ＵＰＦ： 機械翻訳ユーザ辞書の共通フォーマット

亀井真一郎 *1　桧山努 *2　伊藤悦雄 *3　藤井美樹子 *4
平井徳行 *5　斎藤由香梨 *6　高橋雅仁 *7　村木一至 *1
*1 NEC　*2 NEC情報システムズ　*3 東芝　*4 ノヴァ　*5 シャープ　*6 富士通研　*7 九州松下

kamei@ccm.cl.nec.co.jp, hiyama@ats.nis.nec.co.jp, etsuo@sp.tokyo-sc.toshiba.co.jp,
aka@nova.co.jp, nnd6@isl.nara.sharp.co.jp, yukari@ling.flab.fujitsu.co.jp,
takahasi@mmm.kme.mei.co.jp , muraki@ccm.cl.nec.co.jp

1 はじめに

本稿では、アジア太平洋機械翻訳協会（ＡＡＭＴ）に加盟している機械翻訳（ＭＴ）メーカが中心となって１年半にわたって行なってきた機械翻訳ユーザ辞書の共通フォーマット（Universal PlatForm; UPF）設定活動について報告する。

機械翻訳システムを有効に活用するためには、ユーザ毎に必要な用語を「ユーザ辞書」として蓄積し、システムがあらかじめ用意している「システム辞書」と合わせて使用する必要がある。しかし辞書作成は一般に人的・時間的・金銭的コストがかかる作業であり、個々のユーザ一人一人が辞書を充実させるには限界がある。

この問題の解決策として、個人が個別に蓄えた辞書データを流通させ、相互利用することが考えられる。現在、国内では２０数社が機械翻訳システムを商品化しているが、各システムは独自のユーザ辞書フォーマットを設定しており、異なる機種間で辞書データを交換することはできなかった。ＭＴシステムの機種の違いを超えて各社のユーザ辞書データの流通・相互利用を促進することができれば、各人がユーザ辞書を作成するコストが大幅に削減できる。このことによりＭＴの利用が促進され、ひいては日本人の外国語文書受発信が促進される。

このような環境整備の具体的活動として、ＡＡＭＴでは、平成８年度から９年度にかけて、情報処理振興事業協会（ＩＰＡ）の創造的ソフトウエア育成事業の予算補助を受け、各社のＭＴシステムに共通のユーザ辞書記述フォーマット UPF の開発と、辞書データを蓄積・流通させるための一般アクセス可能な電子環境（ＷＥＢのホームページ）の整備とを行なってきた。ユーザ辞書共通フォーマットの仕様は、各社の実システムによる実証評価を経て、平成９年度末にＦＩＸして一般公開する予定である。

UPF: Sharable Formats of MT User Dictionaries
Shin-ichiro Kamei, Tsutomu Hiyama, Etsuo Itoh, Mikiko Fujii, Tokuyuki Hirai, Yukari Saitoh, Masahito Takahashi, and Kazunori Muraki

2 機械翻訳におけるユーザ辞書充実の有効性

一般に、機械翻訳システムを有効に活用するには各ユーザがシステムに関与する必要がある。ユーザの関与の方法としては１）プリエディット、２）ポストエディット、３）ユーザ辞書作成の３種が代表的である。

プリエディットとは、翻訳の入力文をシステムが処理できる形に（短く）する作業である。この作業により一次翻訳の品質は向上するが、プリエディットの方法はシステム毎に異なり、有効な方法を習得するのに熟練を要するという難点がある。

ポストエディットとは、システムの出力結果の翻訳文を手で修正する作業である。最終的な翻訳品質を向上させるためには不可欠な作業であるが、ポストエディットを行なうには、原文と訳文の両方を比較し理解する必要があり、高い言語能力と時間的コストが必要とされるという難点がある。

ユーザ辞書作成とは、ユーザがＭＴを使用する分野の専門用語や、ユーザ個人の必要とする用語を辞書の形で充実させる作業である。この言語知識ベースの充実作業は、ＭＴの実運用には必須の作業であるが、その分野の用語に対して２カ国の知識を持っていなければならず専門性が高いという難点がある。また辞書作成の方法はシステム毎に異なるので、辞書作成方法の習得には熟練を要する。また一般に辞書作成は時間的労力のかかる作業である。このような難点はあるが、ユーザ辞書は、前述のプリエディット・ポストエディットとは異なり、知識が蓄積されて再利用可能であるという大きな利点がある。

