There are 72225 entries in Xiandai Hanyu Cidian. We only selected the disyllabic words which results in 44645 entries, including 43653 words that only have one entry and 829 words that have two entries. We further excluded words that are variant characters corresponding to another standard characters, such as 丁宁 which is the variant of 叮咛, which lead to 606 homonyms and 43413 non-homonyms. For these non-homonyms, there are 8891polysemes and 34522 monosemes.

In order to make a comparison with previous simplified Chinese datasets and extend these datasets, we then check if these words are also included in existing dataset, such as Chinese lexical database (Sun et al., 2018), MELD-SCH (Tsang et al., 2018; Xu & Li, 2020; Xu et al., 2022), and AoA word database (Xu et al., 2021).

To prepare the lexical entries for subsequent psycholinguistic analyses, we conducted a structured data-cleaning procedure on the original *Xiandai Hanyu Cidian* XML-style definitions. Using a series of custom scripts (as the results shown in the *XXX\_pos\_sense\_example* pipeline), we retained only linguistically meaningful information which include the headword, POS label, each distinct sense definition, and all illustrative examples (including those separated by “｜” or black-circled numbers such as “❶”, “❷”). At the same time, we removed extraneous information that is not directly relevant to lexical semantic comparison, such as all XML tags (<entry>…</entry>, <hw>…</hw>, <pinyin>…</pinyin>), HTML-style notes, variant-character references, hyperlinks, as well as purely formatting symbols or redundant duplicate lines. We further standardized punctuation by replacing vertical bars “｜” with full stops to separate examples and by keeping standard Chinese sentence-final punctuation and quotation marks only when they enclosed genuine Chinese words.

Through these cleaning steps, we optimized the structure of each entry. In particular, we extracted the POS label (e.g., 名, 动, 形), segmented multiple senses of a given homonym/monoseme/polyseme into separate records, and linked each sense to its own set of examples. Additional columns were generated in the cleaned table so that for every row we could separately access the headword, corresponded POS, sense, and example. For instance, in all the “*XXX\_pos\_sense\_example*” excel files, the newly added column “clean1” contains the fully cleaned textual content of each word information. The subsequent column “pos” records all the POS categories associated with that word. The column “sense1” lists all the sense definitions of that word, and the column “example1” contains all the example sentences corresponding to each sense of that word. (Note: different POS categories, senses, and examples are each recorded in new separate rows.)

Meanwhile, we temporarily removed the cross-references that appear after “另见” in the original dictionary entries, while also retaining for the time being the special usage labels such as 〈口〉 (colloquial), 〈方〉 (dialectal), and 〈书〉 (literary) and similar category markers.

To provide a general overview of the lexical characteristics after cleaning and restructuring the *Xiandai Hanyu Cidian* entries, we then performed various statistical analyses. The results were compiled into an excel file named “*XXX\_data\_preprocess*”, which contains three separate sheets summarizing complementary aspects of the dataset:

**1st sheet: *pos\_stats:***

The first sheet aggregates information on the grammatical categories assigned to all words. Specifically, each POS labelled in column “*pos\_category*” (e.g., 名, 动, 形, 数量词, etc.) or special full combined POS string (e.g., 名；方位词; which was treated as a single distinct category). For each such category we calculated the column “*pos\_number*”: the total number of words exhibiting that POS pattern, and the column “*pos\_words*”: the list of all the corresponding words. This sheet can capture the overall distribution of POS in the curated vocabulary.

**2nd sheet: *sense\_example\_grid*:**

The second sheet summarizes all internal semantic structure, combined POS categories (more than one POS category only for some words), and corresponded number of examples of each word. For every word we calculated the number of distinct senses it has (see column “*sense\_category\_number*”), and further grouped the words by their POS combinations, for instance, words that function both as nouns and verbs (e.g., “上任” could be both “动” and “名”; see column “*pos\_combination*”). We also calculated the total number of examples recorded respectively (see column “*example\_number*”). Next, we summarized the total number of words falling into that cell (see column “*all\_number*”) and also the complete list of those words (see column “*all\_words*”). This sheet highlights how multiple senses of homonym/monoseme/polyseme words and usage examples distribute across various grammatical categories.

**3rd sheet: *blank\_examples*:**

The third sheet summarizes all words whose cleaned record lacks any example sentences. Each row records the word itself (see column “*headword*”) and the corresponding sense definition (see column “*sense\_no\_example*”) that is missing an example (the corresponded Chinese sentence), with a flag “blank” in the “*example\_NA*” column. A summary line at the bottom of column “*total\_summary*” indicate the total number of such sense entries. This sheet helps identify coverage gaps in the dictionary’s sentence example data.