二言語間の対訳文構造照合ＧＵＩ “V-Para” の作成[*]

中塚幸毅[†] , 宇津呂武仁[†] , 北村美穂子[‡] , 松本裕治[†]

[†] 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座
[‡] 沖電気工業株式会社

1 はじめに

機械翻訳での翻訳知識を獲得するうえで、表層のレベルだけではなく、単語間の依存構造も含めた構造レベルで対訳文の対応づけを行なう構造照合は、非常に有用な技術である [4, 6]。本稿では、二言語間の係受構造の構造照合の結果を視覚化し、その結果を視覚的に修正することを目的とした構造照合ＧＵＩアプリケーション “V-Para” の作成について述べる。

第 2 節において、対訳文の構造照合について、構造照合の結果のスコア (得点) の計算手法を中心に述べる。

第 3 節において、今回作成した V-Para のシステム構成について述べる。第 4 節において、 V-Para で使用できる主要な機能について述べる。

最後に、第 5 節において、今後の V-Para の改良の予定について述べる。

2 二言語間の対訳文の構造照合

2.1 概要

機械翻訳において使用する翻訳知識の獲得の自動化は、機械翻訳において、非常に重要な問題である。

最近では、機械可読な二言語対訳コーパスが大量に出現するようになった。これらのコーパスを利用した研究として、対訳コーパスにおける文や単語の対応づけの研究 [2, 6] や、対訳文の構文構造の対応づけに関する研究 [1, 3] などが現在までに行なわれてきた。

とくに、対訳文において、単語のような表層のレベルだけではなく、単語間の依存構造も含めた構造レベルの対訳文の対応づけを行なうことを構造照合と呼ぶ。構造照合の結果は、対訳辞書、翻訳の用例集や、用例に基づく機械翻訳の翻訳知識として利用することができる点で非常に有用である。また、類似性の高い、適切な照合結果を選ぶことにより、単言語文の構文構造の曖昧性の解消なども実現できる。本稿のシステムでは、とくに松本、北村らが提案した構造照合の手法 [1, 4] を採用している。

構造照合のおおよその流れを図 1 に示す。まず、二言語間の対訳コーパスにおいて、すでに文単位に対応づけが施された対訳文において、それぞれの言語の文について、構造的な曖昧性を含む係受けを示す文の依存構造を求め、あらかじめ対訳コーパスより統計的に求めた二言語間の単語の類似度 (Kay [2] の手法に基づく) と、お互いの依存構造の類似性をもとにして、二言語間の依存構造の部分木の対応づけを求める。

構造照合についての具体的なアルゴリズムなどの詳細については、松本及び北村の手法についての文献 [1, 4, 5, 6] を参照されたい。

---

[*] Graphical User Interface of Structural Matching of Parallel Sentences : *V-Para*

