Pós-Graduação em Ciência da Computação

JAYR ALENCAR PEREIRA

Uma Gramática Semântica Baseada em *Colourful Semantics* para Comunicação Aumentativa e Alternativa

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

Recife
2020

JAYR ALENCAR PEREIRA

Uma Gramática Semântica Baseada em *Colourful Semantics* para Comunicação Aumentativa e Alternativa

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciências da Computação.

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

Orientador: Prof. Dr. Robson do Nascimento Fidalgo

Recife
2020

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

P436g Pereira, Jayr Alencar
Uma gramática semântica baseada em *colorful semantics* para comunicação aumentativa e alternativa / Jayr Alencar Pereira. – 2020.
98 f.: il., fig., tab.

Orientador: Robson do Nascimento Fidalgo.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2020.
Inclui referências.

1. Processamento de sinais. 2. Reconhecimento de padrões. I. Fidalgo, Robson do Nascimento (orientador). II. Título.

006.4 CDD (23. ed.) UFPE - CCEN 2020 - 108

Jayr Alencar Pereira

“Uma Gramática Semântica Baseada em Colourful Semantics para Comunicação Aumentativa e Alternativa”

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

Aprovado em: 04 de março de 2020.

BANCA EXAMINADORA

_____

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

_____

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

_____

Prof. Dr. Robson do Nascimento Fidalgo
Centro de Informática/UFPE
(**Orientador)**

Dedico este trabalho à minha esposa Luana, aos meus filhos, e aos meus pais e irmãos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, aquele com quem, segundo J.R.R Tolkien, somos co-criadores, por tudo que ele permitiu que acontecesse na minha vida durante esses dois anos.

À minha família. Minha esposa Luana, que está sempre do meu lado em todos os momentos, por todo o apoio e confiança. Minhas filhas Teresa e Catarina, que a cada dia dão um novo sentido à minha vida, me ensinando coisas que eu jamais sonhei. Minha mãe, que me apoiou e me incentivou desde o princípio. Meu pai, com quem eu aprendi que a vida é dura, mas é bela. Meus irmãos, pela paciência e apoio nesse período. E também aos meus sogros e cunhados, por todo o suporte e ajuda.

Aos meus professores, do jardim da infância até aqui. Especialmente ao meu orientador Robson Fidalgo, que confiou em mim na realização desse trabalho e que me faz crer que eu posso ir mais longe. Aos professores do Programa de Pós-graduação em Ciências da Computação do Centro de Informática da UFPE, que me fizeram ver o mundo de uma nova perspectiva.

Aos meus amigos. Especialmente aqueles que fiz durante esse período, e aqueles cujos laços de amizade se fortificaram. Francisco (Champs), Júnior, Elias, Natália, Edson, Magno, Renatto, Nitiele, Diego, Dayse, Pedro Esaú, etc.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro dado através da bolsa de pós-graduação.

À empresa Techmetria, nas pessoas dos seus diretores Roberto Fagundes e Nelise Cunha, por todo o apoio e acolhimento que me deram quando cheguei no Recife.

À Igreja Católica Apostólica Romana, especialmente a Comunidade Aliança de Misericórdia e a Renovação Carismática Católica. Que me ensinaram o amor pelo estudo e busca do conhecimento. Onde eu aprendi os sentidos da irmandade, caridade, esperança e fé.

Finalmente, a todos aqueles que de alguma forma contribuíram com este trabalho.

RESUMO

Os sistemas de Comunicação Aumentativa e Alternativa (CAA) são ferramentas importantes para sujeitos com distúrbios de comunicação (e.g., pessoas com Transtornos do Espectro Autista, Síndrome de Down e paralisia cerebral). Esses sistemas possibilitam a comunicação por meio de frases telegráficas (e.g., sem preposições, artigos e conjugação de verbos) que são construídas a partir da seleção de figuras com legendas. Para garantir uma boa comunicação, os sistemas de CAA devem dar suporte à construção de frases com sintaxe e semântica adequadas. Isto é, esses sistemas devem prover facilidades para a construção de frases telegráficas cujas palavras estejam dispostas corretamente e expressem ideias coerentes (e.g., Eu comer bolo ontem). Diversos estudos propõem maneiras de apoiar a construção de frases compreensíveis em sistemas de CAA. Contudo, esses estudos não fornecem apoio visual e semântico para este fim. Neste trabalho, um sistema de pistas visuais (i.e., cores) e semânticas (i.e., perguntas como Quem? Fazendo? O quê? Para quem? Como? Onde? Quando?) é usado como base para a construção de uma Gramática Semântica (GS) para dar suporte à construção de frases telegráficas sintática e semanticamente bem formadas. Inicialmente, realiza-se uma Revisão Sistemática da Literatura, a fim de levantar trabalhos que propõem e utilizam GS no contexto de CAA. Essa revisão fornece um panorama de como essas bases são usadas e quais as metodologias empregadas na sua construção. Em seguida, constrói-se a GS por meio de um método automático com base nos seguintes materiais: (i) frases extraídas de um conjunto de documentos de texto; (ii) uma gramática simples com pistas visuais e semânticas, chamada *Colourful Semantics*; e (iii) um vocabulário controlado para CAA construído com o uso de análises qualitativas e quantitativas. As etapas desse método consistem em: 1) seleção e extração de sentenças a partir dos documentos de texto; 2) análise semântica das sentenças usando técnicas de Processamento de Linguagem Natural (PLN); e 3) remoção das eventuais redundâncias presentes na GS. Por fim, a GS construída é avaliada a fim de medir a sua eficiência com relação à construção de frases com sentido. Essa avaliação é feita por meio da reconstrução de frases telegráficas extraídas de um *corpus* específico para CAA. A métrica utilizada é uma precisão modificada, na qual a frase de referência é comparada palavra a palavra com a frase reconstruída. Como resultados, obtêm-se uma GS com relações de predicado-argumento, que possibilitam a busca de palavras durante a construção de frases, com precisão média de 90% na reconstrução de frases telegráficas. A GS proposta fornece uma estrutura semântica, que pode ser usada por sistemas de CAA como base para o suporte visual e semântico à construção de frases telegráficas com sintaxe e semântica adequadas.

Palavras-chaves: Comunicação Aumentativa e Alternativa. Gramática Semântica. Processamento de Linguagem Natural. Ontologias.

ABSTRACT

Augmentative and Alternative Communication (AAC) systems are essential tools for subjects with communication disorders (e.g., people with Autism Spectrum Disorders, Down's Syndrome, and cerebral palsy). These systems enable communication through telegraphic phrases (e.g., without prepositions, articles, and conjugation of verbs) by selecting figures with captions. CAA systems must support the construction of phrases with appropriate syntax and semantics to ensure proper communication. That is, these systems must provide facilities for the construction of telegraphic phrases whose words are arranged correctly and express coherent ideas (e.g., I eat cake yesterday). Several studies propose ways to support the construction of understandable phrases in CAA systems. However, these studies do not provide visual and semantic support for this purpose. In this work, a system of visual (i.e., colors) and semantic (i.e., questions such as Who? What Doing? What? To Whom? What Like? Where? When?) is used as a basis for the construction of a Semantic Grammar (SG). The propose of the GS is to support the construction of syntactic and semantically well-formed telegraphic phrases. Initially, we carried out a Systematic Literature Review in order to survey works that propose and use GS in the context of CAA. This review provides an overview of how these bases are used and what methodologies are used in their construction. Then, we built the GS by using an automatic method based on three materials: (i) sentences extracted from a set of text documents; (ii) a simple grammar with visual and semantic clues, called Colorful Semantics; and (iii) a controlled vocabulary for CAA. The steps of this method consist of 1) selecting and extracting sentences from text documents; 2) semantic analysis of sentences using Natural Language Processing (NLP) techniques; and 3) removal of any redundancies present in the GS. Finally, we evaluated the constructed GS by measuring its efficiency in the construction of meaningful sentences. This evaluation is done through the reconstruction of telegraphic phrases extracted from a specific corpus for CAA. The metric used is a modified precision, in which the reference phrase is compared word by word with the reconstructed phrase. As a result, a GS with predicate-argument relations is obtained, which makes it possible to search for words during sentence construction, with an average accuracy of 90% in the reconstruction of telegraphic sentences. The proposed GS provides a semantic structure, which can be used by CAA systems as a basis for visual and semantic support for the construction of telegraphic phrases with appropriate syntax and semantics.

Keywords: Augmentative and Alternative Communication. Semantic Grammar. Natural Language Processing. Ontologies.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Ontolex-lemon	*Lexicon Model for Ontologies*
AD	Análise de Dependência
AM	Aprendizagem de Máquina
ARASAAC	Portal Aragonês de Comunicação Aumentativa e Alternativa
AVC	Acidente Vascular Cerebral
biLSTM	*Bidirectional Long Short-Term Memory*
BLEU	*Bilingual Evaluation Understudy*
BNC	British National Corpus
BPMN	*Business Process Model and Notation*
CAA	Comunicação Aumentativa e Alternativa
CHILDES	*Child Language Data Exchange System*
CS	*Colourful Semantics*
DL	*Deep Learning*
DSL	Desambiguação de Sentido Lexical
FE	*Frame Element*
FN	*FrameNet*
GS	Gramática Semântica
IA	Inteligência Artificial
JSON	*JavaScript Object Notation*
LSTM	*Long Short-Term Memory*
PLN	Processamento de Linguagem Natural
POS	*Part Of Speech*
PP	Pergunta de Pesquisa
PreMOn	*Predicate Model for Ontologies*
RAPT	*Renfrew Action Picture Test*
REN	Reconhecimento de Entidades Nomeadas
RPS	Rotulação de Papéis Semânticos
RSL	Revisão Sistemática da Literatura
SKOS	*Simple Knowledge Organization System*
SN	Sintagma Nominal

StArt	*State of the Art through Systematic Review*
SUpO	*Supper Upper Ontology*
SV	Sintagma Verbal
TEA	Transtornos do Espectro Autista
UD	*Universal Dependencies*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta esta dissertação de mestrado, destacando seu contexto, motivação, objetivos e perguntas de pesquisa, bem como a delimitação do escopo e a contribuição esperada. Por fim, é apresentada a estrutura dos capítulos subsequentes.

1.1 CONTEXTO E MOTIVAÇÃO

A Comunicação Aumentativa e Alternativa (CAA) (ASHA, 2019a) é uma tecnologia assistiva, que tem o objetivo de dar suporte ao desenvolvimento das habilidades funcionais de comunicação de indivíduos com dificuldades na fala, por exemplo: pessoas com paralisia cerebral, microcefalia, Transtornos do Espectro Autista (TEA) e sequelas de Acidente Vascular Cerebral (AVC). Segundo Garcia e Mansur (2006), a "comunicação funcional é a habilidade de receber e transmitir mensagens, de modo efetivo e independente de acordo com as exigências do contexto ambiental". O desenvolvimento dessas habilidades promove a independência e a inclusão social de pessoas com problemas de comunicação (ASHA, 2019b), sejam elas adultas ou crianças. Para crianças, sistemas de CAA são uma ferramenta importante para garantir uma educação inclusiva. Nesse contexto, é importante que esses sistemas sejam intuitivos e trabalhem com imagens (e.g., fotos ou pictogramas) que representam um objeto, pessoa, lugar ou conceito. Na Figura 1 apresenta-se um exemplo de frase construída com pictogramas extraídos do Portal Aragonês de Comunicação Aumentativa e Alternativa (ARASAAC) (PALAO, 2019). Geralmente, as frases construídas em sistemas de CAA são telegráficas, semelhantes à ilustrada na Figura 1. Isto é, frases formadas apenas por palavras chaves (i.e., substantivos, verbos, adjetivos e advérbios), sem palavras de ligação (i.e., conjunções, preposições e artigos) e conjugação de verbos. Segundo Franco et al. (2018), sistemas de CAA devem prover pistas que ajudem na construção de frases telegráficas que sejam sintática e semanticamente adequadas. Essas pistas podem ser visuais (e.g., cores, setas, etc.) e/ou semânticas (e.g., perguntas, palavras de exemplo, etc.). Para isso, esses sistemas devem ser providos de bases de conhecimento linguístico, que forneçam informações de como as palavras se relacionam em linguagem natural.

Uma Gramática Semântica (GS) (BURTON, 1976) é uma base de conhecimento na qual as palavras e os conceitos que elas denotam estão conectadas por propriedades léxico-semânticas de hierarquia e de predicado-argumento. As propriedades de hierarquia definem a estrutura semântica da GS a partir das relações de hiperonímia e hiponímia entre os conceitos (e.g., mamífero é hiperônimo de gato, pois é mais amplo, e é hipônimo de animal, pois é menos amplo). Enquanto que as propriedades de predicado-argumento definem a estrutura gramatical da base a partir dos argumentos (e.g., agente, tema, atributo) que

Figura 1 – Exemplo de frase construída com pictogramas

Fonte: do autor

cada palavra com características de predicado (i.e., verbos e substantivos) pode ter. Por exemplo, na frase "João comeu pipoca" o predicado verbal "comeu" tem dois argumentos: "João" como agente e "pipoca" como tema, que formam a estrutura *agente+verbo+tema*. Em uma GS, os argumentos dessa estrutura gramatical são preenchidos pelos elementos da estrutura semântica (i.e., conceitos). Segundo Burton (1976), isso permite que o conhecimento léxico-semântico de hierarquia (e.g., gato é-um mamífero) seja usado no processo de análise e construção de sentenças em linguagem natural, evitando a ambiguidade gramatical que pode existir ao utilizar apenas classes gramaticais (e.g., substantivo).

Diversos estudos propõem o uso de GSs como base para dar suporte à construção de frases compreensíveis em sistemas de CAA (MCCOY; PENNINGTON; BADMAN, 1998; MARTÍNEZ-SANTIAGO et al., 2015; NETZER; ELHADAD, 2006). Contudo, as GSs utilizadas nesses trabalhos não levam em consideração pistas visuais e semânticas, por exemplo: cores para orientar a disposição dos pictogramas na frase, ou perguntas (e.g., quem? fazendo o que? onde? quando?) para guiar a seleção de pictogramas.

1.2 OBJETIVOS E PERGUNTAS DE PESQUISA

O objetivo geral desta pesquisa é propor uma GS para sistemas de CAA, na qual as relações de predicado-argumento são baseadas em um sistema de pistas visuais (cores) e semânticas (perguntas), a fim de dar suporte à construção de frases telegráficas sintática e semanticamente bem formadas. Para alcançar esse objetivo geral, tem-se os seguintes objetivos específicos:

- Levantar o estado da arte sobre o uso de GSs em sistemas de CAA;
- Construir uma GS a partir de um vocabulário controlado CAA.
- Avaliar automaticamente a GS a fim de testar a sua expressividade para construir frases telegráficas com sentido.

Para atingir esses objetivos, este trabalho é orientado pelas seguintes Perguntas de Pesquisa (PP):

- PP-1: Como GSs são usadas em sistemas de CAA?
- PP-2: Como utilizar um sistema de pistas visuais (i.e., cores) e semânticas (i.e., perguntas como Quem?, Como?, Onde?, etc.) para construir uma GS para o domínio de CAA?
- PP-3: Qual é a expressividade da GS proposta para construir frases telegráficas com sentido?

1.3 DELIMITAÇÃO DO ESCOPO

O escopo deste trabalho é delimitado a 3 aspectos:

1. **Idioma** – dada a indisponibilidade de recursos de Processamento de Linguagem Natural (PLN) para a língua portuguesa, este trabalho é desenvolvido para a língua inglesa. Dentre esses recursos se destacam: (i) a ausência de uma base lexical robusta e concisa, como a *WordNet*; e (ii) a ausência de analisadores semânticos e sintáticos de acesso livre e gratuito. A construção desse tipo de recurso está além do escopo deste trabalho.
2. **Público alvo** – crianças que possuem comprometimento de fala, cognitivos e/ou físicos e que não possuam deficiência visual. Em relação ao domínio da linguagem escrita, o público desta proposta pode ser alfabetizado ou não. No caso de uma criança alfabetizada, o comprometimento cognitivo pode comprometer o uso da linguagem escrita e, nesse caso, a CAA é vista como um recurso complementar (i.e, facilitador da comunicação). No caso de uma criança não alfabetizada, a CAA é considerada um recurso alternativo para a comunicação porque é baseado em símbolos pictográficos.
3. **Nível de linguagem** – é considerada a linguagem telegráfica, composta por frases que contam apenas com palavras chaves (i.e., adjetivos, advérbios, verbos e substantivos), sem palavras de ligação (i.e., artigos, preposições e conjunções) e sem conjugação de verbos.

1.4 CONTRIBUIÇÃO

A contribuição deste trabalho de mestrado é a proposta de uma Gramática Semântica para CAA construída a partir de pistas visuais (i.e., cores) e semânticas (i.e., perguntas) e de um vocabulário específico do domínio de CAA para crianças (FRANCO, 2020), que fornece uma base para a construção de frases telegráfica com sintaxe e semântica adequadas.

1.5 ESTRUTURA DO DOCUMENTO

O restante dos capítulos deste trabalho estão estruturados da seguinte maneira:

Capítulo 2 – Fundamentação Teórica: no qual os fundamentos teóricos necessários para apoiar este trabalho são introduzidos.

Capítulo 3 – Trabalhos Relacionados: no qual os trabalhos relacionados a este trabalho são discutidos.

Capítulo 4 – Materiais e Métodos: no qual os materiais e métodos utilizados na construção da Gramática Semântica proposta neste trabalho são apresentados.

Capítulo 5 – Gramática Semântica: no qual a Gramática Semântica proposta é especificada e avaliada.

Por fim, o **Capítulo 6 – Conclusão**: no qual as considerações finais sobre os principais tópicos abordados neste trabalho, bem como as indicações de trabalhos futuros são apresentados.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos que servem de base para esta pesquisa, com as definições necessárias para a compreensão deste trabalho. Na Seção 2.1 são apresentados os aspectos da Semântica de *Frames*, que são a base para a concepção de Gramáticas Semânticas (GS), que, por sua vez, são abordadas na Seção 2.2. Na Seção 2.3 aborda-se o Processamento de Linguagem Natural (PLN), destacando tarefas que podem ser usadas para a extração do conhecimento linguístico que pode ser útil para a construção de GSs. Por fim, na Seção 2.4 aborda-se a Comunicação Aumentativa e Alternativa (CAA), apontando aspectos importantes para o desenvolvimento de sistemas de CAA que dão suporte à construção de frases com sentido.

2.1 SEMÂNTICA DE FRAMES

A Semântica de *Frames* (FILLMORE et al., 2006; FILLMORE, 1977; FILLMORE, 1967) é uma teoria linguística que conecta a semântica linguística ao conhecimento enciclopédico (i.e. conceitos) com a ideia central de que a compreensão do significado de uma palavra requer acesso a todo o conhecimento essencial relacionado a essa palavra. Segundo essa teoria, cada palavra evoca um *frame* semântico que representa o conceito específico ao qual a palavra se refere. Um *frame* semântico, por sua vez, é um conjunto de declarações que fornecem "características, atributos e funções de um conceito, e suas interações características com coisas necessárias ou tipicamente associadas a ele." (ALLAN, 2001). Por exemplo, na frase "Ele comeu o peixe rapidamente", a palavra "comeu" denota o *frame* que carrega o seu significado e seus atributos, que são definidos por papéis semânticos: a pessoa que comeu (i.e., o agente), o objeto que foi comido (i.e., o tema) e a forma como a ação se deu (i.e., maneira). A Seção 2.1.1 apresenta uma visão geral sobre esses papéis, sua constituição e formas de uso.

2.1.1 Papéis Semânticos

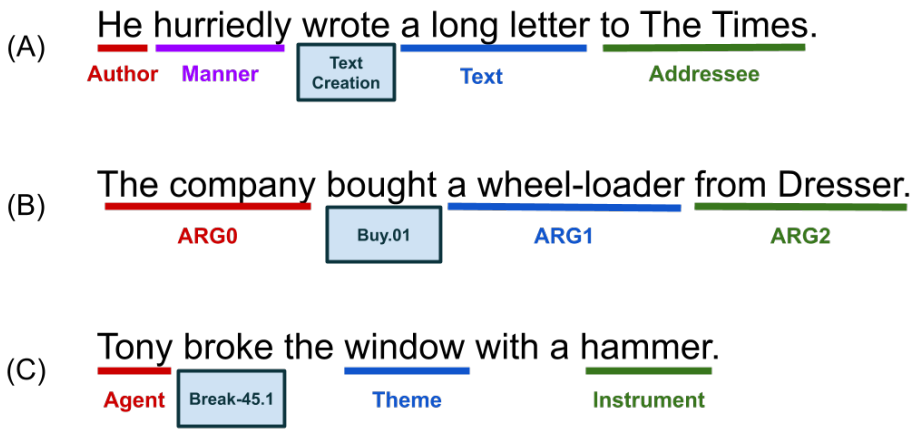
Um papel semântico (a.k.a. papel temático) é um relacionamento que determina o papel desempenhado por um complemento de um dado predicado em uma frase. Como no exemplo citado anteriormente (i.e., "Ele comeu o peixe rapidamente"), no qual o pronome "Ele" desempenha o papel de agente em relação ao predicado "comeu". Esses papéis se aplicam também a predicados nominais. Como na frase "Meu carro é preto", na qual o adjetivo "preto" é um modificador do substantivo "carro". Pelo uso desses papéis é possível construir estruturas semânticas de predicado-argumento a partir da análise de frases em linguagem natural. Essas estruturas são úteis para que um computador possa entender o que está escrito na frase, e para a extração de conhecimento.

Não há um consenso sobre quais são os papéis semânticos existentes, apesar de diversas propostas científicas apresentarem conjuntos que variam na quantidade e na especificidade. Dowty (1991), por exemplo, sugere que existem apenas dois papéis semânticos: *Proto-Agent* e *Proto-Patient*. Já Fillmore (1971), apresenta uma lista com nove papéis: *Agent*, *Experiencer*, *Instrument*, *Object*, *Source*, *Goal*, *Location*, *Time*, e *Path*. Além disso, em alguns sistemas, cada *frame* tem os seus papéis específicos, que são definidos de acordo com a ação denotada pelo *frame* (e.g., *Seller* e *Buyer* para o *frame Buy*) (GILDEA; JURAFSKY, 2002). Segundo Gildea e Jurafsky (2002), os papéis mais abstratos (e.g., *Proto-Agent* e *Proto-Patient*) são geralmente propostos por linguistas, que estão mais preocupados em explicar generalizações entre verbos e o comportamento sintático de seus argumentos. Eles levam em conta questões que envolvem o processo de aquisição da linguagem por um ser humano, por exemplo. Enquanto os papéis mais específicos são frequentemente propostos por cientistas da computação, que estão mais preocupados com os detalhes de como argumentos de verbos específicos são preenchidos. No entanto, as três implementações computacionais da Semântica de *Frames* mais populares atualmente utilizam conjuntos diferentes de papéis semânticos.

Na *FrameNet* (BAKER; FILLMORE; LOWE, 1998), por exemplo, cada *frame* tem papéis semânticos específicos, que são chamados de *Frame Element*s (FEs). Por exemplo, o *frame Ingestion*, que é denotado por verbos como *eat*, *drink*, etc., tem como FEs principais *Ingestor* e *Ingestibles*, que lhes são restritos, ou seja, não são usados por outros *frames*. Além disso, são definidos ainda FEs secundários, como *Manner* e *Location*, que são comuns a vários *frames*. Considerando apenas os FEs principais dos mais de 1200 *frames* anotados na *FrameNet*, a base conta atualmente com mais de 10 mil papéis semânticos diferentes (FRANENET, 2020). Um exemplo de frase rotulada com esses papéis é apresentado na Figura 2-A. O *PropBank* (BABKO-MALAYA, 2005), por sua vez, usa papéis mais generalistas, que consistem em rótulos formados por letras e números para indicar os argumentos de um *frame* semântico. São eles: ARG0, ARG1, ARG2, ARG3, ARG4 e ARG5 (BABKO-MALAYA, 2005). Assim como na *FrameNet*, o *PropBank* diferencia os papéis semânticos dos argumentos adverbiais, que também são representados por rótulos (ARG-MNR, ARG-TMP, ARG-LOC, etc.). Na Figura 2-B é ilustrado um exemplo de frase anotada extraída do *corpus* do *PropBank*. Como um meio termo, a *VerbNet* (SCHULER, 2005), utiliza papéis semânticos que são considerados mais canônicos (BABKO-MALAYA, 2005). Esses papéis não são nem tão especificistas como os da *FrameNet*, nem tão generalistas como os do *PropBank*. São um total de 21 papéis semânticos (cf. Tabela 1), que são semelhantes àqueles nove defendidos inicialmente por Fillmore (1971). Um exemplo de frase rotulada com os papéis semânticos da *VerbNet* é mostrado na Figura 2-B .

A escolha sobre qual conjunto de papéis semânticos usar depende dos objetivos envolvidos em sua aplicação. Para Gildea e Jurafsky (2002), o uso de papéis semânticos mais especificistas, como os da *FrameNet*, adiciona uma semântica maior aos *frames*, pois cada

Figura 2 – Exemplos de frases anotadas com os papéis semânticos de (A) FrameNet, (B) PropBank e (C) VerbNet.



Fonte: do autor

papel passa a carregar um significado único. Ainda segundo o autor, isso possibilita que não apenas verbos sejam tratados como predicados, mas também substantivos. Assim, um *frame* denotado pela palavra *ball* pode contar com papéis semânticos relacionados aos seus atributos, como *size*, por exemplo. Por outro lado, os papéis mais abstratos são também mais universais e mais comuns em analisadores semânticos, como o SLING (RINGGAARD; GUPTA; PEREIRA, 2017), por exemplo. No entanto, o uso de um conjunto não exclui necessariamente o uso de outro, dado que Palmer (2009) realizou um mapeamento entre os papéis semânticos dessas três principais bases[1]. A partir desse mapeamento é possível determinar, por exemplo, qual dos papéis semânticos da *VerbNet* (mais canônicos) cada rótulo do *PropBank* (ARG0, ARG1, etc.) representa para cada *frame*. De todo modo, a escolha do conjunto de papéis semânticos a ser utilizado é essencial, especialmente quando a aplicação utiliza um vocabulário e uma gramática controlada.

[1] <https://verbs.colorado.edu/semlink/>

Tabela 1 – Papéis semânticos da VerbNet

Papel	**Descrição**
Actor	Usado para algumas classes de comunicação quando ambos os argumentos podem ser considerados simétricos (pseudo-agentes).
Agent	Geralmente um sujeito humano ou animado. Usado principalmente como agente volitivo, ou seja, que age por vontade própria.
Asset	Usado para a alternância da soma de dinheiro, presente em classes relacionadas à compra ou aquisição (e.g., buy)
Attribute	Um atributo de Patient/Theme refere-se a uma qualidade de algo que está sendo alterado.
Beneficiary	A entidade que se beneficia de alguma ação.
Cause	Usado principalmente por classes envolvendo verbos psicológicos e verbos que envolvem o corpo.
Destination	Ponto final do movimento, ou direção para a qual o movimento é direcionado.
Source	Ponto inicial do movimento. Geralmente introduzido por uma frase preposicional de origem (geralmente encabeçada por 'de' ou 'fora de').
Location	Destino, origem ou local não especificado, em geral, introduzido por uma frase preposicional de localização ou caminho.
Experiencer	Usado para um participante que está ciente ou está experimentando algo.
Extent	Usado para especificar a faixa ou o grau de alteração, como em "O preço do petróleo subiu (10%)".
Instrument	Usado para objetos (ou forças) que entram em contato com um objeto e causam alguma alteração neles.
Material	Ponto de início da transformação.
Product	Resultado final da transformação.
Patient	Usado para participantes que estão passando por um processo ou que foram afetados de alguma forma. Os verbos que explicitamente (ou implicitamente) expressam mudanças de estado têm Patient como seu objeto direto usual.
Predicate	Usado para verbos com um complemento predicativo.
Recipient	O alvo da transferência. Usado por algumas classes de verbos de mudança de posse, verbos de comunicação e verbos que envolvem o corpo.
Stimulus	Usado por verbos de percepção, para eventos ou objetos que provocam alguma resposta de um experimentador. Esse papel geralmente não impõe restrições.
Theme	Usado para participantes em um local ou passando por uma mudança de local.
Time	Usado para expressar tempo.
Topic	Tópico de verbos de comunicação para lidar com Theme/Topic da conversa ou transferência de mensagem.

2.2 GRAMÁTICAS SEMÂNTICAS

Uma GS é uma base de conhecimento computacionalmente interpretável, que combina a estrutura gramatical de predicado-argumento da Semântica de *Frames* com o conhecimento enciclopédico (i.e., conceitos) de um vocabulário. Assim, os dois componentes básicos de uma GS são: 1) a gramática, que é definida pelo conjunto de papéis semânticos escolhido para a base; e 2) a semântica léxica de hierarquia que está presente na taxonomia dos conceitos denotados pelas palavras de um vocabulário específico.

Uma gramática é um conjunto de regras que determinam o uso correto de um idioma.

Essas regras definem como as unidades básicas (i.e., palavras) devem se unir para constituírem unidades maiores (i.e., frases). Segundo a Teoria Gerativa de Chomsky (CHOMSKY, 1986; CHOMSKY et al., 1988; CHOMSKY, 1993; CHOMSKY, 2014a; CHOMSKY, 2014b), as regras fundamentais de uma gramática são aquelas que dizem respeito à sintaxe. Elas são divididas em dois grupos: base e transformação. As regras de base são aquelas que definem as relações gramaticais entre os elementos que constituem as estruturas profundas das frases. Esses elementos são representados por símbolos categoriais, que podem também ser chamados de categorias não terminais, ou de *slots*. Por exemplo, na frase "O menino caiu" existem dois símbolos categoriais: Sintagma Nominal (SN) e Sintagma Verbal (SV). As relações gramaticais que podem acontecer entre esses símbolos são: sujeito, objeto direto, objeto indireto, complemento adverbial, etc. No caso do exemplo, existe uma relação de sujeito entre o sintagma verbal e o nominal. Essa estrutura profunda (SN+SV) é transformada em uma estrutura de superfície pelas regras de transformação, que utilizam o léxico (i.e., vocabulário) do idioma para buscar por informações que caracterizam cada palavra (e.g., classe gramatical, gênero, número, padrões morfológicos). Essas informações são usadas para fazer as transformações necessárias em cada sintagma (e.g., conjugação de verbo) para transformá-lo em linguagem natural.

