Breaking up the Kanji

Jeroen Hoek MA BEng

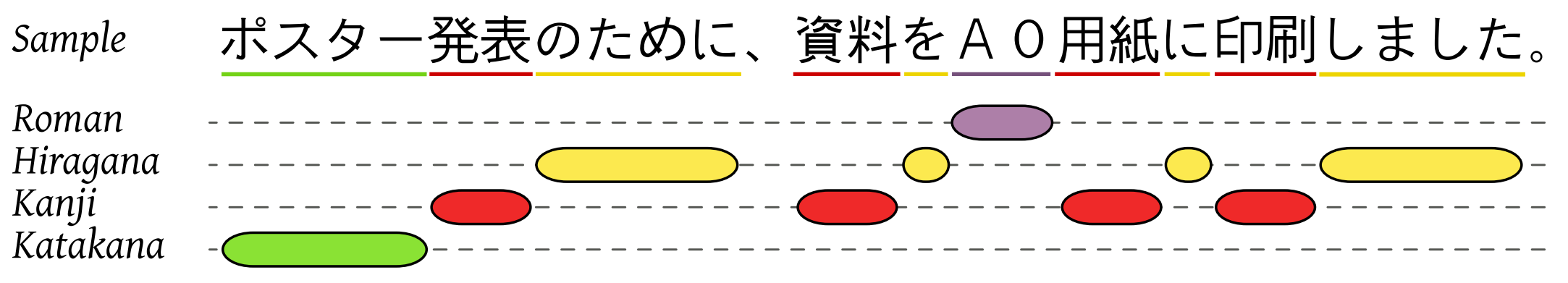
Summary

In this research I present the design for a novel method of **searching** for, **exploring**, and **entering kanji** (the numerous logographs used in Japanese) in a digital environment. By allowing direct interaction with kanji and its **components**, the learner can utilise his acquired knowledge of kanji **phonology** (or: *"Why are all kanji with this component often read like this?"*), kanji component **semantics** (or: *"How to write that one character again? I know it has the 'tree' radical in there somewhere..."*), and the **spatial layout** of components.

Background

The Japanese language

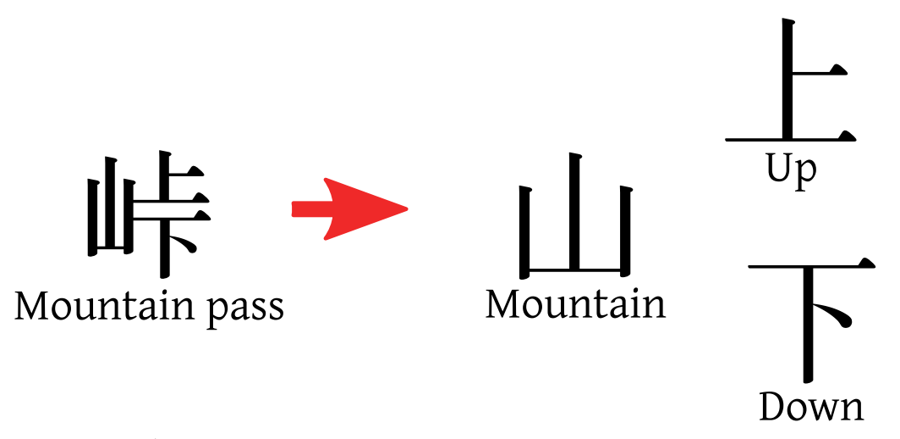
Japanese is written using four scripts; two syllabaries, the Roman alphabet, and the logographs known as **kanji**. These scripts are used in mixed sentences:



"For the poster-presentation, I printed my material on a sheet of A0 paper."

For **basic literacy** mastering nearly **2000** kanji is a necessity; to be able to read newspapers, **3000**. Even with knowledge of all those kanji, encountering unknown, unfamiliar, or downright obscure kanji is part and parcel of learning and using Japanese; up to **6000** kanji can occur in modern-day usage.

