Machine Translation

Harold Somers
Centre for Computational Linguistics
UMIST
PO Box 88
Manchester M60 1QD
England
`Harold.Somers@umist.ac.uk`

Overview

I. Historical perspective

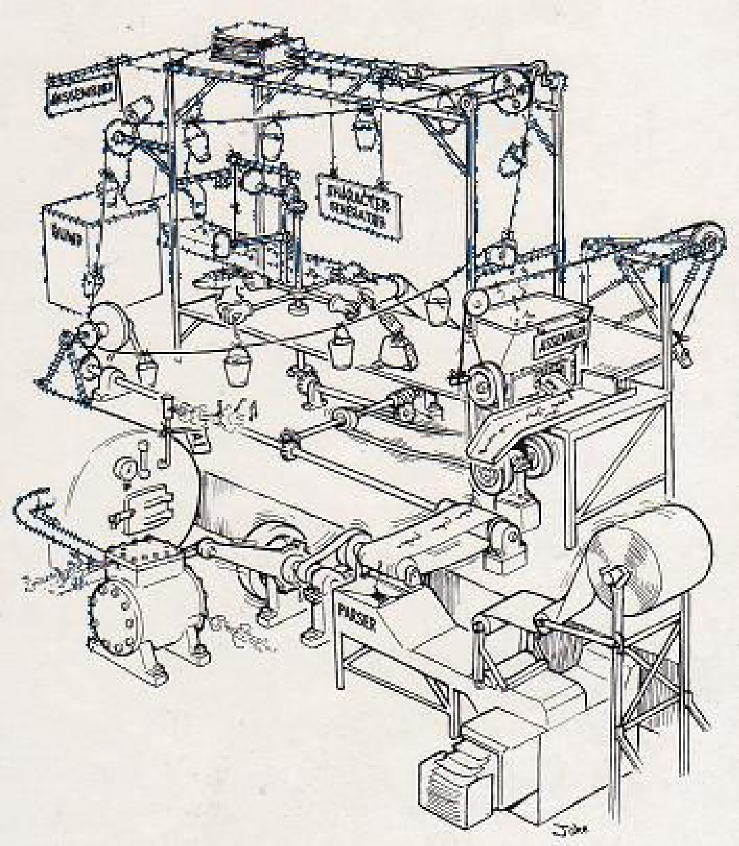
II. Linguistic problems

III. Evaluation

IV. New paradigms

Part I: Historical perspective

1. Historical landmarks
2. The “2nd generation”
3. Modes of use



1. Historical landmarks

- 1933 patents predate invention of computer!
- war-time use of computers in code breaking
- Warren Weaver’s memorandum 1949; cold war
- Early promise of FAHQT

approx. 1955 to 1966

- Difficulties soon recognised:
  - no formal linguistics
  - crude computers
  - need for "real-world knowledge"
  - Bar Hillel's "semantic barrier"
- 1966 ALPAC report
  - "insufficient demand for translation"
  - "MT is more expensive, slower and less accurate"
  - "no immediate or future prospect"
  - should invest instead in fundamental CL research

1966 to 1978

- public MT funding in USA stops
- private research; also in other countries
- “2nd generation approach”: linguistically and computationally more sophisticated
- c. 1976: success of Météo (Canada)
- Systran in use at several sites
- 1978: CEC starts discussions of its own MT project, Eurotra

early 1980s

- MT renaissance (except in US): conferences, journals
- first commercial systems early 1980s
- FAHQT abandoned in favour of
  - "Translator's Workstation"
  - interactive systems
  - sublanguage / controlled input

late 1980s

- Eurotra project at its height
- lots of activity in Japan
- speech MT projects also funded
- emergence of PC
- several systems report “many users”
- despite low quality, users claim increased productivity
- general explosion in translation market

1990s

- Research based on various ideas
- Renewed funding in USA: *“MT might be possible after all”*
- In Europe, Eurotra ends amid huge criticism; despite rising need for translation; little new research; several other projects close: *“MT doesn’t work”*
- In Asia, research moves on to new topics; MT software widely available: *“MT is a success”*

Current situation

- creditable commercial systems now available
- wide price range, many very cheap ($50)
- main platform is PC
- MT available free on WWW
- widely used for web-page and e-mail
- low-quality output acceptable
- but still only a small set of languages covered
- speech translation widely researched

2. The “2nd generation”

- advances in both computer science and formal linguistics
- modular programming styles
  - divide problem up into manageable subproblems
  - separation of algorithms and data
  - “procedural” vs “declarative”
- linguistically sophisticated data structures and translation schema
- linguistic rule-writing formalisms
- depth of analysis

2.1 Algorithms vs. data

- distinguish between generic “translation software” and language-pair-specific data
- linguistic formalisms for lexicons and grammars
- algorithms are “procedural”, data “declarative”

