情報学基礎 64－1
自然言語処理 145－1
（2001．9．10）

機械翻訳システムの評価と基準

田中康仁
兵庫大学
E－mail:yasuhito@humans－kc. hyogo－dai. ac. jp

機械翻訳システムの初期の研究からみると、開発も進み実用化に向けて動き出している。
ここでは機械翻訳システムの評価の基準の移り変りと、機械翻訳システムの今後の発展を考えるにあたって、どのように基準を考えればよいかを考察した。

Evaluation Method of Machine Translation and Related the Measures.

Yasuhito Tanaka
Hyogo University

E－mail : yasuhito@humans－kc.hyogo－dai.ac.jp

As to the initial－stage study on machine translation system, development has benn progressing, and movement toward practical use has started.
Here, we studied the change in standards to evaluate machine translation systems, and how to consider the standards regarding fututre progress of machine translation systems.

〔０〕はじめに

機械翻訳システムの初期の研究からみると、機械翻訳システムは開発も進み実用化に向けて動き出している。

ここでは機械翻訳システムの評価の基準の移り変りと、機械翻訳システムの今後の発展を考えるにあたって、どのように基準を考えればよいかを考察した。

〔１〕従来の基準

評価の方法について

従来からの行われている３つの方法について述べる。

(1) 機械翻訳システムを対象にＡ、Ｂ、Ｃのランクを付けるための評価

これは機械翻訳の紹介記事でよく使われる方法で数文又は十数文を対象として各社の性能を評価するものである。

(2) 機械翻訳システムの利用と使わない方法での費用の分析による評価

機械翻訳システムの初期の開発時期にはこのような議論が行われたが、機械翻訳システムの性能が向上してきた段階ではこうした議論はあまりされなくなってきた。むしろ積極的に利用し、電子化された翻訳情報を作ることが行われている。

(3) ALPACレポートの中にある「理解容易性」と「忠実度」について

これは1996年に出されたALPACレポートの中にある基準で「理解容易性」が５段階、「忠実度」が７段階に分類されている。しかも、評価者は数人の評価者で行うことが考えられている。

これは研究としては大変良い基準であるが、実際に大量のデータに対して適用すると、基準のどの分類に入れてよいかという判断のゆれが発生する。複数人の評価者で行えばさらに判断の個人的ゆれが発生する。また、人手による評価のため、２つの言語に精通した高い教養の持っている人でなければならない。費用がかかるし、大量のデータには不適当である。

〔２〕機械翻訳の開発と基準

機械翻訳システムの開発の初期の段階から概観してみると幾つかの問題点がわかる。

(1) ソフトウェアとしての安定性

機械翻訳システムはコンピュータ上で動くソフトウェアである。このためどのようなデータに対しても、ソフトウェアが何の応答もなく終了することは避けなければならない。システムの安定性と頑強さがなければならない。

(2) 機械翻訳システムと人間の関わり

機械翻訳システムは完全に全ての文を翻訳できるということはない。常に数%から数10%の誤りが発生する。そのため人間との対話が重視されるし、使い易さが求められる。これにより誤りが発生しても人手により補うことで、充分使いものになる。この点の評価は考えなければならない。

次に機械翻訳システム開発者の立場からの評価を考えてみよう。システム開発者にとってはどれだけ短期間に翻訳誤りの例文を多量に集め、それらを基に文法や辞書、システムの構成全体を修正するかという点にある。しかも、例文は例外的なものでなく、ごく普通の文の中に含まれる文で、機械翻訳の誤りを起こす文が重要である。

その結果、機械翻訳の精度が向上すればよいのである。

評価方法として次のようなものを考える。

(1) 単語数列による評価

例えば英日機械翻訳システムを考えるとしよう。この場合英文を単語数列に分類する、そして機械翻訳システムにかけ、翻訳誤りを修正する。単語数が少ないものは定形文であるし、単文である。これらは修正しやすい。修正しやすいものから順次複雑な文の修正をすべきである。これは大変便利な方法で、機械翻訳システム開発者から喜ばれた。又、精度向上に大変貢献した。

