: 1) Schema metrics, which address the design of the ontology, providing indicators of the richness, width, depth, and

<sup>6</sup> OntoQA application available at <https://code.google.com/archive/p/tartir-ontoqa/downloads>

inheritance of the ontology schema; and 2) Instance metrics, which address the way data is placed within the ontology, providing indicators of the effectiveness of the ontology design and the amount of knowledge represented by the ontology. The instance metrics is divided in Knowledge metrics, which describes the ontology as a whole, and Class metrics, which describes the way each class is used in the ontology. OntoQA does not provide reference values for all metrics, only how to interpret them.

#### 4.11 CHAPTER FINAL CONSIDERATIONS

This chapter presented the material and methods of this thesis. Some materials were obtained through a simple search (i.e., corpus and reference ontologies), others through an SLR (i.e., main list and main categories) and others by processing the previous materials (i.e., the corpus considering training, testing, Age and MLU, the super list, the corpus-based lists, the combined list, and the merged category list). Next, we presented the data preprocessing method to compare data independent of word inflections. Finally, we present the methods of the three analysis of this work: 1) the core word analysis, which is based on the *AccumulatedRecall* analysis; 2) the category analysis and vocabulary categorization, which is based on the Delphi Method, and the Lawashe's Method; and 3) the vocabulary organization, which is based on METHONTOLOGY. The next chapter will present the results of this thesis.



## 5 RESULTS

The results reported here will concern: 1) Core Words Analyses (cf. Section 5.1), which answers the first research question; 2) Categories Analyses and Vocabulary Categorization (cf. Section 5.2), which answers the second research question; 3) Core Vocabulary Organization (cf. Section 5.3), which answers the third research question; 4) Comparative Evaluation (cf. Section 5.4), which compares this work with the related works presented in Chapter 3; and 5) Usage Guidelines (cf. Section 5.5), which shows how AAC community can use these results.

### 5.1 CORE WORD ANALYSES

In this analysis we evaluate the recall behavior of lists over corpora in three analyses: the recall analyses over Test Corpora, the recall analyses over Age Corpora, and the recall analyses over MLU Corpora. The first two analyses will be present in Section 5.1.1, while the third analysis will be present in Section 5.1.2. We highlight that the recall analyses begins from the most relevant lemmas of each list (cf. Table 18) and it is calculated by the *AcumulatedRecall* (cf. Equation 4.1).

#### 5.1.1 Recall Analyses over Test Corpora

Figures 40 and 41 present the recall of all 22 core word lists (i.e., 13 *Major Lists*, 1 *Super List*, 4 *Corpus-based Lists*, and 4 *Combined Lists*) over the *Test Corpora*. Figure 40 shows the recall of 13 Major Lists and 1 Super List while Figure 41 shows the recall of 4 Corpus-based Lists and 4 Composed Lists. Moreover, since recall lines are close, we show a detailed area on the bottom side of each figure. Table 20 summarizes the maximum recall score of each list.

Figure 40 shows that the maximum score of Major Lists ranges from 20.320% (Major List M (SHIVABASAPPA; PEÑA; BEDORE, 2017);  $n = 30$ ) to 68.297% (Major List F (TREMBATH; BALANDIN; TOGHER, 2007);  $n = 227$ ) while the Super List achieves its maximum score of 76% starting on  $n = 402$ , that is, the remain 200 lemmas of Super List do not considerably increase its recall score (0.805%). Notice that a large number of lemmas does not mean greater recall, for example: 1) List M (SHIVABASAPPA; PEÑA; BEDORE, 2017) achieves 20.32% with 30 lemmas while List E (BANAJEE; DICARLO; STRICKLIN, 2003) achieves 29.98% with 24 lemmas; and 2) List L (DECKERS et al., 2017) achieves 33.57% with 56 lemmas while List J (BOENISCH; SOTO, 2015) achieves 38.483% with 44 lemmas. We highlight that, *Major List A* (MCGINNIS; BEUKELMAN, 1989) and *Major List D* (MARVIN; BEUKELMAN; BILYEU, 1994) have the same list size and the same recall behavior, so its curves are overlapped. In short, Major List F (TREMBATH; BALANDIN; TOGHER, 2007)

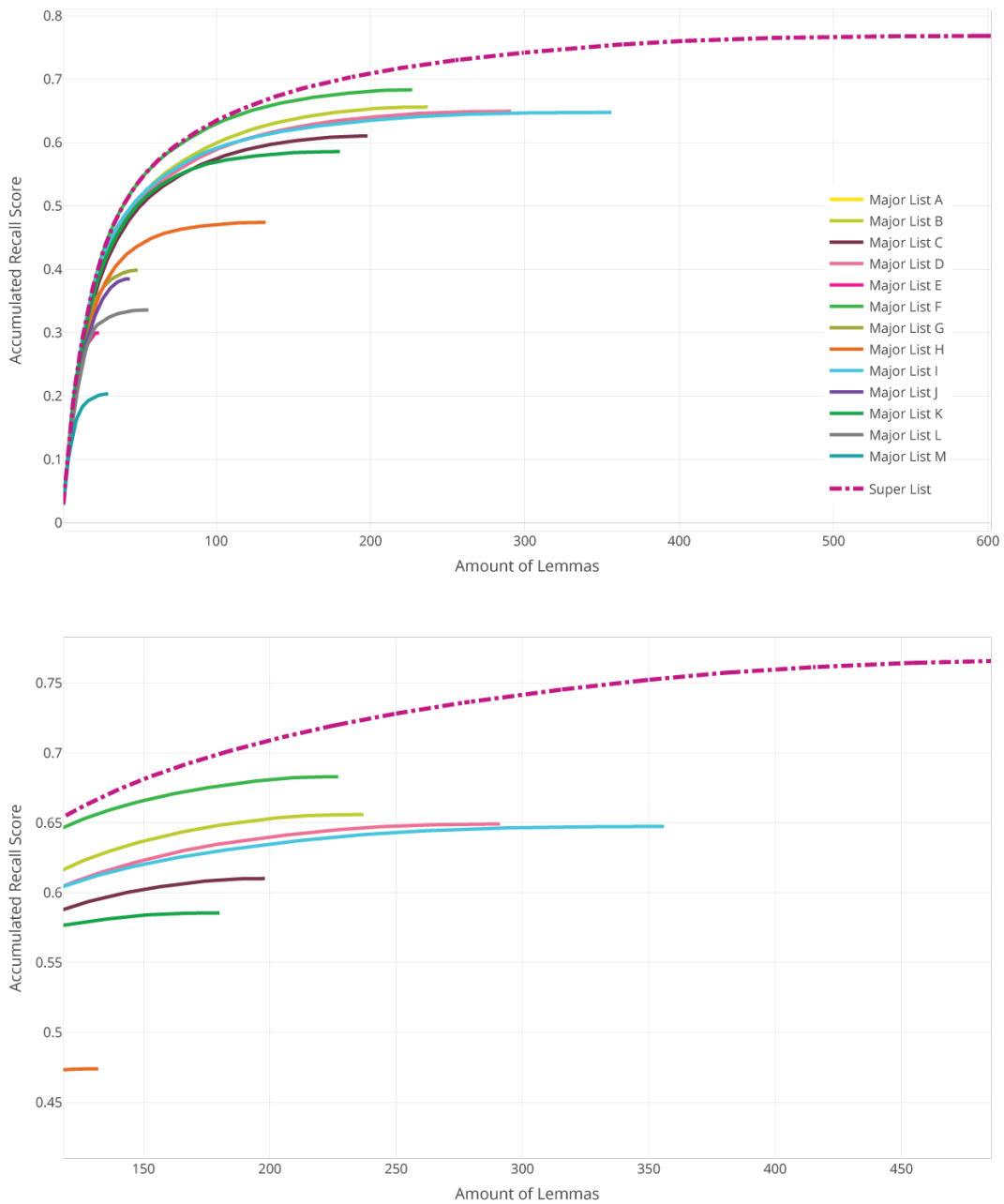


Figure 40 – Overview and detailed view of Major Lists and Super List recall over Test Corpora.

(68.297%;  $n = 227$ ) and Major List B (BEUKELMAN; JONES; ROWAN, 1989) (65.584%;  $n = 237$ ) have the better recalls among Major Lists. However, these results are not greater than Super List that achieves 70.9% with only 200 lemmas.

Figure 41 shows that the maximum score of Corpus-based Lists ranges from 37.770% (Corpus-based TF-IDF List;  $n = 466$ ) to 37.895% (Corpus-based TF List;  $n = 470$ ) and the maximum score of Combined Lists ranges from 76.815% (Combined TF-IDF List;  $n = 617$ ) to 76.847% (Combined DF List;  $n = 623$ ). None of the Corpus-based Lists recalled all the 500 lemmas as established in Section 4.5.2, as well as none of the Combined Lists

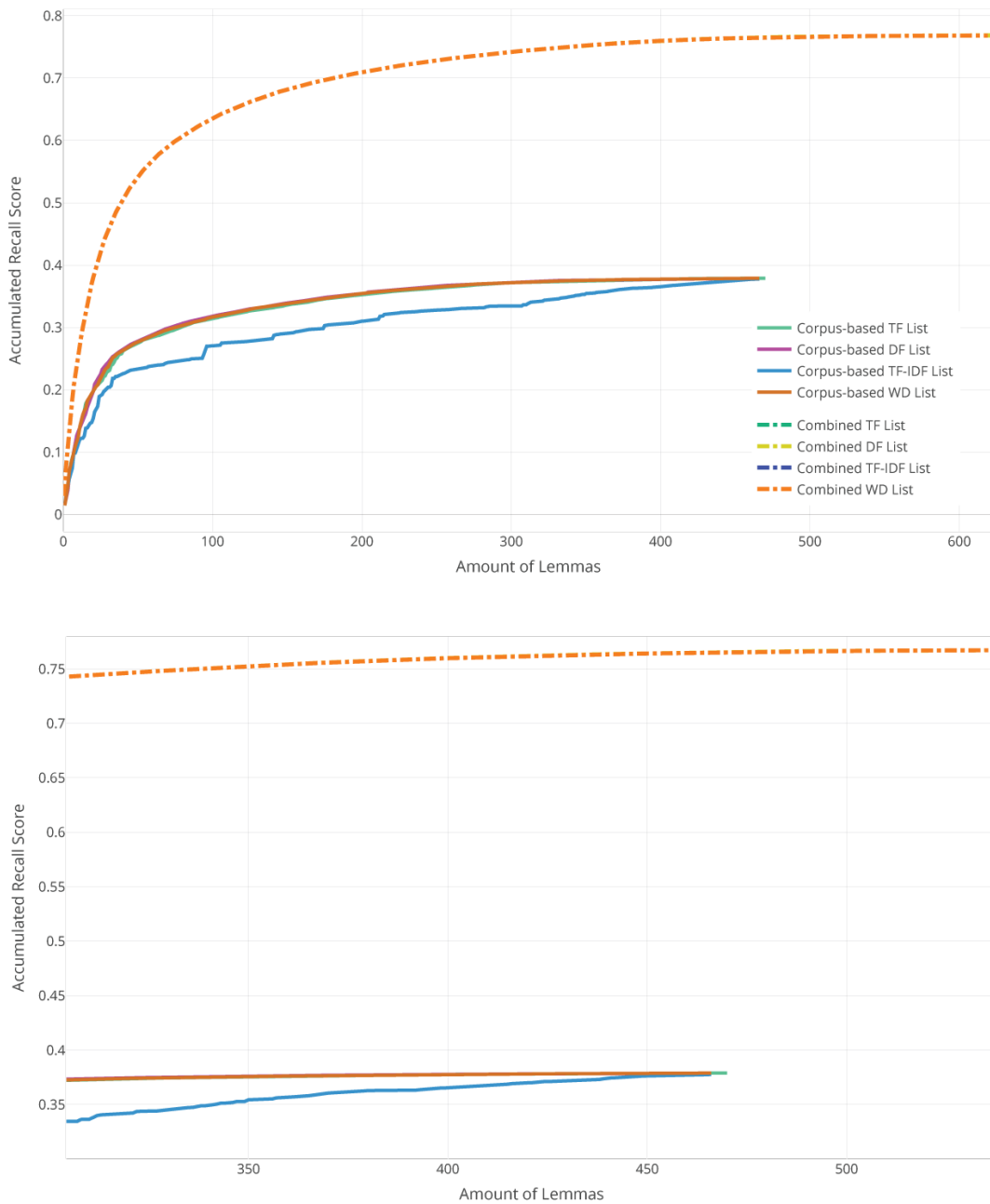


Figure 41 – Overview and detailed view of Corpus-based Lists and Composed Lists recall over Test Corpora.

(cf. Section 4.5.3) recalled all its lemmas in the Test Corpora.

Notice that the four Combined Lists and the Super List have approximately the same recall (i.e., ~76.8% of children's utterance), however, the Super List achieves this recall with fewer words. Moreover, concerning the maximum recall of Corpus-based Lists and Combined Lists, we can notice that the NLP metrics used to generate the Corpus-based Lists does not bring relevant words for the childish communication; whereas the words came from Super List (i.e., the relative complement of Super List added to the Corpus-based Lists) produced a higher recall. Therefore, we selected the 602 lemmas of the Super

List as our *Selected Core Word List* that was recalled in the Test Corpora as our *Selected Core Word List* because:

1. the Super List achieves its maximum score of 76% starting on  $n=402$ , that is, the remain 200 lemmas of Super List do not considerably increase its recall score (only 0.805%);
2. the other 200 considered words provides more variability for communication and more information to discover new categories;
3. the other 151 non-considered words were not used in any of the 506 documents Test Corpora (cf. Table 11), which means that these words are not common for the target group.

Table 20 – Maximum recall and number of lemmas of each lists over Test Corpora.

List	Maximum Recall	List	Maximum Recall
Major List A	64.917% (n= 291)	Major List L	33.570% (n= 56)
Major List B	65.584% (n= 237)	Major List M	20.320% (n= 30)
Major List C	61.005% (n= 198)	Super List	76.805% (n= 602)
Major List D	64.917% (n= 291)	Corpus-based TF List	37.895% (n= 470)
Major List E	29.980% (n= 24)	Corpus-based DF List	37.868% (n= 466)
Major List F	68.297% (n= 227)	Corpus-based TF-IDF List	37.770% (n= 466)
Major List G	39.880% (n= 49)	Corpus-based WD List	37.891% (n= 466)
Major List H	47.394% (n= 132)	Combined TF List	76.819% (n= 618)
Major List I	64.737% (n= 356)	Combined DF List	76.847% (n= 623)
Major List J	38.483% (n= 44)	Combined TF-IDF List	76.815% (n= 617)
Major List K	58.549% (n= 180)	Combined WD List	76.819% (n= 619)

### 5.1.2 Recall Analyses over Age Corpora and MLU Corpora

For this analysis we take the list with the better recall over Test Corpora– the *Selected Core Word List*, that is, the Super List – and calculate its recall over Age and MLU Corpora. Figure 42 presents the recall of the *Selected Core Word List* over Age Corpora, and Figure 43 presents the recall over MLU Corpora. Moreover, since recall lines are close, we show a detailed area on the bottom side of each figure. Finally, Table 21 summarizes the maximum recall score of the *Selected Core Word List* over each corpus.

