インターネット上の協調対訳辞書構築プロジェクト "SAIKAM"

ウッティチャイ アムポーンアラムウェート <vuthi@rd.nacsis.ac.jp>
相澤彰子 <akiko@rd.nacsis.ac.jp>
大山敬三 <oyama@rd.nacsis.ac.jp>
学術情報センター

本稿ではインターネット上に分散する多数の執筆者による協調的な日タイ対訳辞書構築を支援する環境「SAIKAM」の設計および実装について述べる。「SAIKAM」は多言語ウェブインターフェース、オンライン型の編集ツール、用例ナビゲーターを含む統合的な辞書環境で、インターネットを利用することにより、時間や場所の制約を超えた執筆作業を可能にしている。本稿では「SAIKAM」の特徴的な機能として (1) 既に公開されている日英及びタイ英辞書データを利用して執筆者の負担を軽減する辞書初期化機能、(2) 新規登録および校正すべき語のリストをシステム側で選別し、執筆者に提示するための編集機能、(3) 全文検索エンジンによる用例ナビゲーション機能について報告する。

An Internet–Based Dictionary Development Project: "SAIKAM"

Vuthichai Ampornaramveth <vuthi@rd.nacsis.ac.jp>
Akiko AIZAWA <akiko@rd.nacsis.ac.jp>
Keizo Oyama <oyama@rd.nacsis.ac.jp>
National Center for Science Information Systems

In this paper, an ambitous on-line Japanese–Thai dictionary development project called "Saikam" initiated by a group of Thai professionals and students in Japan is discussed. The project provides an on–line collaborative integrated environment to support development of Japaese–Thai dictionary on the Internet. Developers from all over the World can connect to the centralized dictionary database and update the content anytime at their convenience using standard web browsing tools. Some technical efforts have been made to enable trilingual editing environment on the existing WWW tools. Also, indices on large–scale Japanese text corpus were created to assist in selection of frequently–used words, and provide word usage navigating features to students of Japanese language.

1 研究背景・動機

日本語を勉強する外国人にとって、日本語と母国語の間の対訳辞書の存在は不可欠であるが、言語によっては、語彙、用例ともに限定された小規模な辞書しか入手できない場合も多い。例えば、タイ語の場合は表 1 にまとめたように現時点で出版されているタイ日辞書は小・大規模に渡って多数あるにもかかわらず、ほとんどの日タイ辞書は日常的な語に焦点を絞った小規模なものに過ぎない。辞書の内容も対訳語の並びのみで、説明的な意味や用例などを備え、論文作成などの目的で使用できるレベルの辞書は存在しない。そこで我々はインターネット環境を利用した日タイ対訳辞書システム「SAIKAM」(サイカム) [1] の構築を進めている。

大規模な辞書構築のためには多くの労力と時間が必要であるが、インターネットを利用することにより、日本滞在期間を終了し、タイへ帰国した後でも執筆に参加できるなど、時間や場所の制約を超えた執筆作業が可能になる。また、サーバへの検索要求やインターネット上での出現頻度を参照にすることで、検索要求頻度が高い語や頻出語を優先的に登録更新するなど、辞書そのものの柔軟性を高めることができる。

2 「SAIKAM」のシステム概要

「SAIKAM」では、図 1のようにインターネット上に分散する多数の編集者による辞書入力のための作業環境を提供する。辞書執筆の進行と同時に一般のユーザが検索インターフェースを通じて最新のデータにアクセスし辞書を参照することも可能である。このように、辞書編集、登録された語の検索、用法ナビゲーションのための検索エンジンを同一サーバ上で実現すること

表 1: 日 ↔ タイ辞書 (括弧内の数値は概算である)