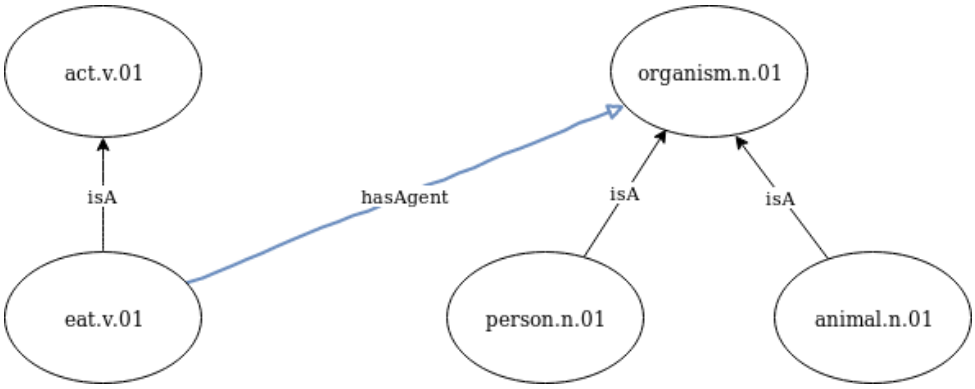
Em uma GS as categorias não terminais de uma gramática são formadas com base em conceitos semânticos (e.g. Pessoa, Ação, Animal, Atributo), e não em classes gramaticais (e.g. Substantivo, Verbo, Adjetivo). Isso permite que o conhecimento semântico (e.g. gato é-um animal) seja usado no processo automático de análise ou de criação de sentenças (BURTON, 1976). Assim, uma GS é mais abstrata que uma gramática sintática, pois representa o conhecimento ao nível de conceitos e não de palavras. Isto é, esta não carrega informações morfossintáticas das palavras, como classe gramatical, gênero, número, etc., mas representa informações semânticas a respeito de como o conceito denotado por uma determinada palavra se relaciona com os outros conceitos presentes na base. Por exemplo, em uma GS, um conceito que denota a ação "comer" tem os seus argumentos que apontam a entidade que come (i.e., agente) e o objeto que é comido (i.e., tema). E pode ter ainda conceitos que se relacionam com ele de forma hierárquica. O conceito de "ingerir", por exemplo, poderia estar um nível acima. Esse tipo de organização do conhecimento é usado por sistemas comerciais como *Alexa*[2], *Cortana*[3], *Siri*[4], etc., e também na academia por proposta que envolvem o PLN: sistemas de respostas a perguntas, interação homem-máquina por texto, sistemas de buscas semânticas, dentre outros. Para Burton (1976), a principal vantagem do uso dessa forma de representar o conhecimento é que ela permite que restrições semânticas sejam usadas para fazer previsões durante o processo de análise ou construção automática de sentenças. Isso reduz o número de alternativas que precisam ser checadas, além de reduzir a ambiguidade gramatical que pode existir ao utilizar apenas

[2] Assistente virtual da *Amazon*
[3] Assistente virtual da *Microsoft*
[4] Assistente virtual da *Apple*

classes gramaticais (e.g., substantivo).

Em síntese, uma GS é uma modelagem semântica de um vocabulário controlado, na qual o conhecimento enciclopédico (i.e., conceitos) está conectado por relações léxico-semânticas de hierarquia e de predicado-argumento. As relações de hierarquia são aquelas que definem a estrutura taxonômica da GS. Elas permitem que as funções de um determinado conceito sejam herdados por seus conceitos "filhos". A constituição dessa estrutura depende do domínio ao qual a GS é direcionada. Assim como a nomenclatura utilizada nas relações, que pode variar de um simples *isA* (é-um), para relações herdadas de bases léxicas como a *WordNet* (MILLER, 1995), que usa relações de hipônimos[5] e hiperônimos[6]. Já as relações de predicado-argumento, definem quais e que tipos de argumentos cada conceito da GS pode ter. São essas relações que definem a parte gramatical da GS, ou seja, que definem qual é a gramática usada na base de conhecimento. Estas podem ser funções sintáticas ou papéis semânticos. A escolha sobre qual usar também depende da finalidade da GS. Entretanto, papéis semânticos, como o nome sugere, carregam uma semântica maior (JAWORSKI; PRZEPIÓRKOWSKI, 2014). Além disso, funções sintáticas se aplicam apenas a verbos enquanto papéis semânticos podem também ser usados para substantivos (cf. Seção 2.1.1). Na Figura 3 apresenta-se um diagrama de uma GS de exemplo, com relações de hierarquia (*isA*) e de predicado-argumento (*hasAgent*).

Figura 3 – Exemplo de Gramática Semântica



Fonte: do autor

2.2.1 Representação Computacional

Uma forma computacional de representar o conhecimento de uma GS é através de bases de conhecimento em forma de ontologias. Segundo Gruber (1993), uma ontologia é uma especificação explícita de uma conceitualização, que, por sua vez, é uma visão simplificada e abstrata daquilo que se deseja representar. Assim, essas bases são artefatos computacionais compostos por vocabulários de conceitos, suas definições e suas possíveis

[5] Um conceito menos abrangente que o seu hiperônimo
[6] Um conceito mais abrangente que os seus hipônimos

propriedades (GUARINO, 1998; GUIZZARDI, 2000). Geralmente, os conceitos presentes nesses vocabulários estão organizados em taxonomias, que consistem na visão simplificada do conhecimento de um dado domínio de aplicação. Assim, segundo Freitas (2003), uma ontologia pode ser considerada como uma materialização do nível de conhecimento de um dado domínio.

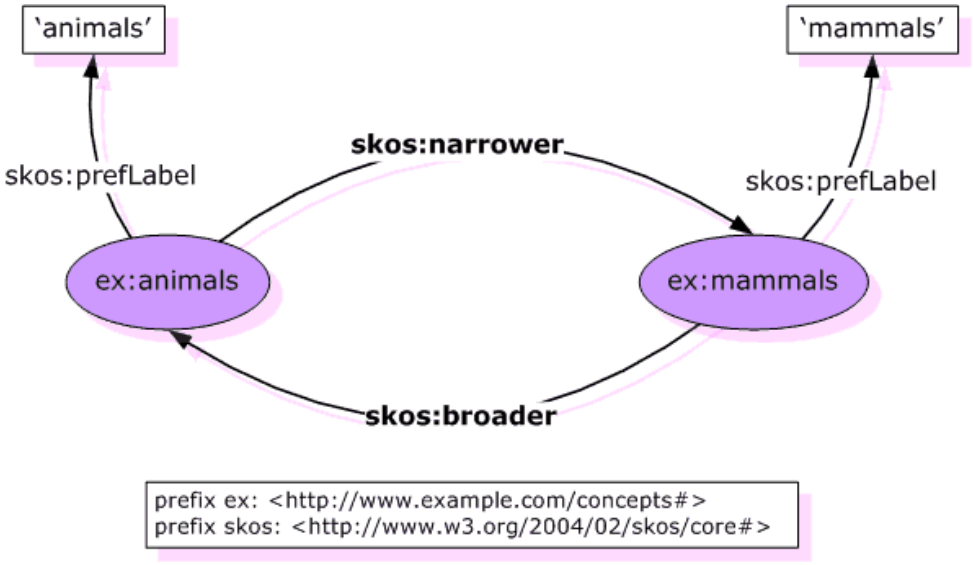
No caso das GS, o que se deseja representar é o conhecimento enciclopédico (i.e., conceitos e taxonomia) e o conhecimento linguístico (i.e., relações de predicado-argumento). Para isso, é possível usar modelos como o *Simple Knowledge Organization System* (SKOS) (ISAAC; SUMMERS, 2009). O SKOS fornece um modelo para expressar a estrutura básica e o conteúdo de esquemas de conceitos, como tesauros, esquemas de classificação, listas de assuntos, taxonomias e outros tipos semelhantes de vocabulários controlados (ISAAC; SUMMERS, 2009). Esse modelo é usado em bases de conhecimento amplamente utilizadas, como *WordNet*[7] (MILLER, 1995), *BabelNet*[8] (NAVIGLI; PONZETTO, 2010) e *DBpedia*[9] (AUER et al., 2007), como uma referência para a representação de esquemas de conceitos. O SKOS também é usado como base para outros modelos, como o *Lexicon Model for Ontologies* (*Ontolex-lemon*), no qual este é utilizada para a representação de conceitos léxicos. Na Figura 4 é ilustrado um exemplo de uso do SKOS, no qual dois conceitos (*animals* e *mammals*) contam com relacionamentos de hierarquia: *narrower* (i.e., menos amplo) e *broader* (i.e., mais amplo). Já para as relações de predicado-argumento, não existe um padrão estabelecido. Talvez pelo fato de papéis semânticos, que são geralmente usados para esse fim, não serem padronizados (cf. Seção 2.1.1). Corcoglioniti et al. (2016), no entanto, propôs uma extensão ao *Ontolex-lemon*, que possibilita a representação dessas relações. Essa extensão é chamada de *Predicate Model for Ontologies* (PreMOn), e sua estrutura básica é mostrada na Figura 5. Com esse modelo é possível criar *frames* semânticos como instâncias da classe *SemanticClass*, e papéis semânticos como instâncias da classe *Semantic Role*. A propriedade *pmo:semRole* é usada para indicar qual papel semântico pertence a cada *frame*. Não é possível indicar qual conceito preenche cada papel semântico, mas o modelo pode ser alterado para isso ou para outros fins.

2.2.2 Métodos de Construção

A construção de GSs pode ser feita de forma manual ou automática. A construção automática é geralmente baseada na análise de documentos de textos para a extração de relações de predicado-argumento. Segundo Zelle e Mooney (1993), construir GSs automaticamente requer a resolução de três questões: A primeira diz respeito ao **conjunto de papéis semânticos** usados na GS. Esse conjunto depende do domínio ao qual a base é direcionada, e do ***corpus* ou analisador semântico** utilizado na sua extração. Já a segunda questão

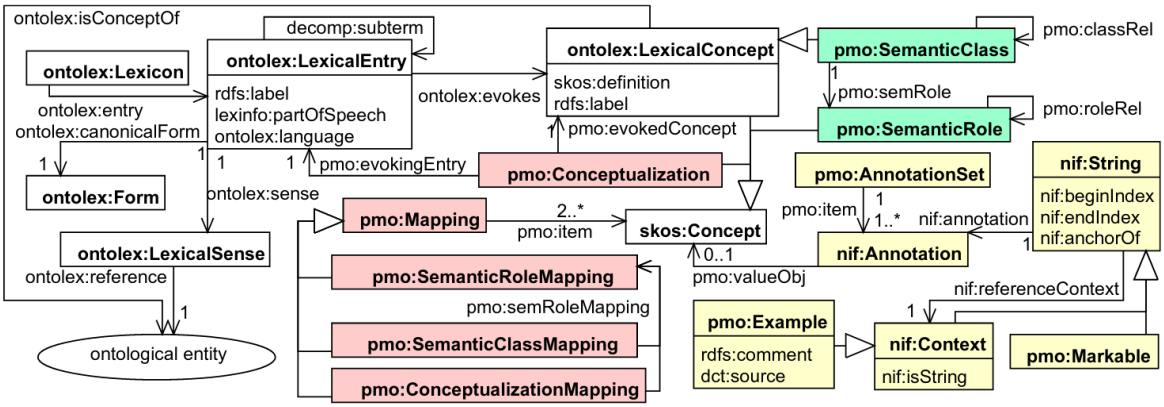
[7] <https://wordnet.princeton.edu/>
[8] <https://babelnet.org/>
[9] <https://wiki.dbpedia.org/>

Figura 4 – Exemplo de uso do SKOS



Fonte: Miles e Brickley (2005)

Figura 5 – Visão geral do PreMOn



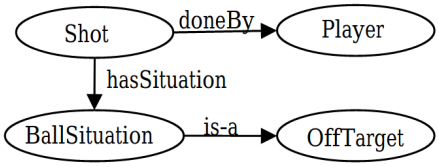
Fonte: Corcoglioniti et al. (2016)

levantada pelo autor, diz respeito às restrições semânticas que são específicas de cada domínio. Essas restrições estão relacionadas à forma com a qual o conhecimento presente na GS está organizado, ou seja, a **taxonomia** usada. Por fim, a terceira questão se refere à importância da capacidade da GS em generalizar o conhecimento linguístico e enciclopédico a partir de um conjunto de treinamento. Isto é, a GS precisa ser capaz de possibilitar a criação e/ou análise de sentenças que não foram analisadas durante o seu processo de construção. Essa questão depende diretamente do **método de construção** usado.

Nas questões levantadas por Zelle e Mooney (1993), quatro pontos se destacam: 1) **o conjunto de papéis semânticos usado** – que como já mencionado, é o que define as possíveis relações predicado-argumento entre os conceitos presentes na GS. Em alguns casos, os papéis semânticos da *FrameNet* são usados (MARTÍNEZ-SANTIAGO et al., 2015;

GRUZITIS; DANNÉLLS, 2017; GRUZITIS; PAIKENS; BARZDINS, 2012). Já em outros casos, papéis mais específicos ao domínio são utilizados, como no trabalho de Kuptabut e Netisopakul (2016), por exemplo, que é voltado ao domínio de futebol (cf. Figura 6); 2) **o *corpus* ou analisador semântico usado** – é o que serve como base de aprendizado para a GS. No trabalho de Martínez-Santiago et al. (2015), por exemplo, o *corpus* da *FrameNet* foi usado como base. Esse *corpus* é anotado com os padrões de valência (i.e., estrutura semântica) das frases. O uso desse tipo de base de treinamento faz desnecessário o uso de analisadores semânticos. Os analisadores só são úteis quando o *corpus* de treinamento não está anotado, que pode ser o caso de alguns domínios ou idiomas específicos. Ou quando se pretende usar diversos *corpora* diferentes como base, e esses seguem padrões diferentes de anotação. A Seção 2.3.1 apresenta mais detalhes sobre a tarefa de análise e rotulação de papéis semânticos; 3) **a taxonomia da GS** – é a forma com a qual o conhecimento enciclopédico (i.e., conceitos) da GS está organizado. Nos trabalhos baseados na *FrameNet* (MARTÍNEZ-SANTIAGO et al., 2015; GRUZITIS; DANNÉLLS, 2017; GRUZITIS; PAIKENS; BARZDINS, 2012), a hierarquia de *frames* presente na base é herdada. Contudo, bases como a *WordNet*, *BabelNet* e *DbPedia* também podem ser usadas. Ou até mesmo bases mais específicas de domínios, como as ontologias biomédicas (FAN; FRIEDMAN, 2011), por exemplo. Além disso, o conhecimento obtido a partir de bases como a *WordNet* pode ser moldado para se adequar ao domínio da GS; e 4) **método de construção** – os procedimentos e técnicas usados na construção da GS. As GSs baseadas na *FrameNet*, citadas anteriormente, utilizam métodos semi-automáticos que consistem na derivação da taxonomia, papéis semânticos e padrões de valência da *FrameNet*. Outras abordagens (FAN; FRIEDMAN, 2011; ZELLE; MOONEY, 1993; MURESAN; KLAVANS, 2013), utilizam métodos completamente automatizados, baseados em lógica descritiva, rotulação de papéis semânticos e análises probabilísticas. Esses métodos consistem na extração das relações predicado-argumento a partir do *corpus* de treinamento por meio da análise das amostras de textos. Formalmente, a construção de uma GS consiste em transformar uma dada taxonomia T em um grafo G. De maneira que, dado um *corpus* C, que é composto por frases $f_1, f_2, ..., f_n$, e dado uma gramática (i.e., conjunto de papéis semânticos) Gr, o agente de construção deve extrair de cada frase f_i uma estrutura semântica S_i, que seja de acordo com as regras de Gr e que conte com palavras que denotam os conceitos de T. A estrutura S_i deve ser inserida no grafo G, respeitando a hierarquia estabelecida em T.

Figura 6 – Exemplo de Gramática Semântica com papéis semânticos específicos ao domínio



Fonte: Kuptabut e Netisopakul (2016)

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Como abordado na Seção 2.2.2, a construção automática de uma GS se dá por meio de procedimentos que envolvem a análise de texto para a extração de conhecimento linguístico. Essa análise é feita a partir do PLN, que, segundo Chowdhury (2003), é um campo da Ciência da Computação que explora como os computadores podem ser usados para entender e manipular texto ou fala em linguagem natural para fazer coisas úteis. O seu uso tem o objetivo de facilitar a iteração homem-máquina por meio da linguagem natural. Além disso, esse campo fornece fundamentos teóricos e práticos para aplicações que envolvem: 1) mineração de texto; 2) tradução automática; 3) extração de conhecimento; 4) sumarização de textos, ou seja, criação automática de resumos; 5) geração automática de textos; etc.

O PLN é dividido em sub-tarefas que, segundo Jurafsky (2000), são direcionadas a análises que envolvem: 1) Fonética e Fonologia – conhecimento sobre sons linguísticos; 2) Morfologia – conhecimento dos componentes significativos das palavras (i.e., afixos, radicais, gênero, etc.); 3) Sintaxe – conhecimento das relações estruturais entre as palavras; 4) Semântica – conhecimento do significado; 5) Pragmatismo – conhecimento da relação de significado com os objetivos e intenções do falante; e 6) Discurso – conhecimento sobre unidades linguísticas maiores que uma única expressão. Dentre essas tarefas, as que são úteis para a construção de GSs são as que envolvem os conhecimentos da Semântica e de Sintaxe, e são direcionadas entendimento de linguagem natural ou extração do conhecimento. Por exemplo: **rotulação de papéis semânticos** (Seção 2.3.1), pode ser usada para identificar predicados e seus argumentos em um conjunto de sentenças; **desambiguação de sentido lexical** (Seção 2.3.2), para identificar qual o sentido que cada argumento denota; **reconhecimento de entidades nomeadas** (Seção 2.3.3), para identificar argumentos com nomes de pessoas, organizações, lugares, etc; e **análise de dependência** (Seção 2.3.4), para identificar as relações de dependência entre as palavras.

2.3.1 Rotulação de Papéis Semânticos

A análise das relações semânticas de predicado-argumento entre os constituintes de uma frase é uma das tarefas principais de qualquer sistema de entendimento de linguagem natural. Essa tarefa foi inicialmente inspirada na gramática de casos (FILLMORE, 1967), a qual depois evoluiu para a Semântica de *Frames* (cf. Seção 2.1). Seu objetivo é identificar os argumentos semânticos de cada predicado de uma dada sentença. Esses argumentos, por sua vez, são rotulados com papéis semânticos (cf. Seção 2.1.1). Os analisadores (a.k.a *parsers*) semânticos, são os principais agentes dessa tarefa. É a estes que cabe a identificação dos predicados e a rotulação de seus argumentos com papéis semânticos. Para essa tarefa, abordagens baseadas em regras ou em Aprendizagem de Máquina (AM) são usadas (MERLO; MUSILLO, 2008). Técnicas baseadas em regras preocupam-se em identificar o predicado de uma frase e preencher seus argumentos com o restante das palavras da frase, com base em um conjunto de regras e heurísticas, como no trabalho de McCoy, Pennington e Badman (1998), por exemplo. Já as abordagens baseadas em AM são geralmente baseadas em *pipelines*[10] de análises semânticas e sintáticas que usam agentes treinados a partir de *corpora* anotados (RINGGAARD; GUPTA; PEREIRA, 2017).

De maneira geral, a Rotulação de Papéis Semânticos (RPS) pode ser tratada como uma tarefa de classificação (PALMER; GILDEA; XUE, 2010). Pois, dado um predicado e dado os outros constituintes de uma frase, o *parser* deve indicar qual é papel semântico de cada constituinte a partir de um conjunto predefinido. Assim, considerando uma abordagem baseada em AM, antes de analisar uma frase, um sistema de RPS deve: (i) extrair características (e.g., classe gramatical, *lemma*[11], dependência sintática, etc.) de cada constituinte da frase; e (ii) treinar um classificador de AM com base nas características extraídas. Dessa maneira, o *parser* treinado será capaz de prever o rótulo de cada constituinte de acordo com suas características. A quantidade de constituintes aos quais o *parser* consegue prever os rótulos corretos é o que determina sua eficiência em cumprir essa tarefa. Segundo Palmer, Gildea e Xue (2010), esse tipo de abordagem requer (i) um conjunto de dados de treinamento em forma de *corpus* anotados; e (ii) informações sintáticas das palavras, que podem ser extraídas usando *parsers* sintáticos.

Segundo Sardinha (2000), um *corpus* é um conjunto de dados linguísticos textuais coletados para pesquisar uma língua ou variedade linguística e usados para explorar a linguagem a partir de evidências empíricas. As evidências empíricas que um *corpus* direcionado para RPS contém apontam as relações semânticas entre o predicado e os outros constituintes de cada frase. Segundo Palmer, Gildea e Xue (2010), a maioria dos sistemas de RPS são baseados no *PropBank* (BABKO-MALAYA, 2005). Mas outros *corpus* como *VerbNet* (SCHULER, 2005) e *FrameNet* (BAKER; FILLMORE; LOWE, 1998) também podem ser usados como base de treinamento. Como mencionado na Seção 2.1, esses três

[10] Execução de tarefas em fluxos de trabalho, onde cada tarefa depende de sua antecessora.
[11] A forma mais básica da palavra. O *lemma* da palavra "comendo" é "comer", por exemplo

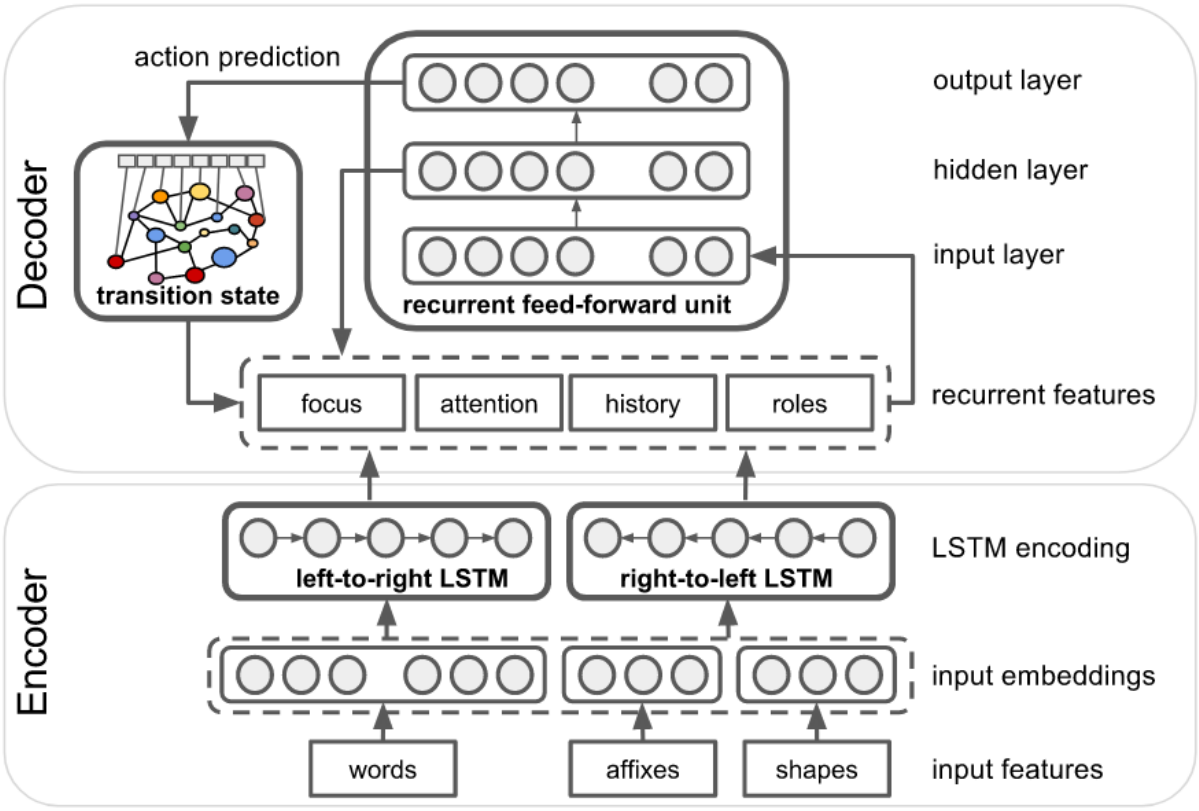
corpus são as principais bases computacionais que implementam a Semântica de *Frames*. A decisão sobre qual delas usar no treinamento de classificadores de RPS depende dos objetivos do classificador, e está estritamente relacionada à decisão de qual conjunto de papéis semânticos utilizar, como abordado na Seção 2.1.1.

A análise sintática das frases extraídas do *corpus* dá ao classificador uma série de *features* (i.e., características) que ajudam a prever qual rótulo usar para cada constituinte de uma frase. Essas informações são extraídas com o uso de *parsers* sintáticos como o de Collins (2003) ou o de Charniak e Johnson (2005), por exemplo. Segundo Palmer, Gildea e Xue (2010), as *features* usadas nos principais *parsers* semânticos são: (i) *phrase-type* – tipo de sintagma (i.e., SN = nominal, SV = verbal); (ii) *governing category* – categoria que governa o sintagma; (iii) *parse tree path* – o caminho da árvore de análise até chegar na palavra; (iv) *position* – a posição da palavra na sentença; (v) *voice* – a voz do verbo (i.e., passiva ou ativa); (vi) *head word* – a palavra que encabeça o sintagma (e.g., o head do sintagma nominal "o homem" é a palavra "homem"); (vii) *Subcategorization* – subcategorização do verbo (i.e., transitivo, intransitivo, ditransitivo); e (viii) *Argument set* – o conjunto de papéis semânticos que aparecem para o verbo em uma sentença.

Em alguns casos, a extração de *features* é evitada devido o uso do *word embeddings*, como no *parser* SLING (RINGGAARD; GUPTA; PEREIRA, 2017), por exemplo. *Word embeddings* é um método que envolve técnicas de AM para mapear palavras, sintagmas e frases para vetores de números reais. Um dos modelos usados para isso é o *Word2vec* (MIKOLOV et al., 2013). Com ele é possível fazer cálculos de similaridade entre palavras e até operações como $King + Woman = Queen$. Ele recebe como entrada para treinamento textos planos em linguagem natural, e um número inteiro que indica a dimensão do modelo. Quanto maior for a dimensionalidade, maior a qualidade do modelo em suas operações. O uso de *features* extraídas por meio de *parsers* ou o uso de *Word embeddings* depende da escolha do método de AM.

O SLING (RINGGAARD; GUPTA; PEREIRA, 2017), por exemplo, usa algoritmos de aprendizagem profunda (i.e., *Deep Learning* (DL)) para a RPS. Segundo seus autores, o uso de DL, ajuda a evitar as limitações comumente presentes em *parses* clássicos que são baseados em *pipelines*. Uma dessas limitações é a dependência existente entre os componentes, que geralmente são *parsers* que tem formatos de entrada e saída de dados não padronizadas. Os autores utilizam uma rede neural recorrente baseada na arquitetura *Bidirectional Long Short-Term Memory* (biLSTM) (HOCHREITER; SCHMIDHUBER, 1997), associada com sistemas de transições (cf. Figura 7). O SLING foi treinado recebendo como entrada um *corpus* anotado com os padrões do *PropBank* e um modelo *word2vec* para a extração de *features*. Em sua execução, o *parser* recebe como entrada uma frase em texto plano (e.g., "*John hit the ball*"), e dá como saída uma estrutura de grafo que aponta o *frame* identificado na frase analisada e seus argumentos (cf. Figura 8).

Figura 7 – Arquitetura da rede neural do SLING



Fonte: Ringgaard, Gupta e Pereira (2017)

2.3.2 Desambiguação de Sentido Lexical

A Desambiguação de Sentido Lexical (DSL) é uma tarefa de PLN que tem o objetivo de identificar o significado das palavras no contexto em que elas estão inseridas e de maneira automática (NAVIGLI, 2009). Por exemplo, a palavra "banco" em português pode ter vários significados, como: "instituição financeira", "edifício onde funciona uma instituição financeira" , "assento estreito e comprido de madeira, ferro ou pedra, com ou sem encosto", etc. Contudo, o significado que ela realmente assume em uma frase depende do contexto, ou seja, do significado da sentença inteira. Para um ser humano, essa tarefa pode parecer um tanto quanto simples. Porém, para uma máquina, essa é uma tarefa complexa, de maneira que ela é considerada um problema *AI-complete* (MALLERY, 1988). Isto é, é uma tarefa cuja solução é pelo menos tão difícil quanto os problemas mais difíceis em Inteligência Artificial (IA), como o Teste de Turing (TURING, 1950), por exemplo. Segundo Navigli (2009), formalmente, a tarefa de DSL pode ser descrita da seguinte maneira: dado um texto T, que é uma sequência de palavras $(p_1, p_2, ..., p_n)$ identifique o mapeamento A de palavras para sentidos. Assim, $A(i) \subseteq Senses_D(p_i)$, onde $Senses_D(p_i)$ é o conjunto de sentidos contidos em um dicionário D para a palavra p_i, e $A(i)$ é o subconjunto de sentidos de p_i que são apropriados no contexto T. Ainda segundo o autor, essa tarefa pode ser categorizada como uma tarefa de classificação, na qual os sentidos das palavras

Figura 8 – Exemplo de saída do SLING

```
{
  :/s/document
  /s/document/text: "John hit the ball"
  /s/document/tokens: [
    {/s/token/text: "John" /s/token/start: 0  /s/token/length: 4},
    {/s/token/text: "hit"  /s/token/start: 5  /s/token/length: 3},
    {/s/token/text: "the"  /s/token/start: 9  /s/token/length: 3},
    {/s/token/text: "ball" /s/token/start: 13 /s/token/length: 4}
  ]
  /s/document/mention: {
    :/s/phrase /s/phrase/begin: 0
    /s/phrase/evokes: {=#1 :/saft/person }
  }
  /s/document/mention: {
    :/s/phrase /s/phrase/begin: 1
    /s/phrase/evokes: {
      :/pb/hit-01
      /pb/arg0: #1
      /pb/arg1: #2
    }
  }
  /s/document/mention: {
    :/s/phrase /s/phrase/begin: 3
    /s/phrase/evokes: {=#2 :/saft/consumer_good }
  }
}
```

Fonte: Ringgaard, Gupta e Pereira (2017)

são as classes e um método automático é usado para associar cada ocorrência de uma palavra a uma ou mais classes, baseado em evidências do contexto (i.e., frase) e de fontes de conhecimento externas.