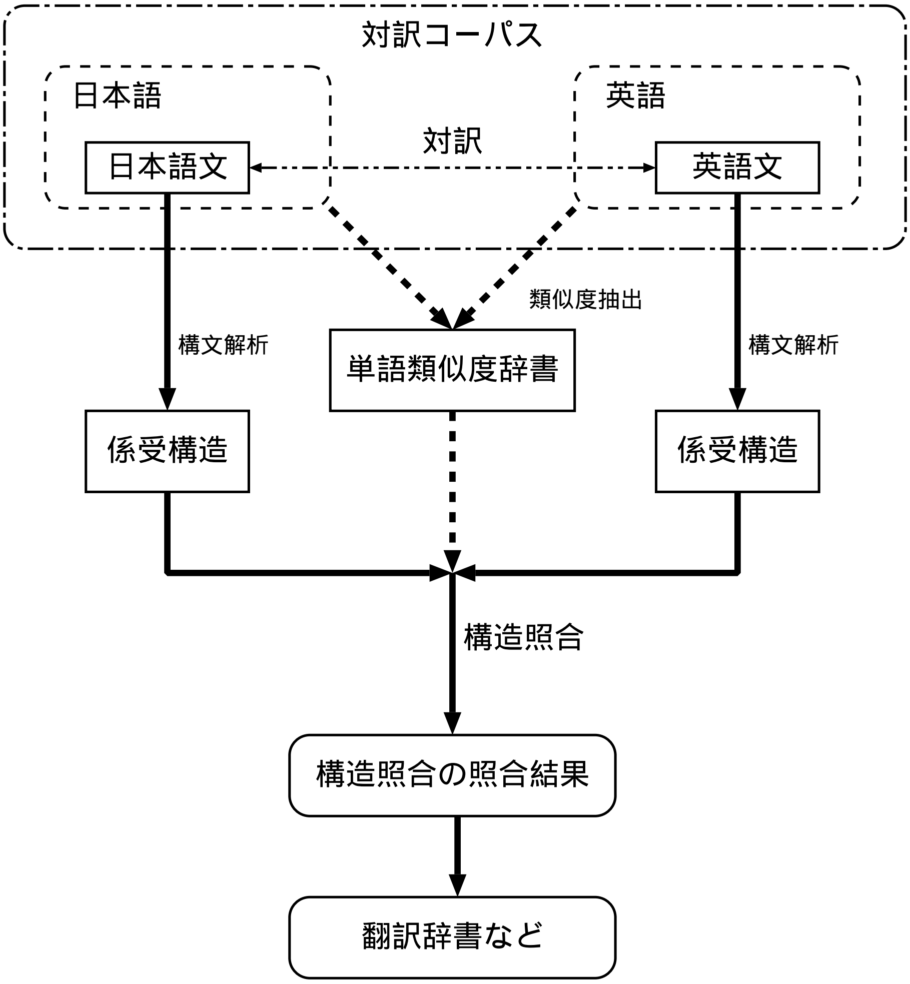


図 1: 対訳文の構造照合の概略

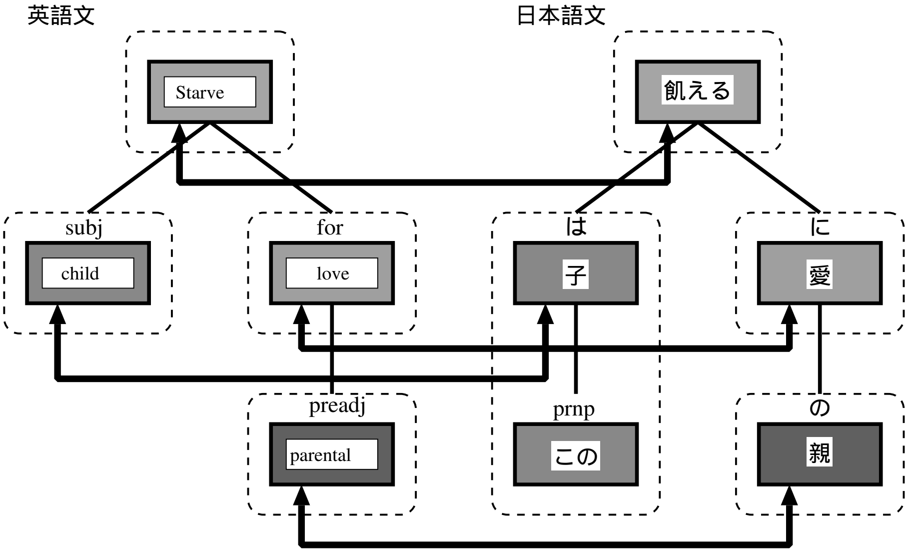


図 2: 対訳文の構造照合結果の例

2.2 構造照合の詳細および得点の算出法

ここでは、構造照合で得られた照合結果について、照合結果の類似度、すなわち、スコア（得点）を求める手法を中心に説明する。

二言語間の対訳文において 2 節で述べた構造照合によって、互いの依存構造の部分木の対応づけを求める。照合結果の具体的な例として、次の日本語、英語対訳文について構造照合を行なった結果を図 2 に示す。なお、ここで、飢える ↔ starve, 子 ↔ child, 親 ↔ parental は、すでに類似度 1.0 で対訳辞書に登録されているものとする。

英語 ：This child starving for parental love.
日本語 ：この子は親の愛に飢えている。

図 2 において、同一の柄と点線で示された部分木が両矢印で結ばれているものが、互いに対応づけられた部分木であることを示す。“parental” と “親”、“starve” と “飢える” のノードが対応づけられているだけでなく、対訳辞書に含まれない “love” と “愛” のノードや、“child” と “この、子” からなる部分木についても対応づけられていることがわかる。

照合結果に対する得点は、まず、照合結果におけるそれぞれの部分木の対応づけに対する得点の総和によって計算される。部分木間の得点の求め方は、基本的に次のような評価基準によって求められる。

- 個々の節点間における単語間の類似度が大きければ大きいほど、より細かい単語間の照合を優先する。
- 個々の節点間における単語間の類似度が小さければ小さいほど、より大きな部分木間の照合を優先する。

実際の部分木の対応のスコア計算は次のようにして行なわれる。

まず、言語Ａ，Ｂでの部分木の節点集合をそれぞれ V_A , V_B とし、簡単のため、言語Ａの部分木に存在するノードは言語Ｂの部分木より少ないと仮定する。

言語Ａでの部分木の節点から、言語Ｂの節点への単射 $p : A \rightarrow B$ によって定義される要素対の集合を $D_p = \{< a, p(a) > | a \in A\}$ とする。そして、言語Ａの平均単語類似度 Sim_{ave} を次の式で求める。

$$Sim_{ave} = \frac{\max_p \left\{\sum_{<a,p(a)>\in D_p} sim(a,p(a))\right\}}{|A|}$$

