# Corpus Lexicography in a Wider Context

**Chen Gafni**
**Bar-Ilan University**
**chen.gafni@gmail.com**

## Abstract

This paper describes a set of tools that offers comprehensive solutions for corpus lexicography. The tools perform a range of tasks, including construction of corpus lexicon, integrating information from external dictionaries, internal analysis of the lexicon, and lexical analysis of the corpus. The set of tools is particularly useful for creating dictionaries for under-resourced languages. The tools are integrated in a general-purpose software that includes additional tools for various research tasks, such as linguistic development analysis. Equipped with a user-friendly interface, the described system can be easily incorporated in research in a variety of fields.

## 1 Introduction

Corpus lexicography, a key component in modern dictionary compilation, has become increasingly powerful and efficient due to the development of various tools such as Word Sketch (Kilgarriff and Tugwell, 2002) and TickBox Lexicography (Kilgarriff et al., 2010). While corpus lexicography deserves further development in its own right, it is worthwhile considering it as an integral part of wider scientific missions. This paper describes several tools for corpus lexicography, whose design takes into consideration their contribution to linguistic research. The tools are integrated in a general-purpose linguistic software, where they can be readily applied in language acquisition studies, psycholinguistics, and other fields of research.

## 2 Preliminaries

The described system is implemented in the *Child Phonology Analyzer* software (CPA; Gafni, 2015)[1], which was built in MS Excel due to its popularity and user-friendly interface.[2] Nevertheless, the concepts behind the system are general and can be implemented in various environments. The software can analyze corpora stored in various file formats, including Excel and plain-text files, as well as several special formats used in linguistic research: Praat's TextGrids, CHAT transcription files, EAF annotation files, and XML schema for TalkBank data. The software converts analyzed corpora into Excel format and adds all analysis products to the Excel file.

### 2.1 Organizing the Data

The described tools ("macros") require that the corpus text be stored in a vector format. The text can be converted to a vector format using CPA's "Data preparation" macro with the "Corpus tokenization" option, which segments the text into words.

Segmentation is performed on the basis of blank spaces and additional word-dividing characters, which can be defined in CPA's "Word dividers" table (Figure 1). There are two types of word dividers, which can be used for separating words even at the absence of a blank space: *punctuation* marks (e.g., comma) are deleted during segmentation, while *final letters* are not (final letters are special letter forms appearing only at word endings. See some examples from Hebrew in Figure 1).

[1] Website: https://chengafni.wordpress.com/cpa/

[2] The code is written in Visual Basic for Applications for MS Excel and is available under the GNU General Public License. A version of this system for LibreOffice Calc is planned to appear in the future in order to free it from dependency on proprietary software.



*Proceedings of Recent Advances in Natural Language Processing*, pages 353–359,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_041

### 2.2 Longitudinal and Multi-Level Corpora

The system was designed especially for analyzing longitudinal data from language acquisition studies. Such corpora typically list utterances made by a child alongside the hypothesized intended target utterances (a corpus containing such paired utterances is a *multi-level* corpus). The age of the child is also recorded for each utterance for the purpose of developmental analysis (Figure 2). Corpus tokenization (see above) processes all the above-mentioned information. This further allows analyzing both the input (target) and output (production) lexicons of the child from a developmental perspective.

### 2.3 Multi-Speaker Corpora

CPA can also handle corpora that contain data from multiple sources ("speakers"), stored in a single or multiple spreadsheets. The sources can be different texts, cross-sectional data from several children (Figure 3), parallel data from elicitation experiments, etc. During corpus tokenization, the data from each speaker is labelled accordingly. This allows building a separate lexicon for each speaker for the purpose of comparative analysis.

## 3 Constructing Corpus Lexicon

The "Construct lexicon" macro takes as input a corpus in vector format and lists the different item types in the vector including the number of occurrences of each item ("Count"). For multi-level corpora, this analysis can be done separately for target and output levels. For multi-speaker corpora, the macro constructs separate lexicons for each speaker. In addition, the macro constructs a general lexicon based on the lexicons of individual speakers (this is done separately for target and output levels). For developmental data, the macro also records the age in which the item is first attempted (for target lexicon) or produced (for output lexicon) and the spreadsheet row index containing the first attempt (Figure 4).