Navigli (2009) afirma que os quatro principais elementos da tarefa de DSL são:

- **A seleção dos sentidos das palavras (i.e., classes)** – levantamento do inventário de sentidos a serem usados na desambiguação. Nesse inventário, o alcance do significado de uma palavra é dividido em vários sentidos. Como a palavra "faca", que pode ter o sentido de ferramenta de cozinha ou de arma, por exemplo, a depender do contexto;

- **O uso de fontes externas de conhecimento** – que podem variar de *corpus*, sejam

eles anotados ou não, a dicionários que podem ser lidos computacionalmente, como tesauros, glossários, ontologias, etc. Um exemplo desses dicionários é a *WordNet* (MILLER, 1995), que codifica conceitos em conjuntos de sinônimos (chamados de *synsets*). A *WordNet* é considerada um padrão para DSL de língua inglesa, devido a sua difusão entre pesquisadores da área;

- **A representação do contexto** – como texto é uma fonte de informação não estruturada, fazer com que ele seja uma entrada adequada para um método automático requer uma formatação. Assim, se faz necessário preprocessar o texto. Para isso, podem ser usados preprocessamentos linguísticos como *tokenization*[12], *part-of-speech tagging*[13], *lemmatization*[14], dentre outros.

- **A seleção de um método de classificação** – a classificação pode ser feita através de métodos supervisionados ou não supervisionados. Abordagens supervisionadas usam técnicas de AM para treinar um classificador a partir de conjuntos de textos rotulados. Isto é, conjuntos de frases anotadas com características das palavras (e.g., *lemma*, classe gramatical, etc.) e com o sentido denotado por cada palavra naquele contexto. Já as abordagens não supervisionadas são baseadas em *corpora* não rotulados e utilizam técnicas como *clustering*[15], por exemplo, para identificar os sentidos das palavras sem treinamento prévio.

A tarefa de DSL pode ainda ser realizada por abordagens que não envolvem IA. São as abordagens baseadas em conhecimento, ou baseadas em dicionários. Seu objetivo é usar o conhecimento de ontologias, dicionários ou tesauros para inferir os sentidos de palavras nos contextos em que elas estão inseridas (NAVIGLI, 2009). Esses métodos geralmente apresentam um desempenho menor que os baseados em AM. Porém, têm a vantagem de conseguir uma cobertura ampla de conceitos, graças ao uso de bases de conhecimento de larga escala como a *WordNet*, por exemplo. Um exemplo desses métodos é o Algoritmo de Lesk (LESK, 1986), que é baseado na sobreposição de sentidos. Isto é, em uma determinada frase o sentido correto de uma dada palavra é aquele que tem maior sobreposição (i.e., mais palavras em comum na frase) em relação aos outros sentidos. Em uma visão formal: dado duas palavras p_1 e p_2, e considerando os sentidos dessas palavras $S_1 \in Senses(p_1)$ e $S_2 \in Senses(p_2)$, considere a Equação 2.1. Onde $gloss(S_1)$ é o conjunto de palavras que definem o sentido S_i da palavra p_i, e $score_{Lesk}$ é valor que cada sentido vai ter para cada palavra. Os sentidos que conseguem os maiores valores são atribuídos às respectivas palavras. Como reportado por Lesk (1986), esse método conseguiu uma acurácia de 50–70% (dependendo da palavra), usando um conjunto relativamente pequeno

[12] Uma etapa de normalização, que divide o texto em um conjunto de *tokens* (geralmente palavras).
[13] Identificação das classes gramaticais (e.g., Substantivo, Verbo, Adjetivo) das palavras.
[14] A redução de variantes morfológicas para sua forma básica (e.g., vivendo → viver, carros → carro).
[15] Agrupamento de dados segundo o seu grau de semelhança.

de sentidos distintos, como os encontrados em um dicionário comum. Banerjee e Pedersen (2003) propuseram uma extensão desse método, de modo a incluir o conhecimento existente nas relações entre conceitos (e.g., hiperonímia, meronímia, etc.). Isso aumentou consideravelmente (16,3%) a precisão do algoritmo de Lesk.

$$score_{Lesk}(S_1, S_2) = |gloss(S_1) \cap gloss(S_2)| \tag{2.1}$$

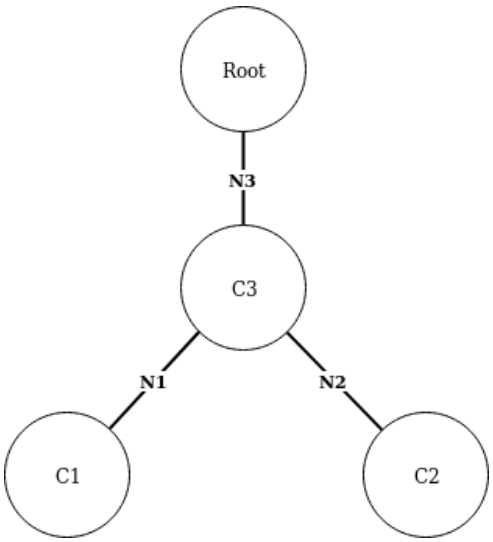
Outra abordagem de DSL que tem seu fundamento em bases de conhecimento é a abordagem baseada em estrutura. Essa abordagem se baseia na hipótese de que as palavras que ocorrem juntas em uma frase devem estar relacionadas em algum grau (PEDERSEN; BANERJEE; PATWARDHAN, 2005). Esse relacionamento pode ser aferido a partir do relacionamento hierárquico entre os sentidos denotados pelas palavras da frase. Assim, essa abordagem usa como base léxicos computacionais, como a *WordNet*, que contam com relações de hierarquia entre seus conceitos. São essas relações que dão o suporte necessário para a aferição da semelhança entre as palavras, que pode ser calculada com métricas de similaridade, por exemplo. Formalmente, essa abordagem pode ser descrita da seguinte maneira: dada uma frase T, composta por palavras $(p_1, p_2, ..., p_n)$, e sendo S_{alvo} o conjunto de sentidos existentes em uma base (e.g., *WordNet*) para a palavra alvo p_{alvo}, o sentido desambiguado de p_{alvo} é aquele que tem maior similaridade acumulada em relação aos sentidos das outras palavras da frase. Essa similaridade pode ser calculada usando métricas como a proposta por Wu e Palmer (1994). Essa métrica leva em consideração a posição dos conceitos na taxonomia em relação à posição do menor nó pai comum a eles, e a distância desse nó pai comum para o nó principal da taxonomia. Ela pressupõe que a semelhança entre dois conceitos é a função do comprimento e profundidade do caminho entre eles e seu pai comum, e entre seu pai comum e a raiz (cf. Figura 9). Assim, a similaridade entre os conceitos $C1$ e $C2$, considerando $C3$ como o nó pai comum mais próximo, é determinada pelo resultado da Equação 2.2. Onde, $N1$ é o número de nós no caminho entre $C1$ e $C3$, $N2$ é o número de nós no caminho entre $C2$ e $C3$, e $N3$ é o número de nós no caminho entre $C3$ e o nó raiz.

$$Sim(C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3} \tag{2.2}$$

O *PyWSD*[16] é uma ferramenta que implementa as abordagens baseadas no algoritmo de Lesk e no cálculo de similaridade. Já as abordagens baseadas em AM são implementadas por ferramentas comerciais, como *supWSD*[17], e por diversas propostas cientificas (YAROWSKY, 2000; HAWKINS; NETTLETON, 2000; VEENSTRA et al., 2000). As implementações baseadas em conhecimento, como o *PyWSD*, são de fácil acesso para pesquisadores e desenvolvedores interessados na área. Enquanto as baseadas em AM, ou são proprietá-

[16] <https://github.com/alvations/pywsd>
[17] <https://supwsd.net/supwsd/index.jsp>

Figura 9 – Exemplo do cálculo de similaridade de Wu e Palmer (1994)



Fonte: do autor

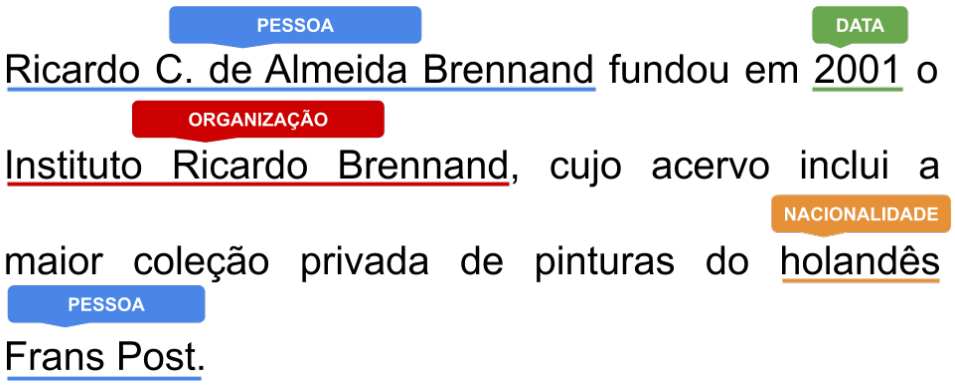
rias e pagas, ou são desenhadas para objetivos específicos dos grupos de pesquisa que as desenvolveram.

2.3.3 Reconhecimento de Entidades Nomeadas

Uma entidade nomeada é um trecho de texto que contém o nome de pessoas, organizações, localizações, etc. (RATINOV; ROTH, 2009). A identificação de referências dessas entidades em texto é considerada de essencial importância para a extração da informação, mineração de dados, entendimento de linguagem natural, etc. Essa tarefa é chamada de Reconhecimento de Entidades Nomeadas (REN), e tem o objetivo de localizar entidades nomeadas mencionadas em textos não estruturados e classificá-las em categorias predefinidas, como ilustrado na Figura 10. Segundo Leaman e Gonzalez (2008), essa tarefa pode ser definida formalmente da seguinte maneira: dado uma sequência de palavras $\mathbf{x} = (x_1, x_2, ..., x_n)$, e um conjunto de categorias (i.e., rótulos) $\mathbf{L}$, determine a sequência de rótulos $\mathbf{y} = (y_1, y_2, ..., y_n)$ de modo que $y_i \in \mathbf{L}$ sendo i para $1 \leq i \leq n$. Assim, esse tarefa pressupõe a existência de uma lista de categorias de entidades, que é usada por uma abordagem de classificação determinada.

Não há um consenso sobre a quantidade ou quais categorias de entidades usar em sistemas de REN. Por exemplo, o primeiro sistema conhecido destinado a essa tarefa (GRISHMAN; SUNDHEIM, 1996) tinha apenas 7 tipos: organização, local, pessoa, data, tempo, dinheiro e expressões percentuais (e.g., 89.0%). Enquanto outros sistemas, como o de Sekine e Nobata (2004), por exemplo, usam mais de 200 tipos. A definição de quais tipos usar, depende do domínio da aplicação. Em aplicações voltadas à bioinformática, por exemplo, é comum encontrar tipos como “proteína”, “tipo de célula”, etc. que são específicos da área e podem não fazer sentidos em outros domínios. Isto é, isso depende

Figura 10 – Ilustração da tarefa de Reconhecimento de Entidades Nomeadas



Fonte: do autor

também do nível de especificidade desejada. Sistemas voltados à bioinformática tendem a ser mais especificistas, e requerem uma gama maior de tipos de entidades (LEAMAN; GONZALEZ, 2008). Enquanto sistemas independentes de domínio usam tipos mais genéricos, como os usados na OntoNotes (HOVY et al., 2006), por exemplo. O projeto OntoNotes é uma iniciativa conjunta de universidades e organizações privadas, que tem o objetivo de organizar recursos linguísticos dos idiomas: inglês, chinês e árabe. Dentre os recursos organizados, estão anotações de entidades nomeadas com 18 tipos (cf. Tabela 2), os quais têm se tornado um padrão em sistemas de REN independentes de domínio (CHIU; NICHOLS, 2016; FINKEL; MANNING, 2009; TKACHENKO; SIMANOVSKY, 2012).

Inicialmente, a classificação feita no REN era realizada por algoritmos baseados em regras (NADEAU; SEKINE, 2007). Mas com o passar do tempo, abordagens que envolvem AM passaram a ser usadas. Os sistemas baseados em regras têm a vantagem de não necessitarem de exemplos de texto anotado para treinamento. Esse tipo de sistema é adequado para idiomas ou domínios nos quais os recursos de textos anotados são inexistentes, ou limitados. Um exemplo de sistema que segue essa abordagem é o proposto por Sekine e Nobata (2004), no qual a classificação é feita com base em um dicionário, seguindo cerca de 1400 regras escritas manualmente. No entanto, segundo Nadeau e Sekine (2007), sistemas baseados em AM se mostram mais eficientes que os baseados em regras. Nessa abordagem, as técnicas mais usadas são as baseadas em AM supervisionada (NADEAU; SEKINE, 2007). Numa visão geral, REN baseada AM supervisionada consiste em anotar dois *corpora* classificando cada entidade nomeada presente nos seus textos, e usar um deles para treinar um algoritmo e outro para testar se esse algoritmo treinado é eficiente em termos de precisão e cobertura. Dentre as técnicas usadas para o treinamento de sistemas de REN está o uso de modelos *Long Short-Term Memory* (LSTM). Essa técnica foi usada para treinar o SLING (RINGGAARD; GUPTA; PEREIRA, 2017), que além de realizar RPS (cf. Seção 2.3.1), também é capaz de reconhecer entidades nomeadas.

Tabela 2 – Categorias de entidades da OntoNotes

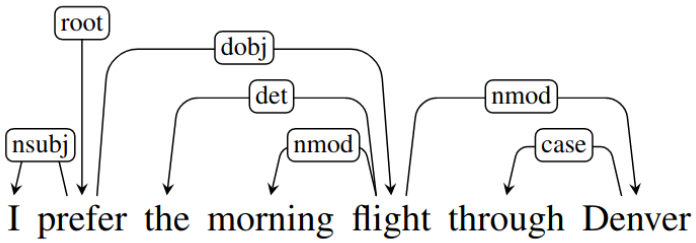
Tipo	**Descrição**
PERSON	Pessoas, incluindo ficcional.
NORP	Nacionalidades ou grupos religiosos ou políticos.
FAC	Edifícios, aeroportos, rodovias, pontes, etc.
ORG	Empresas, agências, instituições, etc.
GPE	Países, cidades, estados.
LOC	Locais que não se classificam como GPE, como cadeias de montanhas, massas de água.
PRODUCT	Objetos, veículos, alimentos, etc. (Não serviços.)
EVENT	Furacões, batalhas, guerras, eventos esportivos, etc.
WORK_OF_ART	Títulos de livros, músicas etc.
LAW	Documentos nomeados transformados em leis.
LANGUAGE	Qualquer idioma nomeado.
DATE	Datas ou períodos absolutos ou relativos.
TIME	Tempos menores que um dia.
PERCENT	Porcentagem, incluindo "%".
MONEY	Valores monetários, incluindo a unidade.
QUANTITY	Medições, em peso ou distância.
ORDINAL	"Primeiro", "segundo", etc.
CARDINAL	Numerais que não se enquadram em outro tipo.

2.3.4 Análise de Dependência

A Análise de Dependência (AD) é a tarefa de extrair a estrutura de dependência que existe entre as palavras de uma frase. Essa estrutura define os relacionamentos gramaticais entre as palavras principais (a.k.a. *head*) e as palavras que as modificam (i.e., dependentes). Na Figura 11 esses relacionamentos são ilustrados com setas que partem das palavras *head* e apontam para as dependentes, com um rótulo indicando o tipo de dependência. Formalmente, a tarefa de AD consiste em mapear uma sentença S, que consiste em palavras $p_0, p_1, ..., p_n$, onde p_0 é a palavra principal, para o seu grafo (i.e., estrutura) de dependências G. Segundo Jurafsky e Martin (2019), uma das vantagens desse tipo de análise é que as relações entre as palavras principais e seus dependentes fornecem uma aproximação das relações semânticas de predicado-argumento, o que possibilita que essa tarefa seja usada em aplicações que envolvem extração de informação. Outra vantagem é o fato de as dependências serem independentes da posição das palavras na frase. Por exemplo, na frase "*She gave me books.*" o objeto direto "*books*" aparece logo depois do objeto indireto "*me*", enquanto na frase "*She gave books to me.*", que tem basicamente o mesmo sentido, o objeto direto aparece logo depois do verbo "*gave*". Essa vantagem é

notada quando a estrutura de dependência é comparada com as árvores de análise, nas quais as unidades gramaticais são tratadas como sintagmas e as relações entre elas não são rotuladas e dependem da ordem em que os sintagmas aparecem.

Figura 11 – Exemplo de estrutura de dependência



Fonte: Jurafsky e Martin (2019)

A estrutura apresentada na Figura 11 é chamada de estrutura de dependência tipada, pois os rótulos usados (e.g., *nsubj*, *dobj*) são extraídos de uma lista fixa de tipos de relações gramaticais (JURAFSKY; MARTIN, 2019). Essa lista também inclui um nó *root*, que marca explicitamente o nó principal de toda a estrutura. Esses rótulos definem o papel que as palavras dependentes desempenham em relação às palavras principais das quais elas dependem. Entre eles, estão funções gramaticais familiares, como sujeito, objeto direto e objeto indireto. Porém, a quantidade e funcionalidade desses rótulos podem variar dependendo do idioma e do domínio da aplicação, em um caso semelhante ao das papéis semânticos, discutido na Seção 2.1.1. No entanto, na computação, existe uma iniciativa que propõe um padrão unificado e universal. Se trata do projeto *Universal Dependencies* (UD) (i.e., Dependências Universais) (NIVRE et al., 2016), que fornece um conjunto de rótulos de relações de dependências que são aplicáveis em diferentes idiomas. O projeto conta com 40 rótulos divididos em três tipos de estrutura: nominais, cláusulas e modificadores. Alguns deles são apresentados na Tabela 3. Esses rótulos são usados em diversos *parsers* de AD, e seu uso é independente de idioma e abordagem.

Segundo Kübler, McDonald e Nivre (2009), as abordagens usadas em *parsers* de AD podem ser divididas em duas classes: baseadas em gramática e orientada a dados. Uma abordagem é considerada baseada em gramática se ela se baseia em uma gramática formal, com regras predefinidas. Essas regras restringem os tipos de sentenças que podem ser analisadas, fazendo com que o alcance dessa abordagem seja limitada. Uma abordagem é classificada como orientada a dados, quando ela faz uso de técnicas de AM para treinar algoritmos a partir de dados linguísticos, como *corpora* anotados, por exemplo. Ainda segundo o autor, as abordagens orientadas a dados se mostram mais eficientes que as baseadas em gramática. Essas abordagens usam técnicas supervisionadas de aprendizagem para tentar resolver dois problemas computacionais de AD. O primeiro deles é um problema de aprendizagem, que é a tarefa de treinar um *parser* a partir de uma amostra

Tabela 3 – Exemplos de rótulos de Dependências Universais

Relações clausais	
csubj	Sujeito Clausal
ccomp	Complemento clausal
xcomp	Complemento clausal aberto
advcl	Modificador clausal de advérbios
acl	Modificador clausal de substantivos
Relações nominais	
nsubj	Sujeito nominal
nmod	Modificador Nominal
dobj	Objeto Direto
iobj	Objeto Indireto
det	Determinador (i.e., artigos)
obl	Nominal oblíquo
nummod	Modificador Numérico
Modificadores	
advmod	Modificador Adverbial
amod	Modificador Adjetival
discourse	Elemento de discurso

representativa de sentenças e suas estruturas de dependência. O segundo é um problema de análise, que é a tarefa de aplicar o *parser* treinado para analisar sentenças não anotadas. A solução desses problemas requer, portanto, (i) uma base de treinamento anotada que seja representativa e (ii) um método de treinamento e análise que seja eficiente.

Na tarefa de AD, *Treebanks* são usados como base para treinamento. Um *Treebank* é um *corpus* de texto anotado com estruturas de dependência. Segundo Jurafsky e Martin (2019) eles desempenham um papel crítico no desenvolvimento e na avaliação de *parsers* de dependências. Geralmente, *Treebanks* são conjuntos de textos extraídos de recursos existentes (e.g., artigos de jornais, blogs, etc.) e anotados manualmente por linguistas, ou automaticamente por *parsers* e depois corrigidos por humanos especialistas. Os principais *Treebanks* de língua inglesa foram construídos seguindo esse método. O *Penn Treebank* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993), por exemplo, foi extraído a partir de seções do *Wall Street Journal*. Atualmente, o projeto UD conta com *treebanks* de mais de 80 idiomas[18].

Em relação aos métodos de treinamento e análise, Kübler, McDonald e Nivre (2009) e Jurafsky e Martin (2019) destacam dois: baseado em transição e baseado em grafo. A AD

[18] <https://universaldependencies.org/>

baseada em transição trata o problema de análise como uma série de decisões que leem palavras sequencialmente de uma lista (i.e., sentença) e as combinam de forma incremental na estrutura sintática, de acordo com sua dependência. Segundo Dyer et al. (2015), nesse método, o número de operações necessárias para construir qualquer estrutura de dependência é linear no comprimento da frase, tornando esse tipo de análise computacionalmente eficiente em relação aos métodos baseados em grafos. Já o método baseado em grafo procura dentro de um conjunto de possíveis estruturas de dependência aquela que melhor se encaixa na sentença analisada. Segundo Jurafsky e Martin (2019), esse método codifica o espaço de pesquisa (i.e., possíveis estruturas) como grafos e emprega métodos extraídos da teoria dos grafos para procurar soluções ideais naquele espaço de pesquisa. A escolha das soluções ideais é dada pela pontuação que cada grafo obtém em relação à sentença analisada. Ainda segundo o autor, esse método se destaca pelo alto desempenho obtido na análise de sentenças longas. Apesar disso, *parsers* baseados em transição são mais comuns (CHOI; TETREAULT; STENT, 2015). Exemplos de *parsers* que implementam essa abordagem são: *ClearNLP*[19], *Redshift*[20], *spaCy*[21], *Yara* (RASOOLI; TETREAULT, 2015), dentre outros. Choi, Tetreault e Stent (2015) sugere o uso do *spaCy* ou do *ClearNLP* para aplicações que precisam trabalhar em alta velocidade. Segundo a análise feita pelo autor, esses dois *parsers* apresentam um bom desempenho na análise de sentenças com menos de 20 palavras.

2.4 COMUNICAÇÃO AUMENTATIVA E ALTERNATIVA (CAA)

A CAA (ASHA, 2019a) é uma tecnologia assistiva importante nas adaptações necessárias para indivíduos com deficiência intelectual e/ou dificuldades na fala (e.g., pessoas com Paralisia Cerebral, Microcefalia, Transtornos do Espectro Autista (TEA), sequelas de Acidente Vascular Cerebral (AVC) e Apraxia). Essas pessoas, podem apresentar limitações na comunicação gestual, oral e/ou escrita, dificultando o desenvolvimento de uma comunicação funcional. Para esses casos, a CAA, a partir da seleção de imagens com legendas, ajuda a: 1) desenvolver a compreensão; 2) reduzir a frustração na tentativa de se comunicar; 3) ter um poder maior de escolha; e 4) expressar sentimentos e opiniões. A CAA pode ser dividida em três categorias: *no-tech*, *low-tech*, e *high-tech* (COOK; POLGAR, 2014). *no-tech* diz respeito à interpretação de expressões faciais e movimentos voluntários, como a língua de sinais, por exemplo. *low-tech* utiliza ferramentas básicas, como livros e painéis de exibição com imagens e frases para ajudar no processo de comunicação (ELSAHAR et al., 2019). Já a CAA de alta tecnologia (*high-tech*) engloba o uso de dispositivos eletrônicos como computadores, *tablets*, *smartphones*, etc.

[19] <www.clearnlp.com>
[20] <https://github.com/syllog1sm/redshift>
[21] <https://spacy.io/>

Geralmente, os sistemas CAA de alta tecnologia são dotados imagens (e.g., fotos e pictogramas), que representam palavras ou expressões, frequentemente usadas na construção de telegráficas (i.e. que contam apenas com palavras chaves, sem o uso de conjunções, preposições, artigos, conjugação de verbos, etc.). Segundo Franco et al. (2018), esses sistemas devem fornecer o suporte necessário para que o usuário crie mensagens telegráficas que sejam compreensíveis. Isto é, forme frases com sintaxe e semântica que favoreçam a compreensão do usuário e dos seus parceiros de comunicação. Para dar esse suporte, um sistema de CAA precisa de: (i) um vocabulário controlado que seja expressivo o suficiente para garantir a comunicação funcional do usuário em seu dia-a-dia, e que seja organizado de maneira a facilitar a busca por imagens a partir de suas legendas (cf. Seção 2.4.1); e (ii) um conjunto pistas visuais (e.g., cores) e semânticas (e.g., perguntas) para ajudar o usuário a construir frases bem estruturadas e com sentido (cf. Seção 2.4.2).

2.4.1 Vocabulários Controlados

Um vocabulário controlado é um conjunto organizado de palavras e expressões usadas para indexar informação (HARPRING, 2010). No contexto de CAA, vocabulários controlados têm a função de organizar palavras e expressões a serem usadas em um sistema de CAA. Nesse contexto, é importante que o vocabulário usado esteja de acordo com o nível de desenvolvimento da comunicação de quem vai utilizá-lo (BROWN, 1973). Segundo Balandin e Iacono (1999), vocabulários controlados podem ser categorizados em dois grupos: *core* e *fringe*. Essa divisão é feita de acordo com sua constituição e utilidade. Um vocabulário *core*, consiste em um pequeno conjunto de palavras com duas propriedades distintas: seu alto nível de comunalidade (i.e., as palavras são compartilhadas entre muitos usuários) e sua alta frequência de uso por muitos usuários. Já um vocabulário *fringe*, geralmente é extenso no número de suas palavras constituintes, muda frequentemente e é altamente individualizado, ou seja, tem baixo nível de comunalidade entre os usuários. Segundo o Centro de Estudos sobre Alfabetização e Deficiência da Escola de Medicina da Universidade da Carolina do Norte (2018), como o uso de substantivos é altamente determinado por seu contexto, eles geralmente são considerados como vocabulário *fringe*, enquanto verbos, pronomes e artigos são considerados como vocabulário *core*. No entanto, a seleção de quais palavras constituem um vocabulário de um usuário de CAA é geralmente feita por um mediador (i.e., pai, educador ou terapeuta), que as escolhe de acordo com as necessidades do usuário. Normalmente, sistemas de CAA contam com vocabulários iniciais que se assemelham a vocabulários *core*, para facilitar a tarefa dos mediadores. Segundo Franco et al. (2018), a existência de um conjunto de palavras pre-definidas e que possam ser editadas, é uma característica importante para um sistema de CAA robusto.