No two kanji are the same, but the majority (90-93%) of daily use kanji can be split up into distinct, recurring, distinguish-able components, often kanji in there own right:

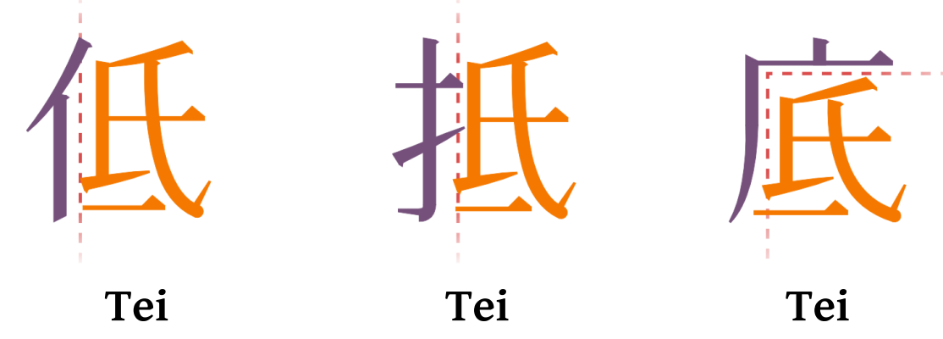


The kanji 峠 (*mountain pass*) split up into its components.

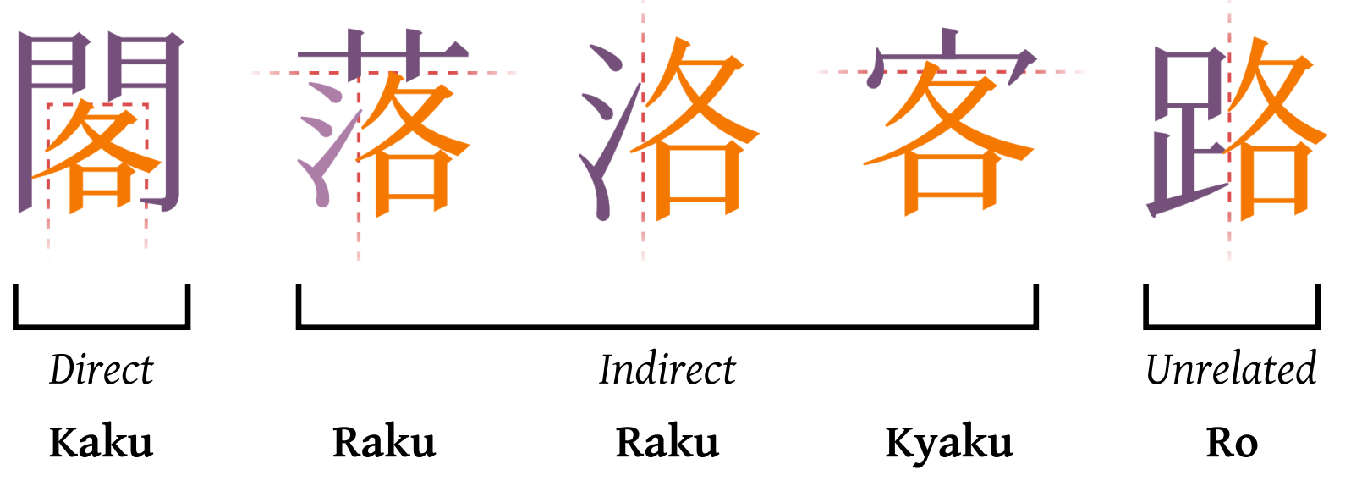
Readings

Kanji were imported during periods of cultural borrowing from China; along with their meaning and reading. Adapted to the Japanese pronunciation these became the **on-readings**. The **kun-readings** came from the words already present in Japanese.

From a cognitive standpoint, the important distinction between the two is that on-readings can often be surmised from a **phonetic component**; most kanji can be split up into a semantic, and a phonetic component. Roughly **80%** of kanji in common use are such **semasio-phonetic** characters.