For items containing explicit morphological boundaries (e.g., # in *ha#kelev* 'the.dog' (Hebrew)), the macro creates a list of potential affixes based on word fragments separated by morpheme boundary markers (naturally, the initial list contains both true grammatical affixes and lexical stems). The list of affixes can be used later for analyzing the properties of polymorphemic words in the lexicon.

| Symbol | Type |
|---|---|
| ! | Punctuation |
| , | Punctuation |
| . | Punctuation |
| ן | Final letter |
| ף | Final letter |
| ץ | Final letter |

Figure 1: Word dividing symbols

| Age | Child | Target |
|---|---|---|
| 1;02.24 | paˈpa: | paɸˈpaɸ |
| 1;02.24 | ˈtikta ˈtikta | ˈtiktak ˈtiktak |
| 1;02.24 | ˈpuaχ ˈpuaχ | taˈpuaχ taˈpuaχ |
| 1;03.05 | ʔaˈba | ˈaba |
| 1;03.05 | mma ˈʔaba ˈaba | ˈima ˈaba ˈaba |
| 1;03.05 | mma | ˈima |

Figure 2: Longitudinal language acquisition data from a Hebrew-speaking child

| Child | Age | Target | Output |
|---|---|---|---|
| Child 1 | 1;04.00 | taˈpuaχ | ˈbuaχ |
| | 1;04.00 | paɸˈpaɸ | ˈpaapa |
| Child 2 | 1;04.20 | taˈpuaχ | deˈbaaχ |
| | 1;04.20 | paɸˈpaɸ | ˈpapa |
| Child 3 | 1;05.03 | taˈpuaχ | ˈpu̯ə̯χ |
| | 1;05.03 | paɸˈpaɸ | paˈpaa |

Figure 3: A corpus of language acquisition cross-sectional study. Data from three Hebrew-speaking children

| Word | Count | First attempt age | First attempt row |
|---|---|---|---|
| ˈaba | 80 | 1;02.07 | 30 |
| aˈdom | 7 | 1;05.08 | 917 |
| af | 4 | 1;04.24 | 565 |
| afiˈfon | 3 | 1;04.24 | 620 |
| agaˈla | 1 | 1;04.17 | 505 |
| aˈgas | 9 | 1;03.05 | 128 |

Figure 4: A lexicon for a developmental corpus

## 4 Importing Lexical Information

The corpus lexicon can be turned into a dictionary by adding information describing the various items. The descriptive information is stored in separate columns in the lexicon spreadsheet. Each such column represents some lexical property (e.g., part-of-speech, grammatical gender). In the absence of external dictionaries, the lexical

information needs to be added manually. However, if there is an available resource for the particular language, the "Import dictionary" macro can import the lexical information from the external resource.

This macro receives as input the corpus lexicon and an external dictionary in a table format. The macro copies information from the external dictionary to the lexicon for every lexicon entry found in the dictionary. The macro can be used for importing information about lexical words (Figure 5), as well as grammatical affixes (Figure 6).

The macro was specially designed to handle lexical entries containing explicit morphological boundaries. For such lexical entries (e.g., *le#ˈkof* 'to.monkey' (Hebrew)), the macro first searches the full entry in the words dictionary. If not found, the macro then searches the individual morphemes. If a morpheme is found in the dictionary (e.g., *kof*), the macro imports the information for that entry to the corpus lexicon (Figure 7).

When importing information from an affix dictionary, the macro can use the external list of affixes to remove irrelevant entries from the corpus affixes lexicon (i.e., stems included in the affixes lexicon during its construction).

An affixes dictionary is constructed in a similar way to a words dictionary; it contains a list of affixes with additional columns providing information about these affixes. However, the additional fields have a functional role: they specify how the affix modifies the properties of affixed words. This information can be used for modifying polymorphemic entries in the lexicon (see 5.1).

