TOWARDS LINGUISTIC KNOWLEDGE DISCOVERY ASSISTANTS: APPLICATION TO LEARNING LEXICAL PROPERTIES OF CHINESE CHARACTERS

Georges FAFIOTTE & François TCHEOU
GETA, IMAG (Université Grenoble 1 & CNRS) - BP 53, F-38041 GRENOBLE Cédex 9, France
georges.fafiotte@imag.fr / francois.tcheou@imag.fr

ABSTRACT

It is highly desirable that users of systems which include NLP-based components, ranging from grammar-checkers to MT systems, can access the underlying Linguistic Knowledge Base in a natural and gratifying way. Our research aims at developing such Linguistic Discovery Assistants by merging hyperdocuments, Data Base Management Systems and interpretive adaptive interfaces.

We have followed a stepwise approach to the idea in the context of the discovery and learning of lexical properties of Chinese characters, by developing several prototypes. We see this system as a facet of a broader base including dictionary knowledge.

Keywords
Cooperative Discovery Assistant, Linguistic Knowledge Observatory, Lexical Properties Discovery, Computer Aided Learning, Kanji, Hanzi, Chinese Characters

1. MOTIVATIONS: COMPUTER AIDED DISCOVERY OR LEARNING OF LINGUISTIC KNOWLEDGE

1.1 A need for making linguistic knowledge accessible to the user, in Personal MT

A current trend in Personal Machine Translation tends towards opening to the user the linguistic data that the system is operating upon [1]. Such 'discoverable' environments should allow some free, self-planned, or coached investigation to users, and provide these in a suitable explanatory form with a large part of the linguistic material embodied in the personal lingware: lexical data bases, syntactic patterns or syntactic rules modules, semantic contrastive aspects, etc.

Our work is oriented towards a particular aspect of such 'open lingware': the grasp, eventually the learning, by monolingual writers or editors of a document who are working in a language they know imperfectly, of the lexical properties of the language to be used.

1.2 A new resource different from dictionaries

In the context of lexical properties we may at first consider dictionaries to be a straight response to such a demand. They usually require of the user some premodelled view of the very organization of the lexical data, a pragmatic know-how of their legibility, or real mastery in order for the searcher user to perform a sensible pruning of the available information. This is particularly true with the lexical properties of the languages we are concerned with in this project, Chinese and Japanese.

Users may experience the complexity of the process when, starting from uncertain or incomplete chunks of recollected knowledge, they wish to investigate a word to be ascertained, a nuance to be expressed, an ideogram to be remembered. Such situations clearly demand a reshaping of dictionaries as interactive knowledge bases, and the proposal of components and cooperative interfaces which could offer alternate access schemes to lexical data bases, or views of them [11].

Some integrated systems for Dialogue Based Machine Translation intend to provide the author with the means for interactive consulting of linguistic facts or rules, for instance in the context of lexical or syntactical disambiguation or indirect pre-editing of contextual semantic features, specific to the text to be composed. The LIDIA architecture [2] and the NADIA model [9] certainly illustrate this approach.

1.3 Discovery Assistants, Cooperative Observatories

The development of interactive environments for monolingual writers leads to modelling new functions for documentation, self-documentation, self-learning and management of individualized personal knowledge bases, to be pooled into open encyclopædic 'discovery environments', specific components for NLP systems.

Such technologies as hyperdocuments, multimedia and voiced data bases, adaptive interfaces, and the benefits of Computer Aided Learning techniques may merge to offer solutions in the realm of such 'cooperative observatories' of linguistic knowledge.

Our project has a stepwise approach to the idea, in the case of the lexical properties of the Chinese ideograms.