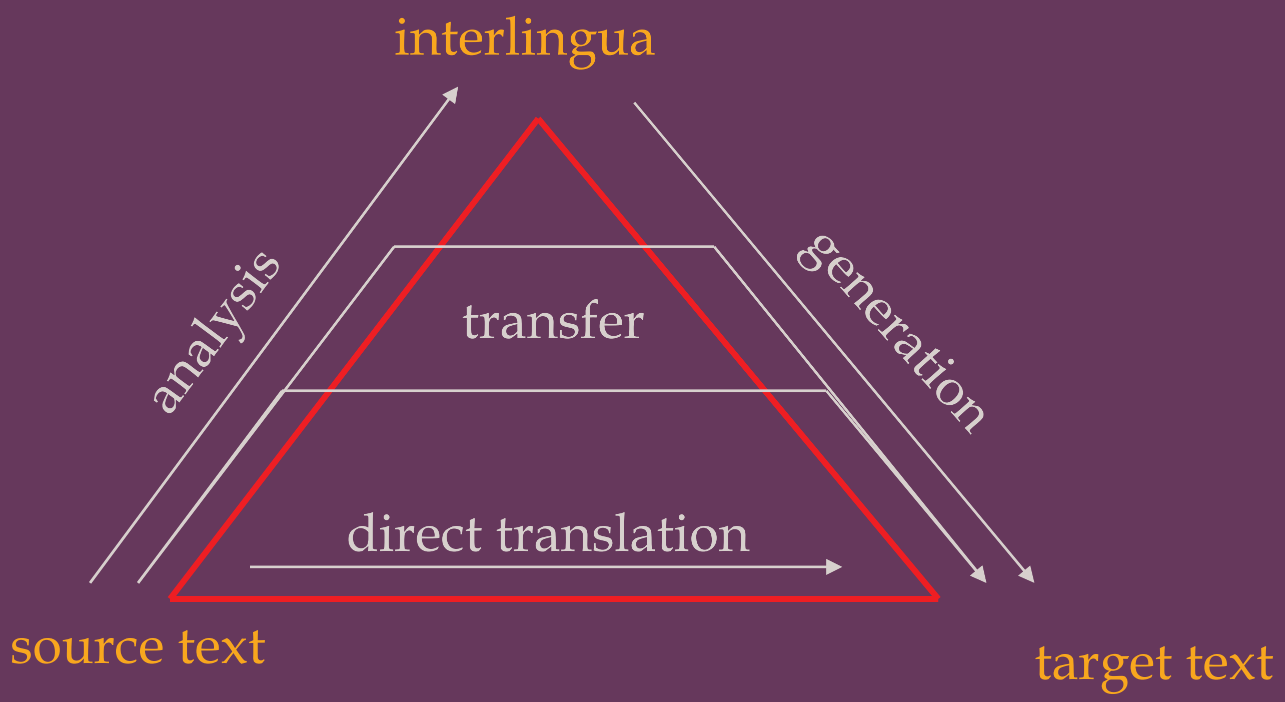
Declarative vs. procedural

- “what” vs. “how”
- e.g. (1) *casa* is feminine:
  - (analysis) with *la* and *nueva*, all “agree” so call it a noun-group
  - (generation) determines the form of article and adjective
- e.g. (2) how to form the plural:
  - (procedural) add -*s*
  - (declarative) *dog* ↔ *dogs*
  - (declarative) STEM = STEM+*s*

2.2 Rule-writing formalisms

- declarative knowledge expressed through linguistic formalisms
- either formalisms familiar to linguists (e.g. PS grammars)
- or custom-built formalisms
- sometimes effectively high-level programming languages

2.3 Depth of analysis



The “Vauquois triangle”

3. Modes of use

- Assimilation vs. dissemination
- Tools for translators
- Use of low-quality output
- Sublanguage and controlled language

- Fully automatic high-quality translation of unrestricted text is not possible, so …
- Not fully automatic = interactive
- Not high quality = can we use low quality?
- Not unrestricted = controlled language or sublanguage

3.1 MT for …

Assimilation	Dissemination
• many SLs, one TL	• one SL, many TLs
• any style	• controlled style
• any topic	• single topic
• partial analysis	• full analysis
• post-editing	• no post-editing
• user is reader	• user is author

3.2 Tools for translators (CAT)

- Humans and computers cooperate
- Which takes the initiative?
- MAHT: human translation using translation tools
- HAMT: MT with human assistance
- Translator’s Workstation may combine elements of all of these