In both corpora (i.e., Age and MLU) the *Selected Core Word List* have the same recall behavior, beginning with a small recall in the Group I, increases until Group IV and then, decreases in Group V and Upper Group, but the recall does not reach lower values than Group II. The recall in Age Corpora ranges from 73.131% (n= 483) in Age I Group to

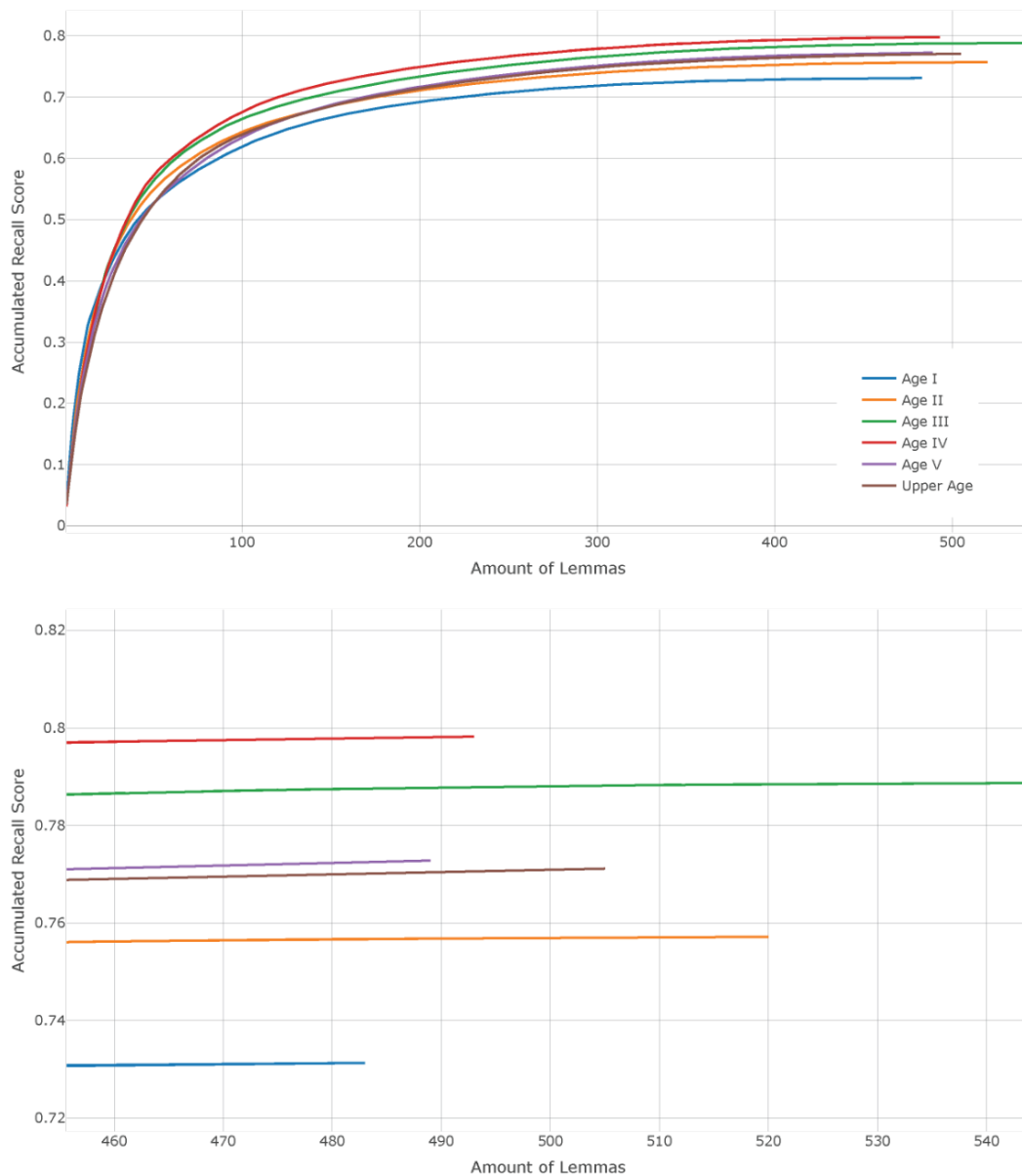


Figure 42 – Recall of the Selected Core Word List over Age Corpora.

79.821% (n= 493) in Age IV Group, while in MLU Corpora ranges from 73.248% (n= 402) in MLU I Group to 78.74% (n= 544) in MLU IV Group. We highlight that none of the groups recalled all the words in the *Selected Core Word List*, which means there are words considered as core by other authors that are related to a specific context in which the list was collected. Notice that the recall is equivalent in both group, Age and MLU, but with a slight difference. In the first three groups, the MLU groups have a higher recall with less words then the Age groups. However, in the last three groups this pattern reverses and the Age groups have a higher recall with less words then the MLU groups.

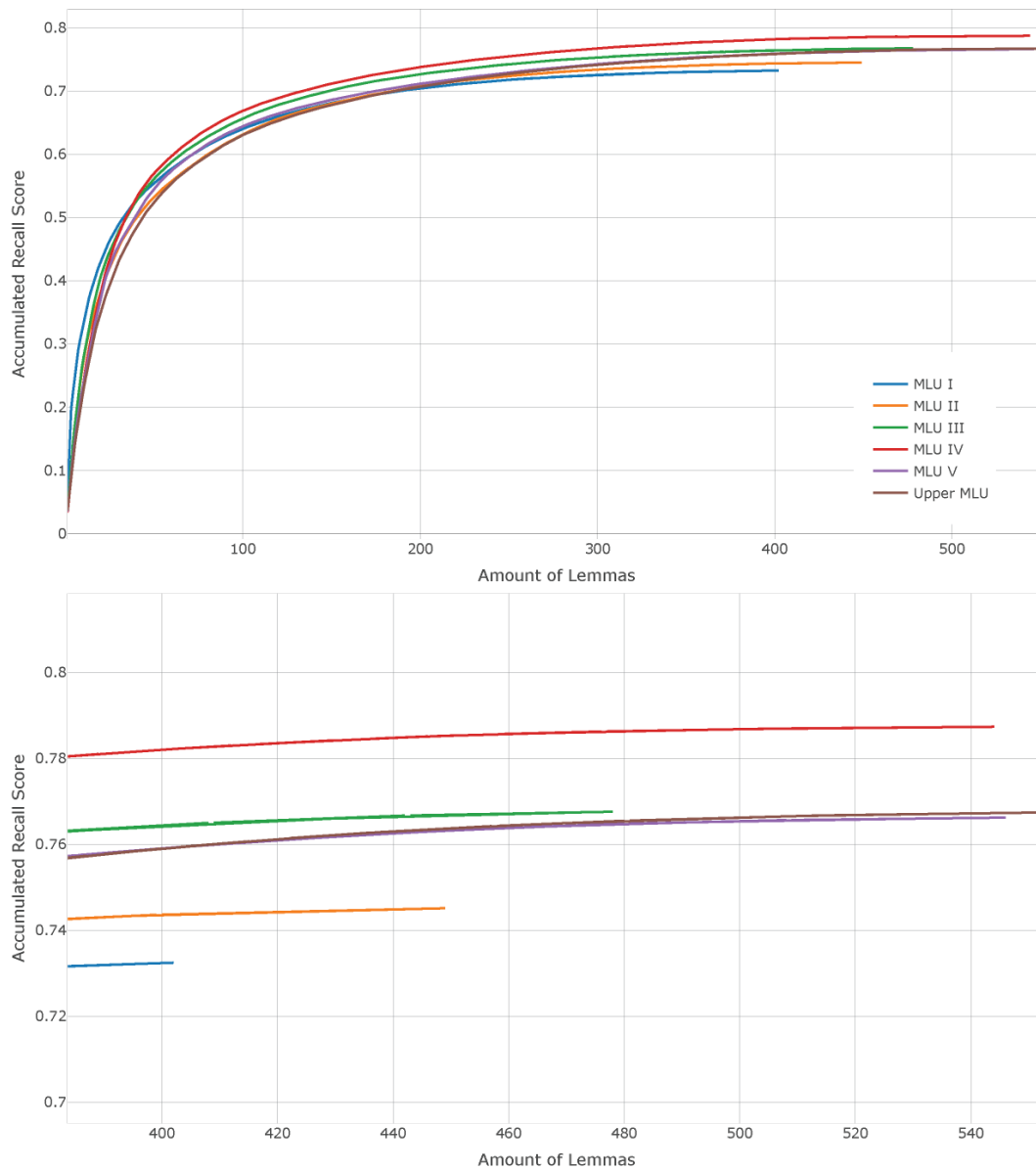


Figure 43 – Recall of the Selected Core Word List over MLU Corpora.

### 5.1.3 Summary of Results

The recall behavior of all lists does not change across different corpora samples. However, the curves seem to be more stretched in some cases (higher recalls), while in others it seems to be more flattened (lower recalls). We highlight the recall behavior during the children development considering Age and MLU. All the lists have a growth starting from lower stages of Age and MLU, achieves a maximum recall in the fourth stage, and then the recall decreases, but never reaches lower values then the second stage. We explain this behavior considering the children development and the usage of new words not included in the *Selected Core Word List*. Despite the MLU is a better metric than Age to analyze the child communication development (cf. Section 4.4.1), in practice, there is no significant

Table 21 – Maximum recall and number of words of the Selected Core Word List over AGE and MLU Corpora.

Recall of the Selected Core Word List			
Age Group	Maximum Recall	MLU Group	Maximum Recall
Age I	73.131% (n= 483)	MLU I	73.248% (n= 402)
Age II	75.716% (n= 520)	MLU II	74.516% (n= 449)
Age III	78.870% (n= 544)	MLU III	76.762% (n= 478)
Age IV	79.821% (n= 493)	MLU IV	78.740% (n= 544)
Age V	77.280% (n= 489)	MLU V	76.625% (n= 546)
Upper Age	77.114% (n= 505)	Upper MLU	76.746% (n= 553)

difference in the recall of the *Selected Core Word List* over the different Age and MLU corpora. This is a good outcome since this list can cover more than 70% of the children communication, independently of the measure used to group the children.

To define our core word list we consider the amount of words and its recall by analyzing recall curves over Test Corpora. When the curve considerably increases by adding new words it seems that these new words have a significant recall over the corpus sample. By contrast, when the curve has a stabilized region the recall of new words does not significantly alter the accumulated recall. For example, List E (BANAJEE; DICARLO; STRICKLIN, 2003) produces an increasing curve achieving 29.980% of children’s utterances with only 24 lemmas. However, a core vocabulary with only 24 lemmas do not bring communication variability for children. Regarding curves with increasing and stabilized regions, List C (FRIED-OKEN; MORE, 1992) recalls 61.005% of children’s utterances with 198 lemmas, while List D (MARVIN; BEUKELMAN; BILYEU, 1994) recalls 64.917% (almost 5% more) with 291 lemmas (93 more). The recall gain of List D comparing to List C is small, but the add of 93 more words give more variability for children communication. That is, we could establish a cut line in 200 words but the stabilized region provide information that can be useful to discover new interest subjects for children and to provide vocabulary variability for them. So, we take the 602 lemmas of the Super List that was recalled in the Test Corpora as our *Selected Core Word List* (cf. Section 5.1.1). We highlight that the recall of Super Lists is almost the same of Combined Lists, but with a fewer words. Moreover, the words with the higher recall of Combined Lists came from the Super List, once the Combined List was generated by the Corpus-based Lists (that have low recall) adding the relative complement of Super List (cf. Section 5.1.1). That is, the words of Super List are more meaningful then the words of Combined Lists.

With these results, we answer the RQ-1 (“*Is it possible to generate a new core word list with better recall than the existing core word lists?*”) showing that the union of the existing core word lists (cf. Section 4.1.1) generates a list with better recall. To the best of our knowledge, there is no other work that compares core word lists regarding the recall

of corpus. So, this study is relevant because allows the conscious comparison and selection of core word lists based on statistic evidence.

## 5.2 CATEGORY ANALYSES AND VOCABULARY CATEGORIZATION

In this analysis we consult specialists to verify the importance of a set of categories (cf. Sections 4.2.1 and 4.6) in an AAC system for children and for the vocabulary organization. This consult is made by using a 2-round-Delphi method, the Lawshe’s method, and a focus group (cf. Section 4.9).

### 5.2.1 Category Analyses

In the 1<sup>st</sup> Delphi round we had 29 respondents, where: 7 are from the Special Education field; 11 are from the Health field; 7 are from the Assistive Technology field; 2 are AAC user relatives; 1 is an AAC user relative which also works in both, the Special Education and the Assistive Technology fields; and 1 is from the Health and Special Education fields. According to Ayre and Scally (2014), the  $CVR_{critical}$  for  $N=29$  is 0.379 and the minimum  $n_e$  is 20. We began the 1<sup>st</sup> round with 60 categories organized into 9 semantic groups (cf. Section 4.9.1) and end up with: 1) 20 categories in the positive consensus area for “*Essential*”, which were mapped to our Useful Categories; 2) 2 categories in the positive consensus area for “*Useful, but not essential*”, which were mapped to our Useful Categories; and 3) 38 categories without consensus in any variable. We highlight that given the redundancy of categories in this round (cf. Section 4.9), there are categories with different names, but the same meaning in both, in consensus area and in the are without consensus. In this phase, we made three types of grouping: 1) Among consensus categories (e.g., *People*, *People/Relationships*, and *People/Professions*); 2) Among non-consensus categories and consensus categories (e.g., *People’s Name*, *People*, *professions and personal pronouns*, and *People*); and 3) Among non-consensus categories (e.g., *Toys*, *entertainment*, *sports*, *instruments and music*, and *Entertainment*). That is, although some categories are considered important for our respondents (i.e., they are not “*Not necessary*”), they are not “*Essencial*” or “*Useful, but not essential*” to be considered as a useful category and can be part of a more general category (e.g., *People’s Name*, *Professions*, and *Relationships* can be part of *People*). Table 22 summarizes these results.



Table 22 – 1<sup>st</sup> Delphi round summarization.

Type	N_categories	Categories
Consensus Categories – “ <i>Essential</i> ” [Useful Categories]	15	Actions/Verbs; Activities and Routines; Articles; Body Parts; Clothes; Foods and Drinks; Locations/Places; Materials; Numbers/Counting; People; Pronouns; Quantity; Questions and Answers; Self-/social-awareness, Emotion/Feelings, and Behavior; Size/Comparison; Social Expressions/Greetings; and Toys.
Consensus Categories – “ <i>Useful, but not essential</i> ” [Useful Categories]	2	Articles; and Materials.
Consensus categories that were grouped in another consensus categories	5	People/Relationships; People/Professions; Question Words; Yes/No Responses; Questions, answers and social expressions.
Non-consensus categories that were grouped in any consensus categories	11	Behavior; Furniture and rooms; People’s name; Comparison; Expressions/Contractions; Home routines and activities; People, professions and personal pronouns; Pronouns and prepositions; School routines and activities; Quantity/Time; Letters and Numbers.
Non-consensus categories that were grouped in other non-consensus category	9	Conjunctions; Furniture and rooms; Prepositions; Sequence; Pronouns and prepositions; Toys, entertainment, sports, instruments and music; Utensils and objects; Quantity/Time; Letters and Numbers.
Suggested Categories [For the 2 <sup>nd</sup> round]	3	Climate; Hygiene; Rooms.
Non-consensus Categories [For the 2 <sup>nd</sup> round]	19	Animals; Colors; Demonstratives; Descriptors (e.g., adjectives and adverbs); Direction/Position; Entertainment (e.g., toys, entertainment, sports, instruments and music); Furnitures; Household items; Letters; Nature; Objects; Particles and connectors (e.g., conjunction, prepositions); Positions and Equipment; School Materials; Shapes; Texture/material; Time/Sequence; Transport; and Weapons.

In the 2<sup>nd</sup> Delphi round, only 15 of the 29 volunteers of the 1<sup>st</sup> round answered the questionnaire. According to Ayre and Scally (2014), the  $CVR_{critical}$  for  $N=15$  is 0.600 and the minimum  $n_e$  is 12. We begun this round with 22 categories (cf. Table 22) and end up with: 1) 2 categories in the positive consensus area for “*Essential*” (*Descriptors* and *Time/Sequence*), which were mapped to the Useful Categories; 2) 2 categories in the

negative consensus area for “*Essential*” (*Particles and connectors*, and *Weapons*), which were excluded from our list; 3) 18 categories without consensus which were mapped to the Useful Categories. We end up this phase with 36 useful categories. Notice that we grouped the *Materials* category from the 1<sup>st</sup> Delphi round into the *Texture/material* category from the 2<sup>nd</sup> Delphi round. Table 23 summarizes this result. Notice that the resulting categories are a join of two different approaches – the Taxonomic and the Schematic approach (cf. Section 2.1.1.2) – because this result combines hierarchical categories (e.g., *People*, *Clothes*) and event schema categories (e.g., *Activities and Routines*).

Table 23 – 2<sup>nd</sup> Delphi round summarization.

Type	N_categories	Categories
Useful Categories	36	Actions/Verbs; Activities and Routines; Animals; Articles; Body Parts; Climate; Clothes; Colors; Demonstratives; Descriptors (e.g., adjectives and adverbs); Direction/Position; Entertainment (e.g., toys, entertainment, sports, instruments and music); Foods and Drinks; Furniture; Household items; Hygiene; Letters; Locations/Places; Nature; Numbers/Counting; Objects; People; Positions and Equipment; Pronouns; Quantity; Questions and Answers; Rooms; School Materials; Self-/social-awareness, Emotion/Feelings, and Behavior; Shapes; Size/Comparison; Social Expressions/Greetings; Texture/material; Time/Sequence; Toys; and Transport.

