TransformELT

UK-China Research Innovation Award (RIA) 2021: Corpus-assisted curriculum and material development

Project Report

Supported by

BRITISH COUNCIL

UK-China Research Innovation Award (RIA) 2021: Corpus-assisted curriculum and material development

Executive summary

This research project was managed by British Council China for the Chinese Basic Education Curriculum and Teaching Material Research Center (BECTMRC). The project sought to identify:

1.) the most commonly used medium to high frequency, age-appropriate language chunks that are presented in the updated 2021 New National English Curriculum (NNEC) (covering Grades 3 to 9) through comparison with commonly used, age-appropriate lexical chunks used by similarly aged native-speaking children in the UK.
2.) prominent gaps in high-frequency language within the New National English Curriculum (NNEC) that can be supplemented or included in future materials revision.

TransformELT, an independent language education consultancy based in the UK, was awarded the grant to conduct the research and publish the results. The principal researcher, James Thomas, created the Corpus of Native Youth English (CONYE23), a corpus of native speaking children's *output*, drawing mainly on the CHILDES database and the age-appropriate sections of the BNC14 Spoken corpus for the language produced by children. Large samples of language *input* written for the NNEC age group, (i.e. 9 to 15), were also collected, as children's language output is strongly influenced by the language they encounter in written texts, when reading both for pleasure and for study purposes. Project outputs have been published in the form of an online database and a 'book of chunks' (with and without metadata), all of which can be accessed at https://transformelt.com/china-corpus-project/.

The research outputs will enable Chinese curriculum designers and materials writers for Grades 3 to 9 to verify their intuitions about the language used by similarly aged native-speaking children in the UK, and also to enhance the NNEC word lists with vocabulary items that will bring their teaching materials into closer alignment with native youth English.

Research process

Two research objectives were set at the outset of this project.

The first was *to identify the most commonly used medium to high frequency, age-appropriate language chunks that are presented in the updated 2021 New National*

English Curriculum (NNEC) covering Grades 3 to 9 (ages 9-15), and to compare them with commonly used, age-appropriate lexical chunks used by similarly aged native-speaking children in the UK. The second objective was to *identify prominent gaps in high-frequency language within the NNEC that can be supplemented or included in future materials revision*.

To address both objectives, a corpus of native-speaking youth English, CONYE23, was created, from which various multi-word units that contain each of the words on the lists were downloaded and then uploaded into a database for further processing and future access. During the life of the project, the output of the corpus research evolved to show the chunks in which the nouns, verbs and adjectives on the NNEC lists are used. The corpus output also enabled us to identify word families whose head words appear on an NNEC list, which simplifies their learning, as well as a significant number of word families not represented on the NNEC lists and which would therefore require more elaborate teaching.

NNEC word lists

The NNEC word lists represent the minimum vocabulary requirement for school-aged children in China learning English. The lists are presented as lemmas (canonical or dictionary forms) without different parts of speech. There are 482 words on the Primary list and 1,628 on the Lower Secondary list. Of these, 1,144 are not on the Primary list. Very many of these words undergo conversion, i.e., they function in more than one part of speech. After downloading the lists of lemmas in all their parts of speech from the CONYE23 corpus, the primary school list has 778 lemmas and the secondary list, which includes the primary list, has 2,326.

	Primary	**Junior secondary**
Raw data	482	1,628
With conversion	778	2,326

Very many of these words also combine with each other to form collocates, chunks and other multi-word units. In the CONYE23 database there are many phrasal verbs, compound nouns and chunks which are combinations of words on the NNEC lists. However, knowing the words *head* and *start* does not ensure knowing what a *head start* is. Knowing *stand* and *with* does not ensure knowing the 'support' meaning of this phrasal verb. Furthermore, the concepts that these items express may well be beyond the conceptual needs of the target age groups.

Corpus of Native Youth English: CONYE23

The Corpus of Native Youth English was created in 2023, using the corpus management tool, Sketch Engine. It contains 53,556,453 tokens which include 43,809,730 words.

As each of the texts was collected, mainly from the internet, they were entered into an online database, and the 5,520 documents were tagged for genre, key stage, provenance (UK / US) and corpus (input / output), school year and school subject, whenever this data was available. In compiling the corpus, Sketch Engine also lemmatised and tagged each word for its part of speech.

Key stages

In analysing the data, the most decisive category of metadata was key stages (KS), divisions in the British education system. The percentages represent their portions of the corpus.

Nursery and Reception Years (3-5 years old).	47.9%
...	
Key Stage 1: Years 1 to 2 (5-7 years old)	3.0%
Key Stage 2: Years 3 to 6 (7-11 years old)	8.2%
Key Stage 3: Years 7 to 9 (11-14 years old)	8.2%
Key Stage 4: Years 10 to 11 (14-16 years old)	13.0%
Key Stage 5: a.k.a. College or Sixth Form.	16.1%

We refer to the first of these as KS0. Due to data collection methods, this occupies the largest proportion of the corpus and revealed features of language such as reduplicatives (word combinations that are duplicates or near duplicates like *ding-dong*) and holophrases (single words used by infants to convey a larger meaning), which are rarely used by older children or adults. The corpus shows that collectively, these very young native speakers have an active and/or passive vocabulary of 22,795 lemmas, of which 6,585 are proper nouns. The proper nouns that children know depend on the people and places in their environments, as well as on their exposure to electronic and print media. KS0 data was not used in the analyses.