Author	Title	# of Words	Price (Yen)
Thai → Japanese			
富田竹二郎	[タイ語辞典]	[30,000]	35,000
松山納	タイ語辞典	20–35,000	40,000
松山納	簡約タイ語辞典	[7,000]	10,000
高橋康敏	タイ語実用辞典	[6,000]	6,000
コサーアリーヤ	タイ・日辞典	22,000	4,757
小此木國満	すぐにつかえる タイ語–日本語辞典	[7,000]	—
Japanese → Thai			
コサーアリーヤ	日・タイ辞典	22,000	4,757
松山納	日タイ辞典	[7,000]	8,000
小此木國満	新すぐにつかえる 日本語–タイ語辞典	7,700	3,800

により、図 2 のような統合的な辞書システムの実現を目指している。

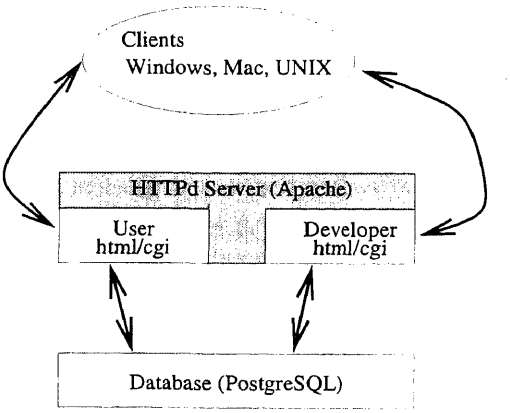

図 1: インターネット上の「SAIKAM」データベース及び http サービス

3 辞書構築のための多言語インターフェース

日タイ辞書検索・編集の作業を行うためには、日本語とタイ語の双方を同一画面上で表示入力する必要がある。本章はこの要求に応えるための多言語インターフェースの実現についてを述べる。

多数の参加者を募る「SAIKAM」のようなプロジェクトにおいてはできるだけ多くのインターネット上の利用者・執筆者が簡単に利用・参加できることが重要なポイントである。ウェブブラウザによるインターフェースは多様なプラットフォームに対応しており、操作方法を熟知している多くの利用者が便利に使えるという利点がある。

そこで「SAIKAM」では特にプラグインソフトをインストールしなくても、一般的なウェブブラウザを介して、システムへのアクセスができるように CGI スクリプトを用いて以下の多言語インターフェースを実現した。

3.1 多言語出力

日本語及びタイ語のフォントをインストールすることにより、ほとんどのシステムにおいて日タイいずれか一方の言語の文字の表示が可能となるが、両言語を同時に表示するために UNICODE などの多言語対応の文章コードを使用しなくてはならない。しかし、UNICODE に対応しているウェブブラウザは一部であり、使用可能なプラットフォームは限定されてしまう。UNICODE を使用しない多言語表示可能なウェブブラウザは OMRON [11] が拡張した WWW Consortium の Arena [10] が存在するが、残念なことに、タイ語の表示が不可能であり [9]、Unix 以外のプラットフォームにも対応していない。