2. PROJECT OUTLINE AND PROTOTYPING SCHEME

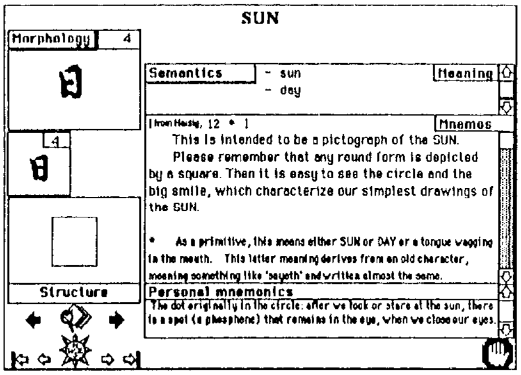
2.1 Computer Aided Chinese Character Learning

The work we report here stems from the initial modelling of an AAOCC system (for 'Apprentissage Assisté par Ordinateur des Caractères Chinois'), intended to provide motivated users with an adaptive environment for the autonomous discovery or review of character properties [4], with a deliberate restriction to a hanzi / kanji subset of characters.

Han-zi simply means Chinese Character, and kan-ji (which alliterates the former, and means the same in Japanese) refers to a small subset of characters that written Japanese almost entirely borrowed from hanzi, and from the combination of which Japanese words are derived. We shall call 'hanzi / kanji' the intersection set, that is the hanzi which also are Japanese kanji.

The conceptual model of the lexical data base schematizes different views and levels of investigation of the material.

- A first alternative for the user is to explore a *language-independent view* (Fig. 1) of the characters (intrinsic morpho-semantic properties of the hanzi / kanji, shared by written Chinese, Japanese and Korean), contrasting with *language-related views* (Fig. 2, Fig. 3).
  These are enriched with groups of other character properties (phonetics, morphological similarities, contextual semantics...), all strongly relevant to one of the three languages of use —presently, only with the core of the Chinese instanciation.



SUN

Morphology 4

日

Semantics - sun
- day

Meaning

[from Heisig, 12 •]

Mnemos

This is intended to be a pictograph of the SUN.
Please remember that any round form is depicted by a square. Then it is easy to see the circle and the big smile, which characterize our simplest drawings of the SUN.

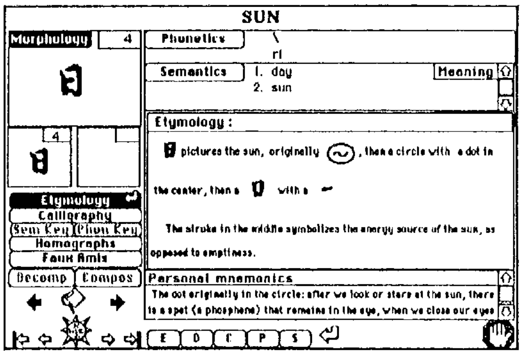
• As a primitive, this means either SUN or DAY or a tongue wagging in the mouth. This latter meaning derives from an old character, meaning something like 'sayeth' and written almost the same.

Structure

Personal mnemonics

The dot originally in the circle: after we look or stare at the sun, there is a spot (a phosphene) that remains in the eye, when we close our eyes.

Fig 1: Language-independent fundamental information for the character SUN (or DAY)



SUN

Morphology 4

Phonetics \ ri

Semantics 1. day
2. sun

Meaning

Etymology :

日 pictures the sun, originally ⊙, then a circle with a dot in the center, then a 日 with a —

The stroke in the middle symbolizes the energy source of the sun, as opposed to emptiness.

Etymology

Calligraphy

Sem Key | Phon Key

Homographs

Faux Amis

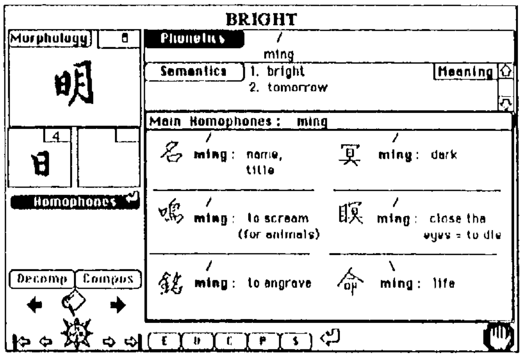
Decomp | Compos

Personal mnemonics

The dot originally in the circle: after we look or stare at the sun, there is a spot (a phosphene) that remains in the eye, when we close our eyes

E | D | C | P | S

Fig 2: Language-dependent complementary information (Etymology) for the character SUN (or DAY)