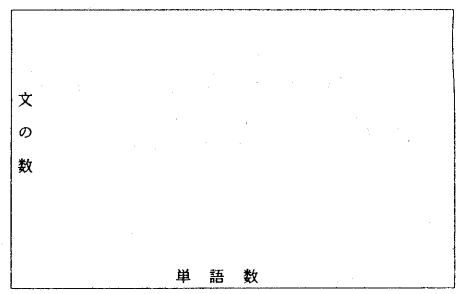
例えば日本ＥＤＲの英文コーパスの単語数別分布を調べると次のようになった。

単語数	1	2	3	4	5	6	7	8	9
文の数	0	19	460	1,889	5,030	6,798	7,791	8,416	8,698

10	11	12	13	14	15	16	17	18	19
9,114	9,018	9,176	9,050	8,815	8,446	8,245	7,466	6,559	5,283

20	21	22	23	24	25	26	27	28	29
3,645	562	464	321	235	142	77	46	23	12

30	31	32	33	合計
2	0	0	1	125,803



これらのデータのうち単語数２～７までのものを評価した。ここでは５段階の「理解容易性」を用いてテストした。

評価結果　Ｃ社の英日翻訳システム

	5	4	3	2	1	合計	平均値
2 単語文	15	1	1	1	1	19	4.47
3 単語文	336	97	16	11	0	460	4.65
4 単語文	1,369	336	157	25	2	1,889	4.61
5 単語文	3,655	808	510	53	4	5,030	4.60
6 単語文	4,742	1,379	595	81	1	6,798	4.58
7 単語文	4,471	2,331	870	118	1	7,791	4.43
合計	14,588	4,952	2,149	289	9	21,987	4.53
%	66.35	22.52	9.77	1.31	0.04	100%	

評点の高いものが良い結果を示す。

費用の関係で７単語以上のものは省略した。

(2) 文法事項別に評価する。

これは興味ある方法である。英文法書には各章ごとに例文がある。この例文を集めて翻訳し、文法事項が正しく処理されているか検討するものである。しかし、例文の翻訳悪りが必ずしも文法事

項と一致しているとは限らない。また多量のデータを集めようとしても文法書に限りがあるので集まらないし、人手によって文法事項を判断し、使用しようとすると大変な費用がかかる。この方法も少し行ってみた。

① 具体的方法

英文法書としては「チャート式シリーズ新英文法」小野経男著数研出版1990年を利用した。約3,000文をこの中から抽出した。この文法書は次のような章立てになっている。

第1章：文	第12章：名詞
第2章：品詞・句・節	第13章：冠詞
第3章：動詞の活用	第14章：代名詞
第4章：動詞の時制	第15章：形容詞
第5章：助動詞	第16章：副詞
第6章：受動態	第17章：比較
第7章：不定詞	第18章：疑問詞
第8章：分詞	第19章：関係詞
第9章：動名詞	第20章：仮定法
第10章：前置詞	第21章：時制の一致と話法
第11章：接続詞	第22章：文の転換

このように個々の章ごとに例文があるためこれを機械翻訳システムに利用し、良い分野と悪い分野を見つけることにする。

② 評価の基準

評価を細分割することも考えられるが、多くの労力を必要とするし、評価のゆれも発生する。そこでここでは3つに評価することにした。

○：文章の意味が正確である。

△：文章の意味はなんとなく分かる。

×：文章の意味が違う(意味がわからない。)

(3) 機械翻訳システムを利用している人による評価

機械翻訳システムを実際に使用している人は翻訳に利用できるか、利用できないかで訳文を見ている。これは機械翻訳システムの開発者の立場とは異なるが、最終的には利用者が満足できるものになり、全体的な精度向上につながればよいのであるし、このような立場の評価者の意見を尊重すべきである。

ここにその基準を示す。

（ｉ）非常に役立つ

（ii）一部ニュアンスが違うが役に立つ

（iii）意味が違っているが、前後関係でわかる。

（iv）一部訳されている。

（ｖ）間違った訳になっている。又は意味不明。

翻訳者がいかに利用できるかという基準が重要である。

次のようなシートを作成して行った。これは模範的な和訳があるため、作業が行いやすかったのと、評価の欄を横に取り○印を入れるようにしたため、前の訳との関係をも考えながら評価できるという良い点があった。評価は微妙な面があるが、ただ単純に評点を記入するよりは良い方法であった。

Seq No.	英　文	和　訳	機械翻訳による日本語訳	評価 5	4	3	2	1
1	How is the economy?	景気はどうですか?	経済はどうですか。	○				
2	The Japanese grew by 4% last year.	日本の経済は昨年4%成長した。	日本の経済は昨年の4%ずつ成長しました。		○			

