JLect

JLect’s Integration of JMdict – A dictionary sorting comparison

By Zachary Read (March 2013)

I have finally given in and decided to integrate Jim Breen’s dictionary (JMdict) into JLect.com’s database. As a result, standard Japanese entries are now suggested in addition to JLect’s own dictionary entries, which still have priority as the dictionary will remain dialect-oriented.

I thought it interesting, however, to compare how other popular online dictionaries sort words and entries pulled from either the JMdict (Jlect and Tangorin) or the EDICT files (all others). Here are a few comparative results:

“sleep”

JLect	Jisho	Jisho (*Common*)	Tangorin	Jiten	Nihongodict	WWWJDIC (*Common*)
寝巻き	寝	寝	睡眠	休	睡眠	痺れる
寝る	寝	眠り	お休み	寝	眠り	二度寝
寝る	寝	睡眠	眠り	休む	お休み	寝る
休む	眠り	安眠	寝	昏睡	寝	眠る
寝かす	スリープ	熟睡	スリープ	熟睡	–	金縛り
浅い	睡眠	寝苦しい	目やに	寝る	–	寝苦しい

“cow”

JLect†	Jisho	Jisho (*Common*)	Tangorin	Jiten	Nihongodict	WWWJDIC (*Common*)
牛	乳用牛	牛乳	雌牛	牛	牛	牛舎
卑劣	ミノ	搾乳	牛	牛舎	雌牛	牛乳
子牛	牝牛	牛舎	乳牛	牛乳	牝牛	搾乳
弱虫	雌牛	牛	牛舎	搾乳	–	子牛
卑怯	牛乳	–	褐毛和種	海牛	–	牛
不器用	搾乳	–	牛小屋	滑革	–	臆病者

† *see addendum section*

In the first search for the word “sleep”, it becomes clear that there are some significant differences between these online dictionaries. In the first seven results, JLect’s dictionary seems to favour verbs as opposed to nouns, which are clearly favoured by Jisho, Tangorin and Nihongodict. Both Jiten and WWWJDIC mix up the results, but Jiten seems to sort by character length, and I assume WWWJDIC just gives out the first database results, though I could be wrong.

In the second example, some interesting differences arise. Both JLect and WWWJDIC use a MySQL “LIKE” search query, which gives out fuzzy results for the word “coward”. While this may not be desirable here, since a “cow” and a “coward” aren’t exactly related, such a feature might come in handy for finding compound words, like “land” and “landmass”. However, though JLect here prioritizes the right word, while WWWJDIC doesn't, the WWWJDIC provides far more cow-related results than JLect does. In the other dictionaries, 牛 clearly comes up in the first few results, except in Jisho's standard search.

So far it seems that the other dictionaries are providing somewhat better results, but consider the following search.

"blue"

JLect	Jisho	Jisho (*Common*)	Tangorin	Jiten	Nihongodict	WWWJDIC (*Common*)
紺	ブリュー	青空	青	紺	ブルー	ブルー
青	青色	濃紺	青い	青	青	ハギ
重い	青色	水色	青色	水色	青い	青雲
青い	青空	コバルトブルー	蒼然	青い	青色	青空
水色	青天	青雲	ブリュー	青雲	蒼然	労働者
青空	蒼空	ブルーカラー	重い	青空	–	ブルーカラー

Since these types of dictionaries are mostly used by language learners, I would expect the term 青い to be amongst the top results, but this is not the case for either Jisho or the WWWJDIC, even with the "common" feature selected.

Considerations

Comparing all three examples, it might be easy to conclude that, ignoring JLect, Tangorin, Jiten and Nihongodict have some of the best search filters. However, when searching for a verb such as "to sleep", Tangorin falls a little short and only gives the common verb 寝る as the seventh result, while Jisho gives it as the eighth result. The WWWJDIC is even more worrying on this front since it doesn't provide any results for the search string "to sleep".

That said, Jiten and Nihongodict are on the mark for verb searches, but I would have to concede that Nihongodict generally does a slightly better job, considering the results of other searches such as "to eat". However, Nihongodict does appear to have an issue with its results: when searching "to sleep", it suggests an array of verbs, including 寝る; but if the user presses the "enter" key, only two verbs show up.

JLect's sorting pattern

Since JLect implements the JMdict file and not the EDICT file, which are organized differently, the sorting implications are a little different. JLect simply loads the entire file, reorders all entries by the length of the first kana (*reb*), and then sorts words of equal length by the number of priority information (*ke_pri*) it has labelled. I considered sorting words by their "nf" priority number, but unfortunately, this caused words such as 労働者 "blue-collar worker" to appear at the top of a search for "blue", while also forcing the term 青い "blue; green" down quite a bit, which was not desirable.

In the long run, I plan to introduce a simple “up and down” voting system, so that users could eventually decide for themselves what words should have higher priority. In order to offset potential abuse, I might consider keeping track of entry views, or add an administrative feature so that I can manually set the priority of certain words. This, of course, is still up in the air.

For now, however, it needs to be clearly noted that JLect’s priority is placed on providing its own dictionary results, which focus on the Japonic languages and dialects of Japan, before suggesting results from the JMdict. As a result, searching for “cow” will give you the Kagoshima and Miyazaki term べぶ way before the standard Japanese term うし.

Addendum

Since first writing this article, a few changes have been made to JLect’s database queries, effectively changing some of the result outputs. Short words now default on fulltext search, so searching for “cow” will no longer give “coward” as a result. The output now gives the following:

JLect
牛
子牛
牛舎
搾乳
乳牛
ミノ

For words of two letters or less, the search defaults on Kana, though suggestions are provided for obtaining English-only results. This is because most one- or two-letter English words are prepositions or pronouns that are difficult to look up in a bilingual search, so the suggestions help guide the user to a more meaningful search key. For example, searching for “of” would provide a link towards the search “possessive”, which then gives results for the particles の, が, ケ and つ, among others. Such a feature could prove useful in other bilingual dictionaries such as Tangorin.

References

- JLect: http://www.jlect.com/
- Jisho: http://jisho.org/
- Tangorin: http://tangorin.com/
- Jiten: http://jiten.net/
- Nihongodict: http://www.nihongodict.com/
- WWWJDIC: http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C