BRIGHT

Morphology 8

明

Phonetics / ming

Semantics 1. bright
2. tomorrow

Meaning

Main Homophones : ming

名 / ming : name, title

冥 / ming : dark

鳴 / ming : to scream (for animals)

瞑 / ming : close the eyes = to die

銘 / ming : to engrave

命 \ ming : life

Homophones

Decomp | Compos

E | D | C | P | S

Fig 3: Language-dependent complementary information (Chinese Homophones) for the character BRIGHT

- In addition, the user may differentiate between two levels of information: a main view condensing *fundamental properties* (Fig. 1) and a second *complementary view* with more advanced lexical characteristics (Fig. 2, Fig. 3).

We willingly refer to J.W. Heisig's work [7], which emphasizes the role of a corpus of predefined mnemonic labels attached, one to one, to kanji and possibly to some subcharacter morphological components. Such mnemonic marking will supplement the etymology, while enhancing the user's «imaginative memory» and strongly relaying visual memory. We also strongly invite the user to personalize his knowledge base, through adding his own mnemonics or imaginative productions.

2.2 Basic functionalities

The lexical base initially modelled provides a good coverage of character properties [3]:

- using one view of the character base, the user may explore language-free morphology (pictogram, stroke number, overall structural vignette, semantic radical, confusing similarities...) and universal semantics,
- on the other hand —and on the other view— an author can discover language-relevant morphological properties (phonetic component in its structural valency, homographs, positional variants for compound characters, use in composition...), language-tied phonetics (written and voiced pinyin and tone, homophones...), language-related semantics.

A very detailed set of structural vignettes (similar in spirit to Halpern's patterns [6]) is proposed with a digital coding, and validated over some 1500 kanji in the Heisig progression. They may deeply improve learner recall of the overall structure of a character. We have for instance, with the grey tint giving the position of the semantic key:

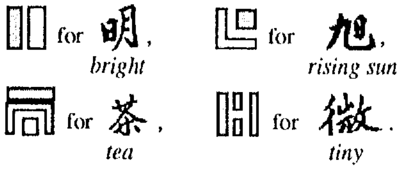


Fig 4: Examples of structural vignettes

Such patterns may also be invoked by the user for the study of compound characters to be derived from a given kanji, when recalling some particular structural model.

2.3 Multiple prototyping development scheme

The first prototype implemented the conceptual model of the property base as a highly structured hypertext, using HyperCard on Macintosh. We considered this platform a good trade-off between a valuable interaction management system, and a temperate framework for expressing the object-oriented and reusable view of the base.

We then adopted a twofold development scheme, with both *incremental prototyping* in HyperCard, for surface multimodal data and for the interface functional layer of the application, and *parallel prototyping* on different development platforms, to explore different structuring and searching methods for the character base.

Three tracks were experimented to express *character search modalities* in more refined interactive ways:

- first a Data Base Management System approach, through two variants: (1) object-oriented modelling of the base (in LOOPS on a Xerox workstation), (2) a relational scheme on a standard medium (an Oracle encapsulation in Hypercard on the Macintosh);
- (3) various sketches of a knowledge base were prototyped in Prolog with a simplified user interface.

3. A DATA BASE ORIENTED MODEL, WITH FLEXIBLE EXPORATION

The prototype may here administer different users (learners, didacticians) with data protection, manage a standard static character property base, maintain session journals and user profiles, tracing their work in the base.

3.1 Multicriterial search of Chinese Characters

There are two main access schemes to characters.

- *Direct designation* is based on a simple selection of the character icons on character boards. The surface organization of the base in character series or lessons matches here structural and pedagogical motivations initially expressed by a didactician (Fig. 5).

Multicriterial search (MCS) allows the user, starting from a partial description of the character, from tentatively discriminant properties he may recall, to refine or focus his request. Partially erroneous demands should be managed adequately by the system (while suggesting default or alternate determinations for dubious proposals, or suppressing irrelevant ones).

The elements remembered or evoked by the learner are put forward in a criteria array organized in 3 lexical property subclasses: potential characteristics about the sought character itself, about its semantic key, and about its phonogramme if it exists (the phonetic marker component, very often structurally present). See Fig. 6. The grid presents main discriminant criteria (left) and secondary characteristics (right). In the MCS of complex characters, we put some emphasis on using structural vignettes and positional morphology aspects.

