Multiword Expressions in Child Language

Rodrigo Wilkens[1], **Marco Idiart**[2], **Aline Villavicencio**[1]
[1]Institute of Informatics, [2]Institute of Physics
Federal University of Rio Grande do Sul (Brazil)
rodrigo.wilkens@inf.ufrgs.br, marco.idiart@gmail.com, avillavicencio@inf.ufrgs.br

Abstract

The goal of this work is to introduce CHILDES-MWE, which contains English CHILDES corpora automatically annotated with Multiword Expressions (MWEs) information. The result is a resource with almost 350,000 sentences annotated with more than 70,000 distinct MWEs of various types from both longitudinal and latitudinal corpora. This resource can be used for large scale language acquisition studies of how MWEs feature in child language. Focusing on compound nouns (CN), we then verify in a longitudinal study if there are differences in the distribution and compositionality of CNs in child-directed and child-produced sentences across ages. Moreover, using additional latitudinal data, we investigate if there are further differences in CN usage and in compositionality preferences. The results obtained for the child-produced sentences reflect CN distribution and compositionality in child-directed sentences.

Keywords: Multiword Expressions, Compound nouns, compositionality, Language Acquisition

1. Introduction

The increasing availability of psycholinguistic, lexical and ontological resources and of more precise and robust natural language processing tools has enabled the automatic annotation of language acquisition corpora with additional sources of information. In particular, among them resources like WordNet (Miller, 1995) and specialised datasets contain information about Multiword Expressions (MWEs) such as noun compounds (*police car*) (Kim and Baldwin, 2006; Nakov, 2008a), phrasal verbs (*break down*) (McCarthy et al., 2003) and collocations (e.g. *salt and pepper*) (Seretan, 2011; Eryiğit et al., 2011). These resources can be used as basis for annotating MWE occurrence in corpora such as the English portion of the Child Language Data Exchange System (CHILDES) (MacWhinney, 1995), which contains transcriptions of child language data. In this paper we introduce the resulting resource, CHILDES-MWE, which contains almost 350,000 English sentences annotated with more than 70,000 distinct MWEs of various types, including compound nouns and phrasal verbs.[1]

The resulting annotated resource can be used as basis for language acquisition studies. In this paper we use it for examining how MWEs feature in child language, focusing on compound nouns (CNs) in English. Firstly, using longitudinal data, which follow specific children across time, we want to examine the link between the input and output of the children in terms of CNs, verifying if there are differences in the distribution of CNs in child-directed and child-produced sentences. Secondly, as in terms of semantics CNs range from compositional to idiomatic combinations, we want to determine if there is any effect of compositionality in the sentences to which children are exposed and in those they produce.

This paper is structured as follows: in §2 we describe some related work on MWEs and child language; and in §3 we present CHILDES-MWE and describe the annotation process. The materials and methods used for this work are in §4, along with an analysis of the results obtained. We finish with conclusions and future work in §5.

2. Learning Multiword Expressions

In the psycholinguistic literature there has been considerable interest in questions about how people acquire, represent and process MWEs (Bod, 2001; Dahlmann and Adolphs, 2007). Compounding in particular can be seen as a way of *introducing new words into the lexicon* (Gagné and Spalding, 2006), and as being at the interface between morphology and syntax. As a consequence, to understand and produce compounds children need to learn to combine information at different levels of linguistic description. This includes how to order the elements of a compound, where the head is, how to do pluralization in a compound, which combinations are frozen and idiomatic and what their meaning is (Berman, 2011). For the latter in particular, there is a wide range of variation, since while some MWEs are more compositional and their meaning can be derived from their component words (e.g. *access road*), others are more idiomatic and semantically unrelated to their component words (e.g. *eager beaver*).

As MWEs can be lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic, they may be more challenging in terms of learning and require more than knowledge about single words and word-to-word relations for their adequate interpretation (Fillmore et al., 1988). For instance, the meaning of an idiomatic MWEs like *bite the dust* cannot be inferred from the meanings of each of its components literally. This introduces a distinction between what a learner is able to infer from language and what must be explicitly stored, as possibly prefabricated units which are retrieved as a whole when needed (Wray, 2002).