（ここに、 $sim(a,p(a))$ は、単語 a と $p(a)$ との類似度を示す。）すなわち、部分木のノード間の１対１の対応について、すべての対応の組合せを考え、それらのそれぞれの組合せについて単語の類似度の総和を計算し、最も大きい値となるものを言語Ａの部分木の節点数で割ることにより、Ａにおける平均単語類似度 Sim_{ave} が求められる。以上の計算により求められた Sim_{ave} より、部分木間の得点は次の式で求められる。

$$Score = \left(Th + \frac{Sim_{ave} - Th}{|A| + |B| - 1}\right) \cdot |A|$$

（ここに、 Th（$0 < Th < 1$）は閾値であり、は北村の実験結果 [4] より、 0.15 とし、 $|A|, |B|$ は、言語Ａ，Ｂの部分木の節点数とする。）

そして、照合結果におけるそれぞれの部分木間の得点の総和によって照合結果の得点が計算される。

ここで、図 2 の照合結果について得点計算の具体例を示す。まず、“child” と “この、子” からなる部分木の対応について考える。この対応の要素対の集合は、 $\{< child, この >\}$, $\{< child, 子 >\}$ の２通りが考えられる。しかし、 $sim(child, 子) = 1.0$, $sim(child, この) = 0.0$ より、$\max_p \left\{\sum_{<a,p(a)>\in D_p} sim(a,p(a))\right\} = 1.0$ となる。よって、

$$Sim_{ave} = \frac{1.0}{1} = 1.0$$

$$Score = \left(0.15 + \frac{1.0 - 0.15}{1 + 2 - 1}\right) \cdot 1 = 0.575$$

同様に “parental” と “親”、“starve” と “飢える”、“love” と “愛” のノードについて得点を計算すると、それぞれ 1.0, 1.0, 0.0 となるので、以上より、この照合結果の得点は、

$$Score_{total} = 0.575 + 1.0 + 1.0 + 0.0 = 2.575$$

となる。

3　システム構成

本節では、 V-Para のシステム構成について述べる。まず、システム構成の概略図を（図 3）に示す。

システムは、主に SICStus-Prolog [7] を用いて書かれた係受け構造の解析および構造照合などを実際に行なう構造照合プログラム paramatch と Tcl/Tk [8] を用いて書かれたＧＵＩ環境インターフェイス paramatch.tcl からなる。そして、 paramatch と paramatch.tcl との間の通信サーバ paramatch_tcl_server により、２つのプログラム間の入出力のやりとりを行なう。

また、 paramatch は日本語の形態素解析に日本語形態素解析システム Juman [9] を使用しており、また係受け構造を解析するために、構文解析システム SAX [10] を使用している。

なお、係受け解析のための単言語文構文解析プログラムや、後述する単語の対訳辞書は、必要に応じて他のものと交換することが可能である。しかし、ＧＵＩインターフェイスとしては対訳辞書の交換は実装しているが、構文解析プログラムや、使用する文法の交換は実装していない。今回のシステムでは、日本語と英語について実装している。

ここで、ＧＵＩインターフェイスの一連の動作の概略について述べる。まず、ＧＵＩインターフェイス paramatch.tcl は、 SICStus-Prolog より立ち上げられた paramatch から paramatch_tcl_server に対する要