A seleção das palavras que compõem esse vocabulário inicial pode ser baseada em propostas científicas de listas de palavras *core* para crianças, como, por exemplo:

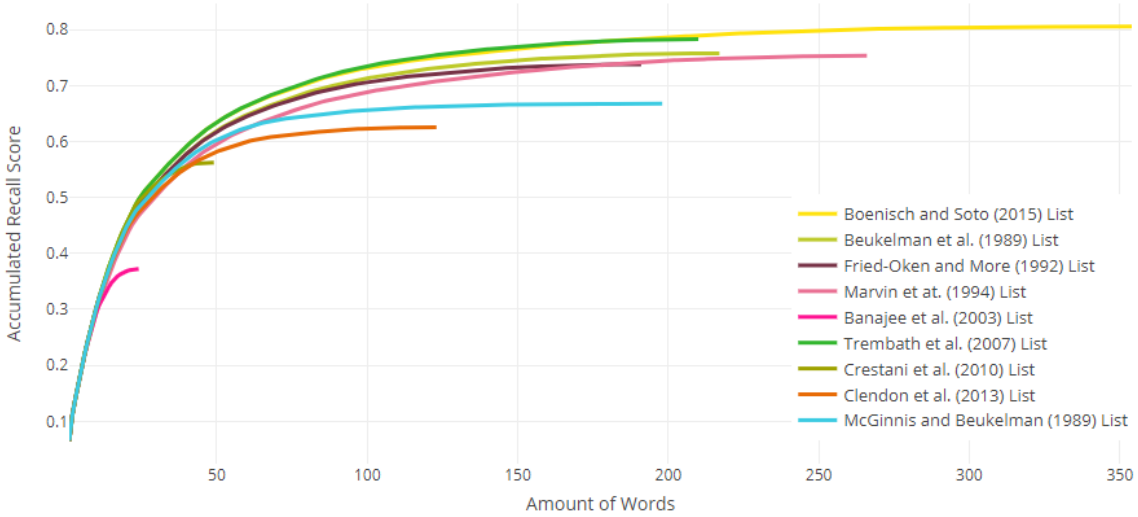
- **A lista de McGinnis e Beukelman (1989)** – composta pelas 329 palavras mais frequentes e distintas que aparecem em amostras escritas de diferentes níveis de comunicação, coletadas de 374 estudantes.
- **A lista de Beukelman, Jones e Rowan (1989)** – abrange 250 palavras, equivalentes a cerca de 85% das amostras de conversa espontânea coletadas de seis crianças em idade pré-escolar sem deficiência e ativas verbalmente, com idades entre 44 e 57 meses.
- **A lista de Fried-Oken e More (1992)** – compreende 206 palavras distintas, produzidas a partir da análise de um conjunto de 90 fontes, que inclui: listas de palavras geradas por pais e médicos de 15 crianças sem fala, diagnosticadas com paralisia cerebral e com idades entre 36 e 71 meses; e amostras de diálogos de 30 crianças com desenvolvimento típico cujos as idades e sexos correspondem às 15 crianças com paralisia cerebral.
- **A lista de Marvin, Beukelman e Bilyeu (1994)** – composta pelas 329 palavras mais frequentes de amostras de conversas espontâneas de 60 crianças sem deficiência, com idades entre 48 a 62 meses, tomadas em dois contextos (casa e pré-escola).
- **A lista de Banajee, Dicarlo e Buras Stricklin (2003)** – formada pelas 25 palavras mais comuns e frequentes de amostras de conversas espontâneas de 50 crianças de 24 a 36 meses.
- **A lista de Trembath, Balandin e Togher (2007)** – contendo 262 palavras distintas, responsáveis por 79,8% das amostras de conversação espontânea de seis crianças envolvidas em atividades pré-escolares e com idades entre 36 e 60 meses.
- **A lista de Crestani, Clendon e Hemsley (2010)** – composta por 173 palavras que equivalem a 80% das usadas por 18 crianças em desenvolvimento típico com idades entre 60 a 68 meses.
- **A lista de Clendon, Sturm e Cali (2013)** – compreende 140 palavras, que representam 70% das amostras escritas por alunos não deficientes, sendo 65 do jardim de infância e 59 da primeira série.
- **A lista de Boenisch e Soto (2015)** – formada pelas 395 palavras mais frequentes usadas por 30 crianças falantes do inglês (20 do ensino fundamental e 10 do ensino médio) com idades entre 84 e 168 meses.

Apesar de terem o mesmo objetivo, não há um consenso entre essas listas que indique qual o melhor conjunto de palavras *core*. Por isso, Franco et al. (2017) realizaram uma série de análises estatísticas para determinar qual dessas listas tem a maior cobertura das palavras faladas por crianças. Para isso, os autores calcularam a cobertura acumulada (cf.

Equação 2.3) de cada lista sobre sentenças extraídas do conjunto de *corpora*[22] do *Child Language Data Exchange System* (CHILDES) (MACWHINNEY, 2014). Onde, $Recall_i$ é o percentual do *corpus* que a i-ésima palavra representa, e n é a quantidade de palavras da lista. Os resultados dessas análises dão evidências que ajudam a decidir qual lista apresenta melhor cobertura sobre as expressões infantis e quais são as melhores palavras para sistemas de CAA direcionados para crianças. Como mostrado na Figura 12, os melhores resultados foram obtidos pelas listas de Boenisch e Soto (2015) e Trembath, Balandin e Togher (2007), que individualmente cobrem pelo menos 80% das falas do CHILDES. A partir dessas análises, Franco (2020) construiu uma lista de palavras consideradas essenciais para sistemas de CAA. São 614 palavras, adquiridas automaticamente a partir das listadas acima e validadas por especialistas através de métodos qualitativos e quantitativos. A lista de Franco (2020) mostrou uma cobertura maior das falas do CHILDES em relação às testadas por Franco et al. (2017) e ilustradas na Figura 12.

$$AcummulatedRecall_n = \sum_{i=1}^{n} Recall_i \tag{2.3}$$

Figura 12 – Cobertura acumulada das listas de palavras *core*



Fonte: Franco et al. (2017)

Outro ponto importante em relação a vocabulários controlados em sistemas de CAA é a forma com que as palavras são organizadas. Nesse ponto, é importante destacar que mentalmente o vocabulário de uma pessoa pode ser organizado de inúmeras maneiras. No entanto, as palavras em um sistema de CAA podem ser organizadas de uma maneira que não é necessariamente compatível com a organização mental do usuário (DRAGER et al., 2003). Segundo Light e Drager (2002), existem pelo menos 5 (cinco) diferentes abordagens para organizar as palavras de um vocabulário. São elas: 1) *Taxonômica* – com

[22] Um *corpus* é um documento que contém amostras textos. *corpora* é a forma plural da palavra

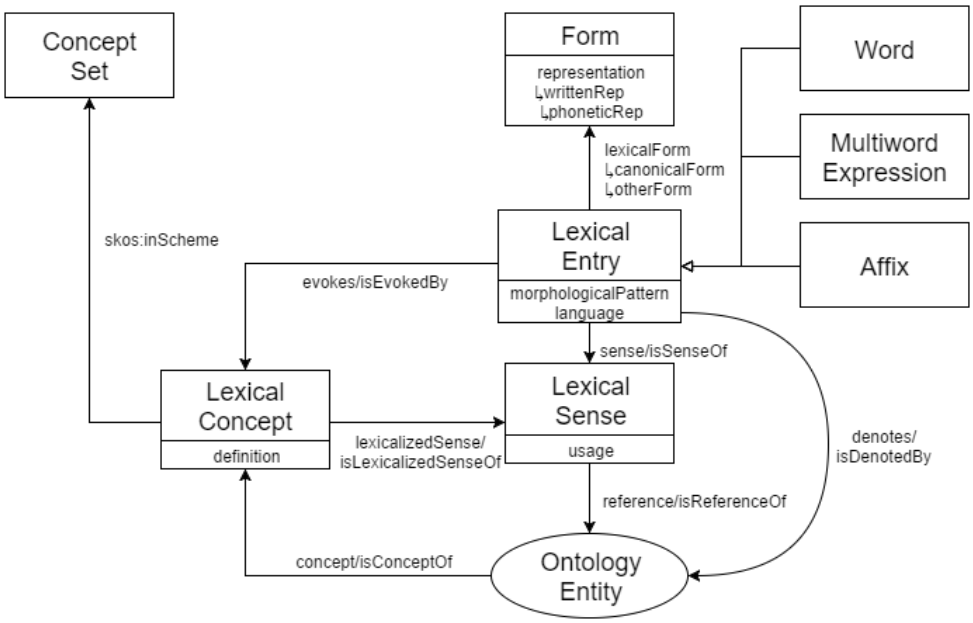
categorias organizadas hierarquicamente (e.g., gato é-um animal, animal é-um ser-vivo); 2) *Esquemática* – baseado em esquemas de eventos (e.g., indo para a cama, tomando café da manhã, escovando os dentes); 3) *Sintático-semântica* – Usa grupos semânticos organizados da esquerda para a direita para facilitar o desenvolvimento da linguagem (e.g., agentes, ações, descritores, objetos); 4) *Alfabética* – Usa a ordem alfabética das palavras; e 5) *Idiossincrática* – usa sistemas organizacionais exclusivos e específicos para cada indivíduo. Segundo Drager et al. (2003), a organização taxonômica exige uma curva de aprendizado menor. Além disso, o uso dessa forma de organização permite que seus usuários aprendam a partir da estrutura conceitual apresentada (CAMPOS; GOMES, 2007).

Computacionalmente, taxonomias podem ser representadas com o uso de ontologias. Como mencionado na Seção 2.2, uma ontologia é uma especificação explícita de uma conceitualização, que, por sua vez, é uma visão simplificada e abstrata daquilo que se deseja representar (GRUBER, 1993). O seu uso possibilita que o conhecimento seja armazenado e recuperado considerando a lógica que existe nas relações e restrições nela estabelecidas. Essa forma de representação do conhecimento é usada em diversas propostas científicas relacionadas a CAA (MARTÍNEZ-SANTIAGO et al., 2015; NETZER; ELHADAD, 2006; HERNÁNDEZ et al., 2014). Além disso, ontologias são utilizadas para a representação de vocabulários controlados de diferentes domínios (KHAN, 2018; FIORELLI et al., 2018; CARVALHO, 2018). O seu uso para esse fim fez com que surgissem modelos como o *Ontolex-lemon* (MCCRAE et al., 2017), que tem o objetivo de fornecer recursos para o desenvolvimento de ontologias léxicas (i.e., vocabulários controlados, tesauros, etc.). Esse modelo inclui a representação de informações morfológicas e sintáticas de unidades lexicais (i.e., palavras ou expressões), assim como uma interface entre sintaxe e semântica, ou seja, a representação do significado dessas unidades em relação a uma ontologia ou taxonomia. Na Figura 13 é ilustrado o núcleo do *Ontolex-lemon*, com suas classes e propriedades principais. Como mostrado na figura, uma *Lexical Entry* (i.e., entrada léxica ou unidade léxica) pode ser uma palavra (*Word*), expressão (*Multiword Expression*) ou um afixo (*Affix*). Além disso, cada *Lexical Entry* pode ter formas (*Form*), denotar (i.e., referenciar) uma entidade de uma dada ontologia, ou evocar um conceito léxico (*Lexical Concept*). A organização taxonômica é feita a partir das relações de hierarquia estabelecidas entre os conceitos léxicos. Sendo assim, cada palavra do vocabulário controlado deve evocar um conceito léxico, e cada conceito léxico deve indicar hipônimos e/ou hiperônimos.

O *Ontolex-lemon* dá suporte também à representação de informações morfossintáticas das palavras. Para isso, a ontologia *LexInfo* (CIMIANO et al., 2011) é usada como apoio. A *LexInfo* é uma ontologia de tipos, valores e propriedades a serem usadas no *Ontolex-lemon*. Ela conta com informações referentes a classes gramaticais (i.e., Substantivo, Verbo, Adjetivo), comportamentos sintáticos das palavras (e.g., verbos transitivos e intransitivos, substantivos possessivos, etc.) e de morfologia (i.e., afixos, prefixos e sufixos). Isso permite, que os vocabulários representados utilizando o *Ontolex-lemon* contem não só com

informações taxonômicas, mas também de morfologia e sintaxe que podem ser úteis na busca por palavras ou na geração automática de textos.

Figura 13 – Núcleo do Ontolex-Lemon



Fonte: Cimiano, McCrae e Buitelaar (2016)

2.4.2 *Colourful Semantics*

O *Colourful Semantics* (CS) é uma ferramenta terapêutica desenvolvida por Bryan (1997) para ajudar crianças com dificuldades de linguagem a desenvolver a construção e o entendimento de frases escritas ou faladas. O objetivo dessa ferramenta é apoiar o desenvolvimento de estruturas sintáticas por meio de um roteiro semântico (HETTIARACHCHI, 2015). Esse roteiro é composto por um sistema de chave de cores associado a perguntas chave (i.e., Quem? Fazendo? O quê? Para quem? Como? Onde? e quando?) que ajudam o indivíduo a entender o papel semântico de cada constituinte de uma frase, como ilustrado na Figura 14. As cores atuam como um suporte visual para indicar a estrutura gramatical de uma sentença. Enquanto as perguntas ajudam a vincular essa estrutura (sintática) ao seu significado (semântica) (LAW et al., 2012).

Diversas outras abordagens utilizam sistemas de chaves de cores como suporte visual à construção de frases com sentido e gramática adequada (FITZGERALD, 1949; GOOSSENS; CRAIN; ELDER, 1994; KALDOR; ROBINSON; TANNER, 2001; LEA, 1965). Porém, segundo Bolderson et al. (2011), o CS se diferencia por identificar os papéis semânticos dos constituintes de uma sentença, que, segundo Bryan (1997), são mais significativos para indivíduos com dificuldades de linguagem do que as funções sintáticas (i.e., sujeito, verbo e objetos). Um papel semântico é uma propriedade que denota a função desempenhada por uma dada palavra em relação ao predicado que ela modifica numa frase (cf. Seção

Figura 14 – *Colourful Semantics*

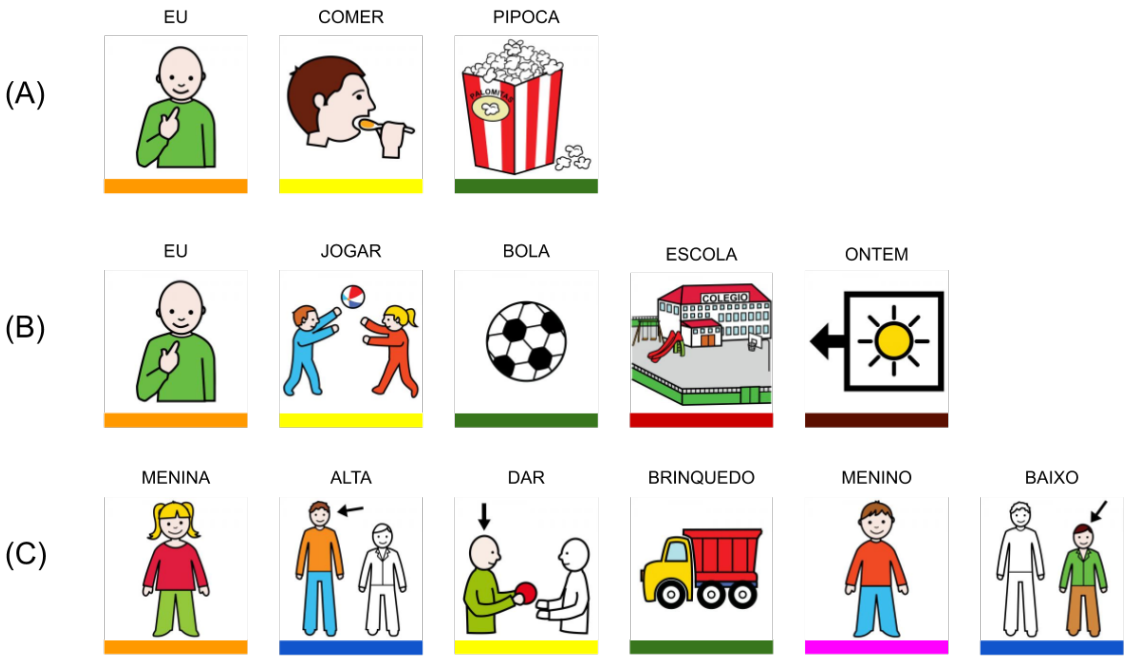
Fonte: do autor

2.1.1). Por exemplo, na frase "O menino comeu pipoca", "menino" é o Agente do predicado verbal "comeu", enquanto "pipoca" é o Tema. No CS são usados os papéis Agente, Tema, Recipiente, Maneira, Descrição, Local e Tempo. Segundo Bryan (1997), esses papéis são associados às cores e perguntas com a intenção de: (i) fazer uma discriminação visual entre cada papel semântico; (ii) estabelecer ainda mais a relação entre a pergunta e o papel semântico; (iii) associar cada tipo de frase a uma sequência visual de cores; e (iv) alertar a criança quando ela omitisse um papel semântico. A autora utilizou essa ferramenta no tratamento de uma criança de 5 anos, que tinha dificuldades no planejamento de sentenças e na ordenação e rememoração de palavras. Os seus objetivos durante esse tratamento eram: 1) ensinar a identificação de papéis semânticos em frases escritas; e 2) incentivar o uso do conhecimento dos papéis semânticos e suas funções para criar frases com as seguintes estruturas de predicado-argumento: a) verbo+agente+tema; b) verbo+agente+local; c) verbo+agente+tema+local; e d) verbo+tema+descrição.

O principal predicado da estrutura gramatical do CS é o verbo (*What Doing?*). Os demais componentes são os argumentos que esse verbo pode ter. O Agente (*Who?*), como nome sugere, é a pessoa, animal, etc. que executa a ação denotada pelo verbo da frase. O Tema (*What?*) é a entidade que recebe a ação. O Recipiente (*To Whom?*) é aquele que se beneficia da ação denotada pelo verbo. É importante destacar que o agente, o tema e o recipiente da frase podem ser tanto substantivos (e.g., menino, homem, casa) como pronome (e.g., eu, ele, você). Local (*Where?*) é um complemento adverbial e indica o local onde a ação está sendo executada, e pode ser preenchido por substantivos (e.g., escola, casa, rua) ou por advérbios (e.g., atrás, dentro, fora, etc.). Isto é, a cor não depende da classe gramatical da palavra, mas do papel semântico que esta desempenha na frase. Portanto, as imagens não podem ter uma cor preestabelecida, uma vez que estas podem desem-

penhar diferentes papeis semânticos. O mesmo acontece com o papel semântico Tempo (*When?*), que também é um complemento que pode ser preenchido por advérbios (e.g., ontem, hoje, amanhã) e por substantivos (e.g., janeiro, sábado, páscoa). *What like?* pode ser tanto um complemento adverbial de maneira, que indica como a ação é executada (e.g., rapidamente, lentamente, etc.), como um descritor (i.e., adjetivo), quando um substantivo é tomado como predicado. Assim, *What like?* pode aparecer mais de uma vez em uma única sentença: depois do verbo, como complemento adverbial; ou depois de agente, tema, recipiente, local ou tempo, como um descritor. Três exemplos de frases telegráficas construídas com base no CS são ilustradas na Figura15. A primeira frase (A) segue uma estrutura de $agente + verbo + tema$. Na segunda (B), os complementos de local (escola) e tempo (ontem) são usados. E na terceira (C) descritos (alta e alto) e um recipiente (menino) são usados.

Figura 15 – Exemplos de frases telegráficas usando o *Colourful Semantics*



Fonte: do autor

Diversos estudos demonstram a eficácia do uso do CS no tratamento de crianças com dificuldades na fala. Na primeira aplicação feita por Bryan (1997), por exemplo, depois de 8 semanas de tratamento, a criança se mostrou capaz de identificar e usar papéis semânticos na construção de frases durante a contação de uma história e depois de alguns meses, também foi possível notar um avanço na construção de frases mais complexas do que as inicialmente ensinadas. No estudo de Bolderson et al. (2011), o CS foi aplicado a 6 crianças com idades entre 5 e 6 anos durante 9 semanas. Depois desse período, notou-se um aumento significativo no tamanho médio das sentenças produzidas por essas crianças e nas métricas extraídas a partir dos testes do *Renfrew Action Picture Test* (RAPT)

(RENFREW, 2016). Apesar disso, não há registros de sistemas de CAA que implementem esse sistema.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos relacionados a esta pesquisa. Na Seção 3.1 é apresentado o método usado para o levantamento desses trabalhos, que consiste em uma Revisão Sistemática da Literatura (RSL). Já na Seção 3.2 os trabalhos são apresentados em detalhes. Por fim, na Seção 3.3 é feita uma análise comparativa entre os trabalhos.

3.1 REVISÃO SISTEMÁTICA DA LITERATURA (RSL)

Uma RSL é um meio de identificar, avaliar e interpretar as descobertas de pesquisa disponíveis relacionadas a uma pergunta de pesquisa, área, tópico ou fenômeno. Segundo Kitchenham e Charters (2007), o principal objetivo da realização de uma revisão desse tipo é reunir evidências sobre as quais basear conclusões a cerca de um tópico. Neste trabalho, o protocolo proposto pelos autores supracitados é usado como base para uma RSL, com o objetivo de levantar trabalhos relacionados a esta pesquisa. Com o apoio da ferramenta *State of the Art through Systematic Review* (StArt) (ZAMBONI et al., 2010). Esse protocolo inclui atividades e passos que definem a realização de uma RSL. Esses passos são agrupadas em três fases principais:

1. Planejamento
    a) Identificação da necessidade de se fazer uma RSL;
    b) Formulação da(s) pergunta(s) a serem respondida(s) pela RSL;
2. Execução
    a) Uma pesquisa abrangente e exaustiva de estudos primários;
    b) Avaliação da Qualidade dos estudos incluídos;
    c) Identificação dos dados necessários ara responder a(s) pergunta(s) de pesquisa;
    d) Extração dos dados;
3. Relatório
    a) Resumo e síntese dos resultados do estudo;
    b) Interpretação dos resultados;
    c) Elaboração do relatório.

3.1.1 Motivação e Pergunta de Pesquisa da RSL

O objetivo principal deste trabalho de mestrado, como definido na Seção 1.2, é propor uma Gramática Semântica (GS) para sistemas de Comunicação Aumentativa e Alternativa (CAA), que dê suporte à construção de frases telegráficas com sentido. Para isso, inicialmente, se faz necessário levantar trabalhos que propõem e utilizam GSs no contexto de CAA, a fim de fornecer um panorama de como essas bases são usadas e quais as metologias aplicadas em sua construção. Assim, o objetivo da RSL apresentada nesta seção é responder a Pergunta de Pesquisa (PP) de número 1, apresentada na Seção 1.2:

> Como gramáticas semânticas são usadas em sistemas de CAA?

3.1.2 Critérios de Inclusão e Exclusão

Segundo Kitchenham e Charters (2007), definir critérios de inclusão e exclusão é essencial para a identificação dos trabalhos que fornecem evidências diretas sobre a pergunta de pesquisa. Além disso, ainda segundo os autores, isso reduz a probabilidade de inserção de viés por parte do pesquisador durante a execução da RSL. Na Tabela 4, são apresentados os critérios de inclusão e exclusão utilizados neste estudo. Vale destacar que são considerados apenas estudos primários, ou seja, que apresentam alguma proposta para a área. Estudos secundários (e.g., revisões da literatura) não são incluídos no estudo. Sendo assim, são elegíveis de inclusão estudos primários que apresentem alguma contribuição sobre o uso de GSs aplicadas a sistemas de CAA. Os estudos que não são elegíveis de inclusão, são artigos secundários (i.e. revisões da literatura), resumos expandidos, não revisados por pares, duplicados, escritos em outro idioma que não inglês ou português, *grey-Literature* (i.e., livros, teses, dissertações), e trabalhos redundantes do mesmo autor.

3.1.3 Bases e Procedimento de Busca

Na Figura 16 apresenta-se o processo utilizado na realização da RSL, assim como o número de trabalhos identificados em cada fase. Na figura, o passo 1 diz respeito à busca por trabalhos, cuja estratégia incluiu bases eletrônicas e sugestão de especialistas. No tocante às buscas em bases eletrônicas, foram incluídas as seguintes bibliotecas digitais: *IEEE Xplore*[1], *ACM Digital Library*[2], *Science Direct*[3] e *Scopus*[4]. Para a realização da busca, foi elaborada uma *String* de busca, conforme sugerido por Chen, Babar e Zhang (2010), com o objetivo de combinar termos de interesse de modo a extrair das bases o maior número de estudos relacionados possível, assim como evitar que estudos não relacionados sejam incluídos nos resultados. Usando a *String* de busca mostrada na Figura 17, e considerando as bases eletrônicas listadas acima, foram encontrados um total de 217 estudos.

[1] https://ieeexplore.ieee.org/
[2] https://dl.acm.org/
[3] https://www.sciencedirect.com/
[4] https://www.sciencedirect.com/

Tabela 4 – Critérios de Inclusão e Exclusão de Estudos da RSL

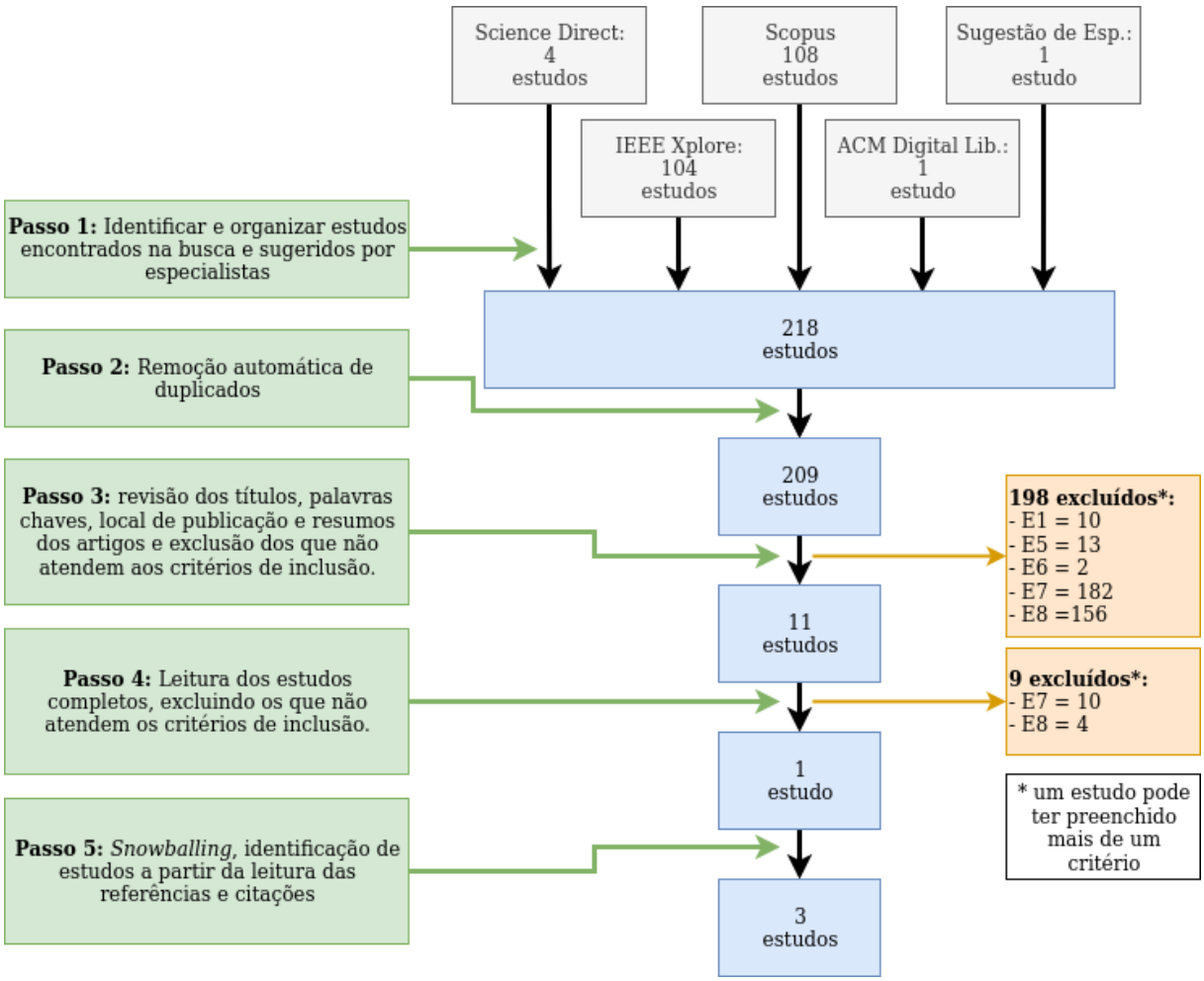
#	**Critérios de Inclusão**
1	Estudos primários
2	Estudos revisados pelos pares
3	Estudos que usam gramáticas semânticas ou semelhantes em CAA
4	Estudos publicados entre 2009 e 2019
#	**Critérios de Exclusão**
1	Estudos secundários
2	Resumos expandidos (*short papers*) com menos de 5 páginas
3	Estudos duplicados (é incluída apenas uma cópia de cada estudo)
4	Não escritos em inglês ou português
5	*Grey-literature*
6	Estudos redundantes dos mesmos autores
7	Estudos que não propõem gramáticas semânticas ou semelhantes para CAA
8	Estudos não relacionados a CAA

No entanto, segundo Kitchenham e Charters (2007), apenas o uso de bibliotecas digitais como base para revisões sistemáticas não é suficiente. Uma das alternativas para isso é a sugestão de trabalhos feita por profissionais ou cientistas especialistas no tema. No caso deste trabalho, um estudo relacionado foi sugerido por um estudante de doutorado especialista em CAA que trabalha na construção de uma ontologia a partir de um vocabulário controlado para CAA. O trabalho sugerido não apareceu nas buscas automáticas, mas foi inserido na lista inicial de trabalhos, para que pudesse passar por todas as outras fases da revisão. Assim, o número total de estudos a serem submetidos à primeira fase da revisão foi de 218.

No passo 2, os artigos duplicados foram removidos automaticamente usando a ferramenta StArt, sobrando 209 estudos. No passo 3, os títulos, palavras-chave, local de publicação e resumos dos trabalhos foram revisados com base nos critérios de inclusão e exclusão. Aqueles que não atendiam os critérios de inclusão e, consequentemente, preenchiam os critérios de exclusão, foram removidos. Esse processo levou à exclusão de 198 trabalhos, restando, assim, 11 para o próximo passo. No passo 4, os artigos restantes foram lidos com foco na introdução e nas conclusões, para avaliar, a partir do seu conteúdo, se eles atendiam aos critérios de inclusão. Nesse passo foram excluídos 10 trabalhos. Os critérios que levaram à exclusão de estudos nos passos 3 e 4 estão no lado direito da Figura 16, enumerados conforme a Tabela 4. É importante destacar que cada estudo pode ter preenchido mais de um critério, por isso a soma dos que preenchem os critérios é maior que o número total de excluídos.