All of the data that was analysed was extracted from a subcorpus of KS1 to 3 which contains 31,023,143 tokens. KS1 was included because this represents the language that native-speaking children already know prior to the age at which Chinese children start learning English. Many of the texts for specific KS or age groups are in mixed categories, as they are recommended for multiple age groups. In these cases, the lowest KS was chosen for the metadata.

Another major division of the corpus is its separation of input and output data. Input refers to the language that children are exposed to and output refers to the language they produce. The following section describes these subcorpora.

Input corpus

This is the corpus of texts created for young native speakers of English, thereby embracing their receptive skills. It includes school subject resources, fiction, the transcripts of films as well as the subtitles of some age-appropriate films. While not every child will read and watch the same things, the basic assumption is that their

authors have reason to believe that the language they choose to use will be comprehensible to their audience. This assumed knowledge is not just knowledge of words but a knowledge of the world that words inhabit. It was a relatively straightforward process for a small team of research assistants to scour the internet for appropriate texts and input them into the corpus database.

Output corpus

It proved more challenging to obtain samples of language produced by young native speakers, to demonstrate their productive use of language. We planned to gather this data by inviting schoolchildren through their teachers to register anonymously on our website, and to paste in texts they were writing as part of their schoolwork as well as any writing they had done on their own. Given the profound concerns around child safety, special care was taken to prevent identification of the children. It was clear that this had to be undertaken on a large scale: one thousand children each submitting one thousand words yields only one million words, which is quite a small corpus. The project appointed a schools liaison officer to contact schools and regional departments of education with a view to garnering their collaboration. UK subject groups on Facebook were also approached: History, Maths, Geography, Physics, Music, Art and Design, Media studies, Computer science. The Literacy Trust and even the Scouts were also approached.

For a number of reasons, this approach to collecting native-speaker children's writing proved impractical. Language produced by teenage bloggers, young presenters of TED Talks and YouTube Kids was subsequently explored as an alternative source, but ultimately this made only a small contribution to the corpus. We therefore decided to turn to appropriate existing data: CHILDES and a subcorpus of the BNC Spoken (2014) to populate the output subcorpus.

Resources

CHILDES

The acronym CHILDES stands for Child Language Data Exchange System, which includes a database of transcripts used for research into child language (MacWhinney and Snow 1990).

The UK sections of the CHILDES English Corpus were downloaded from their website and processed accordingly. There are 11,712,579 tokens from CHILDES in the CONYE23 corpus. The age ranges in CHILDES correspond roughly to those of the key stages. Many thousands of duplicate sentences were deleted, as well as those consisting of one word only. As the table below shows, there was a disproportionate quantity of KS0 data, which the research team decided to keep for future comparisons. However, it was not used in the analyses for multi-word units.

KS0	10,633,106	90.8%
KS1	952,678	8.1%
KS2	70,940	0.6%
KS3	2,463	0.5%
KS4	53,392	0.5%

BNC Spoken

BNC Spoken 2014 (Love et al. 2017) is a corpus of present-day spoken British English, gathered in informal contexts in the years between 2012 and 2016. It contains 10,495,185 words of transcribed content, featuring 668 speakers in 1,251 recordings. Among its metadata is the age groups of speakers, which was crucial for our research. Unfortunately, the two youth age groups are the broad spans of 0 to 10 and 11 to 18 that were established in the first phase of their data collection. In their second stage of data collection, exact ages were stated, so the texts were tagged with age groups:

Up to 3	KS0
From 4 to 6	KS1
From 7 to 9	KS2
From 10 to 12	KS3
From 13 to 17	KS4
The rest	KS5

The relevant sections of the corpus were downloaded and the many sentences containing taboo words were deleted.

From all of this data, some conclusions could be drawn about the vocabulary size of the target groups and their use of the vocabulary. The KS1 subcorpus of CONYE23 contains 17,433 lemmas, though no individual child up to six years could be expected to know all of them. Of these lemmas, 3,429 are proper nouns which will differ for each child, depending on their environment. Excluding proper nouns, there are 9,242 lemmas that occur more than once in the KS1 subcorpus and 4,855 lemmas occurring five or more times, the median being 7,048. The median between 0 lemmas and all the lemmas excluding proper nouns is 7,880. Hence our estimate that English children will start school familiar with around 7,500 lemmas.

The CONYE23 database

This is actually a suite of relational databases, whose home database contains the NNEC primary and junior secondary school words. It facilitates generating lists of the words in each part of speech within these two levels. Beside each word in the Word List database, is a row of buttons: bigrams, collocations, grammar patterns and chunks. Clicking these buttons shows how the word is used syntagmatically in CONYE23. One of the original research questions focussed on chunks alone, but these databases furnish this richer palette of lists. The focus of the linguistic analysis

was on the syntagmatic patterns in which the NNEC words participate. These are bigrams, collocations and chunks. The online database can be accessed via Sketch Engine.