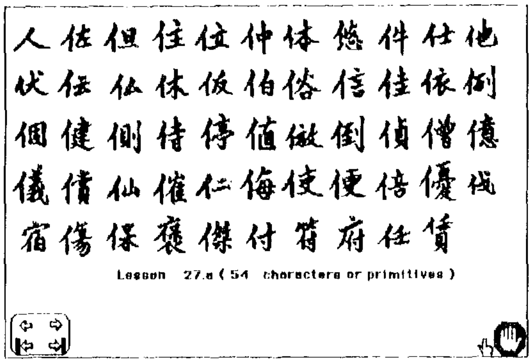
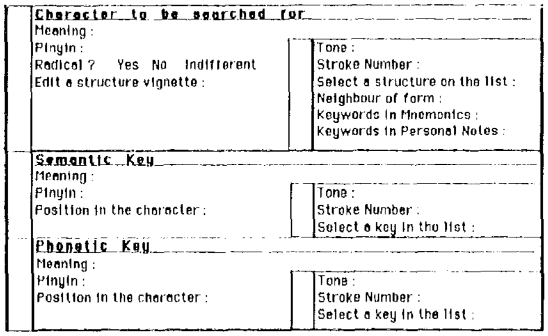

人 佐 但 住 位 仲 体 悠 件 仕 他
伏 伍 似 休 仮 伯 俗 信 佳 依 例
個 健 側 侍 停 値 做 倒 偵 僧 億
儀 債 仙 催 仁 侮 侵 便 倍 優 伐
宿 傷 保 襲 傑 付 符 府 任 貸
Lesson 27.a (54 characters or primitives)

Fig 5: Direct character selection
(series of characters with the semantic key MAN)



Character to be searched for
Meaning :
Pinyin :
Radical ? Yes No Indifferent
Edit a structure vignette :

Tone :
Stroke Number :
Select a structure on the list :
Neighbour of form :
Keywords in Mnemonics :
Keywords in Personal Notes :

Semantic Key
Meaning :
Pinyin :
Position in the character :

Tone :
Stroke Number :
Select a key in tho list :

Phonetic Key
Meaning :
Pinyin :
Position in the character :

Tone :
Stroke Number :
Select a key in the list :

Fig 6: The multicriterial search grid

Results of the search are available as character icons. The user may collect some of them for later study. Direct observation of a character, straightaway, is possible too.

3.2 Evolutive surface structuration for the base, and flexible session planning facility

In its surface design, the standard main character base is originally segmented in lessons, or collections, according to Heisig's view of a pedagogical progression for a methodical discovery of kanji.

On the DBMS driven prototypes, surface reconfiguration of the base is made possible, using in turn both multicriterial search or direct character collecting.

- It may allow *didacticians* to express different views of the intrinsic property base, to restructure character lessons for pedagogical reasons, to elaborate alternate progression schemes (involving for instance use frequency, series with a common semantic key, series with a shared phonogramme, etc). They are here enabled to propose new palettes of 'predefined lessons' with alternate discovery paradigms or mnemonic systems related to various linguistic and cultural views, on which learners may express preferences as well.
- The *learner* himself may build and maintain one or several personal (sub)bases, or collections: series of characters that he selects using coherent criteria, which he plans to explore in future sessions, collections reflecting a personal thematic organisation of the discovery, simple reservoirs of characters built by free picking or rational collecting, through digressive or systematic navigation in the base. The learner might here express, discover, refine, some personal discovery 'customs' according to a cognitive style.

Both types of users are thus allowed (with appropriate access rights to such restructuring resources) to enrich the collection of existing views on the property base, to edit and to reshape predefined lessons or collections into 'personal lessons'.

Along with instant feedback and regular reviewing of the actual work, the system here has some incentives to more intention-driven, self-guided learner activity, through short-term session planning and long-range curriculum self-organization. Case by case spontaneous consultation of lexical information is of course still advised.