### 5.2.2 Vocabulary Categorization

We organize the *Selected Core Word List* (cf. Section 4.8) into our *Useful Categories* (cf. Table 23). We made the initial categorization according to the steps showed in Section 4.9 and presented it to two specialists in linguistics, AAC, and ASD during a 1-section Focus Group. With the orientation of these specialists, we made 5 changes in the initial categorization:

1. Removal of interjections and onomatopoeia – it was removed 12 words (“*Ah*”, “*Boo*”, “*Bum*”, “*Hah*”, “*Huh*”, “*Hum*”, “*Oh*”, “*Ouch*”, “*Um*”, “*Whoa*”, “*Whoop*”, and “*Wow*”);
2. Maintain only infinitive verbs in the *Action/Verbs* category – when any verb inflection appears with a different POS we try to categorize it in another category (e.g., *playing* was categorized in *Activities and Routine*);
3. Grouping word synonyms in the same category – we analyzed words that appear in the same WordNet synset, putted them in order according to its frequency position and separated with a slash, maintaining the position of first appearing (e.g., “78 -

*Hello/Hi/Hey*” means that this concept first appeared in the 78 position with the word “*Hello*”; only for information, “*Hey*” appeared in the position 260 and “*Hi*” in the 399);

4. Change of category names – *Body Parts* was changed to *Body* to encompass words such as “*Blood*” and “*Voice*”. *Article* was changed to *Small Words* to encompass more POS such as conjunctions and prepositions. *Clothes* was changed to *Clothes and Accessories* to encompass words such as “*Watch*”;
5. Combination of some core words to build an expression or a concept – “*Excuse + me*” and “*Thank + you*” in *Social Expressions/Greetings* and “*Hot + Dog*” in *Foods and Drinks*. Notice that there are several combinations among core words and that we only made a few of them to demonstrate this possibility.

We end up this phase with 614 word senses categorized into 36 categories, where 1 category (*Position and Equipment*) had no words categorized, and 12 words remain without categorization (for these words we create a *Miscellaneous* category). The categories of *Action/Verbs*, *Descriptors* and *Time/Sequence* categorized the highest number of words senses: 111, 60, and 33 respectively. By contrast, the categories *Texture/Material* and, *Rooms* categorized 3 word senses, the categories *Hygiene*, and *Household Items* categorized 2 word senses, and the category of *Climate* categorized only one sense. Notice there are words with different senses that are categorized in more than one category, for example the word *Fish* can be an *Action/Verb*, an *Animal* or a *Food*. In Appendix A, Table 32 we show the final categorization of our Core Vocabulary.

During the categorization task we found a lot of overlap among categories. For example, “*Bed*” is a *Furniture*, but is also an *Object*, and “*Blue*” is a *Color* and also a *Descriptor*. With this in mind, we suggest a category hierarchy (cf. Figure 44) that can be used to establish links between concepts and to improve the abstraction learning. The categories in bold (i.e., 1, 30, 34, and 38) were considered in this proposal only for means of organization. On the one hand, in positions 30 and 38 of the hierarchy can be created a supercategory to group subcategories with similar semantics. On the other hand, in position 34 of the hierarchy can be crated some subcategories to specialize more specific content of the *Self-/Social-Awareness*, *Emotion/Feelings*, and *Behavior* category, which is very general. Finally, number 1 in the hierarchy is only to organize the other categories.

### 5.2.3 Summary of Results

In Appendix A, Table 32, we show our Core Vocabulary. From the initial set of categories (cf. Section 4.2) we remain with 36 useful categories based in its relevance for AAC systems used by children. We highlight that categories that were not taken into consensus in the 2<sup>nd</sup> Delphi round were not discarded and were labeled as useful, once that these

Figure 44 – Suggestion of a Category Hierarchy.

1 [ROOT]		
2   Action/Verbs	16   Objects	30   [SOME SUPERCATEGORY]
3   Activities and Routines	17     Household Items	31     Numbers/Counting
4     Hygiene	18     Clothes and Accessoires	32   °– Letters
5   °– Positions and Equipments	19     Furniture	33   Self-/Social-Awareness,
6   Body	20   °– School Materials	Emotion/Feeling, and Behavior
7   Descriptors	21   Location/Places	34   °– [SOME SUBCATEGORIES]
8     Size/Comparison	22     Direction/Position	35   People
9     Colors	23   °– Rooms	36   Pronouns
10     Shapes	24   Nature	37   °– Demonstratives
11   °– Texture/Material	25     Animals	38   [SOME SUPERCATEGORY]
12   Quantity	26   °– Climate	39     Question and Answers
13   Foods and Drinks	27   Entertainment	40   °– Social Expressions/Greetings
14   Time/Sequence	28   °– Toys	41   °– Miscellaneous
15   Transport	29   Small Words	

categories already have been considered as relevant for other authors. We also create a *Miscellaneous* category for those words that do not properly fit into any category.

Concerning the categorization, most of the core words (73.9%; n= 456) were categorized into categories considered as “*Essential*”, specially *Action/Verbs* and *Descriptors*, while 26.1% (n= 161) of them were categorized into categories considered as “*Useful, but not essential*” or that have no consensus in the 2<sup>nd</sup> Delphi round. Considering the first 100 most frequent words, 71 words are categorized into categories considered as “*Essential*” and 28 into categories considered as “*Useful, but not essential*” or that have no consensus in the 2<sup>nd</sup> Delphi round (notice that the sum is not 100 because some words are grouped and others are replicated in other categories – cf. Table 32). *Action/Verbs* is the category that concentrates more words of the first 100 words.

Concerning the categories and its words, we highlight that:

- In general, the categories are not limited to the presented words. Depends on the application or on the audience, these words must be replaced and new words must be added. For example, the whole alphabet may be added in Letters, and other colors may be added in Colors. Another example is the *Social Expressions/Greetings* category that includes the expressions “*Good Morning*” and “*Thank you*”, and can also categorize other expressions that were not considered (e.g., “*Good Night*”);
- New words and concepts can be built by combining core words. The *Position and Equipment* category, which had no words categorized, may be filled with a range of expressions using core words (e.g., “*Sit in the chair*” and “*Turn me in the bed*”) according to the user needs. The *Activities and Routines* category may be filled with expressions that represent user-specific activities and routines (e.g., “*Take the bus*” and “*Listen to rock*”);
- The *Action/Verbs* category, that does not considered the negative verb forms nor phrasal verbs, includes the word “*Not*” to allow the construction of the negative verb forms (e.g., “*Not like*”) and allows the combination of words to compose phrasal verbs (e.g., “*Get over*” and “*Look for*”).

Categories are useful to organize the vocabulary according to the word meaning or its use. For improve this organization, we also suggest an hierarchy that: 1) Establish generalization/specialization links between concepts and improve the abstraction learning – for example, the child learn that *Colors* and *Shapes* are examples of a bigger group named *Descriptors*; and 2) Facilitate the information retrieval when the children vocabulary grows – considering the *Descriptors* category with a lot of instances, the subcategories *Colors* and *Shapes* help the retrieval of “Red” and “Square”, for example.

With these results, we answer the RQ-2 (“Are the already proposed categories useful and sufficient to categorize all the core words of the best list of RQ-1?”) showing that the already proposed categories (cf. Section 4.2.1) are not sufficient to categorize all the words of the *Selected Core Word List*, once we had to insert four other categories (*Climate*, *Hygiene*, *Rooms*, and *Miscellaneous*). However, our results show that the list of categories are useful in the context of children communication.

### 5.3 CORE VOCABULARY ORGANIZATION

In this section, we present the steps for building the ontology, which the computational representation of the Core Vocabulary resulting from Sections 5.1 and 5.2. This section is divided in six parts according to the tasks of the METHONTOLOGY framework (cf. Section 4.10): 1) Ontology specification (cf. Section 5.3.1); 2) Knowledge acquisition (cf. Section 5.3.2); 3) Integration (cf. Section 5.3.3); 4) Conceptualization (cf. Section 5.3.4); 5) Implementation (cf. Section 5.3.5); and 6) Evaluation (cf. Section 5.3.6).

#### 5.3.1 Ontology Specification

The ontology built here (a.k.a. AACOnto) is in the domain of functional communication for children with complex communication needs and aims to organize the common words used by children in functional communication. The ontology is limited to the Core Vocabulary specified in Sections 5.1 and 5.2, and to the chosen reference ontologies (cf. Section 4.3.4). Despite that, the ontology can be evolved to encompass more concepts and realize more functionality. Concerning the formality, the ontology built here is considered as semi-formal (USCHOLD; GRUNINGER, 1996), once it is expressed in an artificial formally defined language. Table 24 summarizes the specification phase.

#### 5.3.2 Knowledge Acquisition

For knowledge acquisition, we used:

- Core word lists for children (cf. Sections 4.1.1, 4.5.1, 4.5.2, and 4.5.3);
- Category lists for organizing children vocabulary (cf. Sections 4.2.1, and 4.6);
- Corpus with children’s utterances (cf. Section 4.3.1, 4.4.1, and 4.4.2);

- Reference ontologies for describing and organizing vocabularies (cf. Section 4.3.4);
- Statistical methods for selecting core words (cf. Section 4.8);
- Qualitative methods with expert consulting for selecting essential categories and organize the vocabulary (cf. Section 4.9).

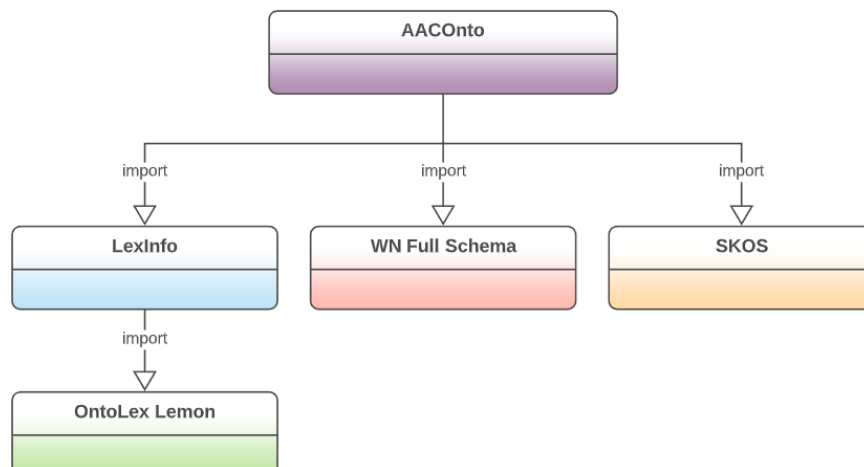
Table 24 – AACOnto Ontology specification

Ontology Specification	
Name	AACOnto
Domain	Functional communication for children with complex communication needs.
Purpose	Create an ontology to organize the core vocabulary used by children in functional communication, and that can be used in AAC systems for a variety of purposes (e.g., communication and education).
Level of formality	Semi-formal
Scope	Contains a set of common and frequent words used by children in functional communication. The scope is limited by analyzing corpus containing children’s utterances.
Sources of Knowledge	Core word lists, category lists, domain specialists, corpus with children’s utterances, and reference ontologies.

### 5.3.3 Integration

Figure 45 shows the four reference ontologies (cf. Section 4.3.4) imported by AACOnto. Three of them (i.e., LexInfo, WN Full Schema, and SKOS) are directly imported, while the other one (i.e., Ontolex Lemon) is an indirect import, once it is already imported by LexInfo.

Figure 45 – Imported ontologies by AACOnto.



The *LexInfo* and the *OntoLex Lemon* are responsible for provide linguistic descriptions about ontology concepts. On the one hand, we will use all the core (cf. Section 4.3.4, Figure 27) of *OntoLex Lemon*, which describes the morphological and syntactic properties of lexical entries. On the other hand, despite *LexInfo*’s expressiveness in describing the elements of ontology with linguistic information, in this version of *AACOnto*, we will use only the morphosyntactic property of POS.

The *WN Full Schema* is responsible for mapping the elements of ontology to a linguistic knowledge database aiming to define the meaning of each lexical entry. For example, by mapping the lexical entry “*fish*” with WordNet synsets, you can define whether “*fish*” is an action, an animal or a food.

The *SKOS* is responsible to create the taxonomic hierarchy that will be seen by users in the AAC user’s Graphical User Interface (GUI). For example, “*Animals*” is a *skos:Collection* that has as member the *skos:Concept* “*Dog*”. These two ontology concepts are linked by the *skos:member* object property.

We emphasize that imported ontologies are only for organizing the concepts and that we do not make any modifications. All the instances, and new object and data properties are created under the *AACOnto* namespace.

### 5.3.4 Conceptualization

We built a Data Dictionary including the domain concepts, its descriptions (i.e., hyponym in taxonomy, and WordNet synset), and type (i.e., class, instance, attribute) with the correspondent class of the reference ontology. Table 25 shows an excerpt of this data dictionary. In this table we show how the description gives a different meaning (i.e., *action*, *animal*, *food*) to the same lexical entry (i.e., “*fish*”), and how the class of the reference ontology modifies the perception of an individual (i.e., *skos:Concept*) and a collection (i.e., *skos:Collection*) in the taxonomic hierarchy.

We also include a graphical representation of the ontology, comprising the classes and relations of reference ontologies and some examples of instances from our Core Vocabulary. Figure 46 shows this graphical representation exemplifying the lexical entry of the noun “*fish*”.

From the graphical representation of the ontology (cf. Figure 46), we build a table with the relations used. An excerpt of the relations – focusing on the relationships between different namespaces (i.e., different reference ontologies) – is shown in Table 26, where relations, inverse relation, cardinality, and domain and range concepts are specified.

### 5.3.5 Implementation: The AACOnto Ontology

The *AACOnto* is the computational representation of the *Core Vocabulary* (cf. Appendix A) for children who uses AAC. This ontology was not built from scratch and imports four reference ontologies (cf. Sections 4.3.4 and 5.3.3).

Figure 46 – AACOnto Schema.

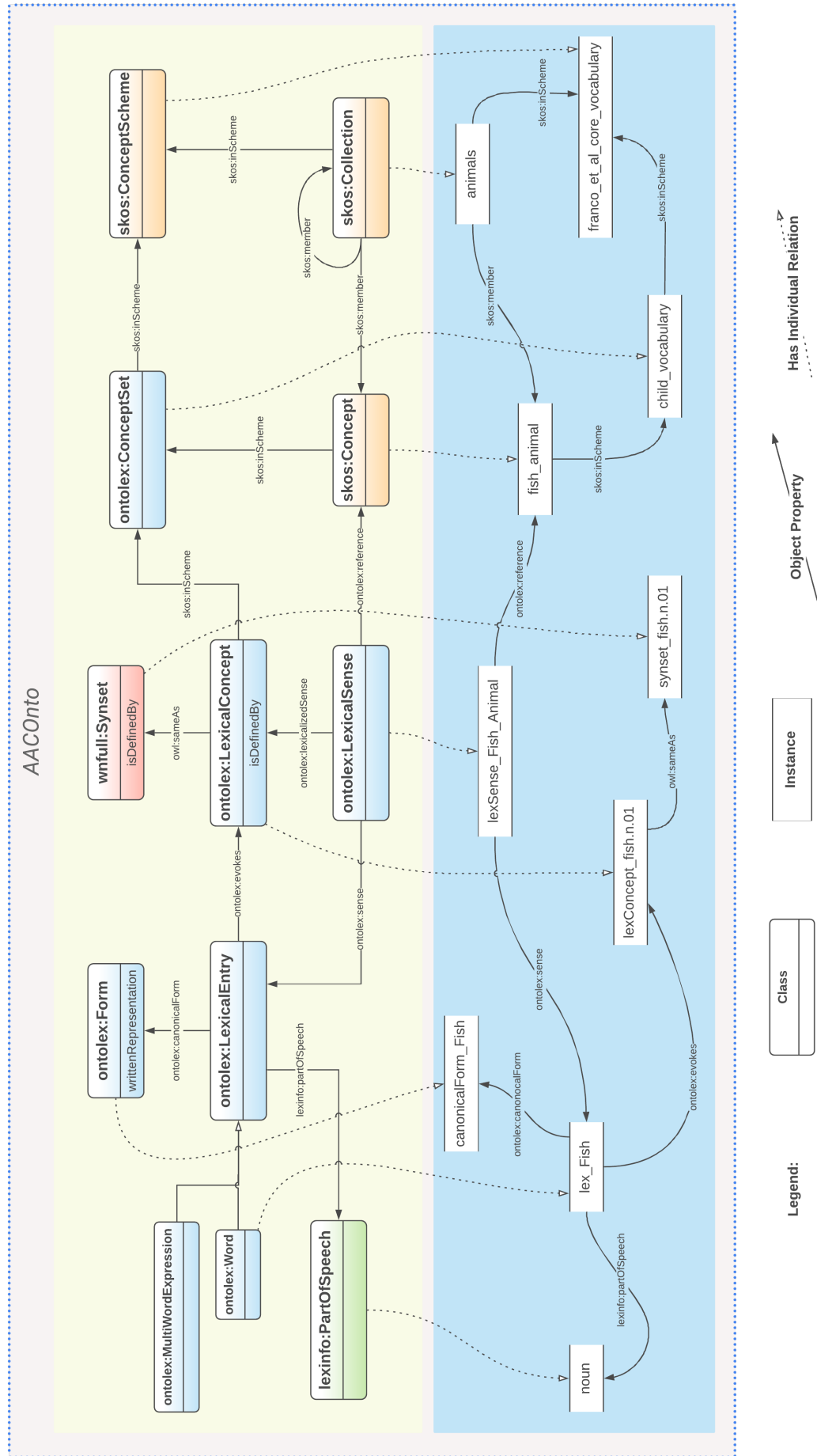




Table 25 – An excerpt of the Data Dictionary for terms classification and descriptions.