Findings

Identifying prominent gaps in the NNEC

In order to identify prominent gaps in the NNEC, it was compared with a word list extracted from CONYE23. There are 43,425 nouns, verbs, adjectives and adverbs in its KS 1–3 subcorpus. The top 400 lemmas, many of which undergo conversion, were selected manually according to their relevance to essential ELT topics such as parts of the body, cultural and sporting activities and technology. The list can be accessed by frequency in CONYE23 and by alphabetical order.

Bigrams

These are pairs of words that occur adjacently in natural language. The high frequency items for each word represent features of word usage that learners need to be aware of. The database of bigrams was created by searching CONYE23 for part of speech pairs, such as noun + verb, verb + noun, adjective + noun, verb + adverb. The lists were downloaded one by one along with their metadata: genre, KS, provenance, corpus. Here, for example, are some of the most frequent adjective + noun pairs from novels which appear in the KS3 list:

> *long time, other side, deep breath, old man, last night, young man, same time, good friend, good thing, old woman, whole thing, next morning, other way, last week, little kid, right hand, next week.*

When looking at the lists of verb + preposition and verb + adverb, a great many phrasal verbs appear. While not all phrasal verbs are contiguous, and they are used much less frequently in the language than single word lexemes, this database shows phrasal verbs widely used by native-speaking youths. These are the most frequent ones from KS3:

> *look at, look like, go on, feel like, look for, get back, look back, turn back, turn away, fall down, run away, get away, look down.*

Compound nouns are an essential word formation process in English. There are 14,610 compound nouns in KS3 where both nouns are on the NNEC lists. Here some of the most frequent:

way home, bedroom door, kitchen table, dining room, night sky, police officer, night air, power plant, stone wall, stone step, police station, pocket watch, car park, bedroom window, school bus, glass door, coat pocket, forest floor, front door, force field, coffee table, day care, dance floor, grandfather clock.

Collocation

The collocation database is based on the grammatical relationships of a word and its collocates. These syntactic relationships, along with grammar patterns, are an invaluable source of data for a lexical syllabus since they facilitate the teaching of clause structures that are governed by the lexicogrammatical properties of words. This accords with the psycholinguistic processes of language production. They instantiate syntactic pairings, such as the nouns that are the objects of a particular verb.

Gramrels (grammatical relations)

When there is enough data, gramrels show the words that follow a specific *preposition* that follows a particular word. These are often the skeletons of chunks. For example,

answer for everything, wait for answer, need for answer, reason for answer, push for answer, search for answer, care about answer, sure about answer, think about answer.

Grammar patterns

Grammar patterns can be thought of as extended colligations and are therefore properties of words. Grammar patterns are not sentence level syntactic structures such as conditionals, passive structures, etc.

The grammar patterns in the database were not derived from the corpus; rather they are a subset of the Collins COBUILD books of the grammar patterns that contain the patterns of many thousands of nouns, verbs and adjectives (Hunston & Francis 2000). The subset contains NNEC words only. This facilitates a systematic approach to vocabulary study that teaches syntactic patterns with the vocabulary in which they function. Students start by saying *someone accepts something for someone,* then they replace the placeholders with concrete references, such as *the captain accepted the prize for her team.*

Chunks

Chunks are by definition, semantically whole and require a different approach to deriving them from corpora. Corpus tools are not currently equipped to identify strings of words that are semantically whole. Nevertheless, using syntagms as an intermediary stage, it has been possible to derive many thousands of semantically whole chunks.

Conclusion

The CONYE23 corpus and database of the multi-word units which instantiate how young native speakers use the words on the Chinese NNEC word lists demonstrate that they use words in the same patterns as adult speakers. This is inevitable as the precise meanings of words in any spoken or written text often result from their co-texts of collocation, colligation and grammar patterns. The majority of the chunks that are handed down to young native speakers remain in the language of adult native speakers.

Some language produced for and by young native speakers is dropped as they mature linguistically, socially, physically and educationally. This includes discourse encompassing activities, games, toys, characters in stories, poems and songs, as well as various manifestations of asserting independence as they grow older.

Project outputs have been published in the form of an online database and a 'book of chunks'. The database is a useful resource that will facilitate the study of existing texts and the writing of new texts to be used in teaching materials. This will enable more linguistically and pedagogically reliable approaches to the study of vocabulary. A well-structured syllabus that teaches collocation, grammar patterns and chunks recycles vocabulary in a way that leads students to grasp the roles of co-text and context in creating meaning.

This study has sought to bring to light the highly patterned nature of the language that young native speakers use. The properties of words which they acquire as they are exposed to language and develop their own use of language have the potential to become enriched content in foreign language teaching syllabuses.

Sarah Mount
Alan Pulverness
James Thomas

July 2023

References

Love, R., Dembry, C., Hardie, A. Brezina, V. and McEnery, T. (2017) *The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations*. International Journal of Corpus Linguistics 22(3), 319-344

MacWhinney, B., & Snow, C. (1990) The Child Language Data Exchange System: An update. *Journal of Child Language, 17*(2), 457-472. doi:10.1017/S0305000900013866