## 5 Analyzing the Lexicon

### 5.1 Morphological Analysis

If the corpus lexicon contains polymorphemic words with explicit marking of morphological boundaries, and an affixes dictionary is available (see 4), the "Morphological analysis" macro can import information from the affixes dictionary to the corpus lexicon.

Each entry in the affixes dictionary should have the following fields (Figure 6): (a) Tier: a name of a field in the words lexicon. For example, a "POS" value in the tier field (stands for "Part-of-speech") indicates that the affix applies to lexical items in a specific lexical category. (b) Condition: a possible value of the lexical field specified in the tier field.

| Word | Gloss | Lemma | POS | Gender | Number |
|---|---|---|---|---|---|
| ˈaba | dad | ˈaba | Noun | M | SG |
| aˈdom | red | aˈdom | Adjective | M | SG |
| baʿt͡sal | onion | baʿt͡sal | Noun | M | SG |
| kof | monkey | kof | Noun | M | SG |

Figure 5: An external words dictionary

For example, a condition value "Noun" indicates that the affix applies to nouns. (c) Function: the name of the lexical field modified by the affix. For example, a value of "Definiteness" indicates that the affix specifies the definiteness value of the hosting word. (d) Value: the value assigned to the lexical field specified in the "Function" field. For example, a value of "Def" indicates that the affix marks the hosting word as being definite.

| Affix | Tier 1 | Condition 1 | Function 1 | Value 1 |
|---|---|---|---|---|
| ha# | POS | Noun | Definiteness | Def |
| le# | POS | Noun | Case | Dative |

Figure 6: An external affixes dictionary

For each affix in the affixes dictionary, it is possible to define multiple feature quadruplets (e.g., "Tier 1", "Condition 1", …, "Tier 2", "Condition 2", etc.). This option is useful for handling affixes that can affect multiple word classes (e.g., nouns and adjectives) or have multiple functions (e.g., express possession and mark tense).

| Word | Count | Gloss | Lemma | POS | Gender | Number |
|---|---|---|---|---|---|---|
| ha#baʿt͡sal | 1 | onion | baʿt͡sal | Noun | M | SG |
| le#ˈkof | 1 | monkey | kof | Noun | M | SG |

Figure 7: Imported lexical information based on the stems of prefixed words

The "Morphological analysis" macro finds lexical entries containing affixes and modifies their properties according to the details of the affix. If an affix modifies a lexical field not defined in the lexicon, the macro adds that field to the lexicon (Figure 8).

| Word | Lemma | POS | Gender | Number | Definiteness | Case |
|---|---|---|---|---|---|---|
| ha#baʿt͡sal | baʿt͡sal | Noun | M | SG | Def | |
| le#ˈkof | kof | Noun | M | SG | | Dative |

Figure 8: Lexical entries of prefixed words after morphological analysis

### 5.2 Lexicon Summary

This macro generates a summary table of the lexicon. The summary table includes a list for each lexical field (e.g., "POS") that specifies the various values of the field (e.g., "Noun", "Verb"). For each value, the list indicates the number of corpus tokens and types. The number of types is the number of items in the lexicon with the relevant value (e.g., the number of noun types), and the number of corpus tokens is calculated from the "Count" field in the lexicon (Figure 9).

| POS | Types | Tokens |
|---|---|---|
| Adjective | 9 | 26 |
| Adverb | 3 | 5 |
| Interjection | 19 | 204 |
| Noun | 134 | 1017 |
| Pronoun | 2 | 16 |
| Verb | 22 | 83 |

Figure 9: Lexicon summary by part-of-speech

## 6 Integrating Lexicons

Efficient integration of information is essential for compiling a dictionary based on data from multiple resources. The "Merge worksheets" macro is a general utility macro that integrates the contents of multiple spreadsheets in a file. Thus, it requires lexicons generated from different corpora to be stored in one file (this can be done either manually or automatically with the "Merge workbooks" CPA macro).

The "Merge worksheets" macro has several operation modes, one of which is designed specifically to integrate lexicon tables. The macro receives as input any number of spreadsheets. It creates a single lexicon[3] containing information from all input lexicons. The merged lexicon contains the union of lexical fields in all input lexicons (i.e., a lexical field will be included in the merged lexicon if it appears at least in one input lexicon).