Concept Name	Synonyms	Hypernym in Taxonomy	Synset	Type
Actions	-	Root	action.n.01	Instance (skos:Collection)
Animals	-	Nature	animal.n.01	Instance (skos:Collection)
Foods and Drinks	-	Root	food.n.01, food.n.02, beverage.n.01	Instance (skos:Collection)
People	-	Root	person.n.01, people.n.01	Instance (skos:Collection)
Fish	-	Action	fish.v.01	Instance (skos:Concept)
Fish	-	Animal	fish.n.01	Instance (skos:Concept)
Fish	-	Foods and Drinks	fish.n.02	Instance (skos:Concept)
Mummy	Mum, Mama, Mom, Mommy	People	ma.n.01, mama.n.02	Instance (skos:Concept)

Table 26 – Examples of relations in ontology.

Domain Concept	Relation	Range Concept	Inverse Relation	Cardinality
wnfull:Synset	<i>sameAs</i>	ontolex:LexicalConcept	<i>sameAs</i>	1:1
ontolex:LexicalEntry	<i>partOfSpeech</i>	lexinfo:PartOfSpeech	-	1:1
ontolex:LexicalSense	<i>reference</i>	skos:Concept	<i>isReferenceOf</i>	1:1
ontolex:ConceptSet	<i>inScheme</i>	skos:ConceptScheme	-	1:N

The AACOnto was implemented in RDF using the *Python* language (PYTHON CORE TEAM, 2019) with the *Owlready2* (LAMY, 2017), a package for ontology-oriented programming in *Python*, in the *PyCharm* platform<sup>1</sup>. To visualize the ontology we use the *Protégé* software (STANFORD CENTER FOR BIOMEDICAL INFORMATICS RESEARCH, 2019), which provides some general metrics of the ontology (cf. Table 27).

We instantiate all the words and categories of the *Core Vocabulary* as instances in AACOnto. With this ontology, we aim to provide only the main concepts, this way we do not create many attributes for instances, only three: 1) *isCoreWord* – a boolean associated to the *skos:Concept* instances that points if the concept came from the *Core Vocabulary* (this information will make sense when other concepts are added to the ontology); 2) *isCoreCategory* – a boolean associated to the *skos:Collection* instances that points if the category came from our *Useful Categories* (cf. Table 23); and 3) *relatedPictogram* – a string associated to the *skos:Concept* and *skos:Collection* instances that contains an URL for an example of pictogram that describes the concept or collection. As mentioned before, the

<sup>1</sup> PyCharm: The Python IDE – <https://www.jetbrains.com/pycharm/>

Table 27 – AACOnto metrics.

Metrics	
Axiom	30047
Logical axiom count	14366
Declaration axioms count	4936
Class count	210
Object property count	176
Data property count	14
Individual count	4524
Annotation Property count	37
DL expressivity	SHOIQ(D)
Class axioms	
SubClassOf	196
EquivalentClasses	110
Individual axioms	
ClassAssertion	4918
ObjectPropertyAssertion	7765
SameIndividual	627

aim of AACOnto is provide knowledge of the main concepts for children communication. The customization will be made in the application that will use AACOnto. For example, the AAC application can have distinct pictograms and legends that reference the same *skos:Concept* in the ontology (cf. Figure 47).

Figure 47 – Example of instances for the concept “Cat” on AACOnto.

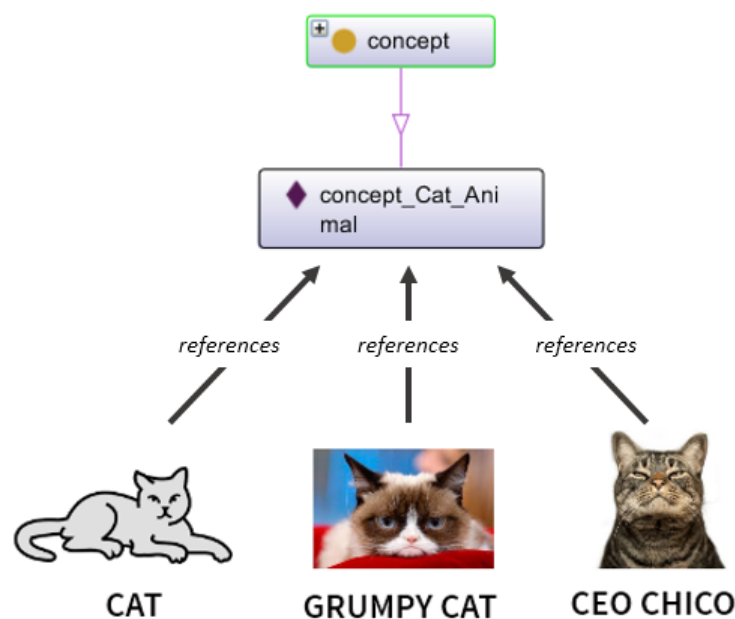


Figure 48 shows an excerpt of the AACOnto centered on the “Cat” concept. In this example, the concept is a noun (i.e., a content word) and is on WordNet, with a specific

synset for this word. For this reason, on AACOnto, there is a specific Lexical Concept that gives the exact meaning of “Cat” (i.e., *cat.n.01* and *animal.n.01*) with a specific Lexical Sense (i.e., *Cat\_Animal*), and the object property *lexicalizedSense* connecting the lexical concept and the lexical sense (i.e., *cat.n.01* is the *lexicalizedSense* of *Cat\_Animal*). In this case, we used those POS provided by WordNet.

Figure 48 – Instance “Cat” on AACOnto.

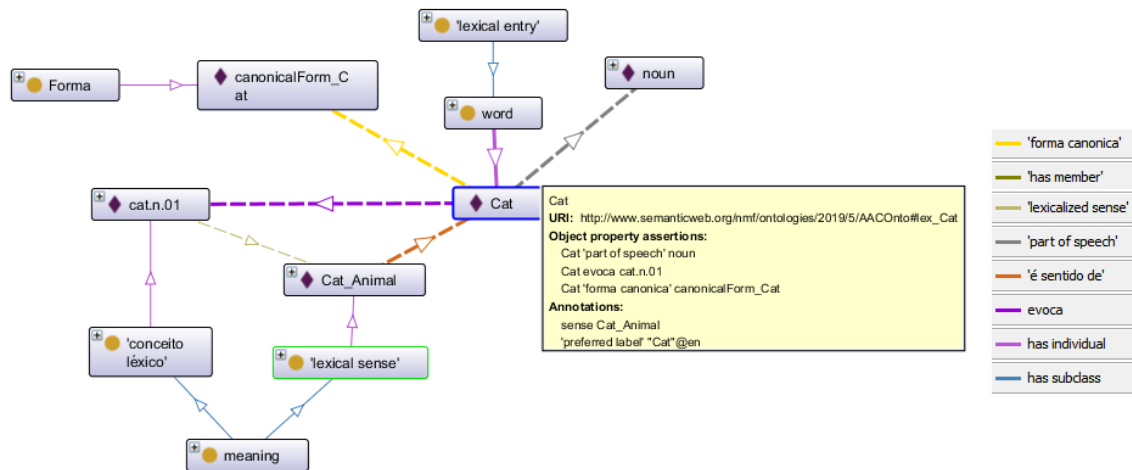
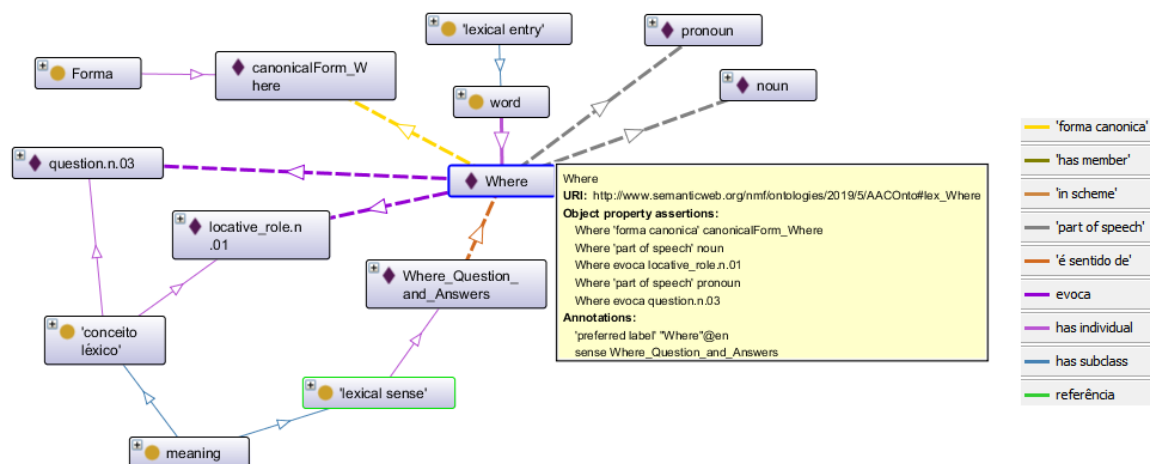


Figure 49 shows an excerpt of the AACOnto centered on the “Where” concept, which is an interrogative adverb (i.e., a structure word) that is not on WordNet. In this case, although we do not have the exact meaning, we used as Lexical Concept those synsets that represent the concept meaning (i.e., *locative\_role.n.01* and *question.n.03*), as Lexical Sense its specific concept (i.e., *Where\_Question\_and\_Answers*), however we do not connect them by the object property *lexicalizedSense*. In this case, we map the non-specific synsets (i.e., those that represent the structure word meaning) to the POS that better represent its meaning, and attribute this POS to the Lexical Entry. We highlight that we made this procedure for all concepts that are not on WordNet.

Figure 49 – Instance “Where” on AACOnto.



### 5.3.6 Evaluation

The AACOnto was evaluated under the metrics of OntoQA (TARTIR; ARPINAR, 2007) (cf. Section 4.10) that is divided into two categories: Schema metrics and Instance metrics. The latter is also divided into Knowledge metrics and Class metrics.

Table 28 shows the obtaining values for the Schema metrics described with its interpretations as follows:

- *Relationship Richness* – the result is a percentage that reflects the diversity of relations in the ontology. An ontology that contains many relations other than class-subclass relations is richer than a taxonomy with only class-subclass relationships. Values near zero show that most of the relations are class-subclass (i.e., *is-a* relations), while values near to 100 indicates that most of relationships are other than class-subclass. Most of AACOnto relationships are other than class-subclass;
- *Attribute Richness* – the result is a real number representing the average number of attributes per class, which generally indicate the quality of ontology design, assuming that the more slots that are defined the more knowledge the ontology conveys. AACOnto has few attributes per class;
- *Inheritance Richness* – the result is a real number representing the average number of subclasses per class. Low values reflect a horizontal ontology, while high values reflect a vertical ontology with large number of inheritance levels. AACOnto is a horizontal ontology.

Table 28 – OntoQA schema metrics for AACOnto.

Schema Metrics	
Relationship Richness	100%
Inheritance Richness	0.0
Attribute Richness	1.0

Table 29 shows the obtaining values for the Knowledge metrics described next:

- *Class Richness* – the result is a percentage that is related to how instances are distributed across classes. The number of classes that have instances is compared with the total number of classes. Low values means that the ontology do not have data that exemplifies all the knowledge in the schema, while high values means that the data in the ontology represents most of the knowledge in the schema. AACOnto instances represent most of the ontology data;
- *Average Population* – the result is a real number representing the average distribution of instances across all classes. Low values reflect that instances are insufficient

to represent most schema knowledge, while high values reflect that instances are sufficient to represent most schema knowledge. AACOnto instances are sufficient to represent most schema knowledge;

- *Instance Coverage* (a.k.a. Cohesion) – considering the ontology as a graph, where nodes represent instances and edges represent the relationships between them, the result of this metric is a real number representing the number of separate components in the ontology. Low values reflect that the ontology is connected, while high values reflect the existence of “islands” in the ontology. AACOnto has some isolated components.

Table 29 – OntoQA knowledge metrics for AACOnto.

Knowledge Metrics	
Class Richness	93.75%
Average Population	266.0
Instance Coverage (a.k.a. Cohesion)	303.35

Table 30 shows the obtaining values for three Class metrics, that we show next, considering the main classes of AACOnto. As *main classes* we consider those classes that subsumes the instances related to the *Core Vocabulary*. Notice that, once AACOnto imports reference ontologies, there are many other classes in the ontology. Besides the three metrics of Table 30, OntoQA also provide values for *Relationship richness* and *Inheritance richness*. The first is a percentage that is related to the quantity of properties in each class of the schema is actually being used at the instance level; and the second is a real number denoting the average number of subclasses per class in the subtree. All the classes of Table 30 have the zero value because: 1) AACOnto only instantiate a few properties; and 2) AACOnto is a horizontal ontology. Next, we describe the three class metrics of Table 30:

- *Direct Instances* – the result is an integer denoting the number of instances in the class without consider the other classes at the subtree. The most populated class in AACOnto is the wnfull:NounSynset;
- *Class Importance* – the result is a percentage that is related to the quantity of instances that belong to classes at the subtree rooted at the current class with respect to the total number of instances. The most important class in AACOnto is the wnfull:NounSynset;
- *Connectivity* – the result is an integer denoting the number of relationships of each class to other classes. A class with a high connectivity plays a central role in the ontology compared to a class with a lower value. The most connected class in AACOnto is the wnfull:NounSynset.

Table 30 – OntoQA Class Metrics.

Class Name	Direct Instances	Class Importance	Connectivity
<i>wnfull:AdjectiveSatelliteSynset</i>	30	0.70%	30
<i>Form</i>	512	12.03%	0
<i>skos:Concept</i>	612	14.37%	4278
<i>ontolex:Word</i>	512	12.03%	1710
<i>skos:Collection</i>	38	0.89%	329
<i>ontolex:LexicalConcept</i>	627	14.73%	2612
<i>lexinfo:PartOfSpeech</i>	11	0.25%	0
<i>ontolex:ConceptSet</i>	1	0.02%	1
<i>wnfull:Synset</i>	0	0.0%	0
<i>ontolex:LexicalSense</i>	646	15.17%	1258
<i>skos:ConceptScheme</i>	41	0.96%	0
<i>wnfull:NounSynset</i>	909	21.35%	2192
<i>ontolex:MultiWordExpression</i>	3	0.07%	9
<i>wnfull:VerbSynset</i>	250	5.87%	521
<i>wnfull:AdverbSynset</i>	21	0.49%	21
<i>wnfull:AdjectiveSynset</i>	43	1.01%	43

### 5.3.7 Summary of Results

In this section, we presented the AACOnto, an ontology for organizing the *Core Vocabulary* (cf. Appendix A) for high-tech AAC systems. We built AACOnto according to the tasks of the METHONTOLOGY framework and with OntoQA metrics (cf. Section 4.10). AACOnto was not built from scratch and reused 4 reference ontologies (i.e., Lex-Info, OntoLex Lemon, WN Full Schema, and SKOS), which bring a consolidate schema for organizing lexicon, taxonomies, and the WordNet data. Our ontology has 210 classes, 176 relationships, and 4524 instances.