我々は、機械翻訳の有効利用に不可欠な上記3種類のユーザ関与の利点欠点を比較し、ユーザ辞書増強の「知識蓄積型」という利点に着目した。この利点を生かし、専門性の問題やシステム毎の差異を減らして、従来にも増してユーザ辞書を機械翻訳システムに役立てることを構想した。それが次節に述べるＵＰＦ構想である。

3 ＵＰＦ構想

ユーザ辞書の「知識蓄積型」という性質を生かして有効利用するために、ユーザ辞書データを広く交換・流通させる仕組みを構想した。アジア太平洋機械翻訳協会（ＡＡＭＴ）に加盟するＭＴメーカが中心となり、ワーキンググループを作って、以下のような環境の開発を行なった。

（1）ＭＴシステムの機種の違いを超えてユーザ辞書データを交換できるようにするためのユーザ辞書共通フォーマット（ＵＰＦ）の設計

（2）共通フォーマット（ＵＰＦ）で記述された辞書を蓄積・流通させるための、一般アクセス可能な電子環境の提供

上記（1）の共通フォーマット設計に際しては、言語学的な厳密さよりも、現実のシステムとの互換性を重視した。また対象をユーザ辞書に絞り、辞書作成の簡便さを重要視した。現実に販売され、ユーザに使用されている複数のＭＴシステム間でそのユーザ辞書の仕様を比較して開発を行なうことで、共通フォーマットが現実のシステムから乖離する危険を回避した。設定したフォーマットは複数のＭＴ製品による実証評価を経てＦＩＸし一般公開する。

さらに共通フォーマットの設定に際しては機械にとっての処理しやすさと共に人間にとっての可読性も重視した。特別なツールなしでユーザが内容を読むことができることを共通フォーマットの条件とし、ＳＧＭＬ的なマークアップ言語で記述することとし辞書のマスタはテキストファイルとした。ただし辞書の構築・参照の効率化のために共通フォーマット専用の辞書エディタは提供する。

共通フォーマットの仕様は一般に公開されるので、共通フォーマット辞書と自社システム独自のユーザ辞書の間の変換用に、各社が独自にコンバータを作成できる。

上記（2）の辞書共有環境は、ＡＡＭＴのＷＥＢホームページ [6] 上に構築した。各ユーザは共有環境に置かれているＵＰＦ形式の辞書を自分のシステムに取り込んで使用できる。また自分の作成した辞書データをＵＰＦ形式に変換して共有環境に置くことで広く他のユーザに提供できる。

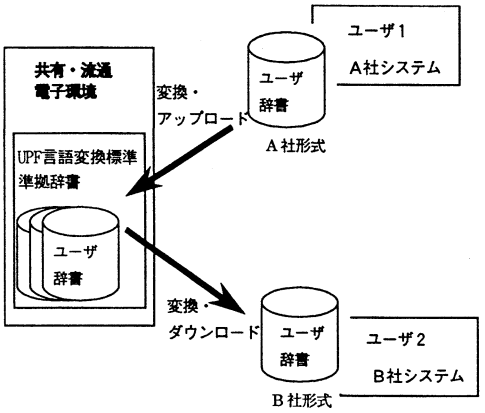

図1　UPFの全体構想

4 基本変換標準と拡張変換標準

共通フォーマットＵＰＦを介すことで、異システムユーザを含む他ユーザとユーザ辞書データが共有・交換でき、個々のユーザの辞書構築コストが軽減できる。そのためには、ＵＰＦから各システムのユーザ辞書への変換（ダウンロード）と各システムのユーザ辞書からＵＰＦへの変換（アップロード）との双方向変換が可能となるようにＵＰＦを設計する必要がある。そこでＵＰＦでは、以下の2種のフォーマットを設定することにした。