3.3 Observing user activity

In our view such a function is essential on the way to interpreting and modelling user activity. On one of the prototypes, the tracing resources provided a first basis

- to elaborate flexible, analytic and synthetic feed-back or witness functions for the user-discoverer interface,
- to build up an information pool about user behaviour and discovery strategies.

While extracting data from the sessions base, we may sketch synthetic session journals, synthetic views on each character (or character property) for a user or for all users.

3.4 Parallel prototyping

The DBMS view and the MCS (MultiCriterial Search) were prototyped on different development platforms.

a) an Object Oriented prototyping (LOOPS)

A COCOA release of the AAOCC project was rewritten in a homogeneous Object-Oriented frame for a corpus of about 100 characters. Results and performances were quite encouraging, though on this small scale model.

b) a classical relational framework (Oracle)

We also adopted, on another prototype (CACAO-4), an integrative scheme merging two functionally specialized environments: HyperCard for multimodal interactive front-end resources, and an Oracle kernel for managing the bases and the user queries. We here aimed at exploring implementation paradigms for larger scale character bases.

The system first configurates entities in an Oracle property base, while extracting relevant data from the HyperCard hyperdocument fields. At query time, arguments sent from the interface layer will generate SQL requests. System response is displayed back to the user, who then collects characters for later use, or directly picks up needed multimedia data on the properties sought.

Though on a still small scale prototype, the relational DBMS scheme eased data security, coherence qualities, as well as some quantitative development aspects.

4. EXPERT ASSISTANCE TO CHARACTER IDENTIFICATION

4.1 Expert System oriented schemes

We prototyped a similar 'scale model' corpus into small knowledge bases (facts, structural and other property rules), providing for deductive and explanatory functions. Data-driven and goal-driven schemes were experimented in a small Expert Assistant for Character Identification.

We try to initiate more interactive multicriterial searches, while coupling very discriminant semantic characteristics (the meanings of the expected character, of its semantic or phonetic keys), less selective indexing (stroke number, pinyin...), with a tentative iconic specification of structural properties.

When efficient criteria are missing, we think such a visual structural recall to be helpful, with or without strong spatial positioning and applied to subcomponents with semantic or phonetic key functions, through a progressive opening or refinement of structural vignettes.

4.2 Cooperative search

Later the user should be able to express preferences regarding the prompting profile or the search strategies adopted by the Discovery Assistant (DA) in cases of underspecified or possibly erroneous queries. The DA could group results or hints, in order to prevent over-stepped talkative dialogues. System explanation, if activated, could justify or illustrate side-hints with details.

In the Annex example, the user tries to recall a character 照, a morphological tree of which is shown below (the semantic keys of the subcomponents being squared) —but he actually knows very little of all this.

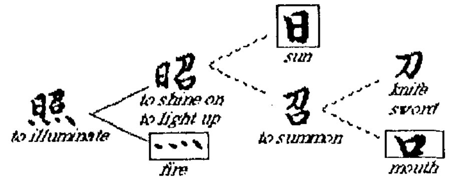


Fig 7: A morphological tree for the character ILLUMINATE

A first scenario (see the Annex) exemplifies a cooperative dialogue with a beginner user whose spontaneous search strategy will favour visual structural characteristics (with no particular attention here to any phonetic unit in the character). The DA repeatedly asks for any semantic recall of another component (a most discriminant criterion), but the user declines. He is therefore asked and helped to visually refine the overall structure, sucessively ⿰▯⿱⿰▯▯▭, ⿱▭▭ then ⿱⿰▯▯▭, finally ⿱⿰▯⿱▭▭▭.

In a second scenario, the user soon proposes (at dialogue step 2) a component *knife* (刀),which he figures to be present —and which really belongs to the investigated character. With the current kanji base, adding this very discriminant element directly produces the unique final solution: *to illuminate*.

In a third scenario, the user erroneously recalls (at step 2) the possible presence of the radical *strengh*, (力), instead of the very similar *knife* radical. The system will first exhibit an empty response, which is correct here. But knowing about this misleading similarity (a 'faux amis' property), it then suggests changing *strength* into *knife*. If the user acknowledges the proposal, the proper character *to illuminate* is reached right away.