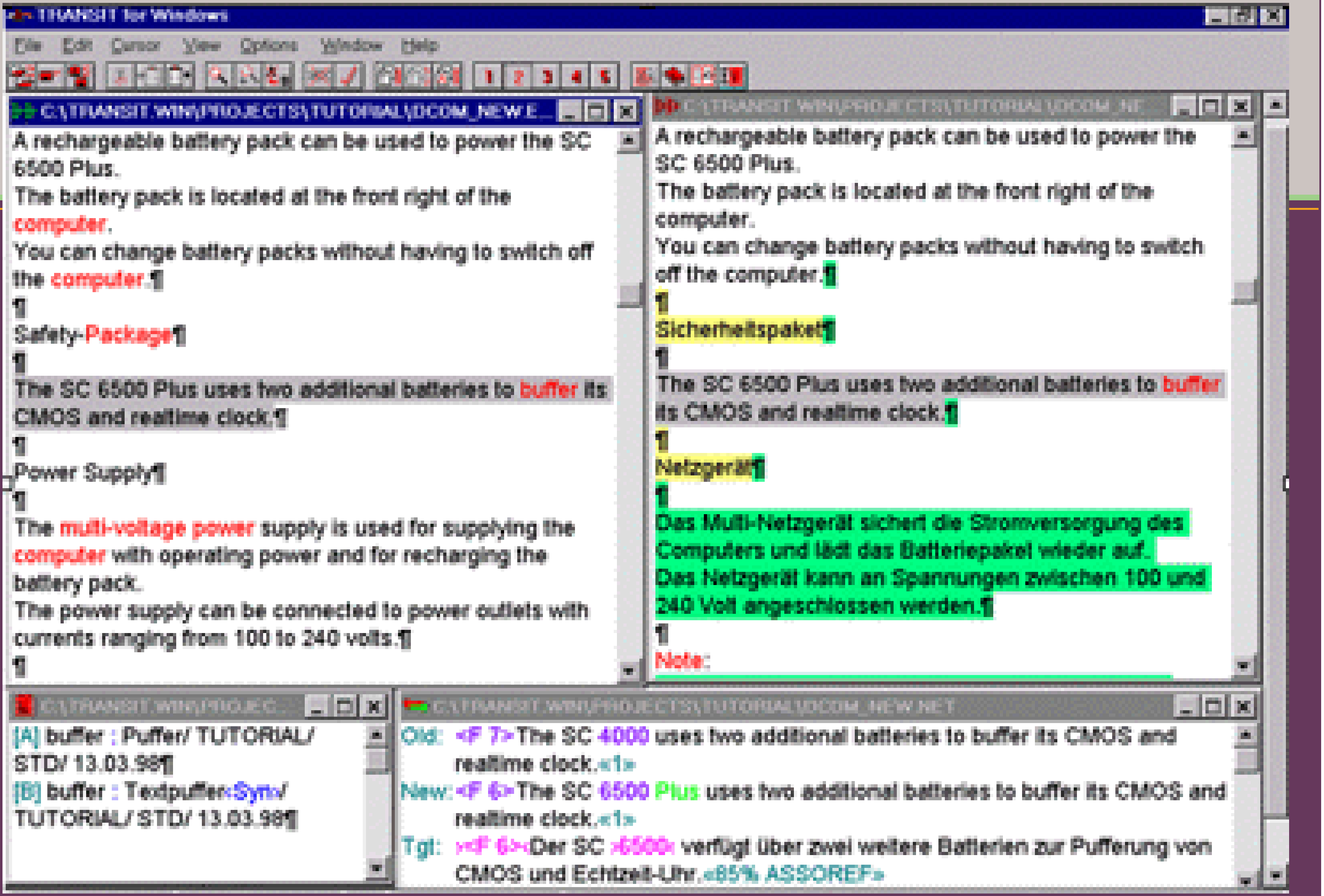
Basic word processing

- range of fonts
- hyphenation tools
- word count
- spelling checkers
- grammar/style checkers
- thesaurus (synonym dictionary)

- though only for certain languages

More sophisticated translators’ tools

- “pre-translation”:
  - automatic lemmatization
  - terminology look-up
  - rough translation (words only, no attempt at structure)
  - Translation Memory
- clever editing tools to make job easier?
  - “translation-oriented editing”


TRANSIT for Windows
File Edit Cursor View Options Window Help
1 2 3 4 5
C:\TRANSIT.WIN\PROJECTS\TUTORIAL\DCOM_NEW.E...
C:\TRANSIT.WIN\PROJECTS\TUTORIAL\DCOM_NE...
C:\TRANSIT.WIN\PROJEC...
C:\TRANSIT.WIN\PROJECTS\TUTORIAL\DCOM_NEW.NET

A rechargeable battery pack can be used to power the SC 6500 Plus.
The battery pack is located at the front right of the computer.
You can change battery packs without having to switch off the computer.¶
¶
Safety-Package¶
¶
The SC 6500 Plus uses two additional batteries to buffer its CMOS and realtime clock.¶
¶
Power Supply¶
¶
The multi-voltage power supply is used for supplying the computer with operating power and for recharging the battery pack.
The power supply can be connected to power outlets with currents ranging from 100 to 240 volts.¶
¶