No passo 5, foi executado um procedimento conhecido como *Snowballing*, descrito por

Figura 16 – Fluxo de seleção de trabalhos



Fonte: do autor

Figura 17 – *String* de busca utilizada na Revisão Sistemática da Literatura

("aac"OR "augmentative and alternative communication") AND ("ontology"OR "ontologies"OR "semantic grammar"OR "predictive semantic grammar"OR "semmantic modelling")

Fonte: do autor

Wohlin (2014) como um complemento na execução de RSLs ou Mapeamento Sistemático da Literatura. Esse procedimento consiste no uso da lista de referências de um artigo para identificar artigos adicionais. O *Snowballing* pode se beneficiar não apenas de olhar para as listas de referência, mas também de um método sistemático de ver onde os artigos são citados. Esses usos de referências e de citações, são chamados respectivamente de *Snowballing* para trás (*backward*) e para frente (*forward*) (WOHLIN, 2014). Com a execução desse procedimento, 2 novos trabalhos foram inseridos nos resultados. Os trabalhos inseridos

a partir do *Snowballing* foram publicados antes de 2009, que é o ano de limite mínimo utilizado nas buscas. No entanto, esses trabalhos se apresentam relevantes para o tema, pois é a partir deles que o estudo do uso de ontologias em sistema de CAA se desenvolve. Sendo assim, a partir da execução da revisão sistemática e do *snowballing*, 3 trabalhos relacionados foram selecionados.

3.2 APRESENTAÇÃO DOS TRABALHOS

Nesta seção os trabalhos relacionados resultantes da RSL são apresentados (*cf.* Tabela 5). A partir das análises desses trabalhos, é possível notar que existe uma ligação cronológica entre eles. Sendo assim, eles são apresentados nas próximas sub-seções em ordem cronológica e tentando responder as seguintes perguntas:

- **P1** – Como a GS é utilizada no trabalho?
- **P2** – Qual o método utilizado para a construção da GS?
- **P3** – Qual o vocabulário usado como base?
- **P4** – Qual teoria, paradigma ou estudo é a base para gramática utilizada na GS?

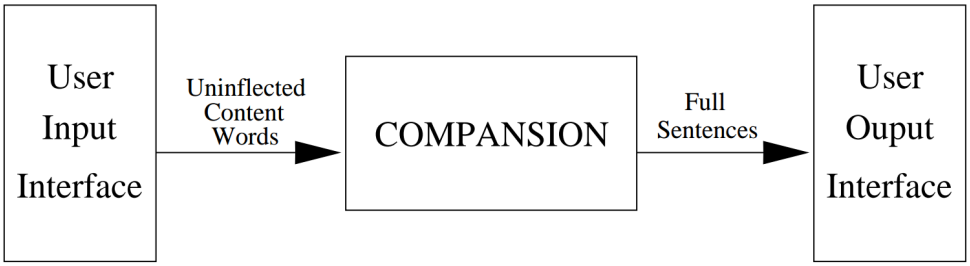
Tabela 5 – Resultados da Revisão Sistemática da Literatura

#	Título	Codinome	Autores	Ano	Método
1	A Semmantic Grammar for Beggining communicators	SUpO	Martínez-Santiago et al.	2015	Revisão Sistemática
2	Using semantic authoring for Blissymbols communication boards	BlissCAA	Netzer e Elhadad	2006	Backward a partir de 1
3	Compansion: From research prototype to practical integration	COMPANSION	McCoy, Pennington e Badman	1998	Backward a partir de 1 e 2

3.2.1 COMPANSION

Com o objetivo de auxiliar o usuário na construção de frases com sintaxe e semântica corretas, McCoy, Pennington e Badman (1998) propuseram o sistema *COMPANSION*. Como mostrado na Figura 18, esse sistema é projetado para atuar no plano de fundo de um sistema de CAA. Sua função é receber frases telegráficas (*COMPressed*) e transforma-las em frases expandidas em linguagem natural (*exPANSION*). De maneira que frases como "*like eat pizza*" sejam expandidas para "*I like to eat pizza*". Para fazer essa expansão, o sistema usa um processamento que é dividido em 3 (três) fases: *Word Order Parser*, *Semantic Parser* e *Translator/Generator*.

Figura 18 – Visão Geral do *COMPANSION*.



Fonte: McCoy, Pennington e Badman (1998)

Das três fases do *COMPANSION*, a que está relacionada com o trabalho apresentado no presente documento é a do *Semantic Parser*. O objetivo desse *parser* é construir a interpretação semântica mais adequada para uma frase a partir do conjunto de palavras de entrada fornecido pelo usuário do sistema. Para fazer isso, o ele tenta decidir quais substantivos são mais adequados para se encaixar nos papéis semânticos do verbo principal da frase. Uma base de dados lexical é usada para dar suporte a esse procedimento. Essa base conta com duas hierarquias: uma para objetos e outra para verbos. A hierarquia de objetos é baseada na *WordNet*, e deriva seus conceitos e classificação taxonômica. Já a hierarquia de verbos, é baseada na Gramática Funcional de Halliday, Matthiessen e Halliday (2014), e estabelece relações de herança entre os verbos. Além disso, informações derivadas da Teoria de Casos Temáticos (*Thematic cases theory*) (FILLMORE, 1967; FILLMORE, 1977) são anexadas à hierarquia de verbos. Essa teoria especifica que existe um pequeno conjunto de papéis semânticos os quais substantivos em uma sentença devem ocupar em relação ao verbo. Assim, nessa base são indicados quais papéis semânticos cada verbo pode ter, mas não são indicados quais objetos podem preenchê-los.

A decisão de quais palavras, que denotam objetos da hierarquia de objetos, podem preencher um dado papel semântico de um dado verbo, é tomada com base em heurísticas de preferência. Essas preferências são de dois tipos: de caso semântico e restrições de caso idiossincráticas. As preferências de caso semântico indicam maneiras possíveis de preencher um papel semântico de um dado verbo com as palavras de entrada. Elas são divididas em três tipos: (i) preferências de preenchimento de caso, que são usadas para indicar que palavra tem preferência no preenchimento de um papel semântico de um dado verbo; (ii) preferências de importância de caso, que indicam quais papéis semânticos são mais importantes para cada verbo; e (iii) preferências de caso de ordem superior, que se destinam a explicar as interações entre os papéis semânticos e seus complementos. Já as restrições de caso idiossincráticas, se diferem das de caso semântico pelo fato de serem inseridas em verbos específicos e não serem herdadas por verbos "filhos". Essas restrições indicam se um verbo é transitivo ou não, ou se um verbo tem, por exemplo, um complemento instrumental.

Os autores do *COMPANSION* não descrevem o método utilizado na construção dessa

base de conhecimento. Isto é, não indicam se as informações contidas na base foram inseridas a partir de um processo manual ou automático, se uma ferramenta foi utilizada para isso, ou outros detalhes. Eles apenas indicam que a hierarquia de objetos foi derivada da *WordNet* e que a de verbos é baseada na Gramática Funcional, mas não deixam claro como se deu essa derivação e construção e nem como as heurísticas de preferência foram estabelecidas. Além disso, eles não deixam claro se um vocabulário específico foi utilizado no sistema. Como visto, o objetivo do sistema é receber frases telegráficas (e.g., eu comer pipoca escola ontem) e expandi-las (e.g., eu comi pipoca na escola ontem). Por isso, ele não utiliza nenhuma gramática específica que baseie a construção da base de conhecimento.

3.2.2 BlissCAA

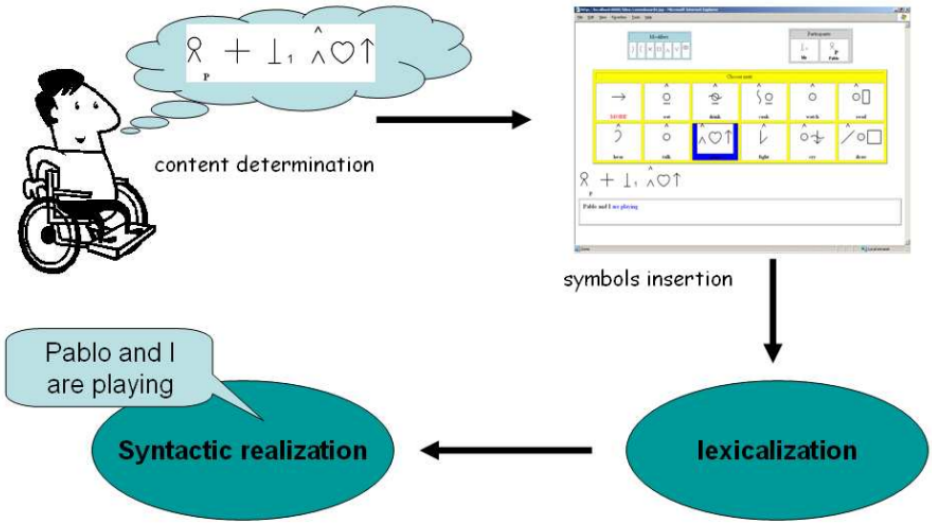
Netzer e Elhadad (2006) propõem um sistema de CAA que se baseia em um processo controlado para a criação de frases com sentido. Nesse sistema, cada passo na seleção de palavras (representadas por símbolos) é controlado por especificações feitas a partir de um *releaser*[5] linguístico. Com isso, a intervenção é feita durante a criação da frase, por meio de sugestões de símbolos (i.e., palavras). Por exemplo, na construção de uma determinada frase, se uma palavra previamente inserida for um verbo que requer um tema instrumental, apenas palavras que denotam instrumentos são apresentados ao usuário. Isso é o que diferencia o trabalho de Netzer e Elhadad do *COMPANSION*, pois a análise semântica é evitada construindo explicitamente uma estrutura semântica enquanto o usuário insere as palavras que constituem a frase. A visão geral da arquitetura desse sistema é ilustrada na Figura 19. Esse sistema conta com uma interface gráfica para a seleção de símbolos e com dois componentes responsáveis pela geração da frase final. O sistema faz uso do Sistema Bliss[6] de símbolos, por isso, neste documento ele é chamado de BlissCAA.

As análises feitas pelo *releaser* durante a construção de uma frase são suportadas por uma ontologia baseada na *WordNet* e na *VerbNet*. Assim como no *COMPANSION*, essa ontologia conta com duas hierarquias, uma para eventos (i.e., verbos) e outra para objetos (i.e., substantivos). Essas hierarquias são compostas por mais de 2200 palavras com relações de herança e predicado-argumento entre elas. A ontologia foi construída a partir de um processo semi-automático, no qual um sentido (i.e., *Synset*) da *WordNet* foi escolhido para cada palavra. Depois, os substantivos foram organizados hierarquicamente por meio da derivação das relações de hipônimos e hiperônimos da *WordNet*. Já para os verbos, as restrições definidas pela *VerbNet* para cada papel semântico de cada verbo, serviram como base para a derivação de relações predicado-argumento na ontologia. De forma que, se um dado papel semântico de um dado verbo for restrito para seres animados, um sentido que denote esse conceito é relacionado a ele como argumento. Assim, uma rede

[5] Utilizado no campo de Geração de Linguagem Natural para gerar textos a partir de representações linguísticas.

[6] O Sistema Bliss é composto por símbolos feitos de formas geométricas que representam palavras.

Figura 19 – Arquitetura do sistema de Netzer e Elhadad



Fonte: Netzer (2005)

de relações predicado-argumento foi construída entre verbos e substantivos. Sendo que, um papel semântico pode ser preenchido por apenas um único conceito e seus descendentes na hierarquia. Por exemplo, o papel semântico Agente, que é uma relação do verbo "*serve*", tem como complemento apenas o conceito *Living* (ser vivo) e os que estão abaixo dele na hierarquia. De maneira que nenhum outro conceito que não seja filho de *Living* pode ocupar o papel de Agente do referido verbo.

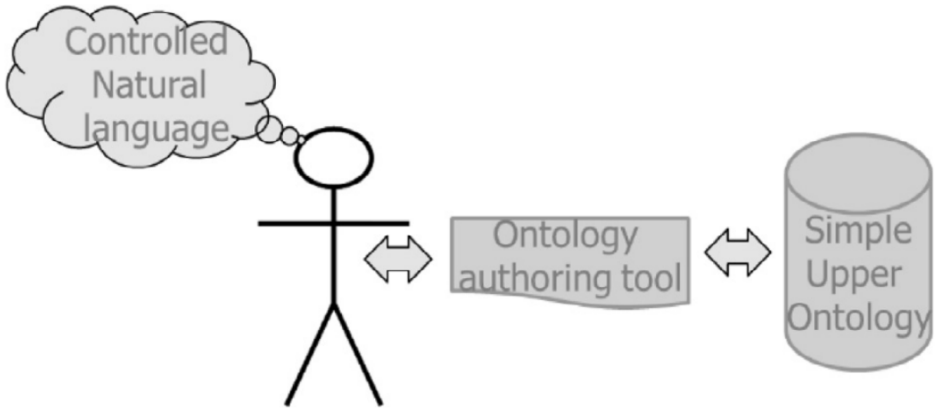
As palavras presentes no vocabulário do BlissCAA foram todas extraídas da *WordNet*. Portanto, esse vocabulário não é baseado em nenhuma teoria ou estudo que aborde o uso da linguagem por usuários de CAA. Outro ponto a ser destacado diz respeito aos tipos de estrutura de frases que podem ser construídas com esse sistema, ou seja, a gramática utilizada. Essa gramática é derivada dos padrões de valência da *VerbNet*. Portanto, não é baseada em estudos ou teorias que abordem as estruturas de construção de frases utilizadas por pessoas com deficiência na fala.

3.2.3 SUpO

Martínez-Santiago et al. (2015) propõem uma GS para comunicadores iniciantes, baseada na *FrameNet* (BAKER; FILLMORE; LOWE, 1998). Essa GS foi projetada para ser usada em sistemas de CAA para auxiliar na construção de frases com sentido e gramaticalmente corretas. Na Figura 20 é apresentada uma visão geral do uso dessa GS, na qual uma ferramenta de autoria é usada para auxiliar o usuário na construção de sentenças. Essa ferramenta faz consultas na GS, que tem o nome de *Supper Upper Ontology* (SUpO).

A SUpO foi construída a partir de um vocabulário controlado que conta com 621 palavras extraídas de um conjunto de atividades de CAA e de educação infantil. Essas

Figura 20 – Visão geral da proposta de Martínez-Santiago et al. (2015)



Fonte: Martínez-Santiago et al. (2015)

palavras foram organizadas pelos autores em uma taxonomia. Além das relações taxonômicas, foram estabelecidas relações de predicado-argumento entre verbos e substantivos, e entre substantivos e adjetivos. A construção dessas relações se deu através de um processo manual que contou com 3 (três) passos:

1. Criação da taxonomia levando em conta as palavras não verbais do vocabulário (e.g., *Toy*, *Food*, *Size*). Para cada elemento dessa taxonomia, os autores buscaram por *frames* na *FrameNet*. Se uma determinada palavra não denotasse nenhum *frame*, um novo era criado para ela. Cada *frame* por sua vez, recebeu a descrição de quais atributos ele possui. Esses atributos são baseados nos *Frame Element*s (FEs) que cada *frame* tem na *FrameNet* (e.g., *Food hasFE Size*);

2. Adição dos verbos à taxonomia. Os autores buscaram na *FrameNet* por *frames* que eram denotados pelos verbos presentes no vocabulário de entrada. Depois eles preencheram os FEs com outros *frames* da taxonomia, levando em consideração a definição (i.e., significado) de cada *frame* e de cada FE; e

3. Remoção de qualquer elemento da taxonomia que não tenha sido transformado em um *frame*.

A SUpO não utiliza nenhuma gramática específica. No entanto, o uso de apenas os FEs principais de cada *frame*, restringe as possibilidades de construção de sentenças. Isso se dá pelo fato de a maioria dos frames da *FrameNet* contar apenas com 2 (dois) FEs principais, que geralmente denotam os papéis semânticos Agente e Tema, ou as funções sintáticas Sujeito e Objeto Direto. Como mencionado na Seção 2.1.1, os FEs que denotam adjuntos adverbais, por exemplo, são tratados como secundários na base. Por outro lado, *frames* cuja classe gramatical padrão é substantivo, podem contar com FEs que apontam para atributos, que devem ser adjetivos. Sendo assim, as estruturas de sentenças que podem ser

construídas com a SUpO são basicamente: i) agente + verbo + tema; ii) agente + verbo + tema + descritor; iii) agente + descritor + verbo + tema + descritor; e iv) agente + descritor + verbo + tema. Essas estruturas, apesar de conterem gramáticas simples, não são baseadas em nenhum estudo ou teoria que apresente ou aborde os tipos de construção de frases usadas por crianças e/ou em contextos de CAA.

3.3 ANÁLISE COMPARATIVA

Nesta seção, os trabalhos relacionados a esta pesquisa são analisados levando em consideração as quatro questões estabelecidas na Seção 3.2. Na Tabela 6 apresenta-se uma visão geral dessas análises. A primeira questão levantada está interessada em saber como GSs são usadas em cada trabalho. Vale destacar, que apenas um dos trabalhos utiliza o termo GS, que é o SUpO de Martínez-Santiago et al. (2015). No entanto, os outros dois trabalhos, apesar de não citarem o termo, fazem uso de ontologias que representam GSs, conforme descritas na Seção 2.2. A ontologia utilizada no *COMPANSION* desempenha um papel diferente das usadas nos outros dois trabalhos. Nesse trabalho, a ontologia é usada como base de conhecimento para um *parser* semântico, ou seja, é usada para a análise semântica das frases. Segundo Netzer e Elhadad (2006), a principal questão em relação a essa abordagem são as lacunas que existem entre palavras de texto telegráfico (e.g., ausência de preposições, conjunções e conjugação de verbo). Segundo o autor, essas lacunas podem dificultar o trabalho do *parser*, que pode classificar um argumento de um tipo no lugar de outro de tipo diferente (e.g., Agente ao invés de Tema). Por exemplo, na frase "*I ran to him*" (Eu corri até ele), que de maneira telegráfica seria "*I run him*", considerando que o usuário pode inserir as palavras em qualquer ordem, o *parser* pode ter dificuldades em identificar *I* como agente e "*him*" como recipiente, devido à ausência da preposição "*to*". Tendo em vista que "*I*" e "*him*" são da mesma classe gramatical. Nos outros dois trabalhos (i.e., SUpO e BlissCAA) a GS desempenha um papel de sugerir palavras. Essas sugestões acontecem conforme a frase vai sendo construída, levando em consideração as palavras que já foram inseridas.

Tabela 6 – Análise dos trabalhos relacionados

Trabalho	**Uso da GS**	**Método de Construção**	**Vocabulário**	**Gramática**
COMPANSION	Análise semântica	Não mencionado	Não mencionado	Não especificada
BlissCAA	Sugestão de palavras	Semi-automático	Baseado na WordNet	Não especificada
SUpO	Sugestão de palavras	Manual	Baseado nas atividades do PECS[7] e em material de educação infantil	Não especificada

A segunda questão diz respeito ao método utilizado na construção da GS usada pelos trabalhos. O *COMPANSION* não descreve como a base de conhecimento usada foi construída, indicando apenas em que ela é baseada. No BlissCAA, a construção se deu por um processo semi-automático de derivação da *WordNet* e da *VerbNet*. Já na SUpO, o processo foi majoritariamente manual, mas com alguns passos automatizados. A construção manual de derivação da *FrameNet* feita por Martínez-Santiago et al. (2015) na SUpO pode possibilitar a inserção do viés do pesquisador na construção das relações predicado-argumento, por exemplo. Isso se dá devido à subjetividade presente na escolha dos *frames* semânticos que podem preencher um FE de outro *frame* específico. Por exemplo, o FE *Ingestibles* do *frame Ingestion* tem como definição "*The Ingestibles are the entities that are being consumed by the Ingestor*". Essa definição é um tanto abstrata, pois entidades (i.e., *entities*) podem ser qualquer coisa. Isso permite que o desenvolvedor da GS interprete segundo os seus pontos de vista, sem levar em consideração como aquele *frame* de fato se comporta em linguagem natural. A definição do FE *Ingestor*, que é citado na definição do FE *Ingestibles*, pode ser usada como um meio de desambiguação. No entanto, dependendo do tamanho do vocabulário controlado, esse tipo de verificação pode ser inviável ou ter um alto custo de mão de obra e tempo. Outro exemplo é o *frame Food*, que engloba palavras relacionadas tanto a bebidas (e.g., *juice*) como a comidas (e.g., *cake*). Martínez-Santiago et al., em seu procedimento, preencheu o FE *Ingestibles* do *frame Ingestion* com o *frame Food*. Isso pode possibilitar a construção de frases como "*[Ingestor] eat water*" ou "*[Ingestor] drink cookies*".

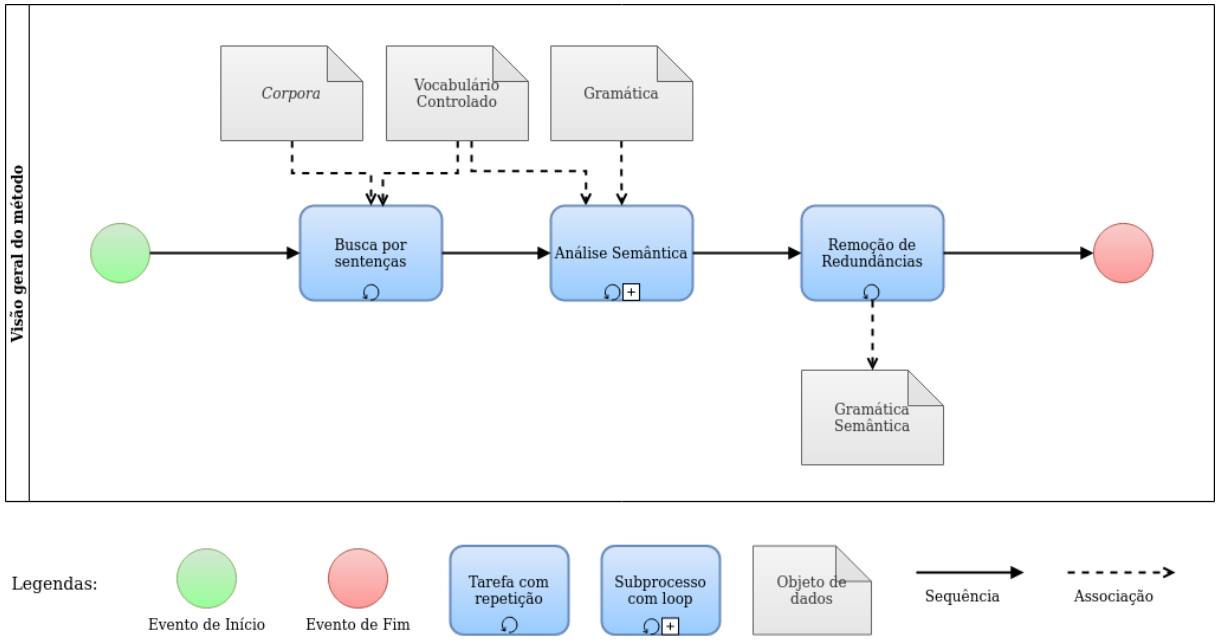
A terceira questão está interessada em saber qual o vocabulário controlado utilizado na construção das GSs apresentadas pelos trabalhos e se esses vocabulários são baseados em algum estudo ou teoria e, se são específicos do contexto de CAA. O vocabulário utilizado no *COMPANSION* não é especificado pelos autores. Já no BlissCAA, o vocabulário usado é baseado nas palavras disponíveis na *WordNet*. É um vocabulário que conta com 2200 palavras. Devido a essa quantidade de palavras, a recuperação de dados desse sistema pode demandar um alto custo computacional. Isso é um ponto negativo quando se pensa em aplicações para dispositivos móveis que tem pouca capacidade de armazenamento e de processamento e que podem não necessariamente ter conexão com a internet. Além disso, esse vocabulário não é voltado para o domínio de CAA, apesar de ser usado para isso, pois a *WordNet* é uma base léxica independente de domínio. O vocabulário usado na SUpO é bem mais reduzido que o do BlissCAA, contando com 621 palavras extraídas de conteúdos voltados para CAA e educação infantil.

A quarta e última questão está interessada na teoria, paradigma ou estudo que foi usado como base para desenvolver a estrutura gramatical da gramática semântica. Nenhum dos trabalhos citados usa qualquer fundamentação desse tipo para justificar os padrões de frases que podem ser construídas a partir do uso da GS proposta.

4 MATERIAIS E MÉTODOS

Neste capítulo são apresentados os materiais e métodos usados na construção da Gramática Semântica (GS) proposta por este trabalho. Na Figura 21 apresenta-se um diagrama *Business Process Model and Notation* (BPMN) com uma visão geral do método usado. Na Seção 4.1 são apresentadas os materiais, com um detalhamento dos seus formatos e padrões. Na Seção 4.2 é apresentado o primeiro passo do método, que trata da seleção e extração de sentenças. Na Seção 4.3 é apresentada a tarefa de Análise Semântica, com um detalhamento sobre suas sub-tarefas. Por fim, na Seção 4.4 é apresentada a tarefa de remoção de redundâncias.

Figura 21 – Visão Geral Método para Construção de Gramática Semântica.



Fonte: Do autor

4.1 MATERIAIS

Os materiais utilizados na construção da GS são três: (i) um vocabulário controlado; (ii) um conjunto de documentos de texto (i.e., *corpora*); e (iii) uma gramática. Eles são detalhados nas próximas subseções.

4.1.1 Vocabulário Controlado

Um vocabulário controlado é um conjunto organizado de palavras usadas para indexar informação de um domínio. Como abordado na Seção 2.4.1, existem diversas listas de palavras consideradas essenciais para sistemas de Comunicação Aumentativa e Alternativa

(CAA) e que podem compor o vocabulário controlado de uma criança. Neste trabalho, é utilizado a lista proposta por Franco (2020), que demonstra uma maior cobertura sobre um conjunto de falas de crianças em estágios iniciais de comunicação, segundo comparação realizada pela autora. Essa lista conta com 614 palavras, que são organizadas em uma estrutura taxonômica, que, por sua vez, é representada computacionalmente em uma ontologia, chamada AAConto. A estrutura básica da AAConto é ilustrada na Figura 22. Essa estrutura é baseada no modelo *Lexicon Model for Ontologies* (*Ontolex-lemon*) e conta com características que são importantes para a construção da GS proposta neste trabalho:

1. as palavras são instâncias das subclasses da classe *lexical entry* do *Ontolex-lemon*. No caso do exemplo da Figura 22, lex_Fish é instância da classe $Word$;

2. cada *lexical entry* tem uma "forma canônica", que aponta o *lemma* da palavra, que por sua vez, é uma instância da classe *Forma*. lex_Fish tem como forma canônica $canonicalForm\_Fish$;

3. cada *lexical entry* aponta a sua classe gramatical. O *Ontolex-lemon* não conta com um conjunto de classes gramaticais predefinidas. Porém, a ontologia *LexInfo* (CIMIANO et al., 2011) é utilizada para, por exemplo, ligar lex_Fish à classe gramatical *noun* (i.e., substantivo);

4. cada *lexical entry* evoca pelo menos 1 conceito léxico. lex_Fish evoca o conceito $lexConcept\_fish.n.01$;

5. os conceitos léxicos contam com relações de hierarquia. Para isso, as relações da *LexInfo*, são usadas. São elas: (a) hipônimo – que aponta os conceitos "filhos" (i.e., menos amplos); e (b) hiperônimo – que aponta o conceito "pai" (i.e., mais amplos);

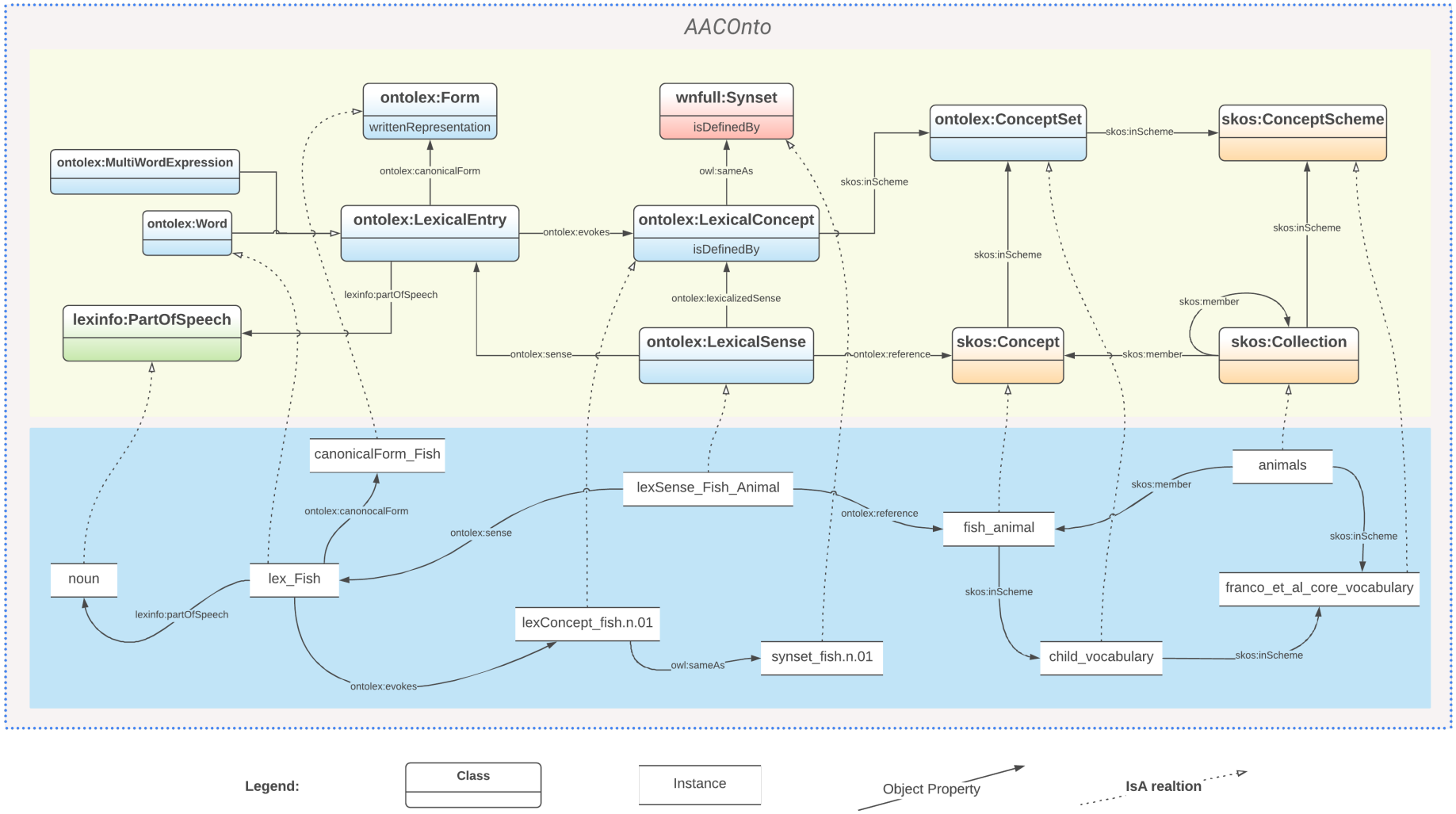
6. os conceitos léxicos da AACOnto são um subconjunto dos *synsets* da *WordNet*. Isso favorece a tarefa de análise semântica, apresentada na Seção 4.3;

7. a ontologia conta com relações entre adjetivos e substantivos (e.g., o adjetivo *big* é um atributo do substantivo *size*). Informação que também é importante na análise semântica.