5. A PROSPECTIVE VIEW

5.1 Functional development methodology

In the context of our prototyping effort, we would like ideally to design the application with a three-fold functional architecture:

- a highly interactive interface layer developed on an appropriate authoring environment, for the surface multimodal representation of the lexical knowlege,
- an object oriented DBMS to express the core of the structural knowledge, to implement efficiently heavy data searching and to structure and update the user history profiles and personal bases,
- a declarative or deductive programming environment or expert system generator, in order to express both the strategy models of a coached or error-compensatory multicriterial character search, and first elements towards typed behavioural profiles and users' discovery strategies.

This could possibly lead to 'client-server' architectures with distributed logical resources in the way of 'white-board' schemes [8]. It seems that the interoperability expected from multiplatform development environements will further such functionally distributed design.

5.2 Cooperative adaptive accompanying interface

To summarize, we intend to develop the first draft of the exploration assistant, towards

- free surface restructuring by the learner of his personal knowledge base, according to a thematic or methodological view he follows,
- personal management —planning, monitoring and reviewing— of sessions or inquiry sequences,
- synthetic or analytical follow-up of the discovery, working on a metaphor of the subbase being explored, with qualitative indicators on the actual navigation,
- production (according to user preferences) of session journals, profile status, global curriculum surveys,
- issuing of personalized written, magnetic or audio documents, for remediation and in-depth work.

5.3 Towards integrated polymorphic or multiple-view Lexical Knowledge Bases

Our system can be viewed as a facet of a broader 'environment for an encyclopædic discovery' with other modes of activity: self-review, semi-tutored lessons, where character thematic 'collections' would drive the discovery.

It would be desirable to be able to find, through different views, in one and the same knowledge base,

- all the information that the Halpern Japanese-English dictionary [7] offers, with the words built from characters, Japanese pronunciation, a sound thesaurus,
- the data of a large 'Chinese-usual language' dictionary,
- character etymology [10], classical, usual calligraphy,
- a language-independent view of the hanzi / kanji, augmented with a progressive and comprehensive proposal of mnemonics in Heisig's style [7], but culturally related to the user's native or usual language,
- the resources and modalities modelled in the AAOCC prototypes, for accompanied hyperdocumental navigation, expert character identification, for the creation and updating of personal subbases or thematic character collections, among other features to appear.

CONCLUSION

We advocate the development of system components for helping authors to access the underlying linguistic knowledge, among others in Personal or DBMT systems.

Such Discovery Assistants (DA) should certainly be highly cooperative, namely show sensible interactivity (within multimodal hyperdocuments and object DBMS frameworks), provide some ways to tentatively adapt to users' mnemonic and cognitive customs, and preferably first be user-tunable: i.e. they could offer means for the users themselves to refine and express their preferences in terms of search strategies (spontaneous, self-planned or coached), their planning intentions for a working sequence, as well as means for an efficient follow-up of their activity. DAs should in our view rather first enhance both user's natural intelligence towards more reflective interaction modes, and user's self-guidance aptitudes.

In the framework of a lexical property base of hanzi / kanji, we have developed, as very first steps, a multiple prototyping of such functions, while exploring object oriented, relational, and deductive (rule-driven) schemes.

We expect progress from patient observation and modelling or user activity, and from the availability of multiplatform software development tools, merging different classes of functionals, heading towards polymorphic or multiple-view knowledge bases.

REFERENCES

[1] **Boitet Ch. (1990)** *Towards Personal MT: on some aspects of the LIDIA project.* Proceedings of Coling-90, 08/90, vol. 3/3, pp30-35.

[2] **Boitet Ch. & Blanchon H. (1993)** *Dialogue-based MT for monolingual authors and the LIDIA project.* Proceedings of NLPRS'93, Fukuoka, Dec 6-7 1993, vol. 1/1, pp208-222 .

[3] **Fafiotte G. (1990)** *A Self-Learning System for Chinese Characters.* Proceedings of COLING 90, Helsinki, Aug 20-25 1990, H. Karlgren ed., ACL, vol. 3/3, pp351-354 .