MWEs are also frequently found in acquisition data, and for verb-particle constructions the types and tokens produced by children seem to be compatible and follow very closely those in language directed to children (Villavicencio et al.,

[1] CHILDES-MWE is freely available at `html://www.inf.ufrgs.br/pln/resource/CHILDESMWE.zip`.

2012a). Moreover, they are present in child-directed sentences and from an early age, with 2-year-old children already understanding a modifier-head relation in compounds (Clark et al., 1985). The same goes for child-produced sentences: Clark (1993) extensive diary data reports that her son used compounding for all his innovative nouns before age 2, and from then to age 4 for over 70% of them. This is also reported by other studies where compounds are used 80% of the time for spontaneous lexical innovations for English-speaking children under age 4, and for older children 63% of the time (Clark, 1993).

This may vary for other languages, but for English-speaking children this preference for compounding may be due to principles of transparency of meaning, simplicity of form (ease of construction) and productivity along with conventionality and contrast that have been used to explain the word formation devices that children adopt to augment their vocabulary (Clark, 2009). For instance, simplicity of word form and semantic transparency feature early on in children's productions (Berman, 2011), with compounds involving base word forms, and a direct link between form and meaning so that particular words in the compound contribute specific parts of the meaning (Clark, 1993; Berman, 2011). Developmentally the acquisition of compounds starts with them being treated as unanalysed monolexemic labels (a single word), then as the juxtaposition of two nouns with no or limited inflection, before full correct inflection, and finally mastery of productive compounding (Berman, 2011). This means that to describe *a person who pushes wagons* the compound would change for example from *wagon boy* to *push wagon, pusher wagon* and finally to *wagon pusher* during acquisition. Therefore, compounding is an important device for children to expand their vocabularies, and several factors play a role such as productivity, transparency and simplicity (Clark, 1993), with the relation between the frequency of compounds in child-directed (CD) and in their acquisition and use in child-produced (CP) sentences in need of further investigation (Berman, 2011). In this paper we examine the link between characteristics of CD and CP for MWEs, via a large-scale corpus investigation, concentrating on compound nouns.

3. CHILDES-MWE

The English CHILDES contains 60 subcorpora (12 from British English and 48 form North American English), with 895,130 word types in 4,845,264 sentences. We extended the data from the CHILDES Verb Construction Database (Villavicencio et al., 2012b) and annotated them with MWE information. The database contains morphological and syntactic information, along with psycholinguistic and distributional information including verb semantic classes, age of acquisition and familiarity. Details about CHILDES-MWE are in Table 1, focusing on children of up to 7 years of age, given the relevance of this period for language acquisition, and their linguistic input (CD) and output (CP).

To automatically annotate MWEs in corpora we used jMWE (Kulkarni and Finlayson, 2011), defining a detector that finds all occurrences of MWEs, as specified in a list of MWE types. We prioritize precision only looking for MWEs whose components occur in the canonical order as consecutive elements without any intervening material, adopting the longest match from Left-to-Right. The list of target MWE types was obtained from lexical resources containing various types of MWEs:

WN WordNet (Fellbaum, 1998) version 3.0 with 155,287 distinct words from which 69,719 are verbal (3,096), nominal (62,410), adverbial (827) and adjectival (3,386) MWEs.

CE Cranberry expressions dataset (Trawinski et al., 2008) containing MWEs whose components cannot be found outside the MWE (e.g. *sandboy* as *happy as a sandboy*).

NC Noun Compounds datasets by Kim and Baldwin (2008) and Nakov (2008b) containing sequences of nouns (e.g. *cheese knife*).

CN Compound Nominalizations (Nicholson and Baldwin, 2008) which are a subclass of compound nouns in which the head noun is deverbal (e.g. *product replacement*).

LVC Light-Verb Constructions dataset (Tu and Roth, 2011) containing expressions where the verb has a light or supporting role and the meaning is mainly derived from the direct object noun like *take a walk*.

VPC Verb-Particle Constructions dataset (Baldwin, 2008) with combinations of verbs and prepositional, adverbial or adjectival particles (e.g. *break down*).