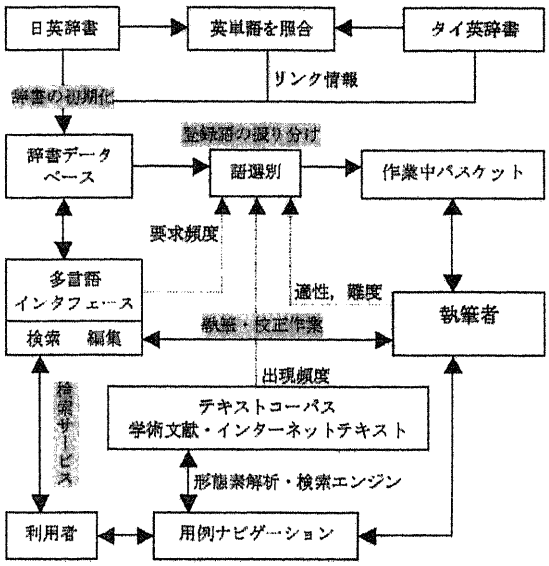

図 2: 「SAIKAM」の全体概念図

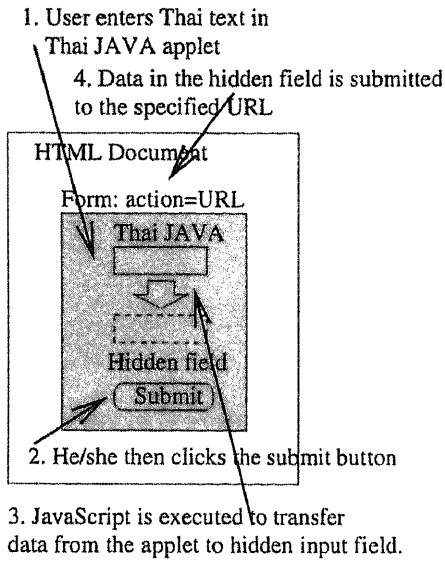

図 3: タイ JAVA Applet によるタイ文字列の入力

また、利用者の使用環境によってはシステムにフォントをインストールする権利を持たない、またはそのシステムに対応するフォント形式が存在しない場合などもある。そこで「SAIKAM」は日本語・タイ語の文字列を GIF 画像の形でブラウザに送る機能を実現している。

3.2 多言語入力

日本語の入力については OS が提供する IME [12] を用いるものとし、タイ語の入力については、タイ語版の OS を利用する場合には OS が提供するタイ語入力機能を、そうでない場合は JAVA Applet を用いるものとする。

タイ JAVA Applet [13] はタイ文字を画像として表示するものであり、一般テキストエディターのように操作できる。JAVA スクリプトと併用することでユーザがこの Applet に入力したタイ文字列は図 3 のような手順で、HTML フォームの他のデータと一緒にサーバへ送信される。

なお、これらの入出力に関わるオプションは利用者のプラットフォームに応じて設定可能項目が決まっており (図 4)、利用時に利用者毎に独立で行なえるようになっている。また選択された入出力方式は利用者側のブラウザに cookie として保存されるので，再び設定する必要がない．[2]

4 辞書データベースの構成

「SAIKAM」における対訳辞書は図 5 のように、1) 日本語単語のリスト (T_JWord)、2) タイ語単語のリスト (T_TWord)、3) 対訳関係にある日本語とタイ語を結ぶリンク情報 (T_Link_JT)、の 3 つの表からなる辞書データベースである。多義語を扱うために、単語の表示と単語の語彙を T_JWord 及び T_JWordinfo に分ける。ちなみに、複数の T_JWordinfo が同じ T_JWord を共有する場合があり、一つの T_JWordinfo のレコードは一つの意味に対応する。T_Link_JT は単語レベルの T_JWord と T_TWord ではなく、語彙レベルの T_JWordInfo と T_TWordInfo を結ぶ。

そして、類義語に対応する T_JWordinfo に同じ JConceptID を与えることによって、類義語の概念グループを生成することになる。単語の用例は別の T_Sample 表に保存されるので、同じ用例を複数の語彙から参照することが可能である。なお、各単語の関わる専門分野は T_Link_FJ に記録する。

T_Staff に登録した辞書の執筆者は各自の作業バスケット T_BasketJ に T_JWord から単語を振り分けることで、作業中の語彙リストを管理する。これによって、同時に複数の執筆者が同じ単語を執筆するような重複作業を回避することができる。