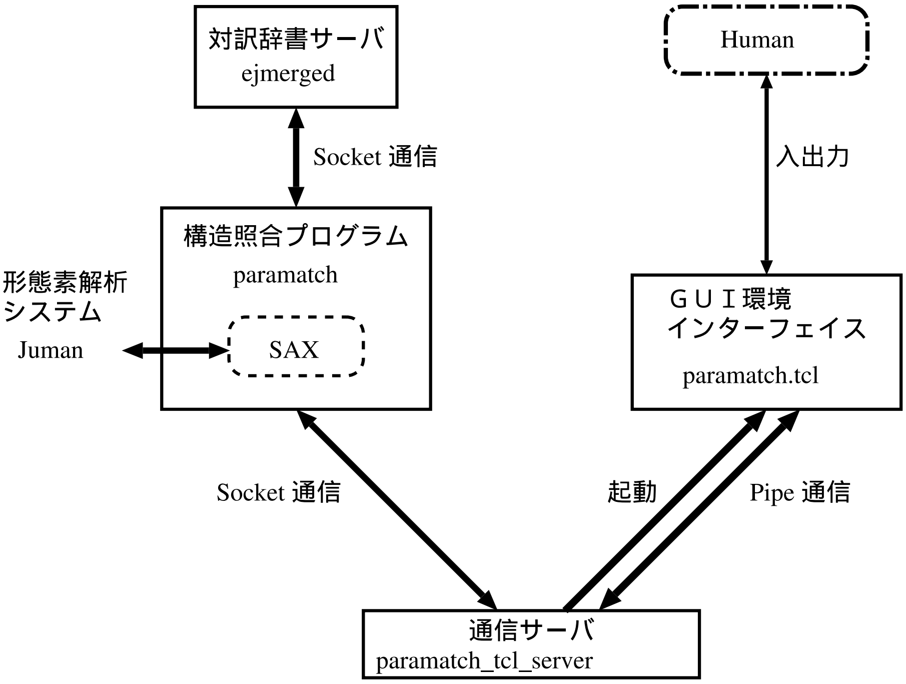


図 3: システム構成

求により、paramatch_tcl_server によって起動される。起動されたＧＵＩインターフェイス paramatch.tcl に対して対訳になった二言語の文を入力すると、paramatch_tcl_server を通じて paramatch に送られ、文の構造解析、および構造照合が行なわれる。

構造照合においては対訳辞書サーバ ejmerged と通信を行ない、二言語間の単語の類似度を獲得する。構造照合が行なわれた結果および、このときに参照された二言語間の単語の類似度は、再び paramatch_tcl_server を介して paramatch.tcl に送られ、 paramatch.tcl で表示される。

4 各機能についての説明

ここでは、 V-Para において実装されているさまざまな機能の概略について述べる。

4.1 コントロールパネル

最初に V-Para を起動した時、図 4 に示すコントロールパネルが現れる。 V-Para での主要な機能の操作は、このコントロールパネルにあるさまざまな入力フィールドやボタンによって行なう。

まず、コントロールパネルの最上部には、対訳文を入力するための “英語:”, “日本語:” と書かれた入力フィールドがある。 (以下、“英語:” フィールド, “日本語:” フィールドと呼ぶ)。また、そのすぐ下に辞書引きサーバおよび、結果保存ファイルを指定するための “対訳辞書サーバ:”, “保存ファイル:” と書かれた入力フィールドがある (以下、“対訳辞書サーバ:” フィールド, “保存ファイル:” フィールドと呼ぶ)。また、その下の部分には、候補選択や、単語類似度表示などの多くの機能を実行するボタン群がある。


英語 : he saw a girl with a telescope.
日本語: 彼は望遠鏡で女の子を見た。
候補: 0 スコア: 3.00000

図 4: コントロールパネル

4.2　対訳辞書サーバ選択

第 3 節のシステム構成で述べたように、構造照合において単語の辞書引きサーバ ejmerged と通信を行ない、二言語間の単語の類似度を検索する。よって、辞書引きサーバを指定することにより、対訳辞書を変更することができる。現在のシステムでは、対訳辞書サーバのあるホストを指定するという方式をとっている。

対訳辞書サーバのあるホストを指定するには、コントロールパネル中の “対訳辞書サーバ” の入力フィールドに対訳辞書サーバのホストマシン名を入力する。

4.3　対訳文入力

対訳になった二言語の文の入力（対訳文入力）を行なう方法について述べる。文を入力する方法としては、コントロールパネルより直接文を入力する方法と、すでに文の対応づけが施された対訳文を集めたテキストより、対訳文を選択する方法がある。

直接、対訳文を入力するには、図 4 にあるように、コントロールパネル上部にある “英語:”，“日本語:” フィールドから対訳になった二言語の文を入力する。

また、ボタン群の “対訳文” ボタンをクリックすると、図 5 左のような対訳文表示ウィンドウが現れる。ここで、ウィンドウ内の “FILE” をクリックすると、図 5 右のような対訳文を集めたファイル、すなわち対訳コーパスファイルを選択するためのセレクタが表示される。ここで、ファイルを選択すると前の対訳文表示ウィンドウに対訳文が現れる（図 5 左）。なお、対訳コーパスファイルは、

言語Ａの文１（言語Ａの文１の対訳）\n
言語Ｂの文１（言語Ｂの文１の対訳）\n
言語Ａの文２（言語Ａの文２の対訳）\n
言語Ｂの文２（言語Ｂの文２の対訳）\n
……
（ここに \n は改行）