The entries in the merged lexicon are sorted alphabetically. If a lexical entry appears in multiple input lexicons, the duplicate entries are merged. The merged entry summarizes token counts from the contributing corpora (e.g., if an item appears 10 times in one corpus and 20 times in another corpus, the merged lexicon will record 30 tokens for that item). In addition, the merged entry will contain the lexical properties collected from all contributing entries. In case of conflicting inputs (e.g., an item is classified as a noun in one lexicon and as a verb in another), the merged entry will indicate all possible values for that property (e.g., Noun / Verb). The merging macro can also add labels indicating the source(s) (i.e., the name of the input lexicon) of each entry.

## 7 Lexical Development

Assessing the size of the child's lexicon is an important part of longitudinal language acquisition studies, from both theoretical and clinical perspectives. In particular, there is evidence that aspects of grammatical development are tightly correlated with vocabulary size (Bates and Goodman, 1997).

The "Lexical development" macro analyzes lexical growth in corpora that record the age of production of every utterance. Using the age of first attempt to produce target words (see 3), the macro divides the child's lexicon into stages of lexical development (Figure 10). The first stage is marked by the acquisition of the first 10 words, the second by a total lexicon size of 50 words, and then an additional 50 words for every subsequent stage (Adam and Bat-El, 2009).

| Age | New words | Total words | Stage | Theoretical boundary |
|---|---|---|---|---|
| 1;02.00 | 5 | 5 | | |
| 1;02.07 | 3 | 8 | | |
| 1;02.16 | 1 | 9 | | |
| 1;02.20 | 1 | 10 | 1 | 10 |
| 1;02.24 | 1 | 11 | | |
| 1;03.05 | 5 | 16 | | |
| 1;03.14 | 5 | 21 | | |
| 1;03.19 | 3 | 24 | | |
| 1;03.25 | 5 | 29 | | |
| 1;04.03 | 9 | 38 | | |
| 1;04.10 | 11 | 49 | 2 | 50 |

Figure 10: Lexical development

Stages of lexical development are aligned with recording sessions, such that if a theoretical stage boundary is reached in mid-session, the actual boundary will be assigned either to that session or to the preceding session (whichever is closer). For

[3] Due to software limitations, the maximal number of entries in a single lexicon (or a lexicon generated by merging a number of lexicons) is 1,048,575. When this limit is exceeded, CPA splits the lexicon over multiple spreadsheets.