（1）基本言語変換標準（Mシート）
全ＭＴシステムのユーザ辞書で取り扱うことができ、ＵＰＦとの間で相互変換（アップロード・ダウンロード）が可能であることを推奨する語彙情報記述形式

（2）拡張言語変換標準（Kシート）
各ＭＴシステムで記述する可能性のあるすべての語彙情報を記述するための形式

現実に利用されている複数のシステムのユーザ辞書で扱える語彙（品詞）には相違があるから、狭い意味で上記の双方向条件を満たすためには、各システムに共通して記述できる語彙（記述可能な語彙の「ＡＮＤ」(= intersection)）だけを対象とする

仕様が必要である。それが基本言語変換標準である。基本言語変換標準を使用することで、効率的に辞書記述を行なうことができる。

一方、そのような「AND」仕様だけでは、詳細・広範な語彙情報の記述を許すシステムが有効活用されないという問題が生じる。そこで各システムで記述される可能性のある仕様を記述するための枠組みが必要である。それが拡張言語変換標準である。

すでに商品化されてユーザに使用されている6種類の異なるMTシステムのユーザ辞書を比較検討することで、上記2種類の共通フォーマット（言語変換標準）の設定を行なった。

5 言語変換標準の概要

5—1 共通仕様の設定準備

言語変換標準の設定に際しては、まず現実の各システムに共通の「記述用語」を設定する必要があった。つまり、品詞のセット、品詞の呼称など用語と定義の統一から作業を開始した。特に日本語の場合、基本となる品詞設定についても学校文法では機械翻訳にとっては不十分であり、準拠すべき標準が存在しないため、各システムはおのおの独自の体系を設定している。このようにシステム毎にユーザ辞書の仕様が異なることが、ユーザにとって辞書作成を困難にしている要因の一つである。具体例としては、「形容動詞」という品詞を独立の品詞として立てているシステムと「形容詞」の下位に位置づけているシステムとが存在した。また活用語の登録単位も、語幹登録、終止形登録の二通りがあった。このような用語・形式の統一を行ない、言語変換標準の設定を行なった。

5—2 基本言語変換標準

基本言語変換標準（Mシート）では取り扱う言語情報を制限した。制限に当てはまらない辞書情報は拡張言語変換標準（Kシート）で記述することを前提にして、基本言語変換標準ではユーザが辞書登録する時の不要な迷いを減らし、頻繁に辞書登録する場合の辞書作成の効率化を図り、データの流通性に重点を置いた。

基本言語変換標準では登録する品詞を以下の品詞に限定した。

日本語：名詞、動詞、形容詞、形容動詞、副詞
英語：　名詞、動詞、形容詞、副詞

また日英の各品詞間の対応にも制限を設けた。日本語と英語とで、名詞は名詞に、動詞は動詞に、形容詞・形容動詞は形容詞に、副詞は副詞に対応する辞書のみが作成できるように制限した。このような制限を加えることで一般のユーザにとっての辞書作成のわかりやすさと効率化を図った。

このような品詞の限定は、MTシステムが商用になってから現在までの10年間に現実にMTユーザが登録した辞書の約9割が固有名詞を含む名詞であるという各社に共通した経験的データに基づいている。なお固有名詞は名詞の下位分類とし、いわゆるサ変動詞は動詞の下位分類とした。また活用語の登録単位は、わかりやすさを重視して終止形に統一した。また用言の格フレームも代表的なパターンに制限することでユーザの不要な迷いを減らし辞書作成の効率化を目指した。

具体的には基本言語変換標準では日本語の場合、以下の情報を記述する。

[名詞]
種類：普通名詞、固有名詞
意味分類：人、組織、その他具体物、
場所、時間、その他抽象物
[動詞]
活用型：一段、五段、
カ変、サ変（含：サ変動詞）
格パターン：が、を、に
格要素意味制限
英語の格パターンへのマッピング情報
[形容詞]
(辞書見出しと品詞以外に記述情報なし)
(格要素は「が」のみに制限)
[形容動詞]
活用型：なだ型、のだ型
(格要素は「が」のみに制限)
[副詞]
(辞書見出しと品詞以外に記述情報なし)