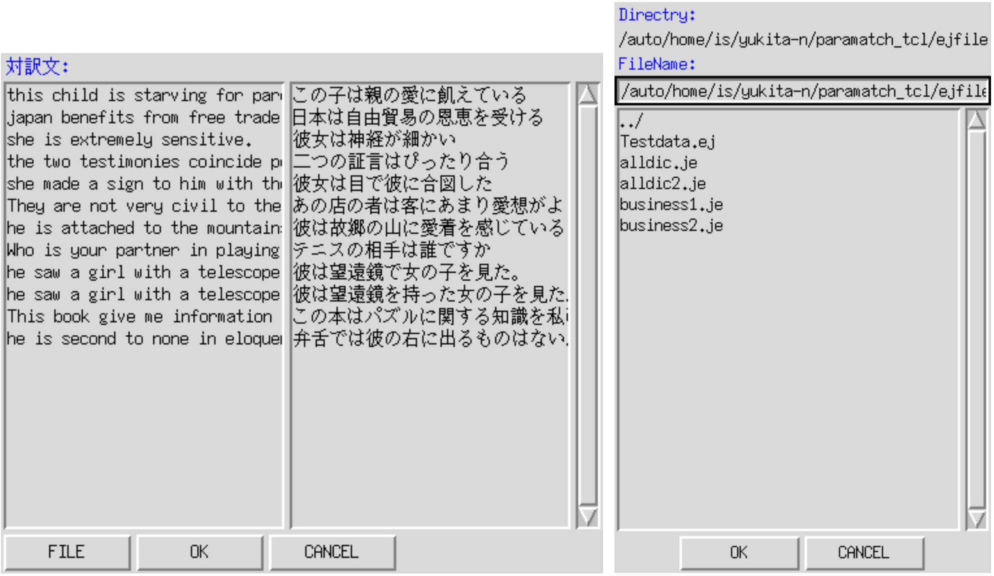
というテキスト形式のファイルである。文の間に空行が入っていても構わない。

そして、表示された対訳文について、入力したい対訳文をクリックすると、その対訳文が “英語:”，“日本語:” フィールドに現れる。

いずれの入力方法においても、このあと、ボタン群中の “入力” ボタンをクリックすると、実際に対訳文が paramatch に送られ、文の構造照合が行なわれる。

4.4　構造照合結果表示

文の構造照合の結果得られた、2 つの言語の依存構造の部分木の対応づけは、図 6 のように表示される。



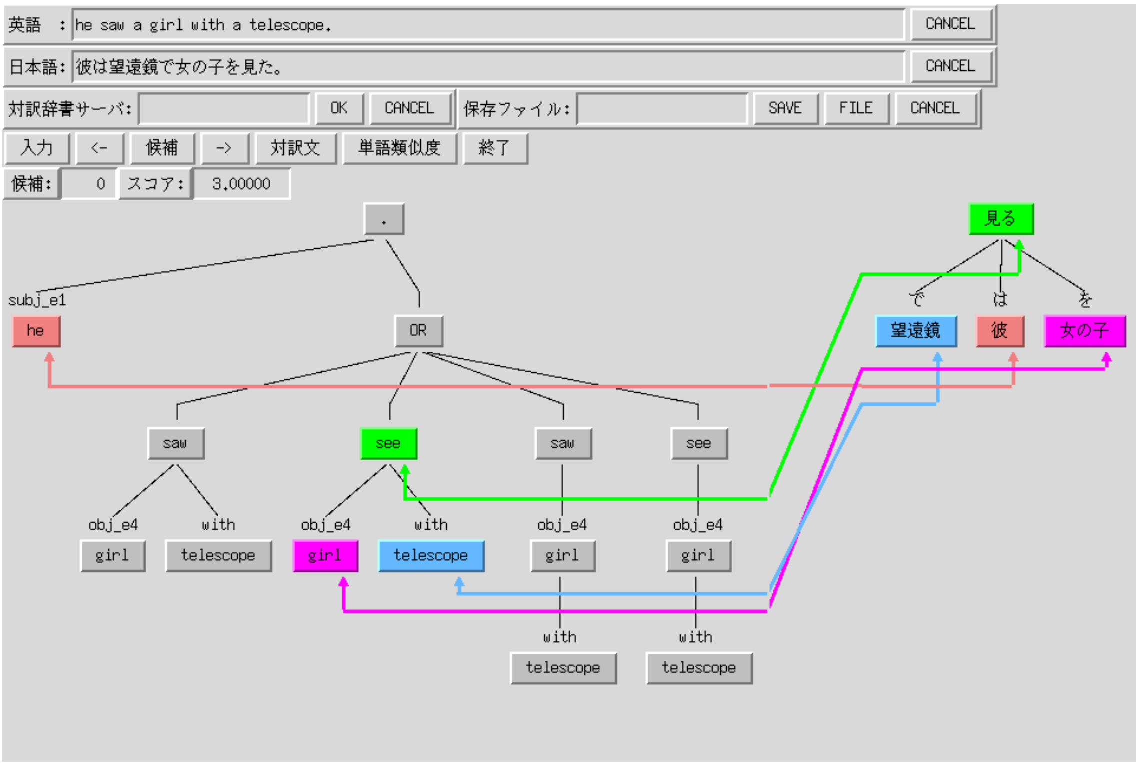
対訳文:

this child is starving for par	この子は親の愛に飢えている
japan benefits from free trade	日本は自由貿易の恩恵を受ける
she is extremely sensitive.	彼女は神経が細かい
the two testimonies coincide p	二つの証言はぴったり合う
she made a sign to him with th	彼女は目で彼に合図した
They are not very civil to the	あの店の者は客にあまり愛想がよ
he is attached to the mountain	彼は故郷の山に愛着を感じている
Who is your partner in playing	テニスの相手は誰ですか
he saw a girl with a telescope	彼は望遠鏡で女の子を見た。
he saw a girl with a telescope	彼は望遠鏡を持った女の子を見た.
This book give me information	この本はパズルに関する知識を私
he is second to none in eloque	弁舌では彼の右に出るものはない.

Directry:
/auto/home/is/yukita-n/paramatch_tcl/ejfile
FileName:
/auto/home/is/yukita-n/paramatch_tcl/ejfil

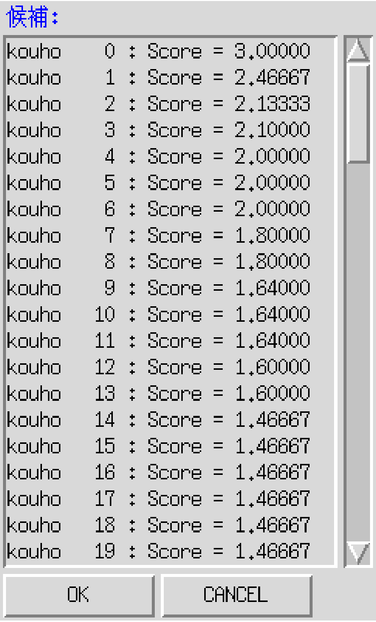
../
Testdata.ej
alldic.je
alldic2.je
business1.je
business2.je

図 5: 対訳文表示 ·対訳コーパスファイル選択ウィンドウ



英語 : he saw a girl with a telescope.
日本語: 彼は望遠鏡で女の子を見た。
対訳辞書サーバ:
保存ファイル:
候補: 0
スコア: 3.00000

図 6: 対訳文の構造照合結果の表示



```
kouho    0 : Score = 3.00000
kouho    1 : Score = 2.46667
kouho    2 : Score = 2.13333
kouho    3 : Score = 2.10000
kouho    4 : Score = 2.00000
kouho    5 : Score = 2.00000
kouho    6 : Score = 2.00000
kouho    7 : Score = 1.80000
kouho    8 : Score = 1.80000
kouho    9 : Score = 1.64000
kouho   10 : Score = 1.64000
kouho   11 : Score = 1.64000
kouho   12 : Score = 1.60000
kouho   13 : Score = 1.60000
kouho   14 : Score = 1.46667
kouho   15 : Score = 1.46667
kouho   16 : Score = 1.46667
kouho   17 : Score = 1.46667
kouho   18 : Score = 1.46667
kouho   19 : Score = 1.46667
```

図 7: 候補選択のウィンドウ

図のように、対応づけされた部分木は、一対の部分木が、同一の色によって色づけされ、互いの部分木の根に当たる部分が、色づけされた両矢印で結ばれる。また、コントロールパネル（図 4 ）左下の “候補:”, “スコア:” には、構造照合結果の候補番号（得点が高い順に 0, 1, 2... と番号が振られている。）と、その結果の得点が表示されている。

4.5 候補選択

二言語間の構造照合による部分木の対応づけにおいて、多くの場合、複数の結果が得られる。これらの部分木の対応づけの複数の候補のうち、最も適切であると思われる候補を選択し、表示する（すなわち、候補選択）方法について述べる。

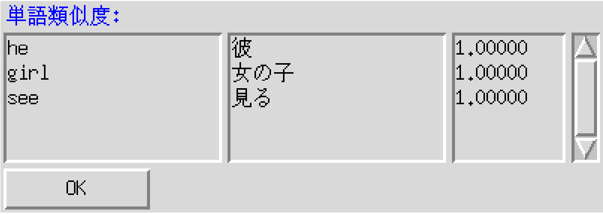
まず、ボタン群中の “–>” をクリックすると、現在表示されている部分木の対応づけの次の候補（現在の候補より得点が低い候補）が表示される。そして、ボタン群中の “<–” をクリックすると、現在表示されている部分木の対応づけの前の候補（現在の候補より得点が高い候補）が表示される。また、“候補” のボタンをクリックすると、すべての照合結果の候補番号とその得点のリストを表示したウィンドウが現れる。（図 7）このリストより、照合結果の候補を選択することができる。