Cases of *to <verb>* (e.g. *to come*, and *to break*) and *<pronoun> <verb>*.[2] were not included. The final list contains 71,888 MWE types characterized as in Tables 2 and 3. CHILDES-MWE contains 347,391 sentences annotated with MWEs, and details about a subset of these corpora are shown in Table 1 and in Figure 1 for children of ages 1 to 7.

Size	# MWE Types
2	59,439
3	9,813
4	1,790
5	846
Total	**71,888**

Table 2: MWE Types per Size

4. Compound Nouns in CD and CP sentences

To investigate the relation between MWEs in the linguistic input and output of children, focusing on compound nouns (CNs), we examined the following hypotheses:

H1 CNs in child-produced sentences follow the distribution found in child-directed sentences across ages.

[2] Listed for recurrent sequences like *you know*, and *I mean*.

CD age	Sents	MWE Sents	MWE Types	MWE Tokens	CN Types	CN Tokens
13-24	96936	93998	21086	105728	2382	5319
25-48	28432	27844	10765	32684	1123	2006
49-60	11405	11126	5743	12990	483	718
61-72	3511	3401	1962	3954	139	215
73-84	3221	3107	2032	3642	115	178
85-96	2197	2162	1073	2442	54	83
Total	**145702**	**141638**	**42661**	**161440**	**4296**	**8519**
CP age	**Sents**	**MWE Sents**	**MWE Types**	**MWE Tokens**	**CN Types**	**CN Tokens**
13-24	38090	35964	10485	38080	1290	2851
25-48	16303	15811	6778	17993	749	1445
49-60	7559	7205	4043	8327	356	538
61-72	2879	2750	1911	3359	113	155
73-84	2862	2770	1893	3361	125	180
85-96	2325	2271	1332	2584	86	140
Total	**70018**	**66771**	**26442**	**73704**	**2719**	**5309**

Table 1: CHILDES-MWE child-directed (CD) sentences and child-produced (CP) sentences per age in months, MWE sentences, types and tokens, and compound nouns (CN) sentences, types and tokens - English Corpora

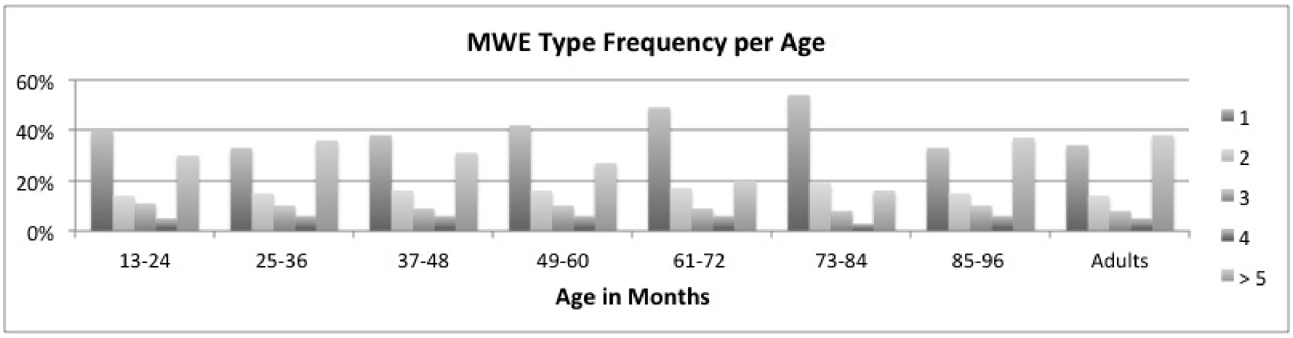

Figure 1: Distribution of MWE Types per Age in Months- from 1 to 5 or more occurrences

Resource	# MWE types
WN	69,517
CE	82
NC	1,328
CN	350
LVC	759
VPC	440
Total	**71,888**

Table 3: Distinct MWE Types in Resources

H2 CN compositionality in child-produced sentences follows the distribution found in child-directed sentences across ages.

We start with an analysis of specific children across different ages, looking at the longitudinal Brown corpus, selecting sentences containing CNs from 3 children, Adam (age = 27 to 58 months), Eve (age = 18 to 27 months) and Sarah (age = 27 to 61 months), Table 4 and Table 5.