A rechargeable battery pack can be used to power the SC 6500 Plus.
The battery pack is located at the front right of the computer.
You can change battery packs without having to switch off the computer.¶
¶
Sicherheitspaket¶
¶
The SC 6500 Plus uses two additional batteries to buffer its CMOS and realtime clock.¶
¶
Netzgerät¶
¶
Das Multi-Netzgerät sichert die Stromversorgung des Computers und lädt das Batteriepaket wieder auf.
Das Netzgerät kann an Spannungen zwischen 100 und 240 Volt angeschlossen werden.¶
¶
Note:

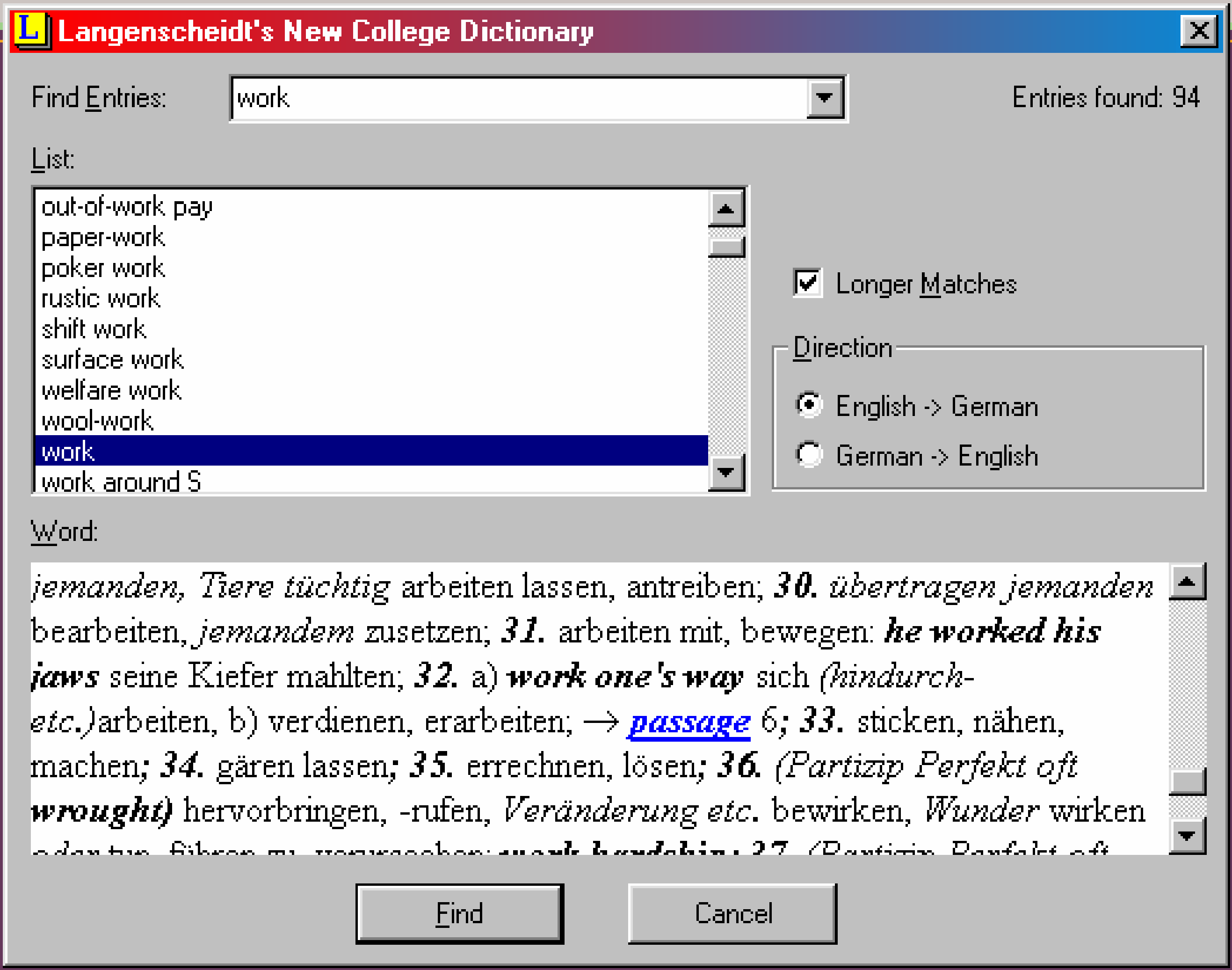
[A] buffer : Puffer/ TUTORIAL/ STD/ 13.03.98¶
[B] buffer : Textpuffer<Sym>/ TUTORIAL/ STD/ 13.03.98¶

Old: <F 7>The SC 4000 uses two additional batteries to buffer its CMOS and realtime clock.<1>
New: <F 6>The SC 6500 Plus uses two additional batteries to buffer its CMOS and realtime clock.<1>
Tgt: <F 6>Der SC 6500 verfügt über zwei weitere Batterien zur Pufferung von CMOS und Echtzeit-Uhr.<85% ASSOREF>

Lexical resources

- Monolingual dictionary
- Bi-, multilingual dictionary
- Thesaurus
- Terminology
- etc.