評価者の作業心理を考慮したことが良かったと考えられる。

(4) 専門分野別の評価

一般的な分野の英日コーパスは約２０万例文程集めた。その結果。多量のデータは機械翻訳システムに有効であるということもわかった。

今後は専門分野別の多量の例文を集め、機械翻訳のテストデータとして使うべきだと考えている。それではどのような分野を考えるべきであろうか。これは利用者がどのような分野を望んでいるかという分析が必要である。富士通のカタログを見ると次のような分野がある。

1、情報処理	14、生物
2、電気・電子	15、〔医学〕生化学
3、物理・原子力	16、〔医学〕薬学
4、機械	17、〔医学〕解剖学
5、工業化学	18、〔医学〕疾患症状
6、プラント	19、〔医学〕精神医学
7、土木建築	20、〔医学〕医療機器
8、金属	21、金融・経済
9、地学・天文	22、法律
10、輸送	23、ビジネス
11、自動車	24、人名・地名
12、軍事	25、環境
13、農林水産	

（富士通、ＡＴＬＡＳＶ６英日・日英翻訳ソフトカタログより）

Ｉ.Ｂ.Ｍのカタログには次のような分野がある。

1）インターネット	5）アート
2）エンターテイメント	6）スポーツ
3）ビジネス	7）科学
4）政治	

さらに１つの専門分野で評価してみた。専門分野といっても、色々な分野が考えられる。ここでは個人信用情報の金融クレジットの分野を考える。

この分野は、日本では古くは質屋とか個人的金融業者（高利貸し）があったが、一般大衆を対象とし、コンピュータを利用したスコアリングに基づいて、個人的に貸し付けるシステムは米国からの技術導入に依存するところが多かった。

	評点と各文数							
単語数	5	4	3	2	1	計	%	評点
4以下	13	5	4	0	4	26	3.08	3.88
5～9	205	107	88	44	57	501	59.36	3.72
10～14	89	62	50	22	19	242	28.67	3.74
15～19	5	8	14	13	3	43	5.09	2.98
20～24	6	8	5	5	0	24	2.84	3.62
25以上	2	3	1	2	0	8	0.95	3.62
合　計	320	193	162	86	83	844	100.00	3.688
%	37.91	22.87	19.19	10.19	9.83	100.00		

全体的な評点は3.688であり、あまり良い結果とはいえない。今後もっとこの分野のテストを行い、精度向上をはからなければならない。

(5)　専門用語による評価

特に科学技術の分野では多くの専門用語が使われている。専門用語も時代と共に名称や概念が少しずつ変化している。

このような専門用語を正確に使うことは重要なことであるが、しばしば曖昧な使用や表記のゆれをもたらしている。次に機械翻訳システムの用語を調べてみる。

ｉ）機械翻訳システムの専門用語

ある企業の機械翻訳のカタログを調べてみると、次のような専門用語があると宣伝している。

専門分野		登録語数
科学技術		170,000
	情報処理	17,800
	電気／電子／通信	16,600
	土木／建築	13,500
	自動車／鉄道／船舶／航空	28,000
	自然科学	24,600
	生物	8,800
	機械	21,700
	工業化学	36,700
ビジネス		60,000
合計		230,000

（A社カタログより）

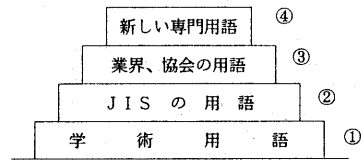
しかし、この程度の専門用語で充分なのだろうか？

このカタログには分野別専門用語が使用されている専門用語の何%をカバーしているという説明はない。

ii）専門用語の階層性について

専門用語は多くの専門家がある特定の分野の概念を表現するために作られたものである。

それゆえ、定義が定まった基礎的なものから、基礎的な専門用語を組み合わせた複合的な専門用語まで幅広く存在する。それらは次のような階層をなしている。



① 学術用語は文部省を中心に各専門分野の人々、学会等が中心となり制定されたものである。２２の分野があり各分野には次のような用語がまとめられている。

1）数学　　12）建築学
2）天文学　　13）船舶学
3）物理学　　14）航空工学
4）気象学　　15）計測工学
5）地震学　　16）原子力工学
6）分光学　　17）歯学
7）化学　　18）論理学
8）動物学　　19）キリスト教学
9）植物学　　20）図書館学
10）遺伝学　　21）土木工学
11）機械学　　22）採鉱冶金学