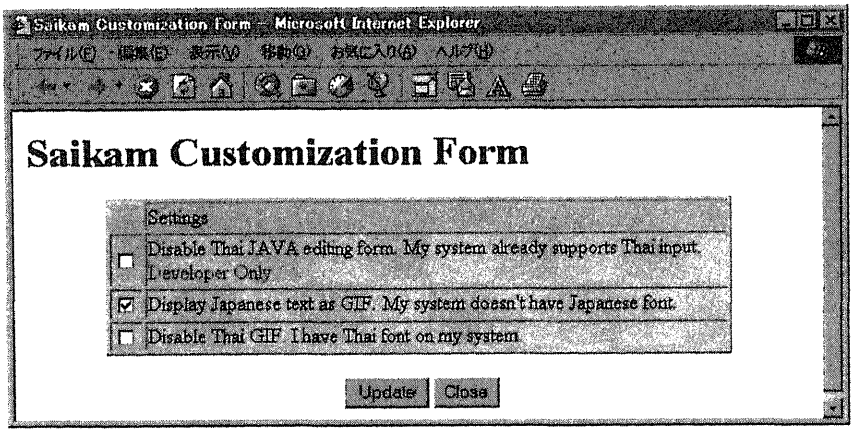

図 4: ユーザが設定可能な入出力に関わる項目

図5には日タイ (右) とタイ日 (左) に対応する表を示すが、現時点では日タイの部分のみ実現されている。

4.1 辞書データベースの初期化

既にインターネット上で公開されている日英対訳 (2.2 万語) [5]、タイ英対訳 (5.1 万語)・タイ単言語 (3.2 万語) [6] の辞書情報を用いて日タイそれぞれの登録語の初期化を行い、さらに、これら 2 つの対訳辞書に含まれる英単語を照合することにより、日タイの対応の初期化を行なった。

以下、英単語の照合による対訳関係の初期化方法をより詳しく説明する。入手した日英辞書とタイ英辞書のテキストファイルには単語及びその語の意味を説明する英文の 2 種から成る単純なデータ形式を持つ。例えば、日英辞書 (総単語数: 22,010) の場合は

だから	/therefore/
復元 [ふくげん]	/restoration (vs) (to original state)/
復習 [ふくしゅう]	/review (vs)/

のように表記されている。ここで、括弧は単語の品詞または追加情報を表す。同様にタイ英辞書 (総単語数: 51,645) の場合は

กลีบ	petaled/petal/a fold
หมู่บ้าน	village
หยาบ	rough/crude

となっている。基本的には共通の英訳を持つ語同士を対応づけるが、タイ英辞書では品詞情報などが記述されていないこと、表記形式に違いがあることなどから、英単語の照合を行なう前に、両辞書の形式を一致させるため、括弧削除などの前処理を行なう必要がある。

4.2 辞書データの前処理

辞書の相違点を考慮して、まず日英及びタイ英辞書に対して次の処理を行なう。

1. 語彙の展開: 多義語を複数の語彙に展開して書く。例:

   困苦 [こんく] /privation/hardship/

   ↓

   困苦 [こんく] /privation/
   困苦 [こんく] /hardship/

2. 品詞などを示す括弧を削除する。例:

   restoration (vs) (to original state) → restoration

   review (vs) → review

3. 不要な空白を除く。例:

   "review_" → "review"
   "get_off" → "get_off"