Machine-readable version of dictionary for human users



Langenscheidt's New College Dictionary

Find Entries: work

Entries found: 94

List:

- out-of-work pay
- paper-work
- poker work
- rustic work
- shift work
- surface work
- welfare work
- wool-work
- work
- work around S

☑ Longer Matches

Direction

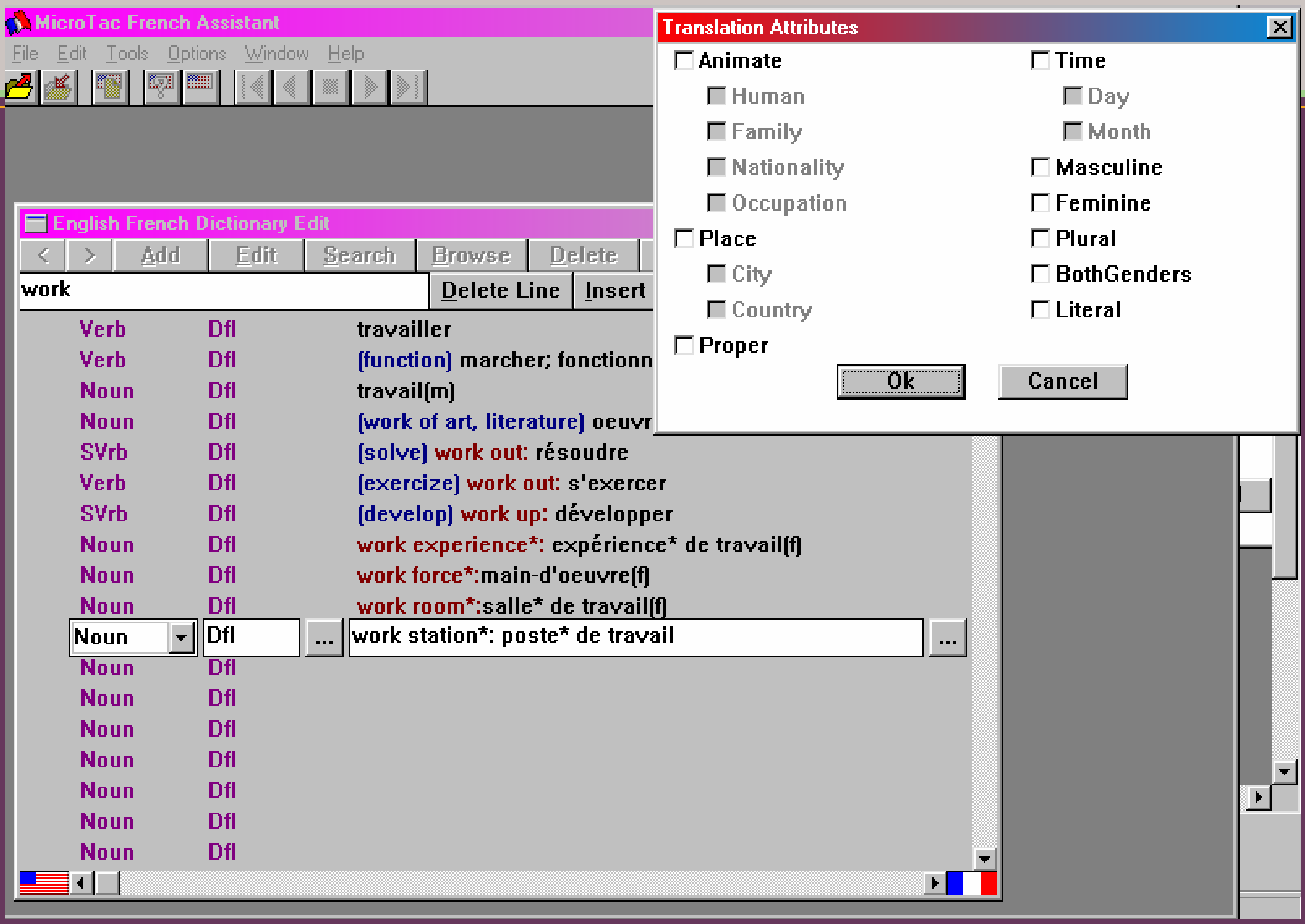
● English -> German

○ German -> English

Word:

jemanden, Tiere tüchtig arbeiten lassen, antreiben; ***30.*** *übertragen jemanden* bearbeiten, *jemandem* zusetzen; ***31.*** arbeiten mit, bewegen: ***he worked his jaws*** seine Kiefer mahlten; ***32.*** a) ***work one's way*** sich *(hindurch-etc.)*arbeiten, b) verdienen, erarbeiten; → ***passage*** 6; ***33.*** sticken, nähen, machen; ***34.*** gären lassen; ***35.*** errechnen, lösen; ***36.*** *(Partizip Perfekt oft* ***wrought)*** hervorbringen, -rufen, *Veränderung etc.* bewirken, *Wunder* wirken oder tun, führen zu, verursachen; ***work hardship***; ***37.*** *(Partizip Perfekt oft*