4.6 単語類似度表示

ボタン群中の “単語類似度” のボタンをクリックすると、二言語の単語間の類似度を表示したウィンドウが現れ（図 8）、構造照合において参照した二言語の単語間の類似度が表示される。

4.7 構造照合結果保存

現在表示されている部分木の対応づけの結果を、ファイルに保存することができる。具体的には、コントロールパネル中の “保存ファイル” の入力フィールドに結果を保存したいファイルを入力しておき、す



単語類似度:

he	彼	1.00000
girl	女の子	1.00000
see	見る	1.00000

図 8: 単語間の類似度表示

ぐ横の “SAVE” のボタンをクリックする。結果は、各言語の全体の依存構造と対応する部分木のペアの集合として、次のような形式でファイルに保存される。

まず、各言語の依存構造を表す木は次のように表現される。

木 ::= {木の根ノード} | {木の根ノード 子木 子木 ...}
木の根ノード ::= {ノードＩＤ ノードラベル 親ノードへの枝ラベル}
子木 ::= 木

ここに、ノードＩＤは、木構造におけるノードに一意に割り当てられた番号及び記号であり、ノードラベルは、実際に木のノードとして表示される単語もしくは、OR ノードを表す ’OR’ であり、また枝ラベルは、ノードから、その親ノードへ伸びる枝に付加されたラベルである。なお、親ノードが木構造の root となる場合等のように、枝ラベルが付加されないときは、枝ラベルは {} となる。

そして、各言語の全体の依存構造を表す木は次のように表される。

L:左側の言語の全体の依存構造を表す木\n
R:右側の言語の全体の依存構造を表す木\n
(ここに \n は改行)

また、構造照合結果は対応する部分木のペアとして、次のように表される。ここで、部分木の表現は前述した依存構造を表す木とフォーマットは同じであり、木構造のうち、部分木のカバーする領域の部分を表現するものとする。

P:左の言語側の部分木１:右の言語側の部分木１\n
P:左の言語側の部分木２:右の言語側の部分木２\n
......
(ここに \n は改行)

ここで、図 6 の場合の構造照合結果は次のように保存される。

```
L:{{1 . {}} {{2 he subj_e1}} {{3 OR {}} {{4 saw {}} {{5 girl obj_e4}}
{{6 telescope with}}} {{7 see {}} {{8 girl obj_e4}} {{9 telescope with}}}
{{10 saw {}} {{11 girl obj_e4} {{12 telescope with}}}} {{13 see {}}
{{14 girl obj_e4} {{15 telescope with}}}}}}
R:{{1 見る {}} {{2 望遠鏡 で}} {{3 彼 は}} {{4 女の子 を}}}
P:{{7 see {}}}:{{1 見る {}}}
P:{{2 he subj_e1}}:{{3 彼 は}}
P:{{8 girl obj_e4}}:{{4 女の子 を}}
P:{{9 telescope with}}:{{2 望遠鏡 で}}
```

(実際は、L: の行は１行)

5 今後の予定

V-Para の今後の開発方針などについて述べる。

今回構築した V-Para のシステム的な問題点を踏まえ、次のような点について検討を加えたいと考えている。

- プロセス間の通信の簡素化

  現在のシステムでは、対訳辞書サーバと通信サーバの 2 箇所でプロセス間のソケット通信を確立させなければならず、プロセス間における通信が非常に複雑となり、軽度のプロセス間の通信の不具合でシステム全体が動作しなくなることが起こる。対訳辞書サーバ及び、通信サーバの通信などをできるだけ省き、プロセス間通信の簡素化をはかる。

- システム全体の簡略化

  システム全体が非常に大きくなっており、また、システムの起動のためには、SICStus-Prolog, Perl, Tcl/Tk など、複数のプログラミング言語を使用する必要があるため、インストールや初期のシステム起動が非常に困難になるという問題がある。ＧＵＩ部である paramatch.tcl を、 Perl による Tk ライブラリである pTk を用いて Perl スクリプト化し、また、構造照合のアルゴリズムを C 言語及び Perl でプログラムすることにより、システム全体を C 及び Perl に統合することができ、システムのインストール等が容易になると考えられる。

- 部分木の対応づけの修正機能の実装

  現在、部分木の対応づけについて、人手によって修正を行なう機能は実装されていない。この機能は、構造照合の段階でほぼすべての部分木の対応の可能性を候補として取り出して表示するために、あまり必要性のない機能であると考えられる。しかし、ユーザ側が自由に部分木を修正することができるように、ごく簡単な修正機能のみを実装する方向で検討したいと考えている。

参考文献

[1] Matsumoto, Y., Ishimoto, H. and Utsuro, T.: Structural Matching of Parallel Texts, *Proceedings of the 31st Annual Meeting of ACL (ACL93)*, pp. 23 – 30 (1993).