これらの学術用語を基礎として、日本工業規格（ＪＩＳ）が制定されている。ＪＩＳは科学技術分野が中心である。

ＪＩＳの専門用語分野は次の通りである。

土木・建築／機械／電気／自動車／鉄道／船舶鉄鋼／非鉄金属／化学／繊維／鉱山／パルプ・紙窯業／医療安全用具／航空／情報技術／その他〈包装、溶接、原子力、放射線（能）、品質管理、音響、照明、他〉

- 各分野で制定された用語規格及び重要個別規格、２８４規格に規定された最新のＪＩＳ用語約７３，０００語を収録。
- 各用語には読み方、対応外国語（英・独・仏語）、定義、慣用語、略号、記号、量記号、単位記号及び図・表を収録。

（日本規格協会カタログより）

最新のＪＩＳ工業用語大辞典には７万３千語収めている。

さらにこれらを基にして業界、協会、団体が専門用語辞書をまとめている。専門用語辞書にまとめる前の段階では、業界の通用語のようなものが確立しつつある。

専門用語がどの程度機械翻訳システムの辞書に登録されているか調べてみた。

専門分野では、専門用語がどの程度正しく翻訳されているかが重要である。

ここでは貿易用語の専門用語約1,000語を用いて評価した。テストに使用した辞書は次のものである。

「基本貿易用語辞典」石田　貞夫　白桃書房

2社の英日翻訳システムを利用した。

テスト結果

B 社

○	210	21.2%
△	175	17.7%
×	604	61.1%
計	989	100.0%

C 社

○	140	14.2%
△	121	12.2%
×	728	73.6%
計	989	100.0%

結果の良い方のB社でも○と△を加えても３８．９％程度である。専門用語の不足が目立つ。重要語は入っているのかもしれないが、あまりにも低い水準である。専門用語の使用頻度を考慮していない面もあるが、それにしても悪い結果である。

科学技術用語を優先していると思うが、もっと経済分野やその他の分野についても力を入れてほしいものである。

英単語１単語の訳語には複数の訳語があり、曖昧さがあるが、英単語２単語以上の専門用語は曖昧さが急激に減るので、どんどん辞書に入力すべきである。

(6) 機械翻訳システムの旧版と新版の評価、副作用の調整

新しい改定版を作ると旧版より良くなっているはずである。しかし、文法規則の追加であるとか、１語で構成されている専門用語等を追加すると、その影響は修正者が期待している以外の所にあらわれることがある。それ故、このような問題点を含む修正の場合は１つの規則を追加するごとに、大規模なコーパスを用いて副作用が出現していないか調べるべきである。我々も幾度か旧版と新版とで翻訳結果が異なるものを見つけた。

① I will help him whatever you may say.
君がなんと言おうとも私はあくまで彼を助ける。
旧版: あなたが何を言っても、私は彼を助けるつもりである。
新版: あなたが何を言うことができるならば、私は彼を助けるつもりである。

② Wherever you may go, you may find examples of his evil doings.
どこへ行こうとも、彼の悪行の例が見つかるでしょう。
旧版: どこに行っても、あなたは彼の悪行の例を見つけることができる。
新版: どこに行くことができるならば、あなたは彼の悪行に関する例を見つけることができる。

新しい版が作られるとその総合結果がどの程度向上したか測定しておき、次にどの程度の向上をめざすか、又どの程度の知識と労力を投入すべきであるか予測することも重要である。

WWWのGreen and Whiteに日本の機械翻訳の製品や評価値がある。これを過去にさかのぼって調べてみると富士通のATLASではこのようになる。

年月日	版	英日評点	日英評点
1997年 7月	4.0	54.5	60.6
1998年 6月	5.0	63.5	65.2
1999年 6月	6.0	77.3	81.7
2000年11月	7.0	83.7	87.5

利用者にはどのような改良と投資がなされたかはわからないが改定版ごとの推移が分かる。これも１つの評価基準である。

〔３〕評価するための各種データやソフトウェアについて

機械翻訳システムの評価は、例文を入れてその結果を人手により調べるというのが普通と考えられてきたが、それでは急激な改良は見込めない。もっと新しい手法を考えるべきである。