MT system’s dictionary


MicroTac French Assistant
File Edit Tools Options Window Help
English French Dictionary Edit
< > Add Edit Search Browse Delete
Delete Line Insert
Translation Attributes
Ok
Cancel

work

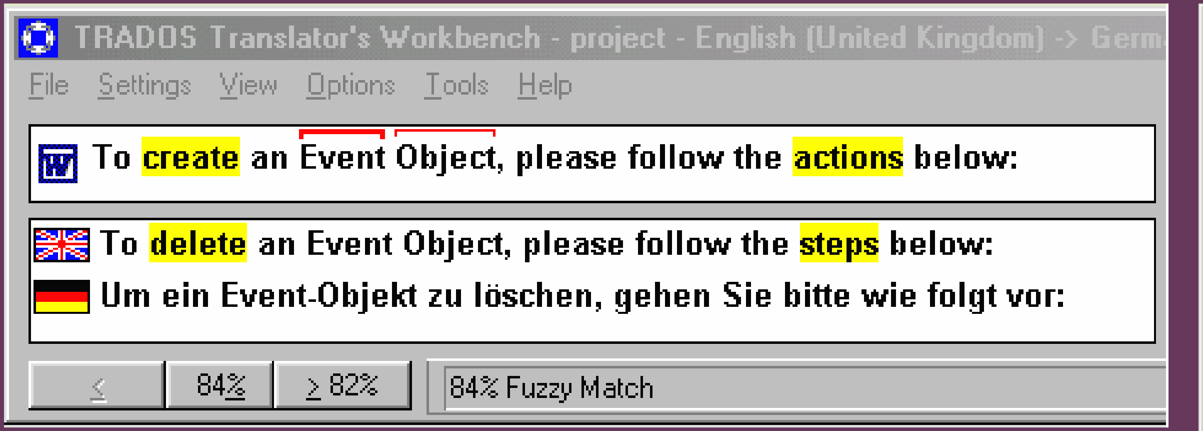
Verb	Dfl	travailler
Verb	Dfl	[function] marcher; fonctionn
Noun	Dfl	travail(m)
Noun	Dfl	[work of art, literature] oeuvr
SVrb	Dfl	[solve] work out: résoudre
Verb	Dfl	[exercize] work out: s'exercer
SVrb	Dfl	[develop] work up: développer
Noun	Dfl	work experience*: expérience* de travail(f)
Noun	Dfl	work force*:main-d'oeuvre(f)
Noun	Dfl	work room*:salle* de travail(f)
Noun	Dfl	work station*: poste* de travail
Noun	Dfl	
Noun	Dfl	
Noun	Dfl	
Noun	Dfl	
Noun	Dfl	
Noun	Dfl	
Noun	Dfl	

□ Animate
□ Human
□ Family
□ Nationality
□ Occupation
□ Place
□ City
□ Country
□ Proper

□ Time
□ Day
□ Month
□ Masculine
□ Feminine
□ Plural
□ BothGenders
□ Literal

Translation Memory

- Database of previous translations
- More or less sophisticated matching algorithm (“fuzzy match”, simple pattern-matching which may incorporate “linguistic “knowledge”)
- But *user* must decide what to do with them



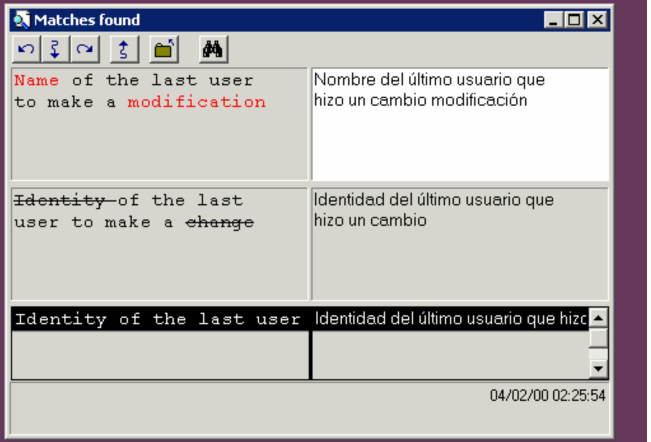
TRADOS Translator's Workbench - project - English (United Kingdom) -> Germ
File
Settings
View
Options
Tools
Help
<
84%
> 82%

To create an Event Object, please follow the actions below:

To delete an Event Object, please follow the steps below:

Um ein Event-Objekt zu löschen, gehen Sie bitte wie folgt vor:

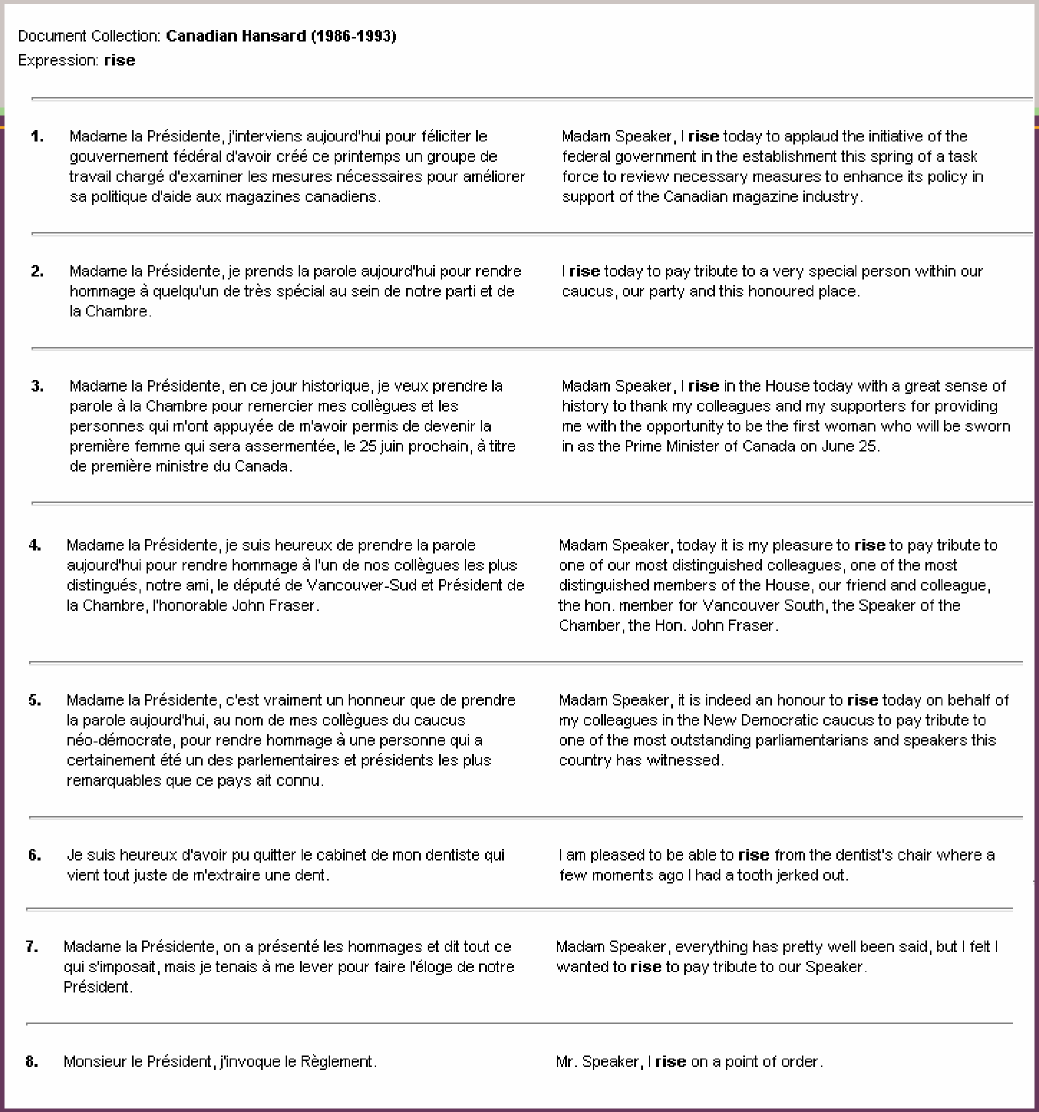
84% Fuzzy Match



Name of the last user to make a modification	Nombre del último usuario que hizo un cambio modificación
~~Identity~~ of the last user to make a ~~change~~	Identidad del último usuario que hizo un cambio
Identity of the last user	Identidad del último usuario que hizc