4.1.2 *Corpora*

Corpora são conjuntos de documentos de textos usados para o Processamento de Linguagem Natural (PLN). O *corpus* usado como material na construção da GS proposta neste trabalho conta com frases que abrangem o conjunto de palavras do vocabulário controlado. É desse documento de texto que são extraídas as relações predicado-argumento entre os conceitos da taxonomia. Os seguintes critérios guiam a escolha do *corpus* usado:

Figura 22 – Esquema da AACOnto.



Fonte: Franco (2020)

1. O *corpus* deve ser suficientemente extenso para garantir uma ampla quantidade de frases nas quais as palavras do vocabulário aparecem;

2. Deve contar com frases sintática e semanticamente corretas;

3. As frases presentes nele devem ser anotadas com pelo menos o *lemma* (i.e., a forma mais básica da palavra) e com o *Part Of Speech* (POS) (i.e., classe gramatical) de cada palavra; e

4. O *corpus* deve ser estruturado de maneira a possibilitar a busca por frases e palavras. Isto é, ele deve dar suporte a consultas que retornem uma lista de sentenças em que uma determinada palavra aparece, por exemplo.

Como a GS proposta é voltada para CAA, o ideal é que os documentos de textos utilizados sejam específicos desse domínio. O *Child Language Data Exchange System* (CHILDES) (MACWHINNEY, 2014) é um conjunto de *corpora* que segue quase todos os critérios listados acima, pois é baseado em falas de crianças, é extenso (mais de 5 milhões de sentenças), conta com frases anotadas e tem uma versão estruturada em banco de dados, que possibilita consultas: *childes-db* (SANCHEZ et al., 2019). No entanto, as frases do CHILDES são transcrições de diálogos entre crianças de diferentes idades e seus parceiros de comunicação em diversos ambientes (e.g., escola, clínica, casa). Isso faz com que algumas dessas sentenças sejam mal formadas, com ausência de palavras de ligação, conjugação de verbos, etc., espacialmente quando a criança falante está em estágios iniciais do desenvolvimento da fala. Logo, esse *corpus* não atende o critério de número 3, o qual é importante pelo fato de que é a partir dessas frases que o conhecimento linguístico é extraído para construir as relações semânticas entre os elementos da GS, e que a má formação das sentenças dificulta essa extração. O AACText (VERTANEN; KRISTENSSON, 2011), apesar de ser do domínio de CAA, é documento que não é anotado, não é estruturado e, além disso, é pequeno (6 mil sentenças), o que faz com que ele seja pouco abrangente.

Assim, na ausência de documentos específicos para o domínio e que atendam os critérios listados acima, é utilizado o British National Corpus (BNC) (ASTON; BURNARD, 1998). O BNC é um *corpus* de texto com mais de 1,5 milhão de sentenças em língua inglesa. O *corpus* abrange o inglês britânico do final do século XX de uma ampla variedade de gêneros, com a intenção de ser uma amostra representativa da língua inglesa falada e escrita da época. O BNC é anotado com *lemma* e POS. Para facilitar as buscas, o conteúdo do *corpus* foi transformado em um banco de dados relacional seguindo a estrutura do *childes-db* (SANCHEZ et al., 2019).

4.1.3 Gramática

Como mencionado na Seção 2.2, uma gramática é um conjunto de regras que determinam o uso correto de um idioma. Esse conjunto de regras define as categorias não terminais

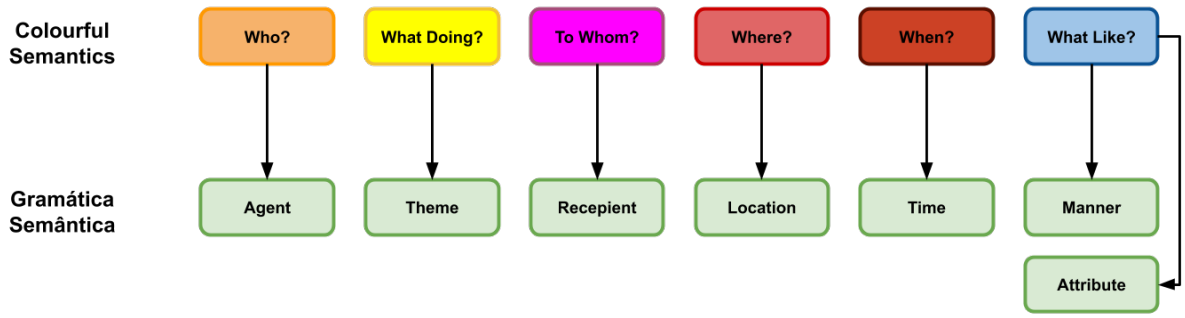
ou *slots* que compõem as estruturas de frases que são possíveis nesse idioma (e.g., sujeito, objeto direto e indireto). Neste trabalho, a gramática proposta pelo *Colourful Semantics* (CS) (cf. Seção 2.4.2) é utilizada como base para o estabelecimento de relações de predicado-argumento entre os itens do vocabulário controlado apresentado na Seção 4.1.1. Essa gramática conta com 7 *slots* que são constituídos por uma combinação de cores (i.e., reforços visuais) e perguntas (i.e., reforços semânticos), e seguem uma estrutura básica de relacionamento baseada em papéis semânticos. A presente seção tem o objetivo de explicar como essa gramática, que é abstrata, é transformada em regras gramaticais mais concretas e computacionalmente interpretáveis, que são usadas no processo automático de construção da GS.

Para transformar o CS em uma gramática mais concreta, dois procedimentos são realizados. O primeiro consiste em um mapeamento conceitual dos *slots* do CS para papéis semânticos canônicos, que seguem os padrões da *VerbNet*, conforme discutido na Seção 2.1.1. Na Figura 23 ilustra-se esse mapeamento, que é baseado na descrição do CS feita por Bolderson et al. (2011) e nas descrições de cada papel semântico da *VerbNet* (cf. Tabela 1). Esse mapeamento é dito conceitual pelo fato de mapear apenas o conceito do *slot* para o do papel semântico.

O segundo procedimento consiste em mapear os rótulos usados pelo analisador semântico para os papéis semânticos da GS. Neste trabalho, o analisador semântico utilizado é o SLING (RINGGAARD; GUPTA; PEREIRA, 2017), que utiliza rótulos baseados no *PropBank*. Como abordado na Seção 2.1.1, o *PropBank* utiliza rótulos genéricos, cujo significado varia de um *frame* para outro. Assim, se faz necessário indicar o que cada rótulo significa para cada *frame*. Para isso, é utilizado o mapeamento realizado por Palmer (2009), que aponta qual o papel semântico que cada rótulo do *PropBank* representa para cada *frame*, considerando o conjunto de papéis da *VerbNet*. No entanto, o conjunto da *VerbNet* conta com 21 papéis semânticos, um número consideravelmente maior que o usado na GS proposta (cf. Figura 23). Por isso, se faz necessário indicar qual papel da GS cada papel da *VerbNet* representa. Para fazer isso, são consideradas as descrições de cada papel semântico da *VerbNet*, conforme detalhados na Seção 2.1.1, mais especificamente na Tabela 1. Na Tabela 7 apresenta-se esse mapeamento, que é representado de forma computacionalmente legível com o uso de um objeto JSON[1].

[1] Um padrão de troca de dados entre sistemas computacionais.

Figura 23 – Exemplo de mapeamento de papéis semânticos.



Fonte: Do autor

Tabela 7 – Mapeamento dos papéis semânticos da *VerbNet* para os do *Colourful Semantics*

VerbNet	***Colourful Semantics***
Actor	Agent
Agent	Agent
Asset	Agent
Attribute	Attribute
Beneficiary	Recipient
Cause	Theme
Location	Location
Experiencer	Agent
Extent	Theme
Product	Theme
Patient	Theme
Recipient	Recipient
Stimulus	Theme
Theme	Theme
Time	Time
Topic	Theme

4.2 BUSCA POR SENTENÇAS

O primeiro passo do método de construção da GS proposta consiste em buscar no BNC sentenças que contam com as palavras do vocabulário controlado. Essas são chamadas de sentenças de referência e são usadas para a extração de relações de predicado-argumento entre os conceitos da taxonomia da AACOnto. Sendo assim, somente palavras de classes gramaticais que podem assumir o papel de predicado são consideradas nessa extração, ou seja, verbos e substantivos. Formalmente, essa tarefa pode ser descrita da seguinte

maneira: dado um subconjunto das palavras do vocabulário V, que inclui apenas verbos e substantivos $(p_1, p_2, ..., p_n)$ e dado um *corpus* C, para cada palavra p_i, é extraído de C um conjunto de sentenças $S(p_i) = (s_1, s_2, ..., s_n)$ nas quais p_i aparece. As informações de *lemma* e POS de cada palavra são usadas para facilitar a busca. Uma vez que, uma palavra pode aparecer de diversas formas (e.g., comi, comeu, comeram) em uma frase, especialmente se for um verbo, e pode pertencer a mais de uma classe gramatical. Por exemplo, a palavra da língua inglesa "*play*" pode aparecer em frases como verbo (e.g., *They **played** long and hard*) ou como substantivo (e.g., *Children learn through **play***).

Três pontos são considerados para a seleção das sentenças:

- **corretude** – as frases extraídas precisam ser sintática e semanticamente corretas. Uma vez que essas serão usadas na análise semântica para a extração de relações predicado-argumento e, como abordado na Seção 2.3.1, analisadores semânticos dependem de frases bem estruturadas para identificar essas relações;

- **tamanho** – não há exatamente uma restrição quanto a isso, mas é natural que frases muito pequenas (i.e., menos de três palavras) contenham pouca informação semântica. Por outro lado, frases muito grandes (i.e., mais de 30 palavras) podem tornar difícil o trabalho dos analisadores semânticos. Sendo assim, são extraídas do BNC apenas frases que tenham entre 3 e 30 palavras.

- **quantidade** –o número de sentenças a serem extraídas para cada palavra é determinado de acordo com a quantidade de relações predicado-argumento que se deseja ter na GS. Quanto maior o número de sentenças, maiores são as possibilidades de relações a serem estabelecidas. No entanto, um número muito alto pode aumentar as hipóteses de erros e ruído. O número de ocorrência das palavras buscadas nos *corpora* também pode variar, fazendo com que a informação extraída para umas seja mais abrangente do que para outras. Por isso, o número de sentenças a serem extraídas para cada palavra é limitado a 500.

Seguindo os pontos apresentados acima, foram extraídas 145210 sentenças de referência para os substantivos e verbos presentes no vocabulário controlado da AACOnto. No entanto, para algumas palavras nenhuma frase foi extraída, como "*out*", por exemplo. Essa palavra está classificada na AACOnto como substantivo e como verbo. No entanto, no BNC ela é classificada com mais frequência como advérbio (54801 vezes) e preposição (2776 vezes), aparecendo pouco como substantivo (33 vezes) e não ocorre como verbo. *Out* foi o único verbo da AACOnto para o qual nenhuma sentença foi encontrada. Já os substantivos, não foram encontradas frases para palavras com sentido de números inteiros (e.g., *Eight*, *Eleven*, *Fifteen*). Uma vez que, no BNC, estas são classificadas como adjetivos. Contudo, os sentidos denotados por essa palavras na AACOnto são hipônimos do conceito *number.n.*02, que é evocado pelo substantivo "*number*", para o qual 500

sentenças foram encontradas. Isso possibilita que essas palavras herdem as relações de predicado-argumento estabelecidas para *number*.

As sentenças extraídas e as palavras às quais elas estão associadas constituem a saída da tarefa apresentada nesta seção. Elas são formatadas seguindo o padrão ilustrado na Figura 24, no qual cada palavra aponta a sua classe gramatical (pos), o conceito que ela evoca na taxonomia (*evokes*) e a lista de sentenças de referência em que ela aparece (*sentences*). Além disso, as sentenças são anotadas com POS e *lemma* em cada palavra, como mostrado na figura. Essa estrutura serve de entrada para a tarefa de Análise Semântica, que é apresentada na próxima seção.

Figura 24 – Exemplo de saída da tarefa de busca por sentenças

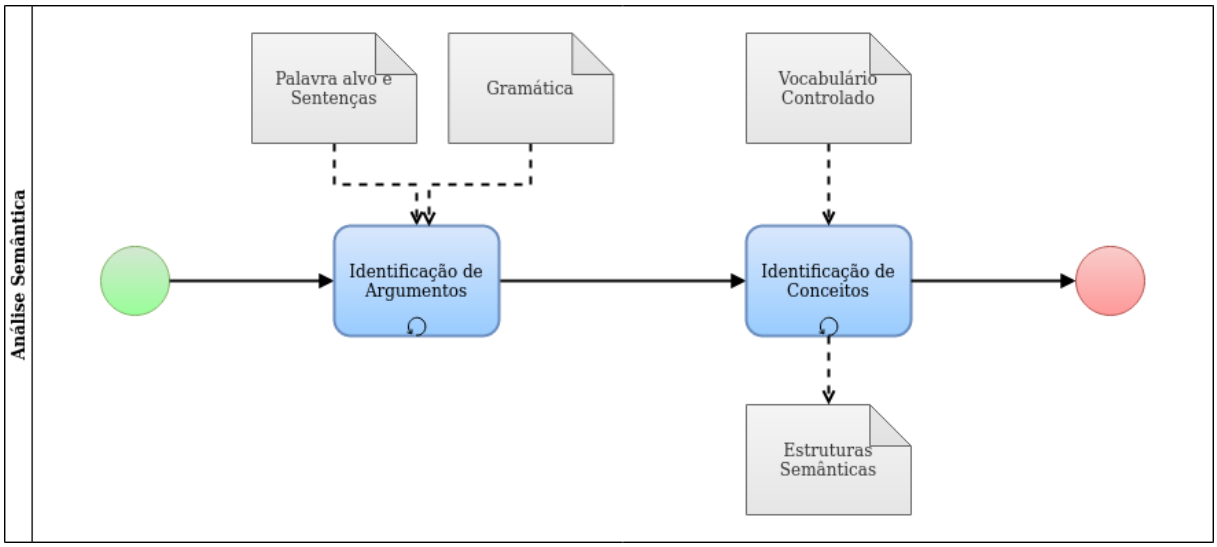

```
[
    {
        "word": "play",
        "pos": "V",
        "evokes": "play.01.v",
        "sentences": [
            {
                "text": "They played hard.",
                "tokens": [
                    {
                        "text": "They",
                        "pos": "PRP",
                        "lemma": "they"
                    },
                    {
                        "text": "played",
                        "pos": "V",
                        "lemma": "play"
                    },
                    {
                        "text": "hard",
                        "pos": "ADV",
                        "lemma": "hard"
                    }
                ]
            },
            ...
        ]
    },
    ...
]
```


Fonte: do autor

4.3 ANÁLISE SEMÂNTICA

Na Figura 25 é apresentado o diagrama BPMN da tarefa de Análise Semântica, que tem o objetivo de construir uma estrutura semântica de predicado-argumento a partir dos conceitos da taxonomia da AACOnto. Essa estrutura é construída por meio da análise das frases de referência extraídas conforme descrito na Seção 4.2 e considerando as regras gramaticais definidas na Seção 4.1.3. Sendo assim, as entradas dessa tarefa são: (i) as sentenças de referência e as palavras-alvo às quais elas são associadas, conforme ilustrado na Figura 24; (ii) as regras que definem o mapeamento dos papéis semânticos usados pelo analisador semântico para os da GS, conforme abordado a Seção 4.1.3; e (iii) a taxonomia dos conceitos do vocabulário controlado, conforme exposto na Seção 4.1.1. Essa tarefa é dividida em duas sub-tarefas: identificação de argumentos, detalhada na Seção 4.3.1; e identificação de conceitos, que é apresentada na Seção 4.3.2. Além disso, detalhes da estrutura semântica dada como saída são apresentados na Seção 4.3.3.

Figura 25 – Visão geral da tarefa de análise semântica.



Fonte: Do autor

4.3.1 Identificação de Argumentos

O objetivo da sub-tarefa apresentada nesta seção é identificar os argumentos que cada predicado (i.e., substantivos e verbos) do vocabulário controlado possui, assim como as palavras que preenchem esses argumentos. Para isso, informações semânticas e sintáticas são extraídas das frases de referência. Essa extração é feita com o uso de Rotulação de Papéis Semânticos (RPS) (cf. Seção 2.3.1) e Análise de Dependência (AD) (cf. Seção 2.3.4).

A classe gramatical de cada palavra-alvo é o que determina como as suas frases de referências são analisadas. Quando a sua classe é verbo, as frases de referência são submetidas

a um *parser* (i.e., analisador) semântico. Esse *parser* tem a função de extrair a estrutura semântica de cada frase, na qual a palavra-alvo desempenha o papel de predicado, e os seus argumentos são apontados com uso de papéis semânticos. O *parser* utilizado neste trabalho é o *SLING* (RINGGAARD; GUPTA; PEREIRA, 2017), que gera uma estrutura de grafo para cada sentença analisada, na qual as relações de predicado-argumento são rotuladas com os papéis semânticos do *PropBank* (BABKO-MALAYA, 2005). Esse *parser* apresenta maior precisão na identificação de predicados e rotulação de argumentos em relação a outros sistemas disponíveis na literatura, como o de Punyakanok, Roth e Yih (2008), por exemplo. Neste trabalho, os papéis semânticos do SLING são mapeados para os papéis utilizados na GS, que são baseados no CS, de acordo com as restrições estabelecidas na Seção 4.1.3.

Quando a classe gramatical da palavra-alvo é substantivo, as frases são analisadas com o uso da AD (cf. Seção 2.3.4). Essa análise gera uma estrutura composta pelas dependências existentes entre as palavras de uma frase. Nessa estrutura, palavras principais são relacionadas às suas dependentes através de rótulos, dentre eles o que indica modificadores adjetivais. São esses modificadores que apontam os atributos (i.e, adjetivos) das palavras-alvo, que são adicionados à estrutura semântica da frase usando como referência o papel semântico *Attribute* da gramática definida na Seção 4.1.3. O *parser* utilizado na AD é o analisador sintático do *spaCy*, que, como abordado na Seção 2.3.4 tem melhor desempenho em sentenças curtas em relação a outros *parsers*.

A saída da sub-tarefa descrita nesta seção são as sentenças de cada palavra-alvo e suas estruturas semânticas de predicado-argumento. Na Figura 26 ilustra-se a estrutura semântica da frase "*Josh is trying to eat the paper*".

4.3.2 Identificação dos conceitos

O objetivo da sub-tarefa apresentada nesta seção é identificar qual dos conceitos da taxonomia que representa o vocabulário de entrada, cada argumento da estrutura semântica ilustrada na Figura 26 evoca. Essa identificação é feita como base nas tarefas de Desambiguação de Sentido Lexical (DSL) e Reconhecimento de Entidades Nomeadas (REN), e no cálculo de similaridade de Wu e Palmer (1994).

Em PLN, a DSL é uma tarefa usada na extração de informação, no entendimento de linguagem natural, dentre outras finalidades. Essa tarefa consiste em identificar o sentido denotado por uma palavra no contexto em que ela está inserida. Neste trabalho, a DSL é usada para identificar qual o conceito denotado por cada argumento da estrutura semântica extraída pela análise abordada na Seção 4.3.1. Para isso, é usada uma abordagem baseada no cálculo de similaridade de Wu e Palmer (1994) (cf. Seção 2.3.2), que é implementada pela ferramenta *PyWSD*[2]. Essa ferramenta indica qual dos *synsets* da *WordNet* cada palavra da frase analisada denota, considerando o contexto no qual ela está inserida

[2] <https://github.com/alvations/pywsd>

Figura 26 – Exemplo de estrutura semântica de uma frase


```
{
    "target": {
        "word": "eat",
        "pos": "VERB",
        "evokes": "eat.01.v"
    },
    "arguments": [
        {
            "role": "Agent",
            "word": "Josh",
            "lemma": "Josh",
            "pos": "PROPN"
        },
        {
            "role": "Theme",
            "word": "paper",
            "lemma": "paper",
            "pos": "NOUN"
        }
    ],
    "sentence": "Josh is trying to eat the paper."
}
```


Fonte: do autor

(i.e., a sentença completa). Assim, essa abordagem pode desambiguar sentidos apenas das palavras que fazem parte da *WordNet* (i.e., substantivos, verbos, advérbios e adjetivos). É importante destacar que, como abordado na Seção 2.3.2, ferramentas que implementam abordagens baseadas em conhecimento, como as do *PyWSD*, têm um desempenho menor em relação às que implementam abordagens baseadas em Aprendizagem de Máquina (AM). No entanto, essas últimas são geralmente softwares proprietários, e fornecem acesso limitado, como a *supWSD*, por exemplo. Enquanto o *PyWSD* é livre e cobre toda a base de conceitos da *WordNet*.

Como ilustrado na Figura 26, as palavras que preenchem argumentos identificados na análise semântica podem também ser substantivos próprios (e.g., Josh). Palavras dessa classe gramatical não estão presentes na *WordNet*. Assim, se faz necessário o uso de técnicas de REN para identificar o conceito denotado por essas palavras, que, geralmente, são entidades nomeadas. Para isso, é utilizado o *SLING* (RINGGAARD; GUPTA; PEREIRA, 2017), que, além de RPS, também é treinado para fazer REN. Tarefa a qual o *parser* é

Tabela 8 – Associação de tipos de entidades e *synsets*

Tipo de Entidade	Synset da WordNet
PERSON	person.n.01
NORP	group.n.01
FAC	structure.n.01
ORG	organization.n.01
GPE	district.n.01
LOC	location.n.01
PRODUCT	product.n.02
EVENT	event.n.01
WORK_OF_ART	work.n.02
LAW	law.n.02
LANGUAGE	language.n.01
DATE	date.n.06
TIME	time_unit.n.01
PERCENT	percentage.n.01
MONEY	medium_of_exchange.n.01
QUANTITY	measure.n.02
ORDINAL	ordinal_number.n.01
CARDINAL	cardinal_number.n.01

capaz de realizar com uma precisão de 85,67%, conforme reportado pelos seus autores. O *SLING* analisa as sentenças rotulando as entidades nomeadas com os categorias do projeto *OntoNotes* (*PERSON*, *ORG*, etc.). Assim, se faz necessário associar cada uma dessas categoria a um conceito que faça parte da base de conhecimento usada, que no caso deste trabalho é a *WordNet*. Essa indicação consiste em associar os tipos da *OntoNotes* a *synsets* dessa *WordNet*. Esse procedimento é feito a partir do mapeamento apresentado na Tabela 8, que foi construído considerando a descrição de cada categoria, conforme Tabela 2, e o significado dos *synsets* relacionados a elas.

Apesar da taxonomia usada como base para a GS ser baseada na *WordNet*, esta conta apenas com um subconjunto de seus *synsets*. Assim, os conceitos identificados pela DSL e pelo REN podem não necessariamente estar presentes na taxonomia. Por isso, se faz necessário indicar qual dos seus conceitos é mais semelhante ao identificado para um argumento específico de um dado predicado. Para isso, é utilizado o cálculo de similaridade de Wu e Palmer (1994), que, como abordado na Seção 2.3.2, consiste em indicar o grau de semelhança entre dois nós de uma taxonomia, considerando a distância de cada nó para o seu pai comum mais próximo, e a distância desse pai comum para o nó principal. Assim, para um determinado conceito C, atribuído a um argumento de um dado predicado, e para cada conceito T_i da taxonomia da GS é buscado na *WordNet* o *synset* S mais baixo na árvore que é hiperônimo tanto de C, como de T_i. Depois, o grau de similaridade entre C e T_i é calculado com a Equação 4.1. Onde, $N1$ é o número de nós no caminho entre C e S, $N2$ é o número de nós no caminho entre T_i e S, e $N3$ é o número de nós no

caminho entre S e o nó raiz da taxonomia. O conceito T_i que tiver maior pontuação nesse procedimento, é considerado o mais semelhante ao conceito identificado para aquele argumento, e é inserido na estrutura semântica da palavra-alvo em questão.

$$Sim(S, C_i) = \frac{2 * N3}{N1 + N2 + 2 * N3} \tag{4.1}$$

Quando a palavra-alvo é um substantivo, o seu complemento é preenchido por conceitos denotados por adjetivos, os quais não são organizados em taxonomia na *WordNet*. Isto é, não contam com relações de hierarquia, que são a base do cálculo de similaridade de Wu e Palmer (1994). No entanto, esses conceitos contam com relações que apontam os conceitos de substantivos dos quais eles são atributos. Por exemplo, o adjetivo *big*, que denota o *synset* $large.a.01$, é um atributo de *size*, que denota o *synset* $size.n.01$. Essas relações possibilitam que os *synsets* de substantivo que estão relacionados aos adjetivos identificados na análise semântica sejam usados no cálculo de similaridade.

4.3.3 Estrutura Semântica de Saída

A estrutura semântica que é dada como saída da tarefa de Análise Semântica é ilustrada na Figura 27. Essa estrutura consiste em um objeto JSON que indica a palavra-alvo (*word*), o conceito que ela denota na taxonomia (*evokes*) e os conceitos que preenchem os seus argumentos (i.e., *Agent* e *Theme*). Cada conceito está acompanhado de um número que indica a frequência com que eles foram identificados para aquele argumento específico. Como mostrado na figura, é possível que haja redundância entre os conceitos. Por exemplo, *person.n.01* está logo abaixo de *organism.n.01* na ontologia, no entanto, os dois aparecem como complementos do argumento *Agent* na figura. Os procedimentos usados para remover essa redundância, são descritos na seção seguinte.

Figura 27 – Estrutura gerada pela análise semântica

```
{
    "word": "eat",
    "pos": "VERB",
    "evokes": "eat.01.v",
    "arguments": {
        "Agent": [
            {
                "concept": "organism.n.01",
                "frequency": 1
            },
            {
                "concept": "pronoun.n.01",
                "frequency": 3
            },
            {
                "concept": "person.n.01",
                "frequency": 2
            },
            ...
        ],
        "Theme": [
            {
                "concept": "food.n.01",
                "frequency": 4
            },
            {
                "concept": "dish.n.01",
                "frequency": 2
            },
            ...
        ],
        ...
    }
}
```

Fonte: do autor

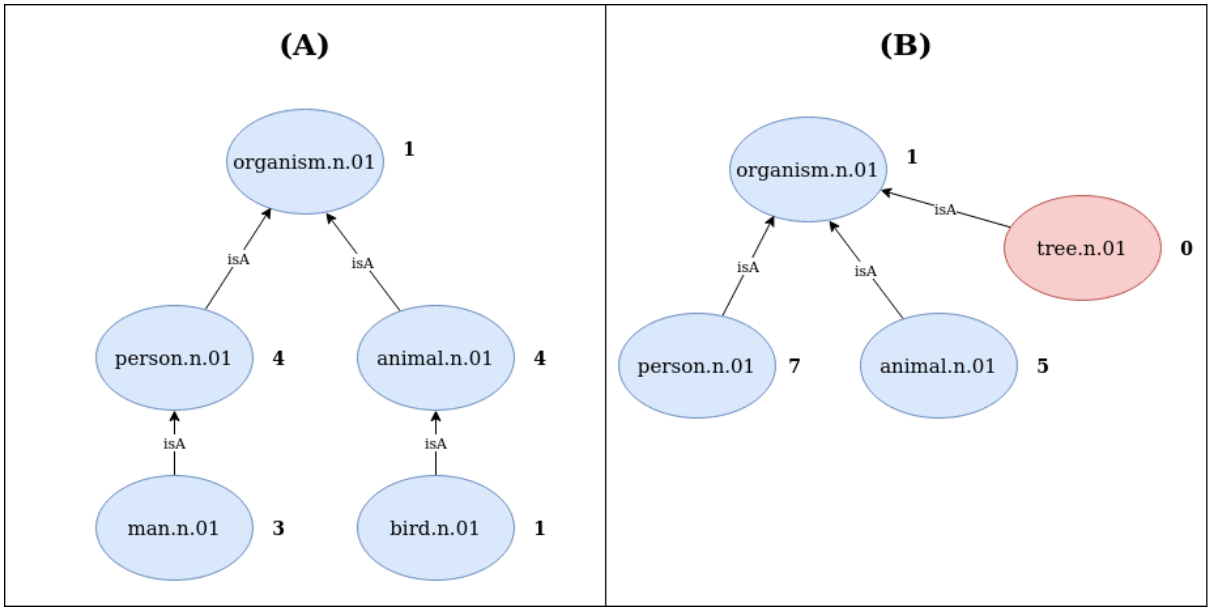
4.4 REMOÇÃO DE REDUNDÂNCIAS

Como mencionado na Seção 4.3.3, pode existir redundância semântica entre os conceitos identificados para um dado argumento de uma dada palavra-alvo. Isso acontece quando dois ou mais conceitos que pertencem ao mesmo galho da taxonomia, ou seja, que têm relação de herança em algum nível, são identificados para um único argumento. Como no exemplo ilustrado na Figura 27, no qual *organism.n.*01 e *person.n.*01, que são do mesmo galho, aparecem como complementos do argumento *Agent*. Essa redundância se caracteriza pelo fato de na GS os conceitos herdarem todas as funções dos que estão acima deles na hierarquia. Assim, remover essas redundâncias significa excluir informações desnecessárias, o que faz com que a GS ocupe menos espaço de armazenamento, o que é importante para sistemas de CAA para dispositivos móveis, além de evitar ruído nas realizações de buscas.