Despite being considered an horizontal ontology by the OntoQA metrics, AACOnto has a part that can be considered as vertical, with many subclasses. The instances under the subtree rooted by *wnfull:Synset* – which, subsumes: *wnfull:NounSynset*, *wnfull:VerbSynset*, *wnfull:AdverbSynset*, and *wnfull:AdjectiveSynset* – rebuilt the full taxonomy of WordNet by using the object properties *hypernymOf* and *hyponymOf*. The WordNet synsets allow the sense disambiguation of the concept; and the full WordNet taxonomy makes possible to perform similarity calculus that can be used for Word Sense Disambiguation (WSD), pictogram suggestion, and pattern identification, for example. This can also be used to make the ontology learn by her own, that is, establish new relationships between concepts, whether old or new.

The main objective of our ontology is to organize the main concepts used by children during communication. We highlight that customization over concepts (e.g., pictogram and legend) are not in the scope of our ontology and must be done in the AAC application that uses AACOnto. Even with a delimited scope, ontologies are incomplete by

nature, as it is impossible to capture all that it is known about the real world in a finite structure (GÓMEZ-PÉREZ; JURISTO; PAZOS, 1995). Seen this, we highlight that AACOnto it is not complete, and it was developed considering the *Core Vocabulary*, and the 4 reference ontologies as input. AACOnto can be evolved using the other classes available in the reference ontologies, adding new instances in the class already used, as well as new semantic database relationships (e.g., the FrameNet frame structure).

Concerning the ontology evaluation, the metrics of OntoQA are useful to describe the ontology, however, to properly evaluate the ontology it is desirable to build an application that uses the ontology. This type of evaluation is underway in the master's dissertation (PEREIRA, 2020) that proposes a semantic grammar using colorful semantics based on AACOnto. Even so, it is important to evaluate these contributions with AAC users.

With these results, we answer the RQ-3 (*“How can AAC vocabulary be computationally organized to allow new functionalities in a high-tech AAC system?”*) showing that ontologies are a feasible way to computationally organize the AAC vocabulary because it make use of semantic knowledge to allow the inference and automatic reasoning about ontological elements. Our ontology is useful because AAC system developers can use it to develop AAC systems with more complex functionalities such as predictive semantic grammar (GRUZITIS; PAIKENS; BARZDINS, 2012) for the construction of meaningful sentences, and the COMPANSION system (DEMASCO; MCCOY, 1992) to expand the telegraphic language into the full and grammatically corrected sentence.

## 5.4 COMPARATIVE EVALUATION

This section analyzes the contribution of this work regarding the four criteria presented in Section 3.2 – remember that each criterion may be fully satisfied (●), partially satisfied (◐) or dissatisfied (○) – in comparison with the related works of Chapter 3, and summarize this evaluation on Table 31:

- *Controlled vocabulary* – in this work, we proposed the *Core Vocabulary*, which is formed by a controlled vocabulary chosen based on evidences of a statistical analysis of different core word lists over children's utterances (cf. Section 5.1.3). This criterion was fully satisfied by this work. Regarding the related works, the work of Nikolova and Cook (2010) (cf. Section 3.3.1) presents a controlled vocabulary based on two resources; the work of Hernández et al. (2014) (cf. Section 3.3.2) uses as controlled vocabulary the result of the analysis of a set of sources; and the work of Martínez-Santiago et al. (2015) (cf. Section 3.3.3) presents a controlled vocabulary with a set of basic words commonly used by preschoolers and children with complex communication needs. However, none of the related works present evidence to proper base their choice. This criterion was partially satisfied by all the related works;

- *Vocabulary categorization* – in this work, we also proposed a set of categories to organize the our controlled vocabulary (cf. Table 23 on Section 5.2.1). This set of categories was chosen after a 2-round Delphi method with expert evaluation. This criterion was fully satisfied by this work. Regarding the related works, the work of Nikolova and Cook (2010) organizes the initial collection of words in the Lingraphica’s hierarchy; the work of Hernández et al. (2014) present a few evidences that may have a hierarchy to organize the concepts; and the work of Martínez-Santiago et al. (2015) organizes the controlled vocabulary into 35 categories. However, none of the related works present evidence to proper base their choice. This criterion was partially satisfied by all the related works;
- *Vocabulary organization* – in this work, we proposed a category hierarchy for low-tech and high-tech AAC systems, and also proposed an ontology (AACOnto) to computational organizes the *Core Vocabulary* to allow complex functionalities in high-tech AAC systems. This criterion was fully satisfied by this work. Regarding the related works, the work of Nikolova and Cook (2010) uses a “semantic network” to organize the vocabulary; the work of Hernández et al. (2014) and the work of Martínez-Santiago et al. (2015) propose an ontology to organize the vocabulary. This criterion was fully satisfied by all the related works;
- *Semantic enrichment* – in this work, we use the semantic relations of WordNet, preserving the full taxonomy of the concepts of our controlled vocabulary. This criterion was fully satisfied by this work. Regarding the related works, the work of Nikolova and Cook (2010) uses WordNet and the Evocation data; the work of Hernández et al. (2014) uses semantics to enrich the ontology with relations among classes; and the work of Martínez-Santiago et al. (2015) uses the FrameNet. This criterion was fully satisfied by all the related works.

Table 31 – Comparative Evaluation with Related Works.

Related Work	Controlled Vocabulary	Vocabulary Categorization	Vocabulary Organization	Semantic Enrichment
Visual Vocabulary for Aphasia (ViVA)	●	●	●	●
User-centric Recommendation Model	●	●	●	●
Simple Upper Ontology (SUPO)	●	●	●	●
<b>Core Vocabulary and AACOnto</b>	●	●	●	●



From Table 31 we can see that, despite the results being close, only this work gives the necessary importance to the basis of AAC communication, that is, vocabulary selection and organization.

## 5.5 USAGE GUIDELINES: HOW CAN OTHERS MAKE USE OF THIS WORK

The results obtained by this work are aimed at three main audiences: mediators, AAC system developers, and AAC users. The first two are directly benefited, while the third is indirectly benefited.

The *Core Vocabulary* (cf. Appendix A) may be used by mediators for defining the vocabulary for both, low and high-tech AAC systems. This may be used as a baseline for defining the AAC vocabulary of a child or a goal to be achieved by children in the early communication stage. In the first case, mediators may use the full version of the *Core Vocabulary* or only a part of it; whereas in the second case, they may start with a small set of words and categories and adding new ones as the child learns. We highlight that the *Core Vocabulary* is not a finished set and new words and categories must be added considering the AAC user preferences and needs. Although the *Core Vocabulary* is not a finished set, with this, mediators may take advantage of the proposed instances (i.e., words) and subjects (i.e., categories) to select the vocabulary. For example, the child may not have interest in “*Doe*” (a core word in our vocabulary) and it may be disregarded in the vocabulary selection. Conversely, since “*Animals*” (a category in our vocabulary) is an important concept for children, the communication partner may consider the inclusion of more interesting animals for the child (e.g., “*Elephant*” or “*Giraffe*”). Apart from words and categories, the proposed category hierarchy (cf. Section 4.10, Figure 44) bring facilities because it makes it easier the information retrieval when the children vocabulary grows, and also improve the abstraction learning (CAMPOS; GOMES, 2007).

By using our *Core Vocabulary* children can: (i) have access to a range of *content words*, *structure words*, and *basic concept words* that allow flexibility in communication; (ii) produce a complete sentence using the instances of the “*Small Words*” category achieving a higher level of language (MARVIN; BEUKELMAN; BILYEU, 1994); (iii) have access to “*Letters*” to write within their AAC system as soon as possible; and (iv) express themselves according to different communicative functions (MC SHANE, 1980). Considering the five major communicative function of McShane (1980) (cf. Table 1), with the words “*No*” and “*Not*” available, the child can use the *Personal* major function for refusal and protesting by combining these words with nouns and verbs. By the same token, with the words of *Questions and Answers* category the user can use the *Conversation* major function. In summary, with our Core Vocabulary it is possible to express:

1. Regulation – using instances of *Social Expressions/Greetings* and *People* (e.g., “*Look*”, “*Hey*”, “*Help me*”, “*Mommy*”, “*Dad*”);

2. Statement – using instances of *Foods and Drinks*, *People*, and *Descriptors* (e.g., “Apple”, “Grandma”, “Big”, “Fast”);
3. Exchange – using instances of *Social Expressions/Greetings*, *Action/Verbs* and *Pronouns* (e.g., “Here”, “Take”, “Your”, “Thank you”, “Okay”);
4. Personal – using instances of *Action/Verbs*, *Question and Answers*, and *Small Words* (e.g., “Put down”, “Leave”, “Eat now”, “No”, “Not want”, “Stop”);
5. Conversation – using instances of *Question and Answers* (e.g., “What?”, “Where?”, “Why?”, “Yeah”, “Sure”).

The AACOnto is the ontology for computational represent our *Core Vocabulary*. AAC system developers may use AACOnto to develop systems with more complex functionalities thanks to the use of ontologies features. Ontologies can “work and reason” with concepts and relationships in ways that are close to the way humans reason about linked concepts. Ontologies relationships provide an easy navigation as users move from one concept to another in the ontology structure. By using ontologies is possible to know the exactly meaning of a word (e.g., “fish” is a noun or a verb?) and make inferences over concepts. Moreover, since AACOnto brings information over the WordNet taxonomy it makes possible to perform similarity calculus that can be used for WSD and pictogram suggestion, for example. This way, the use of AACOnto allow developers to implement more complex functionalities on high-tech AAC systems, such as: predictive semantic grammar, construction of meaningful sentences, COMPANSION system (DEMASCO; MCCOY, 1992), and teaching of new concepts based on previously learned concepts.

## 5.6 CHAPTER FINAL CONSIDERATIONS

This chapter presented the results of this work, which were based on the methods presented in Chapter 4. The results were presented to respond each RQ: Section 5.1 responds RQ-1, Section 5.2 responds RQ-2, and Section 5.3 responds RQ-3. In addition, Section 5.4 shows a comparative evaluation of this work with the related works, and Section 5.5 brings suggestion showing how others can make use of the results of this work. Next chapter presents the conclusion of this work.

## 6 CONCLUSIONS

This chapter presents the conclusions of this work. Section 6.1 presents the final considerations of the main topics of this work. Section 6.2 presents the contributions of this work, including the articles published or under review. Section 6.3 presents the limitations of this work and some indications for future works.

### 6.1 FINAL CONSIDERATIONS

This work presents a solution for the tasks of selection and organization of the vocabulary used by children with complex communication needs on Augmentative and Alternative Communication (AAC) systems. Our solution – that is, the *Core Vocabulary* – were generated based on systematical analyses of words and categories. First, we performed statistical analyses of both, previously proposed core word lists, and new lists systematically generated, over children’s utterances to select the list with better recall. Then, we performed empirical and semantic analyses of category lists with specialists to select a set of useful categories that is sufficient to organize the list with better recall. The *Core Vocabulary* consists of 614 word senses, that covers up to 70% of children’s utterances, categorized into 36 categories. We also proposed an ontology (AACOnto) as a computational representation for our vocabulary. Even with the scope of this work limited to English (cf. Section 1.3), our solution is based on concepts that are not limited to English and can be translated into any language.

With these results, we answered the three Research Questions (RQ) of this work (cf. Section 1.2). The RQ-1 (*“Is it possible to generate a new core word list with better recall than the existing core word lists?”*) was answered on Section 5.1 by statistical analyses, showing that the union of the existing core word lists (cf. Section 4.1.1) generates a list with better recall. The RQ-2 (*“Are the already proposed categories useful and sufficient to categorize all the core words of the best list of RQ-1?”*) was answered on Section 5.2 by empirical and semantic evidences, showing that the already proposed categories (cf. Section 4.2.1) are not sufficient to categorize all the words of the best list of RQ-1. Finally, the RQ-3 (*“How can AAC vocabulary be computationally organized to allow new functionalities in a high-tech AAC system?”*) was answered on Section 5.3, showing that ontologies are a feasible way to computationally organize the AAC vocabulary because it make use of semantic knowledge to allow the inference and automatic reasoning about ontological elements.

## 6.2 CONTRIBUTIONS

This work presents two main contributions:

1. *Core Vocabulary* – composed by a set of words that cover more than 70% of children’s communication, which are categorized into a set of categories considered essential by experts (cf. Sections 5.1 and 5.2). The *Core Vocabulary* may be used by mediators for defining the vocabulary for low and high-tech AAC systems;
2. AACOnto – the ontology made for computational represent our *Core Vocabulary* (cf. Section 5.3). This ontology may be used by AAC system developers to develop systems with more complex functionalities such as construction of meaningful sentences.

Besides these main contributions, we also advance the state of the art because we:

- Make a Systematic Literature Review (Systematic Literature Review (SLR)) of core word list, and perform statistical analyses of these lists, which allow the conscious selection of a core word list based on its coverage of children’s utterances;
- Generate a new core word list with better coverage than the existing lists;
- Make a SLR of AAC vocabulary categorization, and perform qualitative analyses of these categories, which allow the conscious selection of categories based on its relevance for children communication;
- Generate a set of categories cor organizing the content of the new core word list in a flat or hierarchical way.

We highlight that some of these contributions have already been published or are under review process. We also have other publications that contribute to this thesis. All these publications are listed below:

- **Franco, N.**, Lima, A., Lima, T., Silva, E., Lima, R., & Fidalgo, R. (2017). *A Recall Analysis of Core Word Lists over Children’s Utterances for Augmentative and Alternative Communication*. In International Symposium on Computer-Based Medical Systems (CBMS) – This paper presents a set of statistical analysis of 9 core word lists over a corpus with children’s utterances to point the list with better coverage of that corpus.;
- **Franco, N.**, Silva, E., Lima, R., & Fidalgo, R. (2018). *Towards a Reference Architecture for Augmentative and Alternative Communication Systems*. In Brazilian Symposium on Computers in Education (Vol. 29, No. 1, p. 1073) – This paper discusses the major requirements that a robust AAC system must have and presented the first effort towards a Reference Architecture (RA), which is conceptual

and technology-independent, and can be seen as a baseline for instantiating and evaluate concrete AAC software architectures;

- **Franco, N.**, Silva, E., Lima, R., & Fidalgo, R. *Core Vocabulary for Children's Augmentative and Alternative Communication: Words and Categories Selection Using Statistical, and Qualitative Analyses*. Submitted to the ACM Transactions on Accessible Computing (TACCESS) on October 2019 – this paper presents an extension of the first statistical analyses by rising up the amount of core word lists and the corpus samples; moreover, we also show a qualitative analysis to select a set of essential categories. The main contribution of this paper is to propose the Core Vocabulary that will be presented in this thesis.

### 6.3 LIMITATIONS AND FUTURE WORKS

Despite the relevant results, this work has some limitations that can be addressed in future works. We will list them below:

- The target audience – although the analyzes were aimed at children, we believe that the results achieved by this work can also be used in other age groups. For this, future works can analyze the vocabulary effectiveness in these other age groups;
- The data preprocessing – we analyzed the data considering all possible Part of Speech (POS). This way, the recall analysis cannot be accurate with the exact meaning of a word (e.g., “fish” was considered as a noun and a verb; “I” was considered as a pronoun and a letter; and “can” was considered as a verb and an object). This may have introduced some unusual word senses, which is a limitation of our method. Besides, we disregard the contractions and consider only the lemma form of each word. In a future work, we plan to analyze the data specifying the POS, as well as consider the recall of the flexed forms of words (i.e., “ing”, possessive “’s”, “-s” plurals) in each Mean Length of Utterance (MLU) and Age range;
- The selected corpora – the corpora used in this analysis is collected long ago and do not consider utterances of children with complex communication needs, only typically developing children. In a future work, we plan to collect the data generated by users of AAC devices to confirm the presented contribution;
- The categories evaluation – we have consulted specialists to evaluate the categories. However, the selection of these specialists may have introduce some bias, for example, they may have thought of the needs of specific people and evaluated the categories concerning these specific needs. In a future work, we plan to consult more specialists to reduce this bias;

- The semantic enrichment – in AACOnto we used only the WordNet database to enrich our ontology. This can bias the ontology to only one database, so add other relations from different semantic databases may increase the quality of the ontology. An underway master’s dissertation (PEREIRA, 2020) is adding knowledge of the frame semantics theory on AACOnto which allows relations between words from different POS;
- The ontology evaluation – the metrics of OntoQA are useful to describe the ontology, the implementation of an AAC application based on the ontology (PEREIRA, 2020) shows its utility. However, it is important to evaluate these contributions with AAC users;
- The data granularity – the data was analyzed word by word, regardless of the meaning of a sequence of words. As a consequence, our analysis disregard the presence of phrasal verbs (e.g., “*move on*” and “*come back*”) and compound words (e.g., “*hot dog*” and “*dinner table*”), for example.