04/02/00 02:25:54

Bilingual concordance



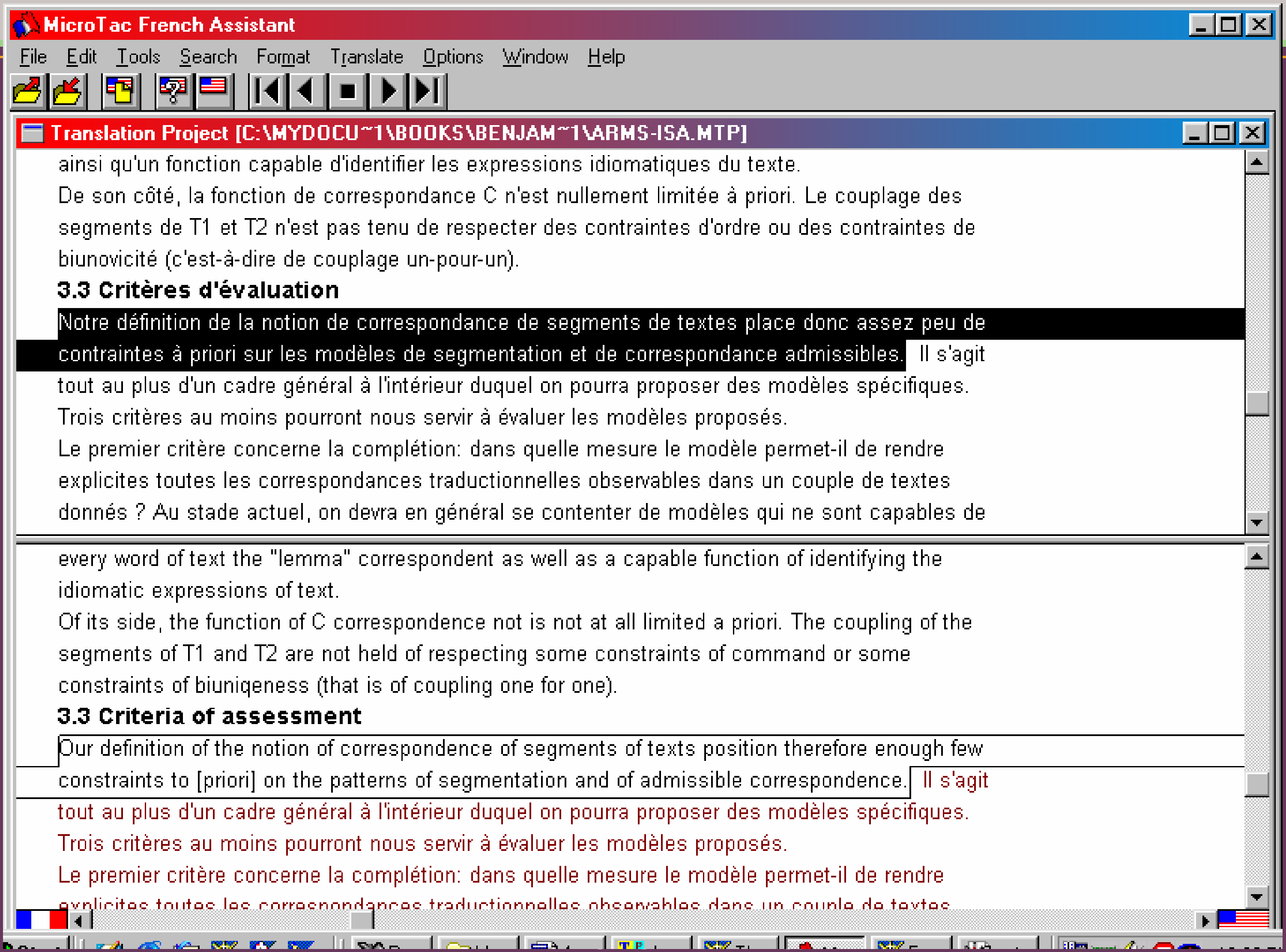
Document Collection: **Canadian Hansard (1986-1993)**
Expression: **rise**

1.	Madame la Présidente, j'interviens aujourd'hui pour féliciter le gouvernement fédéral d'avoir créé ce printemps un groupe de travail chargé d'examiner les mesures nécessaires pour améliorer sa politique d'aide aux magazines canadiens.	Madam Speaker, I **rise** today to applaud the initiative of the federal government in the establishment this spring of a task force to review necessary measures to enhance its policy in support of the Canadian magazine industry.
2.	Madame la Présidente, je prends la parole aujourd'hui pour rendre hommage à quelqu'un de très spécial au sein de notre parti et de la Chambre.	I **rise** today to pay tribute to a very special person within our caucus, our party and this honoured place.
3.	Madame la Présidente, en ce jour historique, je veux prendre la parole à la Chambre pour remercier mes collègues et les personnes qui m'ont appuyée de m'avoir permis de devenir la première femme qui sera assermentée, le 25 juin prochain, à titre de première ministre du Canada.	Madam Speaker, I **rise** in the House today with a great sense of history to thank my colleagues and my supporters for providing me with the opportunity to be the first woman who will be sworn in as the Prime Minister of Canada on June 25.
4.	Madame la Présidente, je suis heureux de prendre la parole aujourd'hui pour rendre hommage à l'un de nos collègues les plus distingués, notre ami, le député de Vancouver-Sud et Président de la Chambre, l'honorable John Fraser.	Madam Speaker, today it is my pleasure to **rise** to pay tribute to one of our most distinguished colleagues, one of the most distinguished members of the House, our friend and colleague, the hon. member for Vancouver South, the Speaker of the Chamber, the Hon. John Fraser.
5.	Madame la Présidente, c'est vraiment un honneur que de prendre la parole aujourd'hui, au nom de mes collègues du caucus néo-démocrate, pour rendre hommage à une personne qui a certainement été un des parlementaires et présidents les plus remarquables que ce pays ait connu.	Madam Speaker, it is indeed an honour to **rise** today on behalf of my colleagues in the New Democratic caucus to pay tribute to