4.3 単語間のリンクの重みの計算

次に英語による語彙が共通している日タイ語間にリンクを決定する。このときに、日単語とタイ単語間のリンクの重みを次のように求める。

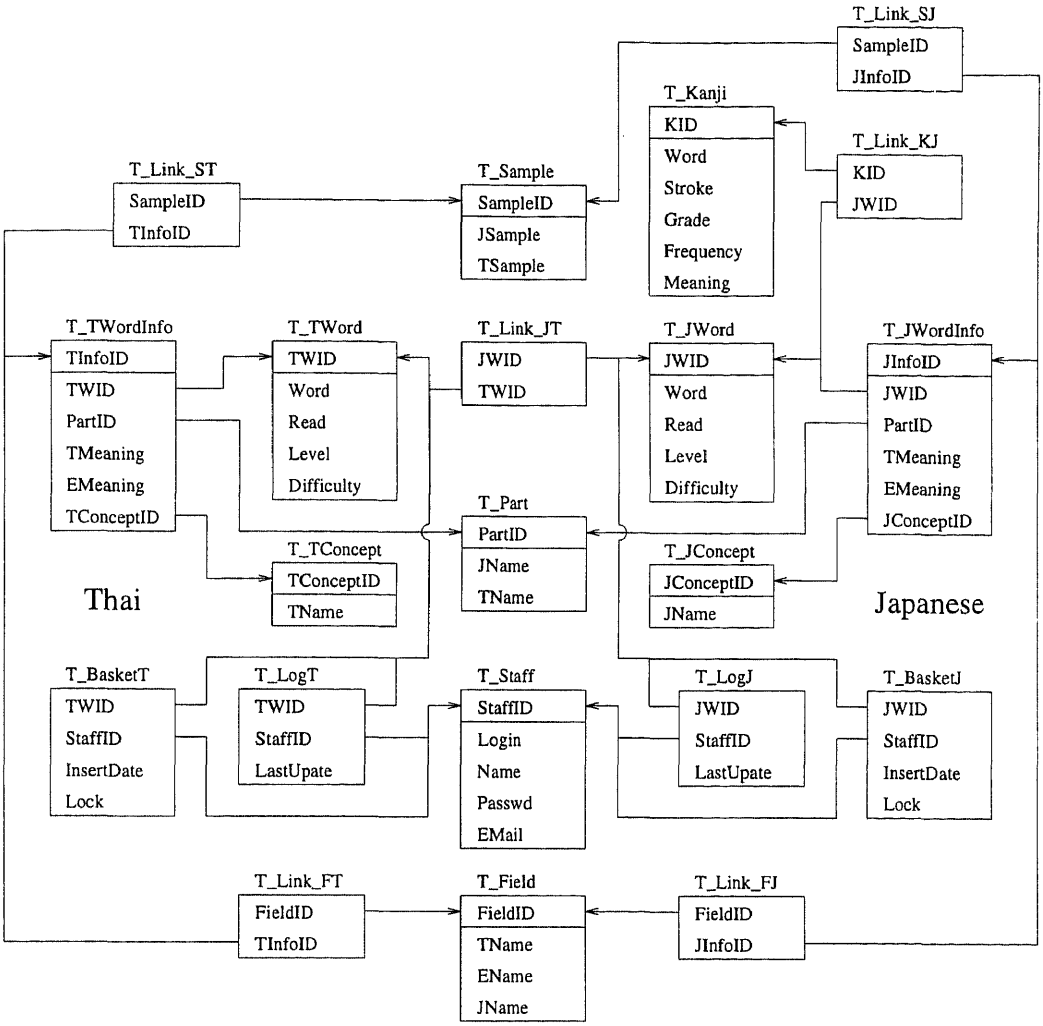

図 5: 「SAIKAM」データベース、 Simplified Entity-Relationship Diagram

1. 英文の説明がそのまま一致するならば

   リンク強度 → 100

2. 動詞を表す “to” の削除により (”to eat” → ”eat”) 一致する場合、

   リンク強度 → 90

3. 記号 (”_!-,’:/), “’s” の削除により一致する場合、

   リンク強度 → 80

4. 英単語を原型に変換することにより [16]:
   ”I am going” → ”I be go”
   ”changing” → ”change”
   ”said” → ”say”
   一致する場合、

   リンク強度 → 70

5. それ以外はそれぞれの定義部分の共有語幹の数から以下のように重みを決定する。

   タイ英辞書: タイ単語の意味を説明する英単語を T1 T2 T3 T4 とした場合
   “タイ単語: T1 T2 T3 T4” ($n_i = 4$)

   日英辞書: 日単語の意味を説明する英単語を J1 J2 J3 とした場合
   “日単語: J1 J2 J3” ($n_j = 3$)

$$\text{リンク強度 (0-50)} = \sum_t \sum_j \frac{2\times(\text{if } T_t==J_j \text{ then } 1)}{n_t+n_j} \times 50$$

特に 2 は日英辞書の記述に固有の処理である。前処理段階における多義語の展開により同一の日タイ対訳の間には複数のリンクが存在する場合もある。そこで、それぞれの語彙に対して得られたリンク強度の和を取り、日単語とタイ単語間のリンクの総重みを計算する。