(1) 大量のバイリンガル又はマルチリンガル、パラレルコーパスを用いる。

１つの改定版の修正に翻訳誤りを起こす文20,000文を準備すると仮定しよう。この機械翻訳システムの精度を７０％とすると20000／(1－0.7) ≒66,666文　約７万のコーパスが必要である。

このようなテストデータを毎年準備すべきである。しかも、作業者の労力を減らすための対訳付のパラレル・コーパスにすべきである。

(2) 分野別のパラレル・コーパスの準備

大量のパラレル・コーパスが必要であることは前に述べたが、これを各専門分野別に準備すべきである。これは専門用語が、どの程度機械翻訳システムの辞書に登録されているかを調べるためにも重要である。

(3) 機械翻訳システムに訳文自動評価システムを追加すべきである。

単語数別ファイルを作り機械翻訳システムを評価する方法は、問題点の抽出が容易で興味ある方法である。

しかし、翻訳結果の評点付けは人間の作業であり、大変労力のかかる作業である。この作業を自動的に行うように機械翻訳システムに自動評価システムを組み込んでおくことが重要である。最終的な判断は人間の作業であるが、ある程度のところまでは機械的に可能である。これにより翻訳結果の大まかなグループ分けが可能である。このようにして作業の迅速化が図られる。

例えば

１）構文解析で曖昧さが減らすことができない。
２）未知語が出現した。
３）意味解析のパターンが無い。
４）専門用語が無いため合成訳を作成した。
５）並列表現の解析がうまく行えなかった。

等、機械処理システムで誤る場所は幾つかある。こ

れらの問題点について重み付けを行い、評点を付けるのも一つの方法である。自動評価システムを作らなければならない。

このような自動評価システムは人間の評価値と少し異なるかもしれないが、改良しなければならない文を大量の訳文の中から早く見つけ出すためには有効な手段である。このようにして問題点を修正し、ほぼ完了した時点で人間の精密な検査を行い、さらなる改良をすべきである。

(4) テスト・データの公開・非公開

テスト・データは非公開で行うか、公開するかは大きな問題である。しかし、テスト結果を公表するにあたってはどうしてそのような結果になったのかを示さなければならない。そのためたとえ非公開にしても少しずつ内容がわかってしまう。

公開するならば毎回のテストごとに内容を変えてゆかねばならない。前回のテスト結果との比較については少し考慮しなければならない。

〔4〕機械翻訳システムの総合評価値

機械翻訳システムの総合評価値が85点であるとか、70点であるとか書かれている。これは注意して見なければ誤解を招く。これらは85%のものが又は70%のものが完全に正しく訳されたわけではない。

例えば次のような評点計算がなされて出されているのである。

	評価点	結果の分布	ランクの得点	得点
評価点の分析	5	65%	1.0	65点
	4	20%	0.7	14点
	3	10%	0.4	4点
	2	5%	0	0点
	1	0%	0	0点
総　　計		100%		83点

実際に使ってみた感じと総合得点が異なると感じるのはこのためである。それ故、評価点1、2のものは極力少なくし、評価点3、4のものは評価点5になるように改良すべきである。これは1つのモデルとして計算してみた。

〔5〕評価と今後の機械翻訳システム開発に向けて

機械翻訳システムがもっと精度向上が図られることを期待して、次のことを提案する。

(1) 各専門分野別の大量のバイリンガル・パラレル・コーパスを準備し、テストすることを望む。

(2) 利用者からの誤り翻訳の例文を収集する方法を確立すること。

このようなことがないと利用者の不満を解消した良い製品開発は望めないし、利用者の不満だけが残る。「我々の知能水準を越えている」と実感する機械翻訳システムの出現は難しい。

(3) 各種辞書の充実

機械翻訳システムに必要な辞書ばかりでなく、関連する用語辞書も重要である。最終的には人間の判断が必要なのであるから、人との対話を行いやすくしなければならない。

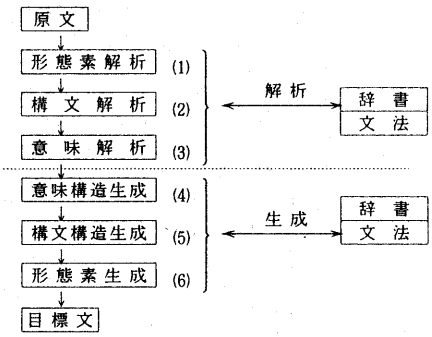
(4) 複雑な処理から単純な処理へ

機械翻訳システムは幾つもの処理段階を経て訳文が出力されるが、多くの処理を通れば通る毎に解析、生成の曖昧さや誤りが発生する。我々人間の処理はもっと単純な処理で訳文を生成している。このことを研究しなければならない。