A remoção de redundâncias é realizada em duas rodadas, utilizando uma estratégia diferente em cada uma delas. Na primeira rodada, a estratégia usada considera a frequência de ocorrência dos conceitos, removendo apenas aqueles que têm frequência menor do que a dos seus eventuais hiperônimos identificados para um dado argumento. Para fazer essa remoção, primeiro é necessário construir uma estrutura taxonômica a partir dos conceitos identificados para cada argumento de uma dada palavra-alvo, como ilustrado na Figura 28-A. Em seguida, a exclusão é feita por meio da análise dessa estrutura, que começa dos nós mais baixos para os mais altos, comparando a frequência de cada nó à do seu pai. Se um nó tiver uma frequência menor que a do seu pai, ele é removido da estrutura e o seu número de frequência é somado à frequência do seu pai. É o caso dos conceitos *man.n.01* e *bird.n.01* do exemplo da Figura 28-A, que são removidos por terem frequências menores que as dos seus respectivos pais (cf. Figura 28-B). Por outro lado, se um conceito tiver frequência igual ou maior que a do seu pai, ele não é removido. Além disso, mesmo se o conceito pai (e.g. *organism.n.01*) tiver frequência menor que os seus filhos, ele não é removido por essa estratégia, pois o nível de importância deste para o argumento ao qual ele está associado precisa ser avaliado.

O nível de importância de um dado conceito para um argumento é definido pela frequência com que ele ocorre. Desse modo, um ponto de corte baseado na frequência precisa ser estabelecido entre os conceitos que são e os que não são importantes. Esse procedimento evita casos como o ilustrado na Figura 28-B, no qual um conceito com baixa frequência de ocorrência (i.e., *organism.n.01*) possibilita a inserção de um conceito que não aparece na lista de conceitos de um argumento (i.e., *tree.n.01*). No entanto, o número de conceitos pode variar de acordo com o argumento e o predicado, assim como sua frequência, tornado difícil o estabelecimento de um ponto de corte global e estático. Assim, esse ponto deve ser estabelecido de maneira empírica, considerando a distribuição de frequências de cada lista de conceitos. Para fazer isso, primeiro é necessário testar se

Figura 28 – Exemplos de redundância semântica.



Fonte: Do autor

essa distribuição é do tipo normal[3]. Para isso, é utilizado o teste de aderência de Shapiro-Wilk (SHAPIRO; WILK, 1965), que testa se uma dada amostra foi extraída a partir de uma população normalmente distribuída. O resultado desse teste é o que determina o método utilizado para estabelecer o ponto de corte. Se ele indicar que a distribuição de frequências segue uma distribuição normal, o teste Z é utilizado, conforme Equação 4.2. Onde, X é o ponto de corte a ser encontrado, μ e σ são respectivamente a média e o desvio padrão das frequências, e Z é o valor da Tabela Z[4] para um dado nível de confiança. Caso contrário, é utilizado o teste *T-student*, conforme Equação 4.3. Na qual, $\bar{x}$ é o ponto de corte a ser encontrado, μ e s são respectivamente a média e o desvio padrão das frequências, n é o número de itens na lista (considerando conceitos únicos), e t é o valor extraído da tabela *T-student* para um dado nível de confiança e um dado grau de liberdade $dof = n - 1$. Como o termo sugere, o nível de confiança é o que determina o quão confiável é o valor de corte obtido. Quanto maior for esse nível, menor será número de conceitos acima do ponto de corte, mas maior será a probabilidade de eles realmente serem importantes para aquele argumento.

Para estabelecer o nível de confiança utilizado neste trabalhos, foram realizados três experimentos, testando os níveis de 95%, 90% e 80%. Para isso foram tomados como base os conceitos *ask.v*.01, *make.v*.01 e *buy.v*.01, que foram escolhidos aleatoriamente da lista de verbos do vocabulário controlado. E os conceitos *eat.v*.01, *drink.v*.01, por denotarem ações básicas para a comunicação funcional. Os complementos rejeitados pelo ponto de corte para cada argumento desses conceitos considerando os três níveis de confiança fo-

[3] Uma distribuição de probabilidade que consiste em uma curva simétrica em torno do seu ponto médio
[4] Tabela da distribuição normal, na qual é atribuído um valor para cada nível de confiança

ram analisados. Com isso, notou-se que com níveis de confiança mais altos (90% e 95%) complementos importantes (e.g., $person.n.01$ e $food.n.01$) eram excluídos de argumentos como Agente e Tema. Enquanto que ao nível de 80% pouco ruído era inserido nos argumentos. Assim, neste trabalho, o nível de confiança utilizado é de 80%.

$$Z = \frac{X - \mu}{\sigma} \tag{4.2}$$

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{4.3}$$

Os conceitos atribuídos a cada argumento, que têm frequência acima do ponto de corte estabelecido pelo procedimento descrito no parágrafo anterior, são submetidos a uma nova rodada de remoção de redundâncias. Nessa segunda rodada, a estratégia utilizada dá preferência aos nós mais altos de cada galho da taxonomia da GS que aparecem na lista de complementos de um dado argumento. Isto é, se conceitos do mesmo galho ($C1 \rightarrow C2 \rightarrow C3$) forem identificados como complementos de um argumento, e não tendo sido eles removidos pela primeira rodada de remoção ou pelo corte por importância, apenas o conceito que está no nó mais alto ($C1$) é considerado.

5 GRAMÁTICA SEMÂNTICA

Neste capítulo são apresentados detalhes da Gramática Semântica (GS) produto deste trabalho. Na Seção 5.1 apresenta-se uma visão geral da GS. Na Seção 5.2 a estrutura de organização da GS é apresentada. Na Seção 5.3 apresenta-se a sua avaliação automática. Por fim, na Seção 5.4 são dados detalhes de como essa GS pode ser usada em aplicações práticas de Comunicação Aumentativa e Alternativa (CAA).

5.1 VISÃO GERAL

A GS construída a partir da execução do método descrito no Capítulo 4, consiste na estrutura semântica do vocabulário controlado usado como entrada (cf. Seção 4.1.1). Essa estrutura conta com relações léxico-semânticas de hierarquia e de predicado-argumento entre os conceitos das palavras que compõem o vocabulário. As relações de hierarquia seguem o padrão da *WordNet* (MILLER, 1995) e são determinadas pelas propriedades de hiperonímia e hiponímia de cada conceito, e são herdadas da ontologia AACOnto (FRANCO, 2020). Já as relações de predicado-argumento foram adquiridas a partir das análises das amostras de textos extraídas do British National Corpus (BNC) (CONSORTIUM et al., 2007). São um total de 4295 relações, que são determinadas pelos papéis semânticos da estrutura gramatical do *Colourful Semantics* (CS) Bryan (1997).

Como abordado na Seção 2.4.2, o CS fornece uma gramática simples, na qual os *slots* são uma combinação entre papéis semânticos, cores e perguntas. Os papéis semânticos em si já carregam uma semântica que pode ajudar na construção de frases, mas a sua associação com cores e perguntas que denotam os seus conceitos, os torna mais significativos e lúdicos para crianças. Pois, as cores servem como pistas visuais, e as perguntas como pistas semânticas, que dão suporte à construção de frases com sentido e gramaticalmente corretas.

Essas pistas visuais e semânticas suportadas pela GS proposta são um diferencial importante em comparação com as GSs propostas e usadas pelos trabalhos relacionados a este (cf. Capítulo 3), pois, como abordado na Seção 2.4.2, estas ajudam no desenvolvimento das habilidades funcionais de comunicação do indivíduo. Além disso, outros pontos merecem destaque: 1) a GS proposta é baseada em um vocabulário controlado específico para CAA, que foi construído por Franco (2020) com o uso de métodos qualitativos e quantitativos; 2) as relações de predicado-argumento estabelecidas entre os conceitos da GS são extraídas a partir de amostras de textos em linguagem natural, o que faz com que a GS seja dita como baseada em evidência de *corpus*; e 3) além de fornecer as pistas visuais e semânticas, o CS, que fornece a estrutura gramatical da GS, é uma ferramenta terapêutica utilizada em contextos clínicos para ajudar indivíduos com dificuldades de

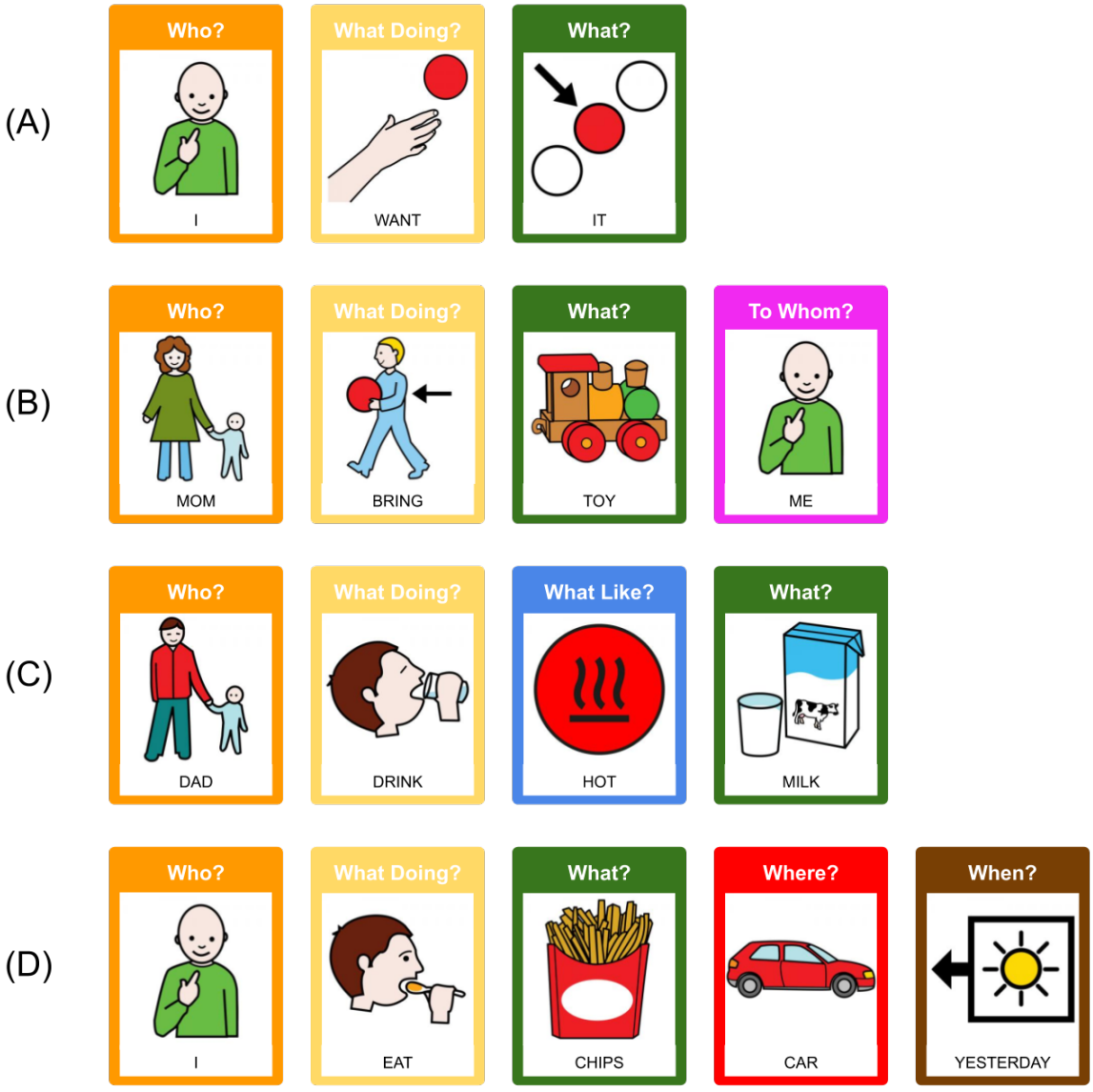
linguagem a construir e entender frases e cuja eficácia é atestada por diversos estudos (cf. Seção 2.4.2);

Exemplos de frases cuja construção é suportada pela GS proposta são apresentadas na Figura 29. O exemplo A consiste em uma frase com estrutura semântica Agente+Verbo+Tema, que é mais comum em estágios iniciais da comunicação. Esse tipo de estrutura pode ser usado para comunicar mensagens simples, como: "*I eat cake*", "*I want water*", "*Mom cook cake*", dentre outras que seguem a estrutura sintática Sujeito + Verbo + Objeto Direto. No exemplo B é adicionado um Recipiente (i.e., *To Whom?*) a essa estrutura, que tem a função sintática de objeto indireto. Já no exemplo C um modificador adjetival (i.e., *What Like?*) é utilizado para o tema (i.e., *What?*). Por fim, no exemplo D, dois complementos adverbiais são utilizados: um de local (i.e., *Where?*) e outro de tempo (i.e., *When?*). Além dessas frases, a GS dá suporte a sentenças que seguem qualquer uma das estruturas possíveis ao CS, por exemplo:

- Agente + Verbo + Maneira;
- Agente + Verbo + Local;
- Agente + Verbo + Tempo;
- Agente + Verbo + Tema + Local;
- Agente + Verbo + Tema + Tempo;
- Agente + Verbo + Tema + Local + Tempo;
- Agente + Verbo + Maneira + Tema + Local + Tempo;
- Agente + Descrição;
- Agente + Descrição + Verbo;
- Agente + Verbo + Tema + Descrição;
- Tema + Verbo + Descrição;
- Tema + Descrição;
- Agente + Verbo + Local + Descrição.

É importante destacar que as imagens utilizadas nos exemplos da Figura 29 servem apenas para ilustrar o uso da GS em um sistema de CAA. Isto é, elas não estão inseridas na GS proposta neste trabalho, que atua apenas no nível das palavras. Geralmente, a associação das palavras com imagens é feita por mediadores (i.e., pais, educadores, terapeutas), que podem usar tanto pictogramas como fotos de pessoas, objetos, animais, etc. Essa associação está além do escopo deste trabalho.

Figura 29 – Exemplo de sentenças baseadas na GS proposta



Fonte: Do autor

5.2 ESTRUTURA

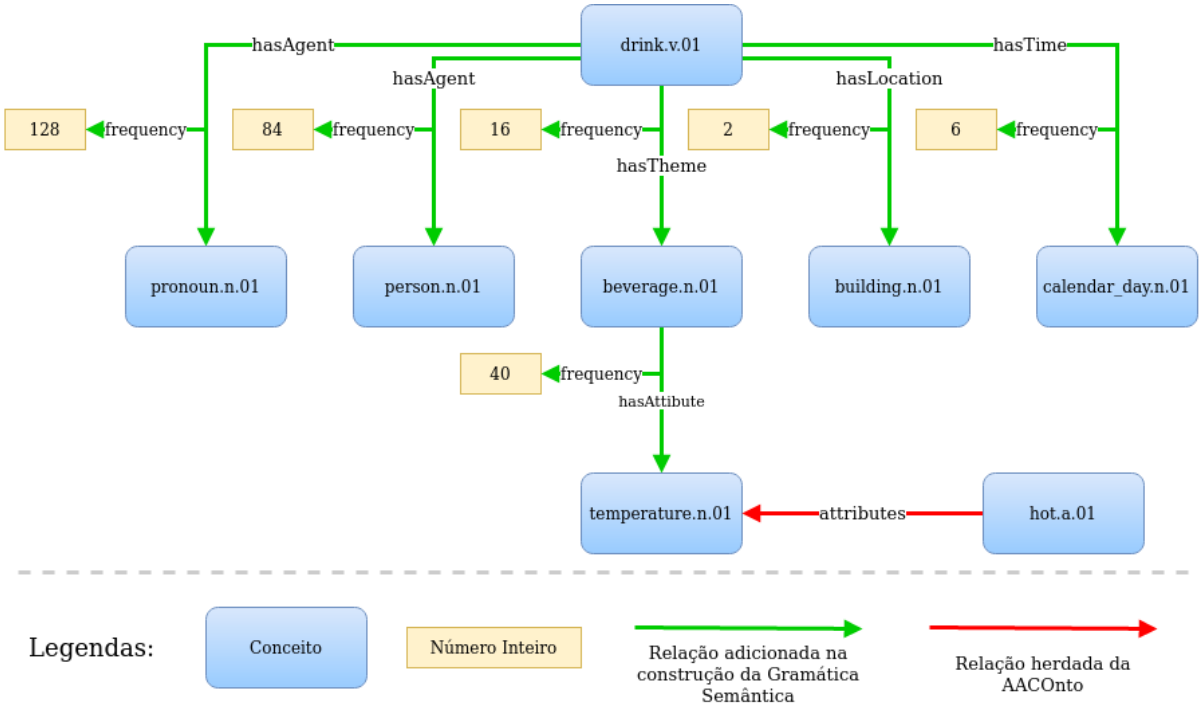
A GS proposta é representada computacionalmente em formato de ontologia, que está disponível para *download*[1] no formato TURTLE[2]. Essa ontologia segue os padrões do modelo *Lexicon Model for Ontologies* (*Ontolex-lemon*) (MCCRAE et al., 2017) para a representação de palavras e seus conceitos, que são a estrutura de base da GS. Como já mencionado, essa estrutura é herdada da AACOnto (FRANCO, 2020), que é um dos materiais dados como entradas no método apresentado no Capítulo 4 (cf. Seção 4.1.1). É a partir dos conceitos

[1] https://bit.ly/2OUlKxC

[2] Uma sintaxe e um formato de arquivo para expressar dados no modelo de dados RDF (Resource Description Framework).

dessa estrutura que são estabelecidas as relações de predicado-argumento da GS. Essas relações são representadas por propriedades inspiradas nos padrões do *Predicate Model for Ontologies* (PreMOn) (cf. Seção 2.2), mas que seguem a nomenclatura dos papéis semânticos utilizados na GS (cf. Seção 4.1.3). São elas: *hasAgent*, *hasTheme*, *hasRecipient*, *hasManner*, *hasLocation*, *hasTime* e *hasAttribute*. Essas propriedades têm como domínio os conceitos denotados por predicados (i.e., ações/verbos e objetos/substantivos), e como imagem os conceitos que são complementos de cada argumento desses predicados. Além disso, elas têm cardinalidade de 1 para n, ou seja, mais de um complemento pode ser atribuído para cada propriedade em cada conceito de predicado. Na Figura 30 apresenta-se um trecho da GS, no qual é possível observar o uso dessas propriedades. Nesse trecho, o conceito do verbo *drink* (i.e., beber) aponta seus agentes, tema, e complementos de lugar e tempo. Ao apontar *beverage.n.01* como tema, por exemplo, a GS inclui todos os outros conceitos que estão abaixo deste na hierarquia (e.g., *milk.n.01* e *alcohol.n.01*). Além disso, o conceito *beverage.n.01* aponta um de seus atributos: *temperature.n.01*. Isso indica que bebidas tem um atributo que pode ser preenchido por um adjetivo de temperatura (e.g., *hot.a.01*).

Figura 30 – Trecho da Gramática Semântica



Fonte: Do autor

As frequências com que cada conceito aparece para cada argumento também são representadas na ontologia. Para isso é usada a propriedade de anotação *frequency*, que recebe um número inteiro. Como se trata de propriedades atribuídas a propriedades, essa anotação é feita através de axiomas, como o mostrado na Figura 31. Essa informação pode ser útil para ordenar os complementos por frequência quando buscas forem realiza-

Figura 31 – Exemplo de axioma que define frequência

```
[ rdf:type owl:Axiom ;
   owl:annotatedSource :drink.v.01 ;
   owl:annotatedProperty :hasAgent ;
   owl:annotatedTarget :pronoun.n.01 ;
   :frequency "128"
 ] .
```

Fonte: do autor

das na ontologia. Especialmente em casos nos quais dois ou mais conceitos aparecem como complementos de único argumento, como no caso dos agentes do verbo *drink* ilustrado na Figura 30. Além disso, essa propriedade pode ser usada em sistemas de CAA para registrar a frequência de uso de palavras como argumentos de um dado predicado. O que é útil não só para a ordenação de palavras, mas também para fins de avaliação do uso da linguagem por parte do usuário.

5.3 AVALIAÇÃO

Nesta seção é apresentado o método utilizado na avaliação da GS, assim como os resultados obtidos.

5.3.1 Método

Segundo Raad e Cruz (2015), existem quatro abordagens para avaliação de ontologias: 1) baseada em "padrão de ouro" – consiste em comparar a ontologia construída com uma ontologia de referência previamente criada; 2) baseada em *corpus* – usada para avaliar até que ponto uma ontologia cobre suficientemente um determinado domínio; 3) baseada em tarefa – usada para medir até que ponto uma ontologia ajuda a melhorar os resultados de uma determinada tarefa; e 4) baseada em critérios – que mede até que ponto uma ontologia ou taxonomia adere a certos critérios desejáveis, que podem ser relacionados a sua estrutura (e.g., consistência, densidade relacional, etc.) ou a questões mais filosóficas (e.g., essência, identidade e unidade). A abordagem usada neste trabalho é a baseada em tarefa, que considera que a ontologia a ser avaliada é direcionada a uma tarefa específica. Essa decisão considera o fato de que a ontologia que representa a GS é direcionada à tarefa específica de auxiliar usuários de sistemas de CAA na construção de frases compreensível e com sintaxe adequada. Assim, é avaliada a sua capacidade de prover aos usuários as

alternativas corretas de palavras para preencher os *slots* do CS, para assim construir frases telegráficas com sentido.

Para dar suporte a essa avaliação são extraídas sentenças do conjunto de *corpora* do *Child Language Data Exchange System* (CHILDES) (MACWHINNEY, 2014), seguindo os seguintes critérios: 1) A frase precisa ter sido falada por uma criança, dado que o CHILDES conta com falas de crianças, pais, terapeutas e outros parceiros de comunicação; 2) Todas as palavras da frase, excluindo preposições, artigos e conjunções, precisam estar presentes no vocabulário controlado que define a GS; e 3) As sentenças devem seguir um dos tipos de estruturas gramaticais utilizados originalmente por Bryan (1997) no CS, que consistem em construções simples e geralmente utilizadas por crianças em estágios iniciais de comunicação:

- **Tipo 1**: Agente + Verbo + Tema (e.g. Eu comi bolo);
- **Tipo 2**: Agente + Verbo + Local (e.g. Eu passeio no parque);
- **Tipo 3**: Agente + Verbo + Tema + Local (e.g. Eu comi bolo na escola);
- **Tipo 4**: Tema + Verbo + Descrição (e.g Eu sou rápido).

A construção de cada uma dessas frases é simulada utilizando como base o conhecimento presente na GS. Para isso, para cada frase da lista extraída do CHILDES são executados os seguintes passos:

1. **Pré-processamento** – Como se trata da reconstrução de frases telegráficas, as sentenças selecionadas precisam passar por um pré-processamento para a remoção de preposições, artigos, conjugação de verbos, etc. Para isso, são utilizadas as informações morfossintáticas de classe gramatical e *lemma* (i.e., forma básica da palavra), presentes no CHILDES. A classe gramatical ajuda a identificar se a palavra é uma palavra-chave para a sentença. São consideradas palavras-chave apenas pronomes, verbos, substantivos, advérbios e adjetivos. Já o *lemma* é utilizado para remover conjugação de verbos e transformar palavras no plural em singular. Assim, esse pré-processamento transforma frases expandidas (e.g., *I ate cookies at school*) em frases telegráficas (e.g., *I eat cookie school*);

2. **Identificação dos conceito das palavras** – Essa identificação se dá pela busca do conceito evocado por cada palavra da frase na GS. Para isso, são considerados o *lemma* e a classe gramatical das palavras. Uma vez que, algumas destas podem pertencer a mais de uma classe. Por exemplo, a palavra *fish* que pode ser um verbo (i.e., pescar) ou um substantivo (i.e., peixe). Tomando como exemplo a frase "*I eat cookie school*", os conceitos identificados são: *pronoun.n*.01 *eat.v*.01 *cookie.n*.01 *school.n*.02;

3. **Identificação do predicado** – Essa identificação é feita por meio da análise das classes gramaticais das palavras. Para sentenças que seguem as estruturas dos tipos 1, 2 e 3, a identificação do predicado consiste em identificar o verbo da frase. Por exemplo, na sentença "*I eat cookie school*" que é do tipo 3, o predicado é o verbo "*eat*", que tem o conceito *eat.v*.01. Já para sentenças do tipo 4, além do predicado verbal, há um predicado nominal (i.e., Tema), que tem como argumento o papel semântico Descrição. Assim, se faz necessário identificar não só o verbo, mas também o substantivo ou pronome que desempenha o papel de Tema;

4. **Preenchimento dos argumentos do predicado** – consiste em testar se as relações de predicado-argumento existentes na frase de referência são cobertas pela GS proposta. Por exemplo, a frase "*I eat cookie school*" tem o verbo "*eat*" como predicado, que tem os argumentos: "*I*" (Agente), "*cookie*" (Tema) e "*school*" (Local). O que é feito é testar, por exemplo, se a relação de Agente existe entre o conceito do predicado (i.e., *eat.v*.01) e o conceito da palavra "*I*" (i.e., *pronoun.n*.01). Se a relação existir, a palavra "*I*", é inserida como Agente em uma nova frase. O mesmo acontece com todas as outras palavras da sentença.

A execução dos passos listados acima cria dois conjuntos de sentenças: (i) de referência – que são as frases extraídas do CHILDES e transformadas em telegráficas; e (ii) candidatas – que são as frases reconstruídas a partir das buscas na GS. As frases candidatas são comparadas às de referência com base em um *score* de precisão modificado, baseado na métrica *Bilingual Evaluation Understudy* (BLEU) (PAPINENI et al., 2002). Esse *score* é calculado com base na Equação 5.1, na qual $candidata_t$ e $referencia_t$ são o número total de palavras nas frases candidatas e de referência, respectivamente.

$$P = \frac{candidata_t}{referencia_t} \tag{5.1}$$

5.3.2 Resultados

Na Tabela 9 apresenta-se o resumo dos resultados obtidos a partir da execução do método apresentado na sub-seção anterior, com a quantidade total de sentenças para cada tipo de frase, assim como a precisão média obtida em cada tipo. Além disso, é apresentado o número de sentenças que foram totalmente reconstruídas (100%) e o número daquelas na qual a precisão da reconstrução foi maior ou menor que 50%. Os resultados mostram que a GS proposta neste trabalho pode dar suporte à reconstrução de frases telegráficas compreensíveis com uma precisão média de 90%, considerando um total de 1246 sentenças extraídas do CHILDES. Isso significa que a probabilidade de um usuário de um sistema de CAA que utiliza a GS como base de dados encontrar os argumentos corretos para os predicados verbais durante a construção de frases é de 90%.

Tabela 9 – Resumo dos resultados da avaliação automática

Tipo de frase	n	Precisão média	100%	>50%	<50%
Agente + Verbo + Tema	1162	0,91	864	285	13
Agente + Verbo + Local	42	0,80	22	13	5
Agente + Verbo + Tema + Local	23	0,86	10	13	0
Tema + Verbo + Descrição	19	0,67	0	19	0
TOTAL	1246	0,90	896	332	18

Em relação aos tipos de estrutura de frases reconstruídas, os três primeiros tiveram uma precisão média igual ou maior que 80%. O tipo 1 (Agente + Verbo + Tema) obteve o maior número de sentenças (i.e., 1162). Isso se dá pelo fato que ele segue uma estrutura comumente utilizada por crianças em estágios iniciais de comunicação. Exemplos de frases telegráficas que seguem essa estrutura são: "*I want it*" e "*Baby need bed*". A GS proposta pode dar suporte à construção de sentenças desse tipo com uma precisão média de 91%. Já para as frases do tipo 2 (Agente + Verbo + Local), a GS dá suporte à construção das 42 sentenças selecionadas do CHILDES com uma precisão média de 80%. Exemplos de frases desse tipo são: "*I eat home*"; "*I go kitchen*"; dentre outras. Para o tipo 3 (Agente + Verbo + Tema + Local), a GS obteve a precisão média de 86%. Esse tipo conta com sentenças como: "*I eat them home*"; "*I hear it car*"; etc. Já nas sentenças do tipo 4 (Tema + Verbo + Descrição) a precisão obtida é consideravelmente menor do que nos outros tipos (67%). Isso se explica por dois pontos: 1) apesar de o papel semântico Descrição estar logo após o verbo, ele está relacionado ao Tema ao invés do verbo. Um exemplo desse tipo de frase seria "*I feel sick*". Onde o adjetivo "*sick*" é atributo de "*I*"; e 2) a maioria das frases selecionadas tinham um pronome como Tema, e nenhuma relação entre o conceito de pronomes (i.e., *pronoun.n.01*) e um atributo foi estabelecida na GS.

Das 1246 sentenças extraídas do CHILDES para a avaliação, a GS proposta foi capaz de dar suporte à reconstrução total de 896 (71,9%). Exemplos dessas frases reconstruídas totalmente são: "*He want book*", "*I go kitchen*", "*We have it home*", dentre outras. A reconstrução total significa que as relações de predicado argumento existentes na frase foram também encontradas na GS. Por exemplo, na frase telegráfica "*duck take bath*", as palavras evocam os conceitos *duck.n.*01, *take.v.*01 e *bath.n.*02, respectivamente, dos quais *take.v.*01 é o conceito evocado pelo predicado verbal da frase. Por seguir a estrutura Agente + Verbo + Tema, as relações presentes na frase são *take.v.*01 *hasAgent duck.n.*01 e *take.v.*01 *hasTheme bath.n.*02. Relações essas que também existem na GS.