*Three kanji sharing the same phonetic component (氐 , highlighted in orange) with the on-reading **tei**.*



*On-readings do not always correspond directly to the phonetic component; readings can also be indirectly related (slight variations), or not related at all. These kanji share the component 各 (**kaku**).*

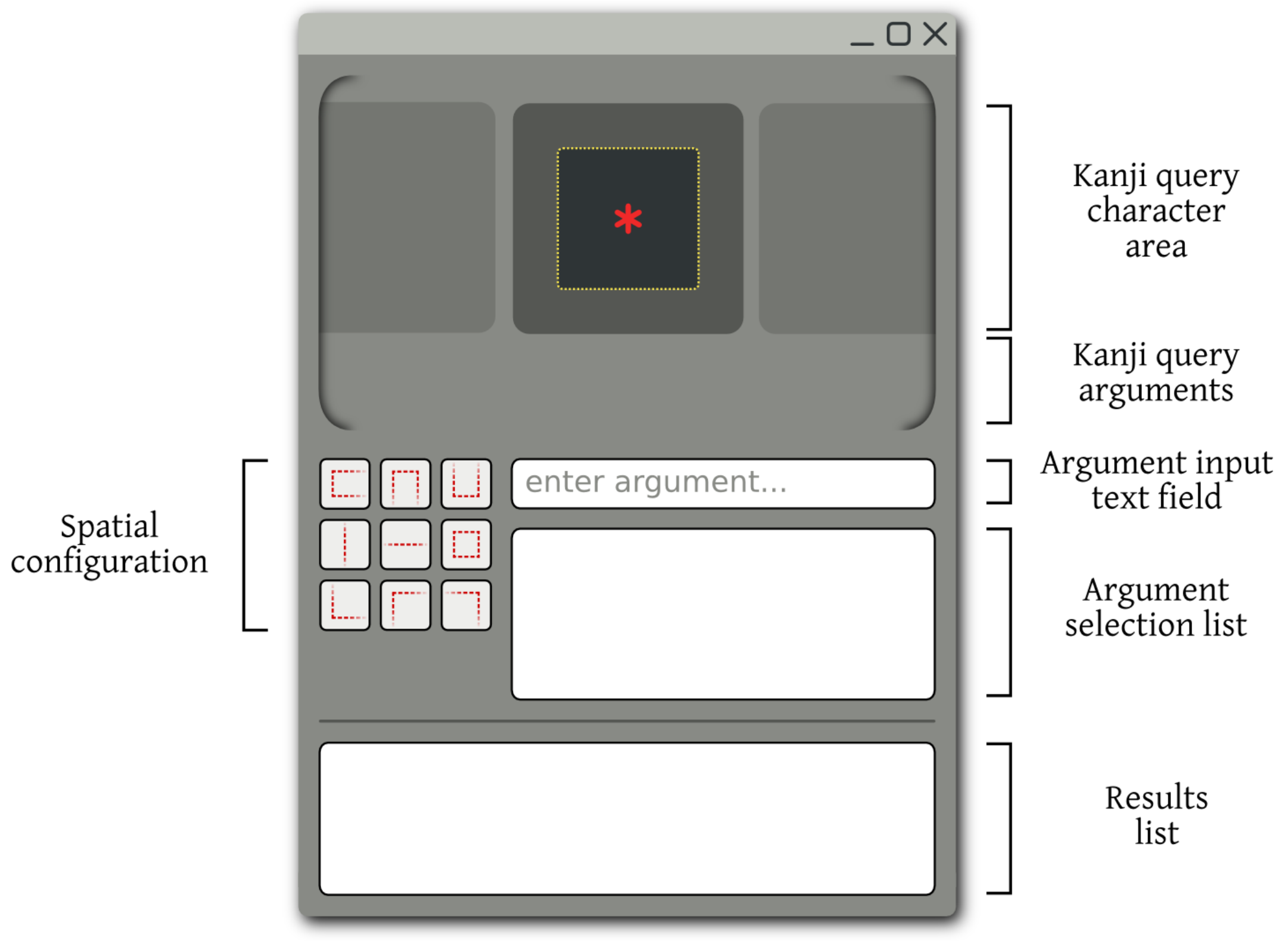
Rationale

Searching for kanji using currently available paper and software dictionaries, depends heavily on the ability to identify the "**radical**" of kanji (in complex kanji a component which acts as a rough semantic classifier), and the number of (brush) **strokes** in the remainder of the character. While sufficient for many cases, the spatial composition of kanji is completely overlooked. Research in the field of cognitive linguistics suggest that we do use, and rely on, the sub-character features of kanji. Can we exploit this fact?

Theoritical footholds

- A good grasp of spatial layout, as well as phonetic and semantic components facilitates the process of learning and retaining kanji.
- Humans process kanji in chunks, not as whole characters (component awareness).
- In the mental lexicon, kanji phonology and orthography are strongly linked.