例えばテンプレートによる解析、訳文生成のシステムをもっと多く使うことが必要である。

(1) 機械翻訳の構造上の問題点

機械翻訳システムは次のような構成要素から成り立っている。



おおまかに6つの構成要素から成り立っている。これらの6つの要素の各信頼性をP_1, P_2, P_3, P_4, P_5, P_6 とする。

原文が目標文まで正確に変換できる信頼性は

$P=P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6$　　である。

ここで〔1〕で評価した結果をあてはめると

$0.70=P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6$　　となる。

P_1, P_2, P_3, P_4, P_5, P_6 の各確立が個別に測定できないので6つの要素が同じ信頼性と仮定する。

$P_1=P_2=P_3=P_4=P_5=P_6$

$0.70=P_1^{\,6}$

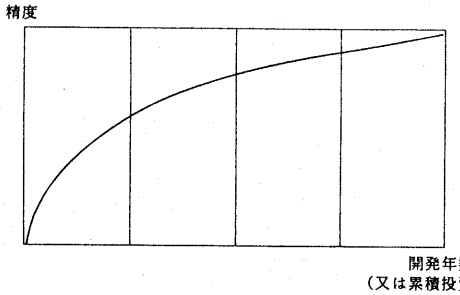
$P_1=\sqrt[6]{0.70} \fallingdotseq 0.942286$

つまり個々の信頼性は平均0.94、%で表示すると94%程度である。これは個々の構成の信頼性はそのプログラムと使用されている文法、辞書によるものである。全体の信頼性を上げるには個々の要素の信頼性を上げなければならない。

実際には個々の確率は異なっている。開発者はどの確率が低いかよく知っている。

低いところを上げる努力をすれば効果的に全体が良くなるのである。しかも、開発費用や期間のかからないものが、改良の最優先順位になる。

精度と開発期間又は累積投資容量との関係は次のようになる。



製品販売部門と協力し、どのように開発コスト、労力を捻出するかが重要である。

〔６〕おわりに

機械翻訳システムの評価方法と評価基準について考えてきた。これら評価基準はただ単なる評価ではなくよりよい製品開発に向けた精度向上でありたい。

最後に、この研究に協力して下さった兵庫大学大学院生中元友子さん、向出敦子さん、英日翻訳システムの訳文評価をして下さった兵庫大学の小泉　毅先生、元岡山県灘崎中学校教諭大賀敏雄先生に感謝の意を表す。

また、機械翻訳システムを提供して下さったメーカーの方々に感謝します。

〔７〕参考文献

(1) 社　日本電子工業振興協会
「自然言語処理システムの動向に関する調査報告書」
平成9年4月

(2) 牧野武則　評価技術　「機械翻訳」Ｂｉｔ別冊
共立出版　1988年9月

(3) 長尾　真　「機械翻訳はどこまで可能か」
岩波出版　1986年6月

(4) 田中康仁　機械翻訳用のテストデータ
情報処理学会第61回（平成12年度後期）全国大会
2－pp127－128　2T－1　2000年10月

(5) 田中康仁　中元友子　機械翻訳システムのテスト
－英文法を利用して－情報処理学会第60回（平成12年前期）全国大会　2－pp102－102　3K－1

(6) 田中康仁　機械翻訳システムの専門用語について
情報処理学会第60回（平成12年前期）全国大会
2－pp103－104　3K－2

(7) 田中康仁　機械翻訳システムの評価と改善
情報処理学会自然言語処理　133－1　pp1－6
1999年9月

(8) 田中康仁　機械翻訳システムの今後について
情報処理学会自然言語処理　137－2　pp9－14
2000年6月

(9) 田中康仁　向出敦子　機械翻訳システムのテスト
－一つの専門分野について－
情報処理学会第63回（平成13年後期）全国大会
（発表予定稿）　2001年10月

(10) 石田貞夫編　「基本貿易用語辞典」白桃書房
昭和51年6月