[2] Kay, M. and Röscheisen, M.: Text-Translation Alignment, *Computational Linguistics, Vol. 19, No. 1*, pp. 121–142 (1993).

[3] Meyers, A., Yangarber, R. and Grishman, R.: Alignment of Shared Forests for Bilingual Corpora, *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 460–465 (1996).

[4] 北村美穂子: 対訳コーパスからの翻訳知識の獲得と機械翻訳への適用, 奈良先端科学技術大学院大学修士論文, (1995).

[5] 北村美穂子, 松本裕治: 対訳コーパスを利用した翻訳規則の自動獲得, 情報処理学会論文誌 *Vol.37, No.6*, pp. 1030–1040 (1996).

[6] 宇津呂武仁, 松本裕治: 対訳辞書および統計情報を用いた二言語対訳テキスト照合, コンピュータソフトウェア *Vol. 12, No. 5*, pp. 414–423, (1995).

[7] Programming Systems Group: Swedish Institute of Computer Science, SICStus Prolog User's Manual, (1995).

[8] Ousterhout, J. K. (西中芳幸, 石曽根信訳): Tcl & Tk ツールキット, ソフトバンク社, (1995).

[9] 松本裕治他: 日本語形態素解析システム JUMAN 使用説明書 version 2.0, 京都大学工学部 (長尾研究室) 奈良先端科学技術大学院大学 (松本研究室), (1994).

[10] 松本裕治他: 構文解析システム SAX 使用説明書 version 2.1, 奈良先端科学技術大学院大学 (松本研究室), (1994).

付録

A　ボタン等の機能一覧

V-Para の機能及び操作の一覧を次に示す。(⟹ ...) で参照すべき図及び節を示す。

A-1 コントロールパネル　(⟹ 図 4, 4.1 節)

入力フィールド	
“英語 :”, “日本語:” (⟹ 4.3 節)	対訳文を入力する。 右隣の “CANCEL” ボタンで入力内容をクリアする。
“対訳辞書サーバ:” (⟹ 4.2 節)	対訳辞書サーバのホスト名を指定する。 右隣の “CANCEL” ボタンでサーバのホスト名をクリアし、 “OK” ボタンでホスト名を実際に指定する。
“保存ファイル:” (⟹ 4.7 節)	構造照合の結果を保存するファイル名を指定する。 右隣の “SAVE” ボタンで結果を保存ファイル保存し、 “FILE” ボタンでファイルセレクタ (⟹ A-5) を開き、 保存ファイルの選択を行なう。 “CANCEL” ボタンでファイル名をクリアする。

ボタン群	
“入力” (⟹ 4.3 節)	対訳文を paramatch に送り、文の構造照合を行なう。
“<–”, “–>” (⟹ 4.5 節)	現在表示されている部分木の対応づけの前の候補 （現在の候補より得点が高い候補）および、 次の候補　（現在の候補より得点が低い候補）を表示する。
“候補” (⟹ 4.5 節)	候補選択ウィンドウ (⟹ A-3) を開き、 照合結果の候補を選択する。
“対訳文” (⟹ 4.3 節)	対訳文表示ウィンドウ (⟹ A-2) を開き、 対訳文を選択する。
“単語類似度” (⟹ 4.6 節)	単語類似度表示ウィンドウ (⟹ A-4) を開き、 単語間の類似度を表示する。
“終了”	V-Para を終了する。

A-2 対訳文表示ウィンドウ　(⟹ 図 5 左, 4.3 節)

ウィンドウ左下のボタン	
“OK”	対訳文を直接クリックした後に押して、対訳文を選択する。 対訳文を直接ダブルクリックしても選択できる。
“FILE”	ファイルセレクタ (⟹ A-5) を開き、 対訳コーパスファイルを選択する。
“CANCEL”	対訳文表示ウィンドウを閉じる。

A-3 候補選択ウィンドウ (⟹ 図 7, 4.5 節)

ウィンドウ左下のボタン	
“OK”	候補番号、スコアを直接クリックした後に押して、候補を選択する。 候補番号、スコアを直接ダブルクリックしても選択できる。
“CANCEL”	候補選択ウィンドウを閉じる。

A-4 単語類似度表示ウィンドウ (⟹ 図 8, 4.6 節)

ウィンドウ左下のボタン	
“OK”	単語類似度表示ウィンドウを閉じる。

A-5 ファイルセレクタ (⟹ 図 5 右)

ウィンドウ左下のボタン	
“OK”	ファイル名を直接クリックした後に押して、候補を選択する。 ファイル名を直接ダブルクリックしても選択できる。
“CANCEL”	ファイルセレクタを閉じる。