	CN Types	CN Tokens	CN Types Brown	CN Tokens Brown
CNs	2342	13828	482	1590
Comp	1392	8550	317	1053
Non-Comp	893	5104	154	516

Table 5: CNs in corpora

Given the lack of resources with compositionality information, to approximate the compositionality of a given CN we use WordNet, assuming that if both the CN and (one or both of) its component words are in the same or a hypernym synset, the CN is compositional. Otherwise it is non-compositional. This simple heuristic is conservative towards compositionality, since the absence of the CN or its components from the relevant synsets may be due to lack of coverage rather than non-compositionality.

For the first hypothesis, we found a high correlation between CD and CP sentences per month in the longitudinal corpora (Spearman correlation = 0.67, $p < 0.01$) confirm-

CD age	Sents	MWE Sents	MWE Types	MWE Tokens	CN Types	CN Tokens
13-24	3043	2822	1629	3038	164	253
25-48	4897	4542	2639	5081	241	350
49-60	3624	3421	2345	3872	159	225
61-72	414	403	276	462	12	16
Total	**11978**	**11188**	**6889**	**12453**	**576**	**844**
CP age	**Sents**	**MWE Sents**	**MWE Types**	**MWE Tokens**	**CN Types**	**CN Tokens**
13-24	1833	1598	872	1657	143	295
25-48	4079	3790	2027	4040	161	276
49-60	3268	3028	1805	3286	106	164
61-72	370	338	227	362	6	11
Total	**9550**	**8754**	**4931**	**9345**	**416**	**746**

Table 4: CHILDES-MWE child-directed (CD) sentences and child-produced (CP) sentences per age in months, MWE sentences, types and tokens, and compound nouns (CN) sentences, types and tokens - Brown Corpus

ing that children tend to follow the distribution in child-directed sentences.

Regarding the second hypothesis, there is a prevalence of compositional tokens in both CD and CP sentences, Figure 2. Moreover, there is a high correlation between the number of compositional tokens (Spearman correlation = 0.63, $p < 0.01$) confirming that children tend to follow the preference for compositional tokens and the distribution found in child-directed sentences.

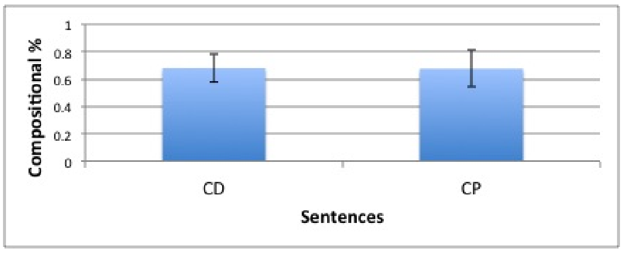

Figure 2: Distribution of Compositional CN tokens in CD and CP sentences - Brown Corpus

These results are compatible with those found by Clark (1993), in that CNs are found in corpora from an early age. Moreover, from an early age children also follow very closely the input they receive in terms of overall CN tokens and compositional (and non-compositional) CN tokens, which are similar to results obtained for verb-particle constructions (Villavicencio et al., 2012a).

5. Conclusion

In this paper we presented CHILDES-MWE, a resource that provides MWE annotation for the English corpora in CHILDES . To ensure higher quality in the automatic annotation, we prioritized precision over recall, focusing on adjacent MWEs in canonical order. The resource contains 347,391 annotated sentences, and is one of the contributions of this work, as the resulting annotation is available to the community and can serve as a basis for language acquisition studies. We investigated the use of MWEs in naturalistic language acquisition data. We focused in the use of CN and compared the production of the children with the input coming from the adults.

The results obtained in the analyses performed are that children tend to follow the CN usage of the adults. The relatively high Spearman correlations indicate that they tend produce more CNs if the hear more CNs. However the present study does not allow us to assert if this happens at the level of types, or if it is an overall effect reflecting the fact that children that are more familiar with CNs are more prone to use them in different contexts, not necessarily reproducing the distribution of types that they hear. Further investigation is planned to determine that.