## REFERENCES

- ALLEN, J. *Natural language understanding*. [S.l.]: Pearson, 1995.
- ASHA. *Augmentative and Alternative Communication*. ASHA, 2019. Available at: <<https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942773>>.
- ASHA. *Definition of Communication and Appropriate Targets*. ASHA, 2019. Available at: <<https://www.asha.org/NJC/Definition-of-Communication-and-Appropriate-Targets/>>.
- ASHA. *Dysarthria*. ASHA, 2019. Available at: <<https://www.asha.org/public/speech/disorders/dysarthria/>>.
- AYRE, C.; SCALLY, A. J. Critical values for lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, Taylor & Francis, v. 47, n. 1, p. 79–86, 2014.
- AZEVEDO, R. R. de; FREITAS, F.; ALMEIDA, S. C. de; ALMEIDA, M. J. S.; FILHO, E. C. de B. C.; VERAS, W. C. Coresec: an ontology of security applied to the business process of management. In: *Proceedings of the 2008 Euro American Conference on Telematics and Information Systems*. [S.l.: s.n.], 2008. p. 1–7.
- BAADER, F.; HORROCKS, I.; SATTTLER, U. *Description logics*. [S.l.]: Foundations of Artificial Intelligence 3, 2008. 135–179 p.
- BAKER, C. F.; FELLBAUM, C. Wordnet and framenet as complementary resources for annotation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Third Linguistic Annotation Workshop*. [S.l.], 2009. p. 125–129.
- BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The berkeley framenet project. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. [S.l.], 1998. p. 86–90.
- BAKER COLLIN F. E FILLMORE, C. J. e. L. J. B. The Berkeley FrameNet Project. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. [S.l.]: Association for Computational Linguistics, 1998. p. 86–90.
- BANAJEE, M.; DICARLO, C.; STRICKLIN, S. B. Core Vocabulary Determination for Toddlers. *Augmentative and Alternative Communication*, v. 19, n. 2, p. 67–73, 2003. ISSN 0743-4618. Available at: <<http://informahealthcare.com/doi/abs/10.1080/0743461031000112034>>.
- BECHHOFFER, S.; HARMELEN, F. V.; HENDLER, J.; HORROCKS, I.; MCGUINNESS, D. L.; PATEL-SCHNEIDER, P. F.; STEIN, L. A. et al. Owl web ontology language reference. *W3C recommendation*, v. 10, n. 02, 2004.
- BERSCH, R. Introdução à tecnologia assistiva: tecnologia e educação. *Porto Alegre*, 2013.

- BEUKELMAN, D.; JONES, R.; ROWAN, M. Frequency of word usage by nondisabled peers in integrated preschool classrooms. *Augmentative and Alternative Communication*, v. 5, n. 4, p. 243–248, 1989. ISSN 0743-4618. Available at: <<http://informahealthcare.com/doi/abs/10.1080/07434618912331275296>>.
- BOENISCH, J.; SOTO, G. The Oral Core Vocabulary of Typically Developing English-Speaking School-Aged Children: Implications for AAC Practice. *Augmentative and alternative communication (Baltimore, Md. : 1985)*, v. 31, n. 1, p. 77–84, 2015. ISSN 1477-3848. Available at: <<http://www.ncbi.nlm.nih.gov/pubmed/25685883>>.
- BOLDERSON, S.; DOSANJH, C.; MILLIGAN, C.; PRING, T.; CHIAT, S. Colourful semantics: A clinical investigation. *Child Language Teaching and Therapy*, Sage Publications Sage UK: London, England, v. 27, n. 3, p. 344–353, 2011.
- BORST, W. N. Construction of engineering ontologies for knowledge sharing and reuse. 1999.
- BOYD-GRABER, J.; FELLBAUM, C.; OSHERSON, D.; SCHAPIRE, R. Adding dense, weighted connections to wordnet. In: CITESEER. *Proceedings of the third international WordNet conference*. [S.l.], 2006. p. 29–36.
- BRACKEN, B. A.; CRAWFORD, E. Basic Concepts in Early Childhood Educational Standards : A 50-State Review. p. 421–430, 2010.
- BRANK, J.; GROBELNIK, M.; MLADENIC, D. A survey of ontology evaluation techniques. In: CITESEER LJUBLJANA, SLOVENIA. *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*. [S.l.], 2005. p. 166–170.
- BRAZAS, M.; BOURKE, R. *An Introduction to the Core Vocabulary Exchange System<sup>(TM)</sup>*. 2016. 106 p. Available at: <<https://www.cvessolution.com/wp-content/uploads/2016/08/Part-1-An-Introduction-to-the-Core-Vocabulary-Exchange-SystemTM-1.pdf>>.
- BROWN, R. *A first language: The early stages*. [S.l.]: Harvard U. Press, 1973.
- BRYAN, A. Colourful semantics: Thematic role therapy. *Language disorders in children and adults: Psycholinguistic approaches to therapy*, Wiley Online Library, p. 143–161, 2003.
- CAMPOS, M. L. d. A.; GOMES, H. E. Taxonomia e classificação: a categorização como princípio. *Encontro Nacional de Pesquisa em Ciência da Informação (VIII ENANCIB)*, 2007.
- CASELLI, M. C.; BATES, E.; CASADIO, P.; FENSON, J.; FENSON, L.; SANDERL, L.; WEIR, J. A cross-linguistic study of early lexical development. *Cognitive Development*, Elsevier, v. 10, n. 2, p. 159–199, 1995.
- CDCP. *What is Cerebral Palsy?* Centers for Disease Control and Prevention, 2019. Available at: <<https://www.cdc.gov/ncbddd/cp/facts.html>>.
- Center for Literacy and Disability Studies. *DLM Core Vocabulary*. 2018. Available at: <<http://www.med.unc.edu/ahs/clds/resources/core-vocabulary>>.
- CHEN, X.; WU, B. Assessing student learning through keyword density analysis of online class messages. *AMCIS 2004 Proceedings*, p. 362, 2004.



- CHOWDHURY, G. G. Natural language processing. *Annual review of information science and technology*, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.
- CIMIANO, P.; BUITELAAR, P.; MCCRAE, J.; SINTEK, M. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 9, n. 1, p. 29–51, 2011.
- CIMIANO, P.; MCCRAE, J.; BUITELAAR, P. Lexicon model for ontologies: Community report. *W3C Ontology-Lexicon Community Group*, 2016.
- CLENDON, S. A.; ERICKSON, K. A. The vocabulary of beginning writers: Implications for children with complex communication needs. *Augmentative and Alternative communication*, Taylor & Francis, v. 24, n. 4, p. 281–293, 2008.
- CLENDON, S. A.; STURM, J. M.; CALI, K. S. Vocabulary Use Across Genres: Implications for Students with Complex Communication Needs. *Language, Speech, and Hearing Services in Schools*, v. 44, n. 1, p. 61–72, 2013. ISSN 1558-9129.
- CLOUTIER, R.; MULLER, G.; VERMA, D.; NILCHIANI, R.; HOLE, E.; BONE, M. The concept of reference architectures. *Systems Engineering*, Wiley Online Library, v. 13, n. 1, p. 14–27, 2010.
- COOK, A. M.; POLGAR, J. M. *Assistive Technologies-E-Book: Principles and Practice*. [S.l.]: Elsevier Health Sciences, 2014.
- CORAZZON, R. *Ontology: Its Role in Modern Philosophy*. 2019. Available at: <<https://www.ontology.co/>>.
- CRESS, C. J.; MARVIN, C. A. Common questions about aac services in early intervention. *Augmentative and Alternative Communication*, Taylor & Francis, v. 19, n. 4, p. 254–272, 2003.
- CRESTANI, C.-A. M.; CLENDON, S. a.; HEMSLEY, B. Words needed for sharing a story: implications for vocabulary selection in augmentative and alternative communication. *Journal of intellectual & developmental disability*, v. 35, n. 4, p. 268–278, 2010. ISSN 1366-8250.
- CRISTANI, M.; CUEL, R. A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS)*, IGI Global, v. 1, n. 2, p. 49–69, 2005.
- CRUTTENDEN, A. Assimilation in child language and elsewhere. *Journal of Child Language*, Cambridge University Press, v. 5, n. 2, p. 373–378, 1978.
- DECKERS, S. R.; Van Zaalen, Y.; Van Balkom, H.; VERHOEVEN, L. Core vocabulary of young children with Down syndrome. *AAC: Augmentative and Alternative Communication*, Informa UK Limited, trading as Taylor & Francis Group, v. 33, n. 2, p. 77–86, 2017. ISSN 14773848. Available at: <<http://dx.doi.org/28431488http://dx.doi.org/10.1080/07434618.2017.1293730>>.
- DEMARIS, A.; SMITH, A. B. Relationships among measures of longest utterances, mlu, age, and number of utterances in child language samples. *Speech, Language and Hearing*, Taylor & Francis, v. 20, n. 2, p. 84–90, 2017.

- DEMASCO, P. W.; MCCOY, K. F. Generating text from compressed input: An intelligent interface for people with severe motor impairments. *Communications of the ACM*, Citeseer, v. 35, n. 5, p. 68–78, 1992.
- DRAGER, K. D.; LIGHT, J. C.; SPELTZ, J. C.; FALLON, K. A.; JEFFRIES, L. Z. The performance of typically developing 21/2-year-olds on dynamic display aac technologies with different system layouts and language organizations. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 46, n. 2, p. 298–312, 2003.
- FALLON, K. A.; LIGHT, J. C.; PAIGE, T. K. Enhancing vocabulary selection for preschoolers who require augmentative and alternative communication (aac). *American Journal of Speech-Language Pathology*, ASHA, v. 10, n. 1, p. 81–94, 2001.
- FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A.; JURISTO, N. Methontology: from ontological art towards ontological engineering. American Association for Artificial Intelligence, 1997.
- FITZGERALD, E. *Straight language for the deaf: a system of instruction for deaf children*. [S.l.]: Volta Bureau, 1949.
- FRANCO, N.; FIDALGO, R.; CUNHA, P.; NÓBREGA, O.; RAMOS, A. Customizing usability heuristics for augmentative and alternative communication systems. In: *Proceedings of the Euro American Conference on Telematics and Information Systems*. New York, NY, USA: ACM, 2018. (EATIS '18), p. 24:1–24:7. ISBN 978-1-4503-6572-7. Available at: <<http://doi.acm.org/10.1145/3293614.3293634>>.
- FRANCO, N.; LIMA, T.; LIMA, A.; SILVA, E.; LIMA, R.; CAVALCANTE, T.; FIDALGO, R. aboard: Uma plataforma para educação inclusiva a partir de comunicação aumentativa e/ou alternativa. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 977.
- FRANCO, N.; SILVA, E.; LIMA, R.; FIDALGO, R. Towards a reference architecture for augmentative and alternative communication systems. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1073.
- FRIED-OKEN, M.; MORE, L. An initial vocabulary for nonspeaking preschool children based on developmental and environmental language sources. *Augmentative and Alternative Communication*, v. 8, n. 1, p. 41–56, 1992. ISSN 0743-4618. Available at: <<http://informahealthcare.com/doi/abs/10.1080/07434619212331276033>>.
- GÓMEZ-PÉREZ, A.; FERNÁNDEZ, M.; VICENTE, A. d. Towards a method to conceptualize domain ontologies. European Coordinating Committee for Artificial Intelligence (ECCAI), 1996.
- GÓMEZ-PÉREZ, A.; JURISTO, N.; PAZOS, J. Evaluation and assessment of knowledge sharing technology. *Towards very large knowledge bases*, p. 289–296, 1995.
- GÓMEZ-PÉREZ, A.; ROJAS-AMAYA, M. D. Ontological reengineering for reuse. In: SPRINGER. *International Conference on Knowledge Engineering and Knowledge Management*. [S.l.], 1999. p. 139–156.

- GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, Elsevier, v. 5, n. 2, p. 199–220, 1993.
- GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, Elsevier, v. 43, n. 5-6, p. 907–928, 1995.
- GRUZITIS, N.; PAIKENS, P.; BARZDINS, G. Framenet resource grammar library for gf. In: SPRINGER. *International Workshop on Controlled Natural Language*. [S.l.], 2012. p. 121–137.
- HALLORAN, J.; EMERSON, M. Lamp: Language acquisition through motor planning. *Wooster (OH): Prentke Romich Company*, 2006.
- HENRY, A. *Belfast English and Standard English: Dialect variation and parameter setting*. [S.l.]: Oxford University Press on Demand, 1995.
- HERNÁNDEZ, S. S.; MANCILLA, D.; MEDINA, J. M.; IREGUI, M. User-centric recommendation model for aac based on multi-criteria planning. In: *International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies and Services*. [S.l.: s.n.], 2014.
- JOHANSSON, R.; NUGUES, P. Using wordnet to extend framenet coverage. In: *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*. [S.l.: s.n.], 2007. p. 27–30.
- JURAFSKY, D. *Speech & language processing*. [S.l.]: Pearson Education India, 2000.
- KAGOHARA, D. M.; MEER, L. van der; RAMDOSS, S.; O'REILLY, M. F.; LANCIONI, G. E.; DAVIS, T. N.; RISPOLI, M.; LANG, R.; MARSCHIK, P. B.; SUTHERLAND, D. et al. Using ipods® and ipads® in teaching programs for individuals with developmental disabilities: A systematic review. *Research in developmental disabilities*, Elsevier, v. 34, n. 1, p. 147–156, 2013.
- KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Engineering*, v. 2, p. 1051, 2007. ISSN 00010782.
- KOVACS, T.; HILL, K. Language samples from children who use speech-generating devices: Making sense of small samples and utterance length. *American journal of speech-language pathology*, ASHA, v. 26, n. 3, p. 939–950, 2017.
- KRACKOW, E.; GORDON, P. Are lions and tigers substitutes or associates? evidence against slot filler accounts of children's early categorization. *Child development*, Wiley Online Library, v. 69, n. 2, p. 347–354, 1998.
- LAMY, J.-B. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*, Elsevier, v. 80, p. 11–28, 2017.
- LAWSHE, C. H. A quantitative approach to content validity 1. *Personnel psychology*, Wiley Online Library, v. 28, n. 4, p. 563–575, 1975.

- LEONARD, L. B.; CAMARATA, S.; ROWAN, L. E.; CHAPMAN, K. The communicative functions of lexical usage by language impaired children. *Applied psycholinguistics*, Cambridge University Press, v. 3, n. 2, p. 109–125, 1982.
- LIGHT, J. “let’s go star fishing”: Reflections on the contexts of language learning for children who use aided aac. *Augmentative and Alternative Communication*, Taylor & Francis, v. 13, n. 3, p. 158–171, 1997.
- LIGHT, J. C.; DRAGER, K. D. Improving the design of augmentative and alternative technologies for young children. *Assistive Technology*, Taylor & Francis, v. 14, n. 1, p. 17–32, 2002.
- LIMA, T.; SILVA, E.; LIMA, A.; FRANCO, N.; FIDALGO, R. aboard: uma plataforma computacional na nuvem para comunicação alternativa e educação inclusiva. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, n. 1, p. 102.
- LOPER, E.; BIRD, S. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.
- LOWE, J. B. A frame-semantic approach to semantic annotation. *Tagging Text with Lexical Semantics: Why, What, and How?*, 1997.
- LUCARIELLO, J.; KYRATZIS, A.; NELSON, K. Taxonomic knowledge: What kind and when? *Child development*, Wiley Online Library, v. 63, n. 4, p. 978–998, 1992.
- LUND, S. K.; LIGHT, J. Long-term outcomes for individuals who use augmentative and alternative communication: Part III—contributing factors. *Augmentative and Alternative Communication*, Taylor & Francis, v. 23, n. 4, p. 323–335, 2007.
- MACWHINNEY, B. *The CHILDES project: Tools for analyzing talk*. Third edit. [S.l.]: Lawrence Erlbaum Associates, Inc, 2000.
- MANCILLA, D.; SASTOQUE, S.; IREGUI, M. Towards ontology personalization to enrich social conversations on aac systems. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *10th International Symposium on Medical Information Processing and Analysis*. [S.l.], 2015. v. 9287, p. 92870Z.
- MANCILLA, D.; SASTOQUE, S.; MENDOZA, J.; IREGUI, M. Conceptualizing a daily activities ontology for an augmentative and alternative communication system. In: *5th Latin American Conference on Networked and Electronic Media (LACNEM-2013)*. [S.l.: s.n.], 2013.
- MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J. R.; BETHARD, S.; MCCLOSKEY, D. The stanford corenlp natural language processing toolkit. In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. [S.l.: s.n.], 2014. p. 55–60.
- MARKMAN, E. M.; COX, B.; MACHIDA, S. The standard object-sorting task as a measure of conceptual organization. *Developmental Psychology*, American Psychological Association, v. 17, n. 1, p. 115, 1981.