以下に基本言語変換標準での記述例を示す。

```
<entry>
<japanese>
  <jentry>      食べる    </jentry>
  <jpos>        動詞      </jpos>
```

```
  <jinfl>        一段      </jinfl>
  <jcase>        が,を     </jcase>
</japanese>
<jetrans>    ((が=主語;名詞句;人;)
             (を=目的語;名詞句;その他の具体物;))
</jetrans>
<english>
  <eentry>       eat              </eentry>
  <epos>         verb             </epos>
  <eheadpron>    consonant    </eheadpron>
  <evpresent>    eats     </evpresent>
  <evpast>       ate      </evpast>
  <evpp>         eaten    </evpp>
  <eving>        eating   </eving>
  <ecase>        svo          </ecase>
</english>
</entry>
```

UPF はさしあたり日本語と英語の２カ国語を対象としているが、その記述形式は上記のように、日本語部分、日本語英語対応部分、英語部分とに分離しており、将来の多言語辞書への拡張への対応を考慮している。

5—3　拡張言語変換標準

拡張言語変換標準（Kシート）では、基本言語変換標準（Mシート）における登録の制限をなくし、MTシステムのユーザ辞書に登録する可能性のある辞書情報はすべて登録可能な仕様とした。

拡張言語変換標準は、あらかじめ設定していない新規辞書情報をユーザが設定して記述できる枠組みを有しているところに特徴がある。新規辞書情報は以下の形式で新規なタグを定義して記述する。

```
<tagdefine>
  <tag_name>      タグ名          </tag_name>
  <tag_descript>  説明            </tag_descript>
  <parent_tag>    親のタグ名      </parent_tag>
  <value_sets>    値のリスト      </value_sets>
  <value_sets_descript>    値の説明
                          </value_sets_descript>
  <tagdefine_comment>    コメント
                          </tagdefine_comment>
</tagdefine>
```

6 おわりに

本稿では、機械翻訳ユーザ辞書の共通フォーマットＵＰＦの概要を報告した。ＵＰＦには、ＵＰＦと各システムとの双方向変換を推奨する基本言語変換標準（Mシート）と、ユーザ辞書に記述しうるすべての情報の標準記述形式である拡張言語変換標準（Kシート）の２種類を設定した。辞書マスタはテキストファイルとし、マークアップ言語で記述することにより、可読性と処理容易性を両立させた。

辞書データをユーザ間で共有し・流通させるための電子環境としては以下を提供する。

（1）ＷＥＢホームページ [6]
（2）ＵＰＦ形式専用辞書エディタ
（3）ＵＰＦ形式の日英・英日辞書
　　（各２万エントリ；ビジネス分野用語）

これら具体的な辞書エントリおよび辞書作成エディタを提供することにより、ＵＰＦの利用方法をユーザが実感し、ＵＰＦが活発に使用されるようになることを図った。実システムでの評価実験を終えて、今年度末１９９８年３月末に最終仕様ＦＩＸおよび一般公開する。このＵＰＦが今後の機械翻訳システムの有効利用に寄与し、ひいては日本人の外国語情報の受発信を促進することを願っている。

参考文献

[1] 亀井 他：商用機械翻訳ユーザ辞書の共通フォーマット設定に向けて　情報処理学会 第５４回全国大会

[2] 伊藤 他：機械翻訳ユーザ辞書の共通フォーマットの設定 －アジア太平洋機械翻訳協会における活動中間報告 － 言語処理学会 第３回年次大会

[3] 赤羽 他：機械翻訳ユーザ辞書データ流通のための共通フォーマット －アジア太平洋機械翻訳協会の活動報告 － 人工知能学会 第１１回全国大会

[4] Kamei et al: “Sharable formats and their supporting environments for exchanging user dictionaries among different MT systems as a part of AAMT activities” MT Summit VI 1997

[5] 桧山 他：機械翻訳ユーザ辞書データ流通・相互利用のための共通フォーマット設定活動 －アジア太平洋機械翻訳協会の活動報告 － 情報処理学会 第５６回全国大会

[6] http://www.jeida.or.jp/aamt/index.html