4.4 リンクの結果

上記のルールによって作成したリンクの重みの分布を下の表に示す。

点数	リンク数
≥ 100	92,790
90–99	19,469
80–89	4,052
70–79	16,305
60–69	21,229
50–59	32,066
40–49	26,350
30–39	210,930
1–29	1,668,564
Total	2,091,755

表より重み 1 以上のリンクは約 209 万本存在し、20,468 語 (93%) の日単語と 43,739 語 (85%) のタイ単語がリンクで結ばれることが分かる。しかし、対応づけを人手により調べた結果、重みが 1–33 のリンクのほとんどは不正確な対応づけであることが分った。そこで、重みが 34 以上の 214,122 本のリンクのみでデータベースを初期化することにした。これらのリンクにより結ばれる語の数は日単語 18,743 語 (85%) とタイ単語 35,737 語 (69%) であった。

5 語の登録と編集作業

5.1 語の登録

「SAIKAM」のデータベースの一貫性を維持するために、一般の執筆者は登録語 (図 5 の T_JWord) を追加・削除することはできない。「SAIKAM」サーバの管理者は検索要求、インターネット上での出現頻度などの情報に基づき優先的に登録すべき語を選択する。

5.2 執筆者への振り分け

「SAIKAM」では、執筆者の適性、作業状況、専門分野及び上記の検索要求、単語の出現頻度などの情報に基づき各執筆者の作業バスケットに登録語を振り分ける。なお選別方式の詳細については現在検討中であり、実装システムでは単純にランダムな登録語の振り分けを行っている。

5.3 執筆および校正

インターネット上の執筆者は、各自の作業バスケット中の単語から適宜単語を選択し、図 6 に示される編集フォームを通じて、品詞選択、初期リンク情報のチェック、新規リンクの追加、概念グループへの割り当て、単語の意味の説明及び用例の執筆などの作業を行う。各登録語は、複数の執筆者による校正を行うものとし、各分野の専門家による最終校正を経て編集作業が終了する。

6 用例ナビゲーション機能

人手により登録される用例は数に限りがあるため、十分に利用者の要求を満足できない場合も考えられる。そこで「SAIKAM」は、大量に蓄積された日本語テキストデータに対して形態素解析 [3] を適用し、さらに全文検索エンジン [4] を介して高速にアクセスして、与えられた語に関連する用例を提示する機能の実現をめざしている。

辞書とコーパスの併用による統合的な辞書環境については文献 [7] などの例があるが、「SAIKAM」では特に日本語を母国語としない外国人に対して柔軟な言語ナビゲーション機能を提供することを目的として、漢字の難しさや語の頻出度情報に基づく文の難易度ランキングを行う。

具体的には、まず、システムが利用者の入力を形態素解析し、入力に含まれる単語毎に単語と品詞のリストを表示する。「私に食べる」など利用者の入力が誤りを含む可能性もあるため、利用者は対話的に個々の項目に単語または品詞を選び検索条件を指定できる (図 7)。検索エンジンはその条件と一致する用例を抽出し、予め定められた難易度基準にしたがって、利用者に分かりやすい順に提示する (図 8)。さらに全文検索エンジンを用いることによって [8]、インターネット上での該当する用例の頻出度情報や、その用例が出現した文書の URL を提示するなどの適用も考えられる。