As frases que não foram totalmente reconstruídas, mas que tiveram precisão igual ou maior a 50% somam 332 (26,6%). E as que tiveram precisão menor que 50% somam 18 (1,4%). Essas 18 frases são apresentadas na Tabela 10, na qual são mostrados os tipos de cada frase, as sentenças originais (i.e., de referência), as suas versões telegráficas, as

Tabela 10 – Frases não reconstruídas

Tipo	**Frase original**	**Frase Telegráfica**	**Frase reconstruída**	***Score***
1	ball sit ball	ball sit ball	sit	0,33
1	bed sit the bed	bed sit bed	sit	0,33
1	bird get way	bird get way	get	0,33
1	blackbirds eat apple	black eat apple	eat	0,33
1	box have a box	box have box	have	0,33
1	cheese say cheese	cheese say cheese	say	0,33
1	doll have a hole	doll have hole	have	0,33
1	doll want the milk	doll want milk	want	0,33
1	egg find egg	egg find egg	find	0,33
1	he sit car	he sit car	sit	0,33
1	my dinner want my dinner	dinner want dinner	want	0,33
1	my money see my money	money see money	see	0,33
1	train come minute	train come minute	come	0,33
2	I sit in a car	I sit car	sit	0,33
2	I sit in back	I sit back	sit	0,33
2	I sit in bath	I sit bath	sit	0,33
2	I sit in chair	I sit chair	sit	0,33
2	I sit on seat	I sit seat	sit	0,33

frases reconstruídas (i.e., candidatas) e o *score* de precisão obtido. Note-se que apenas os verbos foram identificados e inseridos nas sentenças candidatas e nenhum argumento foi preenchido. Isso se dá devido à ausência de relações de predicado-argumento entre os verbos dessas frases e as demais palavras na GS. Por exemplo, o verbo *sit* na GS tem como agente apenas o conceito *person.n.01*, que, considerando os seus hipônimos, é evocado por palavras como *doctor*, *brother*, *dad*, etc. Contudo, nas frases de referência os agentes que aparecem para esse verbo são pronomes (i.e., *I* e *he*) ou objetos (i.e., *ball* e *bed*). É preciso considerar ainda que algumas das frases extraídas do CHILDES não fazem sentido, o que atrapalha na identificação de argumentos. Por exemplo, a frase "*egg find egg*", na qual a mesma palavra aparece como agente e como, enquanto só faz sentido como tema.

Esses resultados demonstram que GS proposta pode dar o suporte necessário para a construção de frases com sentido. No entanto, ainda há pontos que precisam ser melhor trabalhados. Por exemplo, algumas das sentenças que não foram reconstruídas apresentam relações de predicado-argumento que podem ser importantes para a comunicação de um usuário de CAA. A sentença "*black eat apple*", por exemplo, apesar do erro na lematização da palavra *blackbirds*, aponta a relação que deveria existir entre os conceitos do verbo *eat* e do substantivo *apple* ou de um de seus hiperônimos (e.g., *fruit*). Sendo assim, a avaliação da GS serve não só para testar a sua expressividade, mas também para dar dicas de como

melhorar o conhecimento nela representado. No entanto, se faz necessário ainda investigar a eficácia de GS em ambientes reais de uso de CAA. Nesses ambientes é possível coletar informações mais precisas sobre a sua capacidade em ajudar na construção e de frases com sentido, assim como investigar a influência das pistas visuais nesse processo.

5.4 DIRETRIZES DE USO: COMO OUTRAS PESSOAS PODEM FAZER USO DESTE TRABALHO

A GS proposta neste trabalho é direcionada a sistemas de CAA de alta tecnologia. O seu uso pode beneficiar três grupos de pessoas: 1) os desenvolvedores de sistemas de CAA; 2) os usuários desses sistemas; e 3) seus parceiros de comunicação (i.e., familiares, educadores, terapeutas, etc.).

A GS proposta pode ser usada por desenvolvedores como base de dados para sistemas de CAA. Essa base pode ser utilizada de duas maneiras: 1) como um suporte para a construção de frases; e 2) como uma fonte de conhecimento para a análise semântica de frases telegráficas. A GS proposta é concebida com o intuito de dar suporte à construção de frases telegráficas com sentido. Uma vez que ela é baseada em um sistema de pistas visuais (i.e., cores) e semânticas (i.e., perguntas) (cf. Seção 2.4.2) para servir como reforço durante a criação de frases. Por isso, é mais adequado que ela seja usada para este fim. Nesse contexto, a GS serve como uma base de conhecimento que pode ser usada na predição e sugestão de palavras. Por exemplo, se durante a construção de uma frase o usuário preencher o *slot Who?* (i.e., Quem?) com a palavra "*boy*", a GS é capaz de dar o suporte necessário ao preenchimento dos outros *slots* (e.g., *What Like?*, *What Doing?*), através da sugestão de palavras. Contudo, o conhecimento léxico-semântico contido na GS proposta pode ainda ser usado para a análise de frases telegráficas previamente construídas. O que pode ser útil para sistemas de CAA nos quais os usuários são livres para construir sentenças em que as palavras não seguem uma ordem preestabelecida (e.g., *yesterday eat cake I*). Considerando que as palavras que constituem essas frases fazem parte do vocabulário controlado que define a GS, é possível identificar o conceito de cada uma dessas palavras, verificar quais delas podem ser predicados e complementos.

São esses sistemas de CAA, que eventualmente usem a GS, que a fazem útil para pessoas com dificuldades na fala, assim como para os parceiros de comunicação. Como abordado na Seção 2.4.2, o uso de pistas semânticas e visuais ajuda no desenvolvimento das habilidades de comunicação de crianças. Assim, a implementação da GS em um sistema de CAA é útil para: 1) possibilitar a construção de mensagens com sentido; e 2) ajudar o usuário a desenvolver suas habilidades funcionais de comunicação. Além disso, a construção de frases com sentido e sintática e semanticamente corretas favorece o entendimento dos parceiros de comunicação. Contudo, é preciso destacar que ainda é necessário avaliar o impacto do uso desse tipo de aplicação em cenários reais de CAA (cf. Seção 6.4).

6 CONCLUSÃO

Este capítulo tem como objetivo apresentar as considerações finais sobre os principais tópicos abordados neste trabalho, incluindo as contribuições alcançadas e indicações para trabalhos futuros.

6.1 CONSIDERAÇÕES FINAIS

Este trabalho de mestrado propõe uma Gramática Semântica (GS) para Comunicação Aumentativa e Alternativa (CAA), que se baseia em um sistema de pistas visuais (i.e., cores) e semânticas (i.e., perguntas), como uma ferramenta a ser usada como base para a construção de frases compreensíveis em sistemas de CAA. O sistema de pistas utilizado como base é o *Colourful Semantics* (CS) (cf. Seção 2.4.2), que fornece uma estrutura gramatical na qual os componentes (i.e., *slots*) são associados a cores e perguntas como "Quem?" e "Fazendo o quê?". Além disso, um vocabulário controlado voltado ao domínio de CAA e construído por métodos quantitativos e qualitativos é usado como base para a estrutura hierárquica de conceitos da GS. A estrutura gramatical (gramática) foi adicionada à estrutura hierárquica (semântica), por meio da extração de relações de predicado-argumento a partir de amostras de textos extraídas de um *corpus* linguístico. Essa extração deu-se a partir do uso de técnicas de Processamento de Linguagem Natural (PLN) que envolvem a Rotulação de Papéis Semânticos (RPS), Reconhecimento de Entidades Nomeadas (REN), Desambiguação de Sentido Lexical (DSL) e a Análise de Dependência (AD).

A GS proposta está computacionalmente representada por uma ontologia, que foi avaliada através de uma abordagem de avaliação baseada em tarefa. Essa avaliação consistiu em medir quão eficiente é a GS na tarefa de dar suporte à construção de frases telegráficas com sentido. Para isso, foram extraídas sentenças do *Child Language Data Exchange System* (CHILDES), as quais foram transformadas em frases telegráficas, removendo preposições, conjunções, artigos e conjugação de verbos. Depois, sua reconstrução foi simulada a partir do conhecimento representado na GS. A precisão média obtida na reconstrução das 1246 frases extraídas do CHILDES foi de 90%. Esse valor foi obtido por meio da métrica *Bilingual Evaluation Understudy* (BLEU), que compara os n-gramas das frases de referências com os n-gramas das sentenças reconstruídas. Além disso, a GS foi capaz de dar suporta à reconstrução total de 71,9% das frases testadas.

Com isso, as três Perguntas de Pesquisa (PP) feitas por este trabalho (c.f. Seção 1.2) foram respondidas. A PP-1 (Como GSs são usadas em sistemas de CAA?) foi respondida pela revisão da literatura apresentada no Capítulo 3, que mostra os principais trabalhos relacionados a este e como eles constroem e utilizam GSs no contexto de CAA. O Capítulo 4 responde a PP-2 (Como utilizar um sistema de pistas visuais e semânticas para construir

uma GS para o domínio de CAA?), mostrando como o sistema de pistas é utilizado para a construção da GS. Já a PP-3 (Qual é a expressividade da GS proposta para construir frases telegráficas com sentido?) é respondida pela a avaliação automática apresentada na Seção 5.3.

6.2 PRINCIPAL CONTRIBUIÇÃO

A principal contribuição deste trabalho de mestrado é a Gramática Semântica proposta. Ela é construída com base no sistema de pistas visuais e semânticas do *Colourful Semantics* e em um vocabulário controlado específico para sistemas de CAA. Além disso, pelo conhecimento linguístico que ela representa ter sido extraído de amostras de textos, essa GS é uma base de conhecimento baseada em evidência de *corpus*. Ou seja, o que está representado nela é a forma com a qual as palavras se comportam em linguagem natural. A Seção 5.4 dá uma visão geral de como essa base de conhecimento pode ser usada em sistemas de CAA para dar suporte à construção de frases compreensíveis e gramaticalmente corretas.

6.3 LIMITAÇÕES

Contudo, é necessário considerar as limitações deste trabalho:

- A ausência de relações predicado-argumento na GS – essas relações são extraídas de amostras de textos, que podem não necessariamente representar todas as formas de comportamento de uma palavra. Assim, pode ser que nem todos os relacionamentos entre as palavras do vocabulário de entrada sejam inseridos na GS;
- O viés inserido pelo *corpus* utilizado – que é comum a qualquer *corpus*, pois são construídos a partir da transcrição de diálogos ou de documentos de textos. No caso do British National Corpus (BNC), que é utilizado neste trabalho, o fato de ele ser uma compilação da língua inglesa britânica pode ser um fator de interferência. Além disso, a linguagem desse *corpus* é, na maior parte, não coloquial, o que pode implicar na ausência de relações entre palavras que podem ser usadas por crianças e em contextos informais de comunicação;
- O erro inserido pelos analisadores semânticos e sintáticos – o SLING, por exemplo, que é usado como analisador semântico, reporta apenas 87,46% de precisão na rotulação de papéis semânticos. Isso significa que existe 12,54% de probabilidade de as relações predicado-argumento inseridas na GS estarem erradas ou ausentes; e
- A ausência de avaliação em contextos reais de CAA – A GS proposta não foi avaliada por seres humanos em situações reais de utilização de sistemas de CAA.

6.4 TRABALHOS FUTUROS

Para dar continuidade ao trabalho de pesquisa descrito nesta dissertação, lista-se, nesta seção, propostas de trabalhos futuros a serem realizadas:

- Replicar o método utilizado na construção da GS para construir uma base semelhante na qual as palavras sejam em língua portuguesa. Isso requer o levantamento de recursos de PLN de língua portuguesa, como: uma base de conhecimento lexical semelhante à *WordNet*; um sistema analisador semântico; um sistema identificador de entidades nomeadas; um sistema de desambiguação do sentido lexical; e um sistema para análise de dependência. Além de documentos de textos que possam ser usados como fonte de treinamento;
- Implementar a GS construída em um sistema de CAA real;
- Avaliar a aceitação do sistema de CAA que implementa a GS junto a terapeutas e pessoas com dificuldades na fala que usam CAA; e
- Adicionar à GS informações morfossintáticas (e.g., formas, padrões morfológicos, gênero, número, grau, etc.) para serem usadas na expansão (i.e., adição de preposições, conjunções, etc.) de frases telegráficas;

REFERÊNCIAS

ALLAN, K. Natural language semantics. 2001.

ASHA. *Augmentative and Alternative Communication*. ASHA, 2019. Disponível em: <https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589942773\T1\ textsectionion>.

ASHA. *Definition of Communication and Appropriate Targets*. ASHA, 2019. Disponível em: <https://www.asha.org/NJC/ Definition-of-Communication-and-Appropriate-Targets/>.

ASTON, G.; BURNARD, L. *The BNC handbook: exploring the British National Corpus with SARA*. [S.l.]: Capstone, 1998.

AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. Dbpedia: A nucleus for a web of open data. In: *The semantic web*. [S.l.]: Springer, 2007. p. 722–735.

BABKO-MALAYA, O. Propbank annotation guidelines. *URL: http://verbs. colorado. edu*, 2005.

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.

BALANDIN, S.; IACONO, T. Crews, wusses, and whoppas: Core and fringe vocabularies of australian meal-break conversations in the workplace. *Augmentative and Alternative Communication*, Taylor & Francis, v. 15, n. 2, p. 95–109, 1999.

BANAJEE, M.; DICARLO, C.; Buras Stricklin, S. Core Vocabulary Determination for Toddlers. *Augmentative and Alternative Communication*, v. 19, n. 2, p. 67–73, 2003. ISSN 0743-4618. Disponível em: <http://informahealthcare.com/doi/abs/10.1080/ 07434610310001120 34>.

BANERJEE, S.; PEDERSEN, T. Extended gloss overlaps as a measure of semantic relatedness. In: *Ijcai*. [S.l.: s.n.], 2003. v. 3, p. 805–810.

BEUKELMAN, D.; JONES, R.; ROWAN, M. Frequency of word usage by nondisabled peers in integrated preschool classrooms. *Augmentative and Alternative Communication*, Taylor & Francis, v. 5, n. 4, p. 243–248, 1989.

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. Disponível em: <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.

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, 1997.

BURTON, R. R. Semantic grammar: An engineering technique for constructing natural language understanding systems. ERIC, 1976.

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.

CARVALHO, S. M. P. A terminological approach to knowledge organization within the scope of endometriosis: the endoterm project. 2018.

Centro de Estudos sobre Alfabetização e Deficiência da Escola de Medicina da Universidade da Carolina do Norte. *DLM Core Vocabulary*. 2018. Disponível em: <http://www.med.unc.edu/ahs/clds/resources/core-vocabulary>.

CHARNIAK, E.; JOHNSON, M. Coarse-to-fine n-best parsing and maxent discriminative reranking. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 43rd annual meeting on association for computational linguistics*. [S.l.], 2005. p. 173–180.

CHEN, L.; BABAR, M. A.; ZHANG, H. Towards an evidence-based understanding of electronic data sources. 2010.

CHIU, J. P.; NICHOLS, E. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 4, p. 357–370, 2016.

CHOI, J. D.; TETREAULT, J.; STENT, A. It depends: Dependency parser comparison using a web-based evaluation tool. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. [S.l.: s.n.], 2015. p. 387–396.

CHOMSKY, N. *Knowledge of language: Its nature, origin, and use*. [S.l.]: Greenwood Publishing Group, 1986.

CHOMSKY, N. *Lectures on government and binding: The Pisa lectures*. [S.l.]: Walter de Gruyter, 1993.

CHOMSKY, N. *Aspects of the Theory of Syntax*. [S.l.]: MIT press, 2014. v. 11.

CHOMSKY, N. *The minimalist program*. [S.l.]: MIT press, 2014.

CHOMSKY, N. et al. *Language and problems of knowledge: The Managua lectures*. [S.l.]: MIT press, 1988. v. 16.

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, 10 may 2016. *Community Report, W3C*, 2016.

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.

COLLINS, M. Head-driven statistical models for natural language parsing. *Computational linguistics*, MIT Press, v. 29, n. 4, p. 589–637, 2003.

CONSORTIUM, B. et al. The british national corpus, version 3 (bnc xml edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium*, v. 5, n. 65, p. 6, 2007.

COOK, A. M.; POLGAR, J. M. *Assistive Technologies-E-Book: Principles and Practice*. [S.l.]: Elsevier Health Sciences, 2014.

CORCOGLIONITI, F.; ROSPOCHER, M.; APROSIO, A. P.; TONELLI, S. PreMOn: a lemon extension for exposing predicate models as linked data. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 877–884. Disponível em: <https://www.aclweb.org/anthology/L16-1141>.

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.

DOWTY, D. Thematic proto-roles and argument selection. *language*, Linguistic Society of America, v. 67, n. 3, p. 547–619, 1991.

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.

DYER, C.; BALLESTEROS, M.; LING, W.; MATTHEWS, A.; SMITH, N. A. Transition-based dependency parsing with stack long short-term memory. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015. p. 334–343. Disponível em: <https://www.aclweb.org/anthology/P15-1033>.

ELSAHAR, Y.; HU, S.; BOUAZZA-MAROUF, K.; KERR, D.; MANSOR, A. Augmentative and alternative communication (aac) advances: A review of configurations for individuals with a speech disability. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 19, n. 8, p. 1911, 2019.

FAN, J.-W.; FRIEDMAN, C. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. *Journal of biomedical informatics*, Elsevier, v. 44, n. 5, p. 805–814, 2011.

FILLMORE, C. J. The case for case. ERIC, 1967.

FILLMORE, C. J. Some problems for case grammar. Ohio State University. Department of Linguistics, 1971.

FILLMORE, C. J. The case for case reopened. *Syntax and semantics*, v. 8, n. 1977, p. 59–82, 1977.

FILLMORE, C. J. et al. Frame semantics. *Cognitive linguistics: Basic readings*, Berlin & New York: Mouton de Gruyter, v. 34, p. 373–400, 2006.

FINKEL, J. R.; MANNING, C. D. Joint parsing and named entity recognition. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2009. p. 326–334.

FIORELLI, M.; STELLAT, A.; LORENZETTI, T.; TURBATI, A.; SCHMITZ, P.; FRANCESCONI, E.; HAJLAOUI, N.; BATOUCHE, B. *Towards OntoLex-Lemon editing in VocBench 3*. [S.l.]: AIDAinformazioni, 2018.

FITZGERALD, E. *Straight language for the deaf: a system of instruction for deaf children*. [S.l.]: Volta Bureau, 1949.

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.

FRANCO, N. d. M.; LIMA, A. L. de; LIMA, T. P.; SILVA, E. A. da; LIMA, R. J. de; FIDALGO, R. do N. A recall analysis of core word lists over children's utterances for augmentative and alternative communication. In: IEEE. *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on*. [S.l.], 2017. p. 278–283.

FRANCO, N. de M. Vocabulary selection and organization for augmentative and alternative communication of children with speech impairment (unpublished phd's thesis). Universidade Federal de Pernambuco, 2020.

FRANENET. *Current Project Status*. 2020. Disponível em: <https://framenet.icsi.berkeley.edu/fndrupal/current_status>.

FREITAS, F. L. G. de. Ontologias e a web semântica. *Jornada de Mini-Cursos em Inteligência Artificial, SBC*, v. 8, 2003.

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. Disponível em: <http://informahealthcare.com/doi/abs/10.1080/07434619212331276033>.

GARCIA, F. H. A.; MANSUR, L. L. Habilidades funcionais de comunicação. *Acta fisiátrica*, v. 13, n. 2, p. 87–89, 2006.

GILDEA, D.; JURAFSKY, D. Automatic labeling of semantic roles. *Computational linguistics*, MIT Press, v. 28, n. 3, p. 245–288, 2002.

GOOSSENS, C.; CRAIN, S.; ELDER, P. Engineering the classroom environment for interactive symbolic communication. In: *Birmingham, AL, USA: Southeast Augmentative Communication Conference Publications*. [S.l.: s.n.], 1994.

GRISHMAN, R.; SUNDHEIM, B. M. Message understanding conference-6: A brief history. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. [S.l.: s.n.], 1996.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, Elsevier, v. 5, n. 2, p. 199–220, 1993.

GRUZITIS, N.; DANNÉLLS, D. A multilingual framenet-based grammar and lexicon for controlled natural language. *Language Resources and Evaluation*, Springer, v. 51, n. 1, p. 37–66, 2017.

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.

GUARINO, N. Formal ontology in information systems. In: *Proceedings of the first international conference (FOIS'98)*. [S.l.: s.n.], 1998.

GUIZZARDI, G. Uma abordagem metodológica de desenvolvimento para e com reuso, baseada em ontologias formais de domínio. *Programa de Mestrado em Informática–Universidade Federal do Espírito Santo, Vitória, http://wwwhome. cs. utwente. nl/˜ guizzard/MSc*, 2000.

HALLIDAY, M. A. K.; MATTHIESSEN, C.; HALLIDAY, M. *An introduction to functional grammar*. [S.l.]: Routledge, 2014.

HARPRING, P. *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. [S.l.]: Getty Publications, 2010.

HAWKINS, P.; NETTLETON, D. Large scale wsd using learning applied to senseval. *Computers and the Humanities*, Springer, v. 34, n. 1-2, p. 135–140, 2000.

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.

HETTIARACHCHI, S. The effectiveness of colourful semantics on narrative skills in children with intellectual disabilities in sri lanka. *Journal of intellectual disabilities : JOID*, 06 2015.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, v. 9, p. 1735–80, 12 1997.

HOVY, E.; MARCUS, M.; PALMER, M.; RAMSHAW, L.; WEISCHEDEL, R. Ontonotes: the 90% solution. In: *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*. [S.l.: s.n.], 2006. p. 57–60.

ISAAC, A.; SUMMERS, E. Skos simple knowledge organization system primer. *Working Group Note, W3C*, 2009.

JAWORSKI, W.; PRZEPIÓRKOWSKI, A. Semantic roles in grammar engineering. In: *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*. [S.l.: s.n.], 2014. p. 81–86.

JURAFSKY, D. *Speech & language processing*. [S.l.]: Pearson Education India, 2000.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2019. ISBN 9780131873216 0131873210. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

KALDOR, C.; ROBINSON, P.; TANNER, J. Turning on the spotlight. *Speech and Language Therapy in Practice*, 2001.

KHAN, A. Towards the representation of etymological data on the semantic web. *Information*, Multidisciplinary Digital Publishing Institute, v. 9, n. 12, p. 304, 2018.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007.

KÜBLER, S.; MCDONALD, R.; NIVRE, J. Dependency parsing. *Synthesis lectures on human language technologies*, Morgan & Claypool Publishers, v. 1, n. 1, p. 1–127, 2009.

KUPTABUT, S.; NETISOPAKUL, P. Event extraction using ontology directed semantic grammar. *J. Inf. Sci. Eng.*, v. 32, n. 1, p. 79–96, 2016.

LAW, J.; LEE, W.; ROULSTONE, S.; WREN, Y.; ZENG, B.; LINDSAY, G. 'what works': interventions for children and young people with speech, language and communication needs. Department for Education, 2012.

LEA, J. A language scheme for children suffering from receptive aphasia. *Speech Pathology and Therapy*, v. 8, n. 2, p. 58, 1965.

LEAMAN, R.; GONZALEZ, G. Banner: an executable survey of advances in biomedical named entity recognition. In: *Biocomputing 2008*. [S.l.]: World Scientific, 2008. p. 652–663.

LESK, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: CITESEER. *Proceedings of the 5th annual international conference on Systems documentation*. [S.l.], 1986. p. 24–26.

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.

MACWHINNEY, B. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. [S.l.]: Psychology Press, 2014.

MALLERY, J. C. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. In: CITESEER. *Master's thesis, MIT Political Science Department*. [S.l.], 1988.

MARCUS, M.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank. 1993.

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.

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. Disponível em: <http://informahealthcare.com/doi/abs/10.1080/07434619412331276930>.

MCCOY, K. F.; PENNINGTON, C. A.; BADMAN, A. L. Compansion: From research prototype to practical integration. *Natural Language Engineering*, Cambridge University Press, v. 4, n. 1, p. 73–95, 1998.

MCCRAE, J. P.; BOSQUE-GIL, J.; GRACIA, J.; BUITELAAR, P.; CIMIANO, P. The ontolex-lemon model: development and applications. In: *Proceedings of eLex 2017 conference*. [S.l.: s.n.], 2017. p. 19–21.

MCGINNIS, J.; BEUKELMAN, D. Vocabulary requirments 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. Disponível em: <http://informahealthcare.com/doi/pdf/10.1080/07434618912331275186>.

MERLO, P.; MUSILLO, G. Semantic parsing for high-precision semantic role labelling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. [S.l.], 2008. p. 1–8.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

MILES, A.; BRICKLEY, D. Skos core guide. *W3C Working draft*, v. 2, 2005.

MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, ACM, v. 38, n. 11, p. 39–41, 1995.

MURESAN, S.; KLAVANS, J. L. Inducing terminologies from text: A case study for the consumer health domain. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 64, n. 4, p. 727–744, 2013.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, John Benjamins, v. 30, n. 1, p. 3–26, 2007.

NAVIGLI, R. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 2, p. 10, 2009.

NAVIGLI, R.; PONZETTO, S. P. Babelnet: Building a very large multilingual semantic network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 48th annual meeting of the association for computational linguistics*. [S.l.], 2010. p. 216–225.

NETZER, Y.

Semantic authoring for Blissymbols augmented communication using multilingual text generation, 2005.

NETZER, Y.; ELHADAD, M. Using semantic authoring for blissymbols communication boards. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. [S.l.], 2006. p. 105–108.

NIVRE, J.; MARNEFFE, M.-C. D.; GINTER, F.; GOLDBERG, Y.; HAJIC, J.; MANNING, C. D.; MCDONALD, R.; PETROV, S.; PYYSALO, S.; SILVEIRA, N. et al. Universal dependencies v1: A multilingual treebank collection. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. [S.l.: s.n.], 2016. p. 1659–1666.

PALAO, S. *ARASAAC: Aragonese Portal of Augmentative and Alternative Communication*. 2019. Disponível em: <http://www.arasaac.org/>.

PALMER, M. Semlink: Linking propbank, verbnet and framenet. In: GENLEX-09, PISA, ITALY. *Proceedings of the generative lexicon conference*. [S.l.], 2009. p. 9–15.

PALMER, M.; GILDEA, D.; XUE, N. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–103, 2010.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 40th annual meeting on association for computational linguistics*. [S.l.], 2002. p. 311–318.

PEDERSEN, T.; BANERJEE, S.; PATWARDHAN, S. *Maximizing semantic relatedness to perform word sense disambiguation*. [S.l.], 2005.

PUNYAKANOK, V.; ROTH, D.; YIH, W. tau. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, v. 34, n. 2, 2008. Disponível em: <http://cogcomp.org/papers/PunyakanokRoYi07.pdf>.

RAAD, J.; CRUZ, C. A survey on ontology evaluation methods. In: *International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. [S.l.: s.n.], 2015.

RASOOLI, M. S.; TETREAULT, J. Yara parser: A fast and accurate dependency parser. *arXiv preprint arXiv:1503.06733*, 2015.

RATINOV, L.; ROTH, D. Design challenges and misconceptions in named entity recognition. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado: Association for Computational Linguistics, 2009. p. 147–155. Disponível em: <https://www.aclweb.org/anthology/W09-1119>.

RENFREW, C. E. *The Renfrew language scales: Action picture test*. [S.l.]: Speechmark, 2016.

RINGGAARD, M.; GUPTA, R.; PEREIRA, F. C. Sling: A framework for frame semantic parsing. *arXiv preprint arXiv:1710.07032*, 2017.

SANCHEZ, A.; MEYLAN, S. C.; BRAGINSKY, M.; MACDONALD, K. E.; YUROVSKY, D.; FRANK, M. C. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, Springer, v. 51, n. 4, p. 1928–1941, 2019.

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. Disponível em: <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.

SEKINE, S.; NOBATA, C. Definition, dictionaries and tagger for extended named entity hierarchy. In: LISBON, PORTUGAL. *LREC*. [S.l.], 2004.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

TKACHENKO, M.; SIMANOVSKY, A. Named entity recognition: Exploring features. In: *KONVENS*. [S.l.: s.n.], 2012. p. 118–127.

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.

TURING, A. M. Computing machinery and intelligence. *Mind*, v. 59, n. 236, p. 433, 1950.

VEENSTRA, J.; BOSCH, A. Van den; BUCHHOLZ, S.; DAELEMANS, W. et al. Memory-based word sense disambiguation. *Computers and the Humanities*, Springer, v. 34, n. 1-2, p. 171–177, 2000.

VERTANEN, K.; KRISTENSSON, P. O. The imagination of crowds: Conversational aac language modeling using crowdsourcing and large data sources. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.]: ACL, 2011. p. 700–711.

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.

WU, Z.; PALMER, M. Verbs semantics and lexical selection. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. [S.l.], 1994. p. 133–138.

YAROWSKY, D. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, Springer, v. 34, n. 1-2, p. 179–186, 2000.

ZAMBONI, A.; THOMMAZO, A.; HERNANDES, E.; FABBRI, S. Start uma ferramenta computacional de apoio à revisão sistemática. In: *Proc.: Congresso Brasileiro de Software (CBSoft'10), Salvador, Brazil*. [S.l.: s.n.], 2010.

ZELLE, J. M.; MOONEY, R. J. Learning semantic grammars with constructive inductive logic programming. In: *AAAI*. [S.l.: s.n.], 1993. p. 817–822.