The "Kanji Navigator"

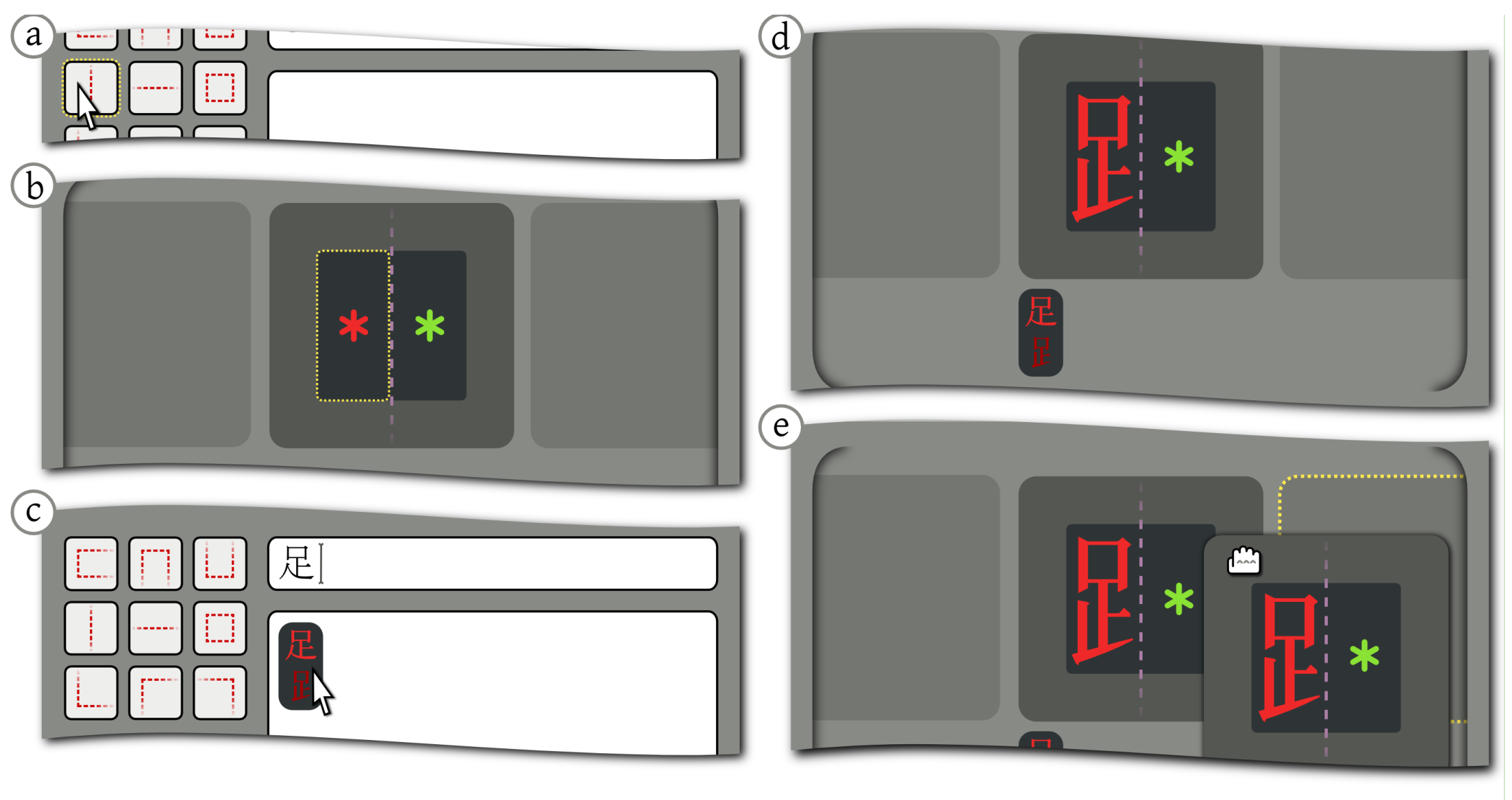


Approach

Explicitly enable interaction with kanji; allow modification of kanji to form new queries for use against a Japanese character or vocabulary dictionary. The concept is best explained with pracitcal examples:

Case 1: Search

An intermediate learner is trying to read a Japanese article, and runs into the two-character word 躊躇 (*hesitation*). He does not know its pronunciation, and while he correctly guesses that the radical for both kanji is 足 (foot), counting the remaining strokes proves difficult. Moreover, he wants to find the meaning of the whole word, so finding one of the kanji still necessitates looking for the whole word in the list of words that feature it. Using a software dictionary equiped with the Kanji Navigator, he may instead opt to follow these steps to find the word:



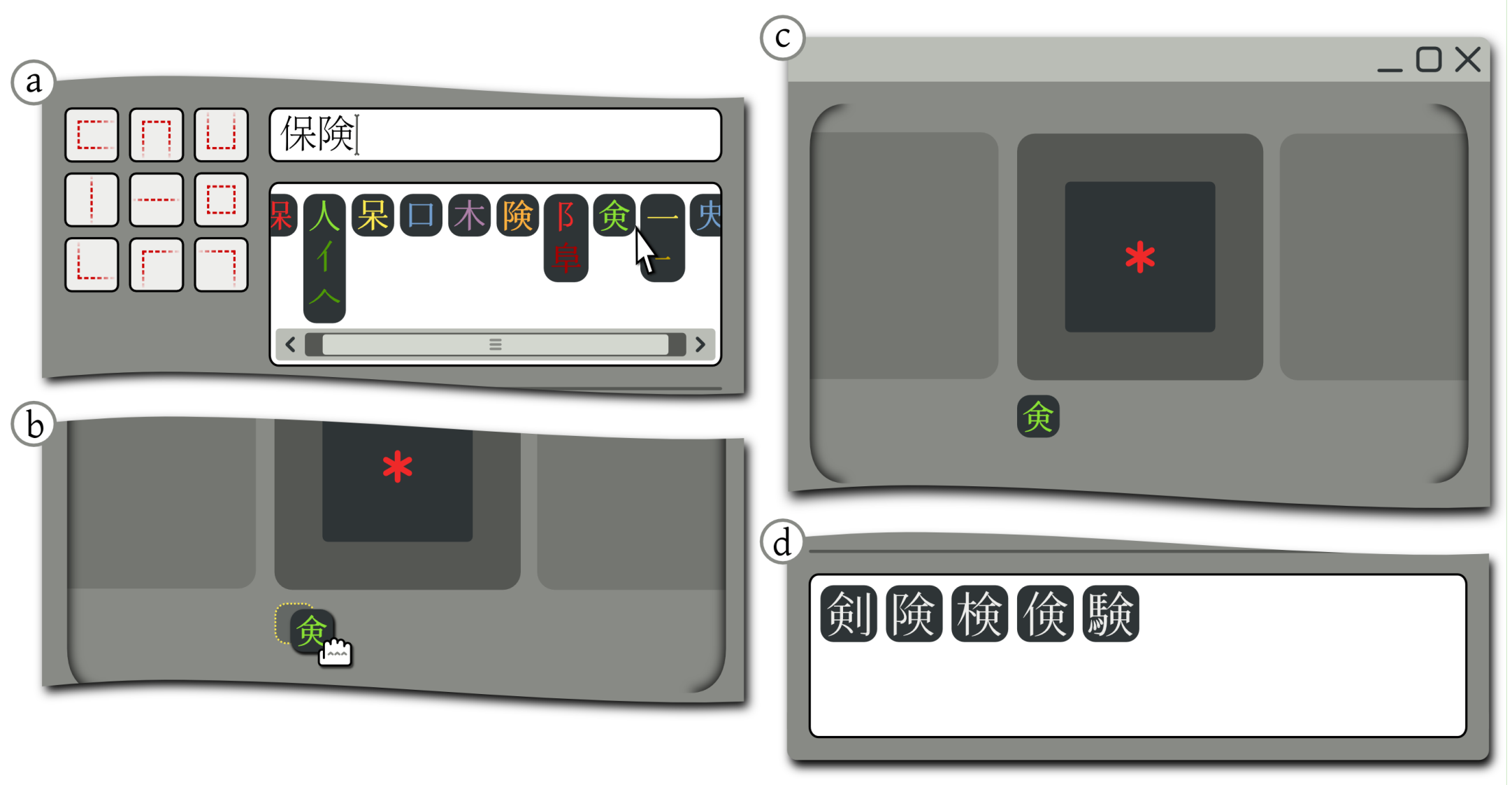
ⓐ ⓑ Split the kanji field in two by selecting the corresponding spatial layout button.