[4] **Fafiotte G. (1990)** *Apprentissage assisté par ordinateur des caractères chinois.* Proceedings of 10èmes Journées Internationales "Les systèmes experts et leurs applications", Avignon, May 28-Jun 1 1990, EC2, vol. 8/8, pp61-70 .

[5] **Giuse D. (1988)** *Computer-Based Intelligent Tutoring for Foreign Languages.* Proceedings of Asia-Pacific Conference on Computers in Education, Shanghai China, Oct 88.

[6] **Halpern J. (1990)** *New Japanese-English Character Dictionary.* Kenkyusha, Tokyo, 1992 p.

[7] **Heisig J. W. (1977)** *Remembering the Kanji I - A complete course on Japanese characters.* Japan Publications Trading Co, Tokyo, 2 vol., 495 p.

[8] **Seligman M. & Boitet Ch. (1994)** *The 'Whiteboard" Architecture: a way to integrate heterogeneous components of NLP systems.* Proceedings of Coling 94, Kyoto, Aug 5-9 1994.

[9] **Sérasset G. & Blanc É. (1993)** *Une approche par acception pour les bases lexicales multilingues.* Proc. of T-TA-TAO '93, Montréal, Sep 30-Oct 1993, Université de Montréal.

[10] **Wieger L. (1972)** *Caractères chinois. Etymologie. Graphies. Lexique.* Kuangchi Press, Taichung.

[11] **Yokoi T. (1993)** *Very Large-Scale Knowledge Bases Embodying Intelligence Space.* Proceedings of KB&KS'93, Tokyo, Dec 1-2 1993, JIPDEC, pp11-20 .

ANNEX An example of cooperative search, through stepwise structural vignette refinement (First scenario)

The user interface was schematized here for a more compact linear reading. Heading numbers indicate dialogue steps. DA stands for Discovery Assistant. 'User:' announces a selection or an entry, 'DA-->' a DA prompting and 'DA...' a DA response. Bold typing shows user's selections or entries. SK and PK respectively stand for Semantic Key and Phonetic Key, two possible functions for some components. Six main positions for a component inside a character are: South, North, West, East, Central and External. Grey tint in the patterns indicates the actual position of a *fire* Semantic Key.

0. DA--> Any component known ? If not, do you have structure proposals ? Any other criterion ?

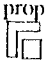
0. User: **Yes** No Yes **No** Yes **No**

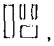
1. DA--> Component...Meaning ? Function ? Position ? Vignette list ? Any other criterion ?

1. User: **fire** **SK** PK **South** Yes **No** Yes **No**

1. DA... Ok, I have several proposals, with different vignettes ...for the global structure

2. DA--> Any other component known ? If not, I have the following vignette proposals ...for the global structure :

2. User: Yes **No** 1 **2** 3  4

2. DA... Ok, I propose 燃 *to burn* ...for 2 , and... I have several proposals ...for 1 .

3. User: Yes **No**

3. DA--> Any other component known ? If not, I have refined structures ...for 1 :

3. User: Yes **No** 1.1 **1.2** 1.3

3. DA... Ok, I have 2 solutions ...for 1.2 .

4. DA--> Any other component known ? If not, I have refined structures ...for 1.2 :

4. User: Yes **No** 1.2.1 **1.2.2**

4. DA... Ok, I propose 照 *to illuminate* ...for 1.2.2 .

Remarks

As a Semantic Key, *fire* may have alternative pictograms such as #1 and #2. With the current kanji base, choosing vignette 3 (#3) at step 2 would lead to #4, and vignette 4 (#5) to #6. At step 3, choosing 1.1 (#7) as a final refined structure would yield #8, #9 and #10, while choosing 1.3 (#11) would bring in #12. At step 4, selecting final structure 1.2.1 (#13) would lead to #14.

火	灬		灰		炭		炎,	災,	焦		点		烈
#1	#2	#3	#4 *ashes*	#5	#6 *charcoal*	#7	#8 *inflammation*	#9 *disaster*	#10 *to char*	#11	#12 *spot*	#13	#14 *ardent*