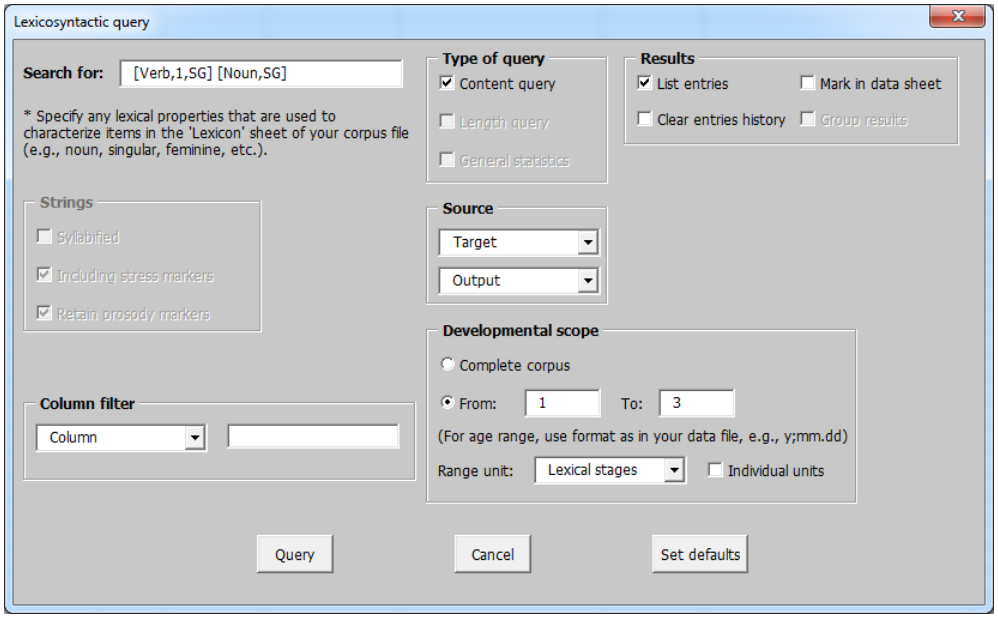


Figure 11: Lexicosyntactic query form

example, if the child has reached 49 cumulative target types at the end of session 1 and 58 cumulative target types at the end of session 2, the theoretical landmark of 50 words will be assigned to session 1. If the lexicon grows rapidly, such that more than one theoretical stage is passed in a single session, the macro will "skip" intermediate stages and assign only the last stage to that session. For example, if the size of the lexicon jumps from 100 words (stage 3) to 200 words (stage 5) in a single session, that session will be marked as the end point of stage 5, skipping stage 4. The macro also provides a more fine-grained account, indicating the number of new words added to the lexicon in every session and the total lexicon size after every session.

By default, lexical development is calculated based on the full list of lexical entries. However, this list is organized by word form (types), such that words that are interrelated via inflectional morphology (e.g., *cat–cats*) are listed as separate entries. Relying on plain surface forms can result in over-estimation of lexicon size. This can be avoided by analyzing lexical development by lemma/lexeme. When this option is chosen, the macro analyzes the lemma field of the lexicon rather than the word field. The lemma field indicates the lemma of each lexical entry (e.g., the lemma of *cat* and *cats* is *cat*). The lemma information can be supplied manually or imported from an external dictionary (see 4). When a lexical entry has no lemma specified, the surface form of the entry will be taken as the lemma.

## 8 Lexical Queries

Once lexical properties are specified in the lexicon, this information can be used to analyze the corpus. CPA has a set of macros that can extract linguistic information from the corpus on various levels of analysis, via a user-friendly query form (Figure 11). One of these macros, "Lexicosyntactic query", queries the corpus at the word and utterance levels. Specifically, "Content Lexicosyntactic queries" can find occurrences of lexical properties and sequences of lexical properties in the corpus. For example, the query [Verb] [Noun] will find all instances of verbs followed by nouns in the corpus. Similarly, the query [Verb,1,SG] [Noun,SG] will find all instances of verbs conjugated in the first person singular followed by singular nouns.

The scope of queries can be constrained by age or stage of lexical development. Thus, for example, it is possible to get all verbs attempted by a child at a given age/lexical stage or range of ages/lexical stages. This option allows for investigation of lexical development at a more fine-grained level.

Queries over single-item sequences (e.g., [Verb,SG]) calculate the number of tokens and types and can also return a list of items that matched the query (Figure 12). Queries over multi-item sequences (e.g., [Verb] [Noun]) do not return

| Query: | [Noun,SG,M] | in: | Target, Reference: Output<br>Age: 1;02.00-1;04.20<br>Unsyllabified tokens<br>Include stress. | |
|---|---|---|---|---|
| **Row** | **Utterance** | **Age** | **Source token** | **Reference token** |
| 2 | 1 | 1;02.00 | ta'puaχ | 'buaχ |
| 3 | 1 | 1;02.00 | ta'puaχ | 'pụạχ |
| 23 | 20 | 1;02.07 | paɻ'paɻ | pa'pa: |
| 30 | 26 | 1;02.07 | 'aba | ba |

Figure 12: Corpus instances of singular masculine nouns (source) paired with the corresponding forms produced by an infant (reference).

specific items, but rather a list of indices of rows in the corpus where such sequences are found.

In addition to lexicosyntactic queries, CPA has similar query macros for analyzing the phonological properties of corpora. These queries, too, can be correlated with lexical development. Additional, more advanced macros can be used to combine queries on different levels of analysis. This allows, for example, to study the interaction between phonological and lexical development.

## 9 Discussion

The described set of tools can help creating resources for under-resourced languages. For example, it was used for creating a lexicon with corpus frequency data for Hebrew (Gafni, 2019), and it is currently being used in an ongoing longitudinal study of phonological development in twins. In addition to building a lexicon for each participating child and assessing lexical development, the system can assist in improving the quality of the transcribed data.