For future work we plan to augment the MWE annotation, since current coverage is determined by the available resources used in the annotation. Moreover cases of MWE tokens involving internal modification or a different word order not listed in the resources will also not be identified (e.g. *hold right on*). We also plan to do a qualitative analysis of the data and use it as basis for evaluating a computational model of language acquisition.

Acknowledgments

We would like to thank to Instituto de Informática for the support in this research. Part of the results presented by this paper were obtained through the project named *Simplificação Textual de Expressões Complexas* sponsored by Samsung Eletronica da Amazonia Ltda., under the terms of Law number 8.248/91. This work was also partly supported by CNPq (482520/2012-4, 312184/2012-3, 551964/2011-1) and FAPERGS.

6. Bibliographical References

Timothy Baldwin. 2008. A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In Grégoire et al. (Grégoire et al., 2008), pages 1–2.

Ruth Berman. 2011. Children's acquisition of compound constructions. In Rochelle Lieber and Pavol Stekauer, editors, *The Oxford Handbook of Compounding*. Oxford University Press.

Rens Bod. 2001. Sentence memory: Storage vs. computation of frequent sentences. In *Proceedings of the 14th Annual Meeting of CUNY Conference on Human Sentence Processing (CUNY-2001)*. Philadelphia, Pennsylvania.

Eve V. Clark, Susan A. Gelman, and Nancy M. Lane. 1985. Compound nouns and category structure in young children. *Child Development*, pages 84–94.

Eve V. Clark. 1993. *The Lexicon in Acquisition*. Cambridge University Press.

Eve V. Clark. 2009. *First language acquisition*. Cambridge University Press, second edition.

Irina Dahlmann and Svenja Adolphs. 2007. Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? In Nicole Grégoire, Stefan Evert, and Su Nam Kim, editors, *of the ACL Workshop on A Broader Perspective on (MWE 2007)*, pages 49–56, Prague, Czech Republic, June.

Gülşen Eryiğit, Tugay Ilbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 45–55. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. May. 423 p.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:501–538, September.

Christina L Gagné and Thomas L Spalding. 2006. Using conceptual combination research to better understand novel compound words. *SKASE Journal of Theoretical Linguistics*, 3(2):9–16.

Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors. 2008. *of the LREC Workshop Towards a Shared Task for (MWE 2008)*, Marrakech, Morocco, June.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In James Curran, editor, *of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 491–498, Sidney, Australia, July.

Su Nam Kim and Timothy Baldwin. 2008. Standardised evaluation of English noun compound interpretation. In Grégoire et al. (Grégoire et al., 2008), pages 39–42.

Nidhi Kulkarni and Mark Alan Finlayson. 2011. jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124. Association for Computational Linguistics.

Brian MacWhinney. 1995. *The CHILDES project: tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *of the ACL Workshop on : Analysis, Acquisition and Treatment (MWE 2003)*, pages 73–80, Sapporo, Japan, July.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Preslav Nakov. 2008a. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 103–117. Springer.

Preslav Nakov. 2008b. Paraphrasing verbs for noun compound interpretation. In Grégoire et al. (Grégoire et al., 2008), pages 46–49.

Jeremy Nicholson and Timothy Baldwin. 2008. Interpreting compound nominalisations. In Grégoire et al. (Grégoire et al., 2008), pages 43–45.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Dordrecht, Netherlands, 1st edition. 212 p.

Beata Trawinski, Manfred Sailer, Jan-Philipp Soehn, Lothar Lemnitzer, and Frank Richter. 2008. Cranberry expressions in English and in German. In Grégoire et al. (Grégoire et al., 2008), pages 35–38.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *of the ALC Workshop on : from Parsing and Generation to the Real World (MWE 2011)*, pages 31–39, Portland, OR, USA, June.

Aline Villavicencio, Marco Idiart, Carlos Ramisch, Vitor De Araujo, Beracah Yankama, and Robert Berwick. 2012a. Get out but don't fall down: verb-particle constructions in child language. In Robert Berwick, Anna Korhonen, Thierry Poibeau, and Aline Villavicencio, editors, *of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 43–50, Avignon, France, April.

Aline Villavicencio, Beracah Yankama, Marco Idiart, and Robert C. Berwick. 2012b. A large scale annotated child language construction database. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 2370–2374.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge, UK. 348 p.