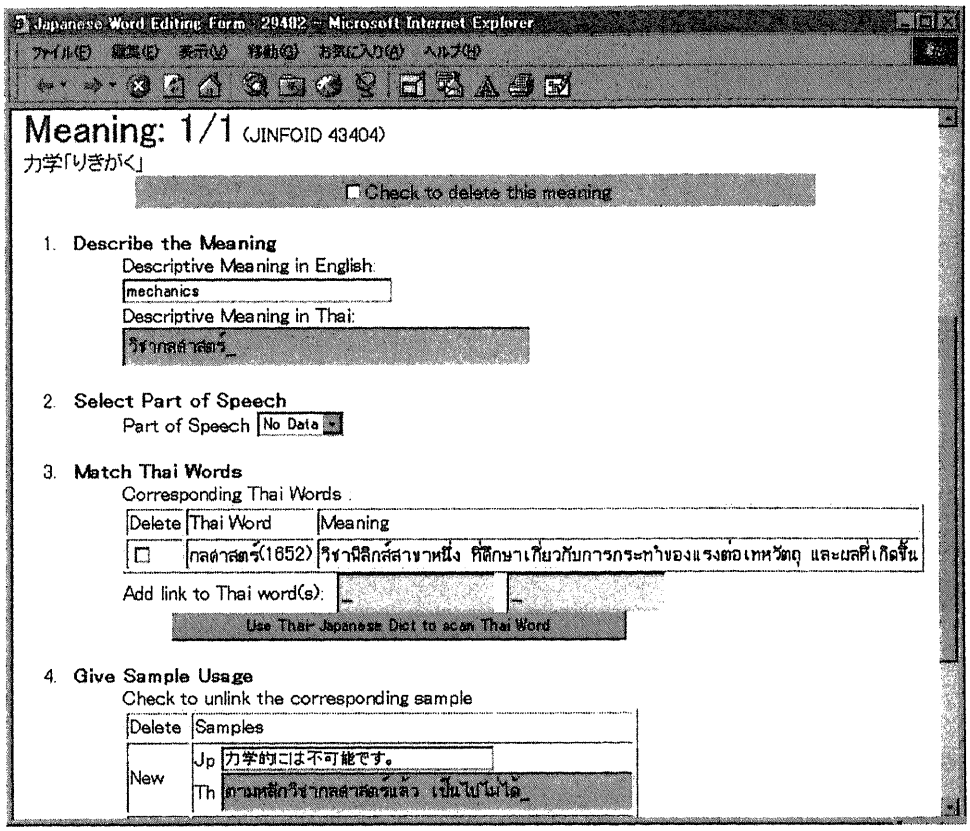

図 6: 日単語編集画面

7 まとめと今後の課題

本稿ではインターネット上の協調的辞書構築システムの概要を述べ、通常のブラウザで日タイ英の三ヶ国語が混在するテキストを入力表示できる多言語インタフェースの実装、RDBMに基づく辞書および編集者管理データベースの設計、日英・タイ英の基本辞書の読込みと英単語の参照による日タイデータベースの初期化法及びその機能を説明した。すでに上記の機能の主要部分を実現し、インターネット上で辞書構築サーバを運用している (http://thaigate.nacsis.ac.jp:8888/)。

執筆作業開始から半年後の現在、編集グループへの参加者は 45 名に上っている。試行システム上で辞書構築に参加する執筆者からの評価により、編集作業において問題となっているのは、ネットワークの制約によりサーバへのアクセス及び適切な日本語用例の入力の難しさであることが分っている。そこで現在、以下のような方向でシステムの拡張を検討している。

- ネットワーク上での負荷分散および執筆環境の整備 : サーバへのアクセスを容易にするため、広域にカバーする複数の分散型データベースサーバ構成に展開する。そして、オフラインでも編集作業が行えるように、パソコン用のクライアントソフトも用意する。
- コーパスを用いた言語習得支援機能の実現 : 日タイ対訳辞書執筆を支援するため、辞書やテキストコーパス、インターネット文書などの大量日本語テキストデータに形態素解析、検索エンジンと言ったツールを適用し、語共起や単語の用例などを取得する言語ナビゲーターツールを開発する。