Given that the transcribed data can contain many errors (typos, misperceptions), it is important to have it validated. Since the amount of transcribed data can be enormous (tens of thousands of tokens) and the transcription task is very time-consuming, it is impractical to have every token transcribed by multiple transcribers. One possibility to check data quality is to have a random subset of the corpus (e.g., 10% of the tokens) be transcribed by more than one transcriber, and calculate inter-transcriber reliability. However, such an approach can help detecting problems in a limited part of the corpus. The tools described in this paper offer a more systematic approach to quality check of transcribed data. In this lexicon-based approach, one goes over the entries in the automatically generated corpus lexicon and looks for suspicious entries. For the lexicon of target words, this mainly involves looking for non-existing words, which likely resulted from typos. For the lexicon of produced forms (output lexicon), quality check mainly involves examining tokens with unusual structure that deviates from the phonology of the ambient language. Thus, in the proposed approach, one estimates the potential of lexical entries to contain errors, and then focuses on suspicious forms. This is more effective than examining corpus subsets randomly.

The described tools can be integrated in any task involving corpus analysis. For example, the CPA software includes an n-gram frequency calculator, which can calculate corpus-weighted mean n-gram frequencies over a list of strings (in this context, n-gram refers to a sequence of letters or phones within words). This is useful for creating controlled sets of stimuli for psycholinguistic experiments.

Finally, it should be acknowledged that there is some overlap between the described system and other existing systems. Well-established systems such as Sketch Engine (Kilgarriff et al., 2014) provide powerful solutions for corpus lexicography, and the CLAN program (MacWhinney, 2000) can be used for studying lexical development.

Compared to these programs, CPA is currently limited in areas such as collocation analysis and POS tagging. On the other hand, CPA has some advantages over these programs. Its unique built-in lexical development tool allows for more comprehensive study of language development, and its querying system allows for combined lexical and phonological corpus analysis. The user-friendly interface enhances user experience and saves the need to learn complex query syntax, as used by the CLAN program. In addition, CPA is distributed as an Excel file. This means that Excel users can perform the various analysis tasks in the natural environment of the data, without the need to install (or purchase) additional software.

To conclude, this paper views corpus lexicography in a wide context of linguistic research. Accordingly, the described tools are integrated in a single, user-friendly system designed to support any task requiring corpus analysis. Future improvements to the current system will include the addition of standard lexicographic functions, such as collocation analysis and morphological analysis that does not require overt marking.

## References


Galit Adam and Outi Bat-El. 2009. When Do Universal Preferences Emerge in Language Development? the Acquisition of Hebrew Stress. Brill's Journal of Afroasiatic Languages and Linguistics, 1(1):255–282.

Elizabeth Bates and Judith C. Goodman. 1997. On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. Language and Cognitive Processes, 12(5–6):507–584.

Chen Gafni. 2015. Child Phonology Analyzer: processing and analyzing transcribed speech. In The Scottish Consortium for ICPhS 2015, editor, Proceedings of the 18th International Congress of Phonetic Sciences., pages 1–5, paper number 531, Glasgow, UK: the University of Glasgow.

Chen Gafni. 2019. General Lexicons of Hebrew: Resources for Linguistic and Psycholinguistic Research (version 1.0).

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. Lexicography, 1(1):7–36.

Adam Kilgarriff, Vojtěch Kovář, and Pavel Rychlý. 2010. Tickbox lexicography. In Sylviane Granger and Magli Paquot, editors, eLexicography in the 21st century: new challenges, new applications. Proceedings of eLex 2009, volume 7, pages 411–418, Louvain-La-Neuve. UCL Presses Universitaires De Louvain.

Adam Kilgarriff and David Tugwell. 2002. Sketching words. In Marie-Hélene Corréard, editor, Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins, pages 125–137. Euralex.

Brian MacWhinney. 2000. The CHILDES Project: Tools for Analyzing Talk. Lawrence Erlbaum Associates Inc, Mahwah, NJ, 3rd edition.