- MARTÍNEZ-SANTIAGO, F.; CUMBRERAS, M. Á. G.; RÁEZ, A. M.; GALIANO, M. C. D. Pictogrammar: an aac device based on a semantic grammar. In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. [S.l.: s.n.], 2016. p. 142–150.
- MARTÍNEZ-SANTIAGO, F.; DÍAZ-GALIANO, M. C.; UREÑA-LÓPEZ, L. A.; MITKOV, R. A semantic grammar for beginning communicators. *Knowledge-Based Systems*, Elsevier, v. 86, p. 158–172, 2015.
- MARTÍNEZ-SANTIAGO, F.; MONTEJO-RÁEZ, A.; GARCÍA-CUMBRERAS, M. Á. Pictogram tablet: A speech generating device focused on language learning. *Interacting with Computers*, Oxford University Press, v. 30, n. 2, p. 116–132, 2018.
- MARVIN, C.; BEUKELMAN, D.; BILYEU, D. Vocabulary-use patterns in preschool children: Effects of context and time sampling. *Augmentative and Alternative Communication*, v. 10, n. 4, p. 224–236, 1994. ISSN 0743-4618. Available at: <<http://informahealthcare.com/doi/abs/10.1080/07434619412331276930>>.
- MCCARTHY, J. H.; SCHWARZ, I.; ASHWORTH, M.; MCCARTHY, J. H.; SCHWARZ, I.; ASHWORTH, M. The availability and accessibility of basic concept vocabulary in AAC software: a preliminary study. *Augmentative and alternative communication (Baltimore, Md. : 1985)*, Informa UK Limited, trading as Taylor & Francis Group, v. 33, n. 3, p. 131–138, 2017. ISSN 14773848. Available at: <<http://dx.doi.org/28597688https://doi.org/10.1080/07434618.2017.1332685>>.
- MCCRAE, J. P.; BOSQUE-GIL, J.; GRACIA, J.; BUITELAAR, P.; CIMIANO, P. The ontolx-lemon model: development and applications. In: *Proceedings of eLex 2017 conference*. [S.l.: s.n.], 2017. p. 19–21.
- MCGINNIS, J.; BEUKELMAN, D. Vocabulary requirements for writing activities for the academically mainstreamed student with disabilities. *Augmentative and Alternative Communication*, v. 5, n. 3, p. 183–191, 1989. ISSN 0743-4618. Available at: <<http://informahealthcare.com/doi/pdf/10.1080/07434618912331275186>>.
- MCGUINNESS, D. L.; HARMELEN, F. V. et al. Owl web ontology language overview. *W3C recommendation*, v. 10, n. 10, p. 2004, 2004.
- MCNAUGHTON, D.; LIGHT, J. *The iPad and mobile technology revolution: Benefits and challenges for individuals who require augmentative and alternative communication*. [S.l.]: Taylor & Francis, 2013.
- MCSHANE, J. *Learning to talk*. [S.l.]: Cambridge University Press, 1980.
- MIKA PETER E AKKERMANS, H. Towards a new synthesis of ontology technology and knowledge management. *The Knowledge Engineering Review*, v. 19, n. 04, p. 317, 2005. ISSN 0269-8889. Available at: <[http://www.journals.cambridge.org/abstract\\_S0269888905000305](http://www.journals.cambridge.org/abstract_S0269888905000305)>.
- MILES, A.; BECHHOFFER, S. Skos simple knowledge organization system reference. *W3C recommendation*, v. 18, p. W3C, 2009.
- MILLER, G. A. WordNet: a lexical database for English. *Communications of the ACM*, v. 38, n. 11, p. 39–41, 1995. ISSN 00010782.

- MIRENDA, P. Toward functional augmentative and alternative communication for students with autism. *Language, speech, and hearing services in schools*, ASHA, 2003.
- MORENO, D.; NARVAEZ, C.; SASTOQUE, S.; GARNICA, G. Computational model based on language development theories for languages learning and training: Vocabulary module. In: IEEE. *2016 XLII Latin American Computing Conference (CLEI)*. [S.l.], 2016. p. 1–12.
- MORGAN, D. L. Focus groups. *Annual review of sociology*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 22, n. 1, p. 129–152, 1996.
- NAVIGLI ROBERTO E PONZETTO, S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, Elsevier B.V., v. 193, p. 217–250, 2012. ISSN 00043702. Available at: <<http://dx.doi.org/10.1016/j.artint.2012.07.001>>.
- NEWELL, A.; LANGER, S.; HICKEY, M. The rôle of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, Cambridge University Press, v. 4, n. 1, p. 1–16, 1998.
- NIKOLOVA, S.; COOK, P. Building semantic networks to improve word finding in assistive communication tools. In: ACM. *Proceedings of the 1st international workshop on Semantic models for adaptive interactive systems*. [S.l.], 2010. p. 19–23.
- NIKOLOVA, S. S.; BOYD-GRABER, J.; FELLBAUM, C.; COOK, P. Better vocabularies for assistive communication aids: Connecting terms using semantic networks and untrained. In: *ACM Conference on Computers and Accessibility*. [S.l.: s.n.], 2009.
- NIMH. *Autism Spectrum Disorder*. U.S. Department of Health and Human Services, 2019. Available at: <<https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml>>.
- PALAO, S. *ARASAAC: Aragonese Portal of Augmentative and Alternative Communication*. 2019. Available at: <<http://www.arasaac.org/>>.
- PANTER, J. E.; BRACKEN, B. A. Validity of the bracken school readiness assessment for predicting first grade readiness. *Psychology in the Schools*, Wiley Online Library, v. 46, n. 5, p. 397–409, 2009.
- PAUL, R. *Language disorders from infancy through adolescence: Assessment & intervention*. [S.l.]: Elsevier Health Sciences, 2007.
- PAURA, A. C.; DELIBERATO, D. Estudo de vocábulos para avaliação de crianças com deficiência sem linguagem oral. *Revista Brasileira de Educação Especial*, Associação Brasileira de Pesquisadores em Educação Especial-ABPEE, p. 37–52, 2014.
- PEREIRA, J. A. Uma gramática semântica baseada em colourful semantics para comunicação aumentativa e alternativa (unpublished master's dissertation). Universidade Federal de Pernambuco, 2020.

- PORTER, G.; CAFIERO, J. M. Pragmatic organization dynamic display (podd) communication books: A promising practice for individuals with autism spectrum disorders. *Perspectives on Augmentative and Alternative Communication*, ASHA, v. 18, n. 4, p. 121–129, 2009.
- PYTHON CORE TEAM. *Python: A dynamic, open source programming language*. [S.l.], 2019. Available at: <<https://www.python.org/>>.
- RAAD, J.; CRUZ, C. A survey on ontology evaluation methods. In: . [S.l.: s.n.], 2015.
- RANTA, A. Grammatical Framework: Programming with Multilingual Grammars. 2011. Available at: <<http://www.grammaticalframework.org/gf-book/>>.
- RHYNER, P. M. P.; BRACKEN, B. A. Concurrent validity of the bracken basic concept scale with language and intelligence measures. *Journal of communication disorders*, Elsevier, v. 21, n. 6, p. 479–489, 1988.
- ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, Emerald Group Publishing Limited, v. 60, n. 5, p. 503–520, 2004.
- ROBILLARD, M.; MAYER-CRITTENDEN, C.; MINOR-CORRIVEAU, M.; BÉLANGER, R. Monolingual and bilingual children with and without primary language impairment: core vocabulary comparison. *Augmentative and alternative communication*, v. 30, n. 3, p. 267–78, 2014. ISSN 1477-3848. Available at: <<http://www.ncbi.nlm.nih.gov/pubmed/24921189>>.
- ROUSSEY, C.; PINET, F.; KANG, M. A.; CORCHO, O. An introduction to ontologies and ontology engineering. In: *Ontologies in Urban development projects*. [S.l.]: Springer, 2011. p. 9–38.
- SARDINHA, T. Corpus Linguistics: history and problematization. *DELTA: Documentação de Estudos em Lingüística ...*, v. 16, n. n.2, p. 323–367, 2000. ISSN 0102-4450. Available at: <[http://www.scielo.br/scielo.php?pid=S0102-44502000000200005&script=sci\\_arttext&tlng=es](http://www.scielo.br/scielo.php?pid=S0102-44502000000200005&script=sci_arttext&tlng=es)>.
- SCHULER, K. K. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.
- SEVILLA, J.; SAMPER, J. J.; HERRERA, G.; FERNÁNDEZ, M. Smart-asd, model and ontology definition: a technology recommendation system for people with autism and/or intellectual disabilities. *International Journal of Metadata, Semantics and Ontologies*, Inderscience Publishers (IEL), v. 13, n. 2, p. 166–178, 2018.
- SEVILLA, J.; ZAPATER, J.; HERRERA, G. An ontology-based recommendation system for people with autism and technology apps: Ontology application for helping persons with autism. In: ACM. *Proceedings of the Euro American Conference on Telematics and Information Systems*. [S.l.], 2018. p. 47.
- SHIVABASAPPA, P.; PEÑA, E. D.; BEDORE, L. M. Core vocabulary in the narratives of bilingual children with and without language impairment. *International Journal of Speech-Language Pathology*, Taylor & Francis, 2017. Available at: <<https://doi.org/10.1080/17549507.2017.1374462>>.

SHULL, F.; SINGER, J.; SJØBERG, D. I. *Guide to advanced empirical software engineering*. [S.l.]: Springer, 2007.

STANFORD CENTER FOR BIOMEDICAL INFORMATICS RESEARCH. *A free, open-source ontology editor and framework for building intelligent systems*. 2019. Available at: <<https://protege.stanford.edu/>>.

STILL, K.; REHFELDT, R. A.; WHELAN, R.; MAY, R.; DYMOND, S. Facilitating requesting skills using high-tech augmentative and alternative communication devices with individuals with autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders*, Elsevier, v. 8, n. 9, p. 1184–1199, 2014.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. *Data & knowledge engineering*, Elsevier, v. 25, n. 1-2, p. 161–197, 1998.

SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; FERNÁNDEZ-LÓPEZ, M. The neon methodology for ontology engineering. In: *Ontology engineering in a networked world*. [S.l.]: Springer, 2012. p. 9–34.

SURE, Y.; STAAB, S.; STUDER, R. On-to-knowledge methodology (otkm). In: *Handbook on ontologies*. [S.l.]: Springer, 2004. p. 117–132.

SUTTON, A.; GALLAGHER, T.; MORFORD, J.; SHAHNAZ, N. Relative clause sentence production using augmentative and alternative communication systems. *Applied Psycholinguistics*, Universidade Federal de Pernambuco, v. 21, n. 4, p. 473–486, 2000. ISSN 01427164. Available at: <<http://www.scopus.com/inward/record.url?eid=2-s2.0-0034550255&partnerID=40&md5=6cda5d364828509ce6886d48857c2c29>>.

TARTIR, S.; ARPINAR, I. B. Ontology evaluation and ranking using ontoqa. In: IEEE. *International conference on semantic computing (ICSC 2007)*. [S.l.], 2007. p. 185–192.

THANAKI, J. *Python natural language processing*. [S.l.]: Packt Publishing Ltd, 2017.

THEAKSTON, A. L.; LIEVEN, E. V.; PINE, J. M.; ROWLAND, C. F. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, Cambridge University Press, v. 28, n. 1, p. 127–152, 2001.

TODMAN, J.; ALM, N.; HIGGINBOTHAM, J.; FILE, P. Whole utterance approaches in aac. *Augmentative and Alternative Communication*, Taylor & Francis, v. 24, n. 3, p. 235–254, 2008.

TOMASELLO, M. *Constructing a language: A usage-based theory of language acquisition*. [S.l.]: Harvard university press, 2009.

TOMMERDAHL, J.; KILPATRICK, C. D. The reliability of morphological analyses in language samples. *Language Testing*, Sage Publications Sage UK: London, England, v. 31, n. 1, p. 3–18, 2014.

TREMBATH, D.; BALANDIN, S.; TOGHER, L. Vocabulary selection for Australian children who use augmentative and alternative communication. *Journal of intellectual & developmental disability*, v. 32, n. 4, p. 291–301, 2007. ISSN 1366-8250.

University of Oxford. *British National Corpus*. 1995. [Http://www.natcorp.ox.ac.uk/](http://www.natcorp.ox.ac.uk/).



- USCHOLD, M.; GRUNINGER, M. Ontologies: Principles, methods and applications. *The knowledge engineering review*, Cambridge University Press, v. 11, n. 2, p. 93–136, 1996.
- W3C. *OWL 2 Web Ontology Language Primer (Second Edition)*. 2012. Available at: <<https://www.w3.org/TR/owl2-primer/>>.
- W3C. *RDF Schema 1.1*. 2014. Available at: <[https://www.w3.org/TR/rdf-schema/#ch\\_type](https://www.w3.org/TR/rdf-schema/#ch_type)>.
- WALKER, V. L.; LYON, K. J.; LOMAN, S. L.; SENNOTT, S. A systematic review of functional communication training (fct) interventions involving augmentative and alternative communication in school settings. *Augmentative and Alternative Communication*, Taylor & Francis, v. 34, n. 2, p. 118–129, 2018.
- WELLS, C. G. *Learning through interaction: The study of language development*. Cambridge, UK: Cambridge University Press, 1981.
- WIENS, V.; LOHMANN, S.; AUER, S. Webvowl editor: Device-independent visual ontology modeling. In: *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks*. [S.l.]: CEUR-WS.org, 2018. (CEUR Workshop Proceedings, v. 2180).
- WILSON, F. R.; PAN, W.; SCHUMSKY, D. A. Recalculation of the critical values for lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, Taylor & Francis, v. 45, n. 3, p. 197–210, 2012.
- WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: CITESEER. *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. [S.l.], 2014. p. 38.
- WOOD, C.; APPLEGET, A.; HART, S. Core Vocabulary in Written Personal Narratives of School-Age Children. *Augmentative and alternative communication (Baltimore, Md. : 1985)*, v. 32, n. 3, 2016. Available at: <<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5247772/>>.
- World Health Organization. *WHO global disability action plan 2014-2021. Better health for all people with Disability*. [S.l.], 2015. 32 p. Available at: <[www.who.int](http://www.who.int)>.
- World Wide Web Consortium. W3C, 2006. Available at: <<https://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/>>.
- YORKSTON, K.; DOWDEN, P.; HONSINGER, M.; MARRINER, N.; SMITH, K. A comparison of standard and user vocabulary lists. *Augmentative and Alternative Communication*, v. 4, n. 4, p. 189–210, 1988. ISSN 0743-4618. Available at: <<http://informahealthcare.com/doi/abs/10.1080/07434618812331274807>>.
- ZARTHA, J. W.; MONTES, J. M.; VARGAS, E. E.; PALACIO, J. C.; HERNÁNDEZ, R.; HOYOS, J. L. Methods and techniques in studies related to the delphi method, innovation strategy, and innovation management models. *International Journal of Applied Engineering Research*, v. 13, n. 11, p. 9207–9214, 2018.