さらにシステムの有効性評価のため、留学生を対象にしたアンケート調査の準備も進めている。

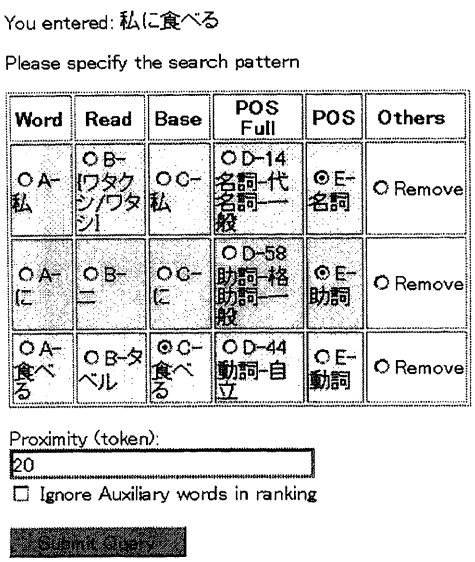

図 7: 検索条件指定画面

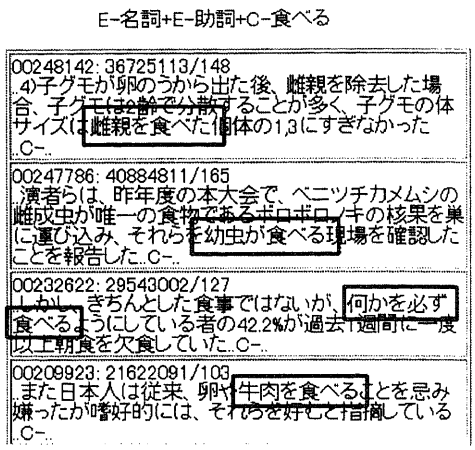

図 8: 検索結果

参考文献

1. *V. Ampornaramveth*, "SAIKAM: An Online Dictionary Development Project", Proc. of the 4th Intl. Workshop on Academic Information Networks and Systems, February 1998, NACSIS Seminar House, Karuizawa, Japan. http://thaigate.nacsis.ac.jp:8888/
2. *V. Ampornaramveth*, "Trilingual WWW Interface to SAIKAM Dictionary Project", Proc. of the 5th Intl. Workshop on Academic Information Networks and Systems, December 1998, AIT, Thailand.
3. 松本裕治他, "日本語形態素解析システム「茶筌」使用説明書", 奈良先端科学技術大学院大学, 1997.
4. "Pat Reference Manual", OpenText Corp.
5. "EDICT: Japanese–English Dictionary", `http://www.dgs.monash.edu.au/%7eejwb/japanese.html`
6. "Resource of Thai Language Processing", `http://www.links.nectec.or.th/thaires/`
7. 春野雅彦, "AIDA: コーパスを利用した適応的辞書環境", 自然言語処理 1997/7/25, 117-124.
8. 大山敬三, "インターネットに適応した全文データベース検索システムの構成", 学術情報センター紀要第 7 号 (1995).
9. 錦見美貴子, 高橋直人, 戸村哲, 半田剣一, 桑理聖二, 向川信一, 吉田智子, "マルチリンガル環境の実現 X Windows/Wnn/Mule/WWW ブラウザでの多国語環境", Prentice–Hall (Japanese) 1996. `http://www.biwa.or.jp/%7etomoko/wnn/`
10. "Arena", `http://www.w3.org/Arena/`
11. "Arena i18n: Extended Arena by OMRON", `http://www.wg.omron.co.jp/%7eshin/Arena-CJK-doc/`
12. "Microsoft Global Input Method Editor", `http://www.asia.microsoft.com/windows/ie/intlhome.htm`
13. "Thai JAVA Applets", `http://thaigate.nacsis.ac.jp/refer/thaijava/`
14. "Thai Computing References", `http://thaigate.nacsis.ac.jp/refer/`
15. 電器通信大学タイ留学生会, "ZzzThai Homepage", `http://www.fedu.uec.ac.jp/zzzthai/`
16. *Yasumasa Someya* <ysomeya@gol.com>, "e_lemma.txt'', September 1, 1998. `http://www.liv.ac.uk/%7ems2928/wordsmith/index.htm`