ⓒ ⓓ Enter 足 as an argument, and drag the argument tile with 足 (and its variant shapes) to the leftside of the now divided character area.

ⓔ Duplicate the "Any kanji with 足 on the left" character query to form a query for "Words with two consecutive kanji with 足 on the left"; only a handful of words match this query.

Case 2: Inquiry

A student notices that both 険 and 検 are read ken in compound kanji words (such as 保険, insurance), and inquisitively, on a whim, decides he wants to study all common use kanji that share the 僉 bit:



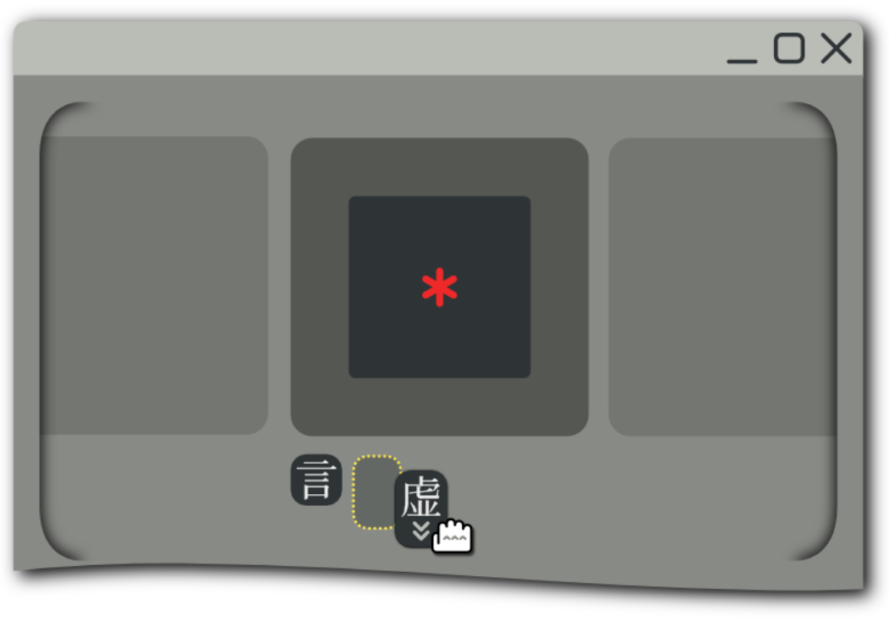
ⓐ Enter a familiar word with one of these kanji, and from the components and component groups found by the software, select the component he suspects is responsible for the **tei** sound.

ⓑ ⓒ Drag the 僉 argument to the kanji character query area, and place it beneath it as an argument for the whole kanji (without a specific spatial argument).

ⓓ Assuming this student (temporarely) limited the software to only return results for the nearly 2000 common use kanji, five kanji are returned; all of which can be clicked for queries against a vocabulary dictionary, or more information on their readings and composition.

Case 3: Rare kanji

Japanese is entered by typing in the pronunciation of a word using a normal Roman-letter keyboard, and selecting the proper word in kanji from a list of matching options. But what if the kanji we want to enter is so rare that it cannot be found using this method, even though the font to display it exists? Or what if we simply do not know the pronunciation? The kanji 譃 (a creative alternative for 偽, *to deceive*) can be found in Natsume Sōseki's famous novel "I am a Cat":



Here, the hypothetical user choose to enter the two components that make up this kanji — simply because those are common kanji he knows how to enter; 言 is part of the very common verb 言う (**iu**, *to speak*), and 虚 may be taken from a compound word such as 虚偽 (**kyogi**, *deception*). No spatial information is entered in this example; specifying these two components is already sufficient to limit the results to just the one kanji sought.

Future work

- Finish and release an experimental free software implementation and encourage feedback from users.
- Conduct controlled experiments with L2-learners (including advanced learners!) comparing this concept with traditional search and lookup tactics, and confirm (or invalidate) the notion that active use of kanji sub-character features is beneficial to the learning process for L2-learners.
- Refine the design and make the software available on multiple platforms.

References & thesis

A full account of this research, along with references can be found in my MA thesis, available on-line:

`http://handle.jeroenhoek.nl/hoek_2009`