## APPENDIX A – CORE VOCABULARY

Table 32 shows the Core Vocabulary produced in this paper, that is, a Core Word list categorized into Useful Categories. Each word is associated with its position in the frequency order (i.e., 1 - *I* and 2 - *That* are the two most frequent words of this vocabulary). The first 100 most frequent words are highlighted in bold. That is, with these 100 words it is possible to recall more than 60% of children utterances (cf. Figures 40, 42, and 43). The words between brackets and in italics encompasses a set of specific words (e.g., personal names, and names of cities and countries) that should to be customized for each child. The words with an asterisk designate words that can also be used as categories (cf. Figure 44). Synonyms are grouped, appear separated by a slash, and are ordered according to its frequency position. The group of synonyms is placed in the position of its first appearing. Expressions between parenthesis are made by a combination of two or more core words.

Concerning the categories and its words, we highlight that:

- In general, the categories are not limited to the presented words. Depends on the application or on the audience, these words must be replaced and new words must be added. For example, the whole alphabet may be added in Letters, and other colors may be added in Colors. Another example is the *Social Expressions/Greetings* category that includes the expressions “*Good Morning*” and “*Thank you*”, and can also categorize other expressions that were not considered (e.g., “*Good Night*”);
- New words and concepts can be built by combining core words. The *Position and Equipment* category, which had no words categorized, may be filled with a range of expressions using core words (e.g., “*Sit in the chair*” and “*Turn me in the bed*”) according to the user needs. The *Activities and Routines* category may be filled with expressions that represent user-specific activities and routines (e.g., “*Take the bus*” and “*Listen to rock*”);
- The *Action/Verbs* category, that does not considered the negative verb forms nor phrasal verbs, includes the word “*Not*” to allow the construction of the negative verb forms (e.g., “*Not like*”) and allows the combination of words to compose phrasal verbs (e.g., “*Get over*” and “*Look for*”);

Table 32 – Core Vocabulary.

<b>ACTION / VERBS</b>				
<b>(n= 111)</b>				
<b>9 - Not</b>	<b>95 - Take/Bring</b>	231 - Could	341 - Paint	470 - Listen
<b>16 - Do</b>	104 - Find/Happen	235 - Fell	343 - Must	484 - Hug
<b>19 - Want/Need/ Wanna</b>	119 - Stuck	245 - Cause	351 - Would	497 - Laugh
<b>17 - Go</b>	117 - Sleep	237 - Sing	347 - Wear	491 - Dig
<b>23 - Can</b>	121 - Say/Tell	248 - Leave	356 - Break/ Crack	501 - Ask
<b>31 - Be</b>	123 - Fall	250 - Move	360 - Show	502 - Fix
<b>33 - Get/Catch</b>	126 - Open	252 - Cry	361 - Step	515 - Hate
<b>34 - Look</b>	128 - Drink	253 - Wash	363 - Kick	520 - Suck
<b>35 - Have</b>	129 - Will	254 - Drive	364 - Fly	526 - Record
<b>36 - Put</b>	136 - Think/ new- line Suppose/Guess	279 - Cut	378 - Close	528 - May
<b>37 - Up</b>	147 - Shall	291 - Wait	384 - Press	529 - Punch
<b>40 - Like/Wish/ Care</b>	161 - Give	295 - Bite	389 - Touch	540 - Turn
<b>41 - Down</b>	164 - Fish	301 - Buy	396 - Call	552 - Win
<b>45 - Out</b>	166 - Read	303 - Stay	412 - Talk	558 - Shoot
<b>48 - Come</b>	173 - Run	304 - Try	418 - Hear	560 - Smash
<b>61 - Know</b>	192 - Jump	307 - Clean	422 - Use	565 - Forget
<b>63 - See/Watch</b>	197 - Help	308 - Ride	432 - Write	577 - Die
<b>69 - Make</b>	200 - Lose/Miss	309 - Tickle	442 - Play	582 - Bet
<b>73 - Need</b>	203 - Stand	319 - Walk	450 - Cook	590 - Store
<b>80 - Sit</b>	208 - Push	322 - Shut	453 - Start	
<b>81 - Eat</b>	222 - Hold/Keep	324 - Hide	455 - Taste	
<b>83 - Please</b>	224 - Hurt	329 - Hit	456 - Remember	
<b>90 - Let</b>	226 - Throw	339 - Pick	461 - Should	
<b>ACTIVITIES AND ROUTINES</b>				
<b>(n= 18)</b>				
117 - Sleep	217 - Dinner	355 - Finish	425 - Party	551 - Telling
157 - Draw	273 - Playing	356 - Break	484 - Hug	596 - Trip
188 - Work	297 - School	388 - Breakfast	499 - Fun	
213 - Bath	309 - Tickle	417 - Lunch	531 - Homework	
<b>BODY</b>				
<b>(n= 12)</b>				
<b>68 - Back</b>	183 - Hair	251 - Bottom	316 - Face	
160 - Hand	187 - Head	264 - Leg	581 - Blood	
167 - Eye	207 - Foot	310 - Finger	589 - Voice	

Table 32 continued from previous page

<b>CLOTHES AND ACCESSORIES</b>				
<b>(n= 10)</b>				
135 - Shoe	196 - Watch	293 - Sock	383 - Clock	457 - Shirt
142 - Hat	249 - Coat	334 - Pant	413 - Pocket	538 - Jacket
<b>DESCRIPTORS</b>				
<b>(n= 60)</b>				
<b>37 - Up</b>	200 - Lost	318 - Happy	403 - High	496 - Used
<b>41 - Down</b>	201 - Better	335 - Hungry	404 - Shape*	498 - Sure
<b>55 - Off</b>	210 - Ready	338 - Light	405 - Favorite	502 - Fixed
<b>59 - Just</b>	227 - Cold	342 - Wrong	437 - Wanted	512 - Kind
102 - Right	241 - Color*	343 - Must	439 - Real	513 - Thirsty
119 - Stuck	259 - Yum/ Yummy/Tasty	348 - Fast	459 - Mean	521 - Opened
145 - Nice	276 - Funny	362 - Bad	465 - Quick	527 - Easy
154 - Broke	277 - Long	368 - Together	466 - Best	534 - Beautiful
155 - Hot	278 - Finished	373 - Different	468 - Pretty	540 - Turned
158 - Naughty	285 - New	377 - Mess/Messy	480 - Great	554 - Mad
169 - Good/Well	307 - Clean	379 - Sick	482 - Cool	591 - Peace
175 - Top	315 - Hard	382 - Old	485 - Fine	597 - Damn
<b>FOODS AND DRINKS</b>				
<b>(n= 27)</b>				
114 - Orange	152 - Dog (Hot Dog)	234 - Food	435 - Meat	564 - Candy
115 - Egg	156 - Milk	257 - Cheese	449 - Butter	587 - Cookie
128 - Drink	164 - Fish	332 - Roll	463 - Corn	598 - Gum
134 - Green	176 - Apple	387 - Sauce	471 - Cracker	
143 - Water	198 - Chip	402 - Bean	529 - Punch	
144 - Cake	199 - Juice	411 - Potato	555 - Peanut	
<b>LOCATION / PLACES</b>				
<b>(n= 23)</b>				
<b>6 - There</b>	178 - Well	294 - Hole	355 - Finish	532 - Flat
<b>30 - Here</b>	188 - Work	297 - School	361 - Step	583 - <i>[Specific Locations]</i>
<b>89 - House</b>	194 - Home	299 - Field	375 - Lift	590 - Store
101 - Way	242 - Outside	305 - Room*	415 - Pool	
137 - Stop	262 - Garage	354 - Farm	474 - Square	
<b>NUMBERS</b>				
<b>(n= 19)</b>				
<b>11 - One</b>	148 - Five	290 - Nine	428 - Twenty	519 - Fifteen
<b>43 - Two</b>	218 - Six	331 - Number	446 - Twelve	543 - Sixty
<b>87 - Three</b>	246 - Seven	345 - Ten	452 - Eleven	576 - Hundred
130 - Four	272 - Eight	427 - Thirteen	490 - Thirty	

Table 32 continued from previous page

<b>PEOPLE</b>				
<b>(n= 23)</b>				
<b>24 - Mummy/ Mum/Mama/ Mom/Mommy</b>	172 - Girl	321 - Doctor	462 - Policeman	550 - Family
<b>52 - Baby</b>	225 - Dad/Father/ Papa	369 - Sister	493 - Mr	593 - Aunt
<b>88 - Man</b>	263 - Friend	398 - Grandpa	494 - Brother	602 - Guy
111 - <i>[Personal Names]</i>	269 - People	434 - Miss	510 - Teacher	
151 - Boy	287 - Grandma	450 - Cook	514 - Kid	
<b>PRONOUNS</b>				
<b>(n= 29)</b>				
<b>1 - I</b>	<b>46 - We</b>	<b>92 - Him</b>	302 - Our	458 - Anything
<b>3 - It</b>	<b>50 - They</b>	<b>94 - Her</b>	370 - Somebody	467 - Everything
<b>12 - You/Ya</b>	<b>53 - Them</b>	<b>100 - His</b>	372 - Their	472 - Someone
<b>22 - My</b>	<b>58 - Another</b>	103 - She	407 - Myself	511 - Ours
<b>28 - He</b>	<b>70 - Your</b>	205 - Something	421 - Everybody	539 - Everyone
<b>29 - Me</b>	<b>74 - Mine</b>	255 - Yours	451 - Somewhere	
<b>QUANTITY</b>				
<b>(n= 18)</b>				
<b>44 - Some</b>	<b>67 - Little</b>	220 - Any	390 - Much	460 - Half
<b>47 - More</b>	<b>72 - Bit</b>	267 - Nothing	394 - Around	495 - Whole
<b>54 - All</b>	159 - Very	284 - Lot	416 - Many	
<b>59 - Just/Only</b>	219 - Piece	331 - Number*	459 - Mean	
<b>QUESTIONS AND ANSWERS</b>				
<b>(n= 11)</b>				
<b>4 - No/Nah</b>	<b>32 - Where</b>	146 - When	365 - Which	
<b>5 - Yeah/Yes/ Yep</b>	<b>66 - Why</b>	174 - How	498 - Sure	
<b>25 - What</b>	138 - Who	221 - Maybe		
<b>SELF-/SOCIAL-AWARENESS, EMOTION/FEELINGS, AND BEHAVIOR</b>				
<b>(n= 19)</b>				
<b>76 - Away</b>	276 - Funny	350 - Tired	498 - Sure	522 - Care
155 - Hot	295 - Bite	379 - Sick	512 - Kind	554 - Mad
224 - Hurt	318 - Happy	392 - Love	513 - Thirsty	591 - Peace
252 - Cry	335 - Hungry	469 - Sad	515 - Hate	
<b>SIZE / COMPARISON</b>				
<b>(n= 13)</b>				
<b>40 - Like</b>	201 - Better	338 - Light	373 - Different	549 - Higher
<b>51 - Big</b>	277 - Long	340 - Same	403 - High	
<b>67 - Little</b>	328 - Bigger	353 - Small	548 - Than	

Table 32 continued from previous page

<b>SOCIAL EXPRESSIONS / GREETINGS</b>				
(n= 18)				
<b>34 - Look</b>	137 - Stop	232 - Sorry	485 - Fine	553 - Whatever
<b>78 - Hello/Hi/Hey</b>	170 - Bye	312 - Really	498 - Sure	597 - Damn
<b>83 - Please</b>	197 - Help	419 - Morning (Good Morning)	509 - Wish	
122 - Okay	230 - Alright	448 - Excuse (Excuse Me)	544 - Thank (Thank You)	
<b>TIME / SEQUENCE</b>				
(n= 33)				
<b>39 - Now</b>	266 - Today	327 - Still	424 - Yesterday	545 - Year
<b>64 - Again</b>	268 - Birthday	330 - Never	438 - Already	547 - Until
113 - Then	280 - Day	358 - After	476 - Before	572 - Hour
162 - Turn	286 - First	376 - Night	478 - While	578 - Easter
209 - Time	288 - Next	381 - Christmas	481 - Sometimes	600 - Second
216 - Minute	296 - Later	409 - Always	525 - Ever	
256 - Yet	313 - Last	419 - Morning	528 - May	
<b>TOYS</b>				
(n= 11)				
<b>38 - Car</b>	118 - Toy	371 - Game	556 - Puppet	
<b>56 - Train</b>	149 - Ball	374 - Swing	573 - Wand	
<b>89 - House</b>	177 - Doll	500 - Goal		
<b>ANIMALS</b>				
(n= 18)				
<b>97 - Doe</b>	150 - Duck	184 - Bear	364 - Fly	517 - Pet
109 - Cow	152 - Dog	211 - Animal	385 - Frog	530 - Bug
120 - Cat	164 - Fish	244 - Spider	433 - Turtle	
133 - Horse	181 - Rabbit/ Bunny	323 - Bird	492 - Ant	
<b>SMALL WORDS</b>				
(n= 17)				
<b>7 - A</b>	<b>21 - To</b>	<b>84 - Too/Also/ As_Well</b>	191 - An	346 - Else
<b>8 - The</b>	<b>57 - For</b>	<b>93 - Because</b>	228 - If	
<b>9 - Not</b>	<b>75 - With</b>	127 - But	282 - By	
<b>13 - And</b>	<b>77 - Of</b>	163 - So	317 - Or	
<b>CLIMATE</b>				
(n= 1)				
123 - Fall				
<b>COLORS</b>				
(n= 9)				
105 - Blue	114 - Orange	134 - Green	238 - White	344 - Brown
106 - Red	131 - Yellow	223 - Black	241 - Color	

Table 32 continued from previous page

DEMONSTRATIVES				
(n= 6)				
2 - That	85 - These	124 - Those		
20 - This	110 - Other	410 - Both		
DIRECTION / POSITION				
(n= 25)				
14 - On	76 - Away	162 - Turn	275 - Into	378 - Close
15 - In	82 - Over	175 - Top	281 - Through	394 - Around
21 - To	86 - At	195 - Under	325 - Left	429 - End
45 - Out	101 - Way	251 - Bottom	352 - Side	507 - Middle
68 - Back	102 - Right	265 - From	367 - Inside	600 - Second
ENTERTAINMENT				
(n= 18)				
65 - Play	273 - Playing	431 - Triangle	533 - Music	575 - Jingle
118 - Toy*	341 - Paint	440 - Jumping	541 - Beat	596 - Trip
173 - Run	360 - Show	464 - Rock	551 - Telling	
258 - Story	426 - Song	504 - Pas	566 - Spiderman	
FURNITURE				
(n= 4)				
116 - Bed	171 - Chair	292 - Table	393 - Seat	
HYGIENE				
(n= 2)				
213 - Bath	261 - Poo			
HOUSEHOLD ITEMS				
(n= 2)				
189 - Cup	516 - TV			
LETTERS				
(n= 4)				
1 - I	7 - A	140 - U	357 - M	
NATURE				
(n= 7)				
190 - Tree	366 - Sun	464 - Rock	537 - Leaf	
239 - Fire	391 - Being	486 - Hill		
OBJECTS				
(n= 13)				
23 - Can	204 - Money	349 - Bin	454 - Bell	592 - Recorder
125 - Door	274 - Saw	383 - Clock	508 - Microphone	
141 - Box	320 - Fence	445 - Stuff	546 - Sign	

Table 32 continued from previous page

<b>ROOMS</b>				
<b>(n= 3)</b>				
305 - Room	400 - Kitchen	559 - Bathroom		
<b>SCHOOL MATERIALS</b>				
<b>(n= 5)</b>				
132 - Book	306 - Tape	580 - Glue/Gum		
298 - Paper	503 - Scissors			
<b>SHAPES</b>				
<b>(n= 5)</b>				
300 - Line	332 - Roll	431 - Triangle	474 - Square	567 - Point
<b>TEXTURE / MATERIAL</b>				
<b>(n= 3)</b>				
143 - Water	406 - Sand	532 - Flat		
<b>TRANSPORT</b>				
<b>(n= 5)</b>				
<b>38 - Car</b>	56 - Train	214 - Boat	215 - Bus	308 - Ride
<b>POSITIONS AND EQUIPMENT</b>				
<b>(n= 0)</b>				
<b>MISCELLANEOUS</b>				
<b>(n= 12)</b>				
139 - Thing	311 - Name	430 - Actually	488 - Word	
206 - About	312 - Really	445 - Stuff	523 - Even	
265 - From	395 - Stick	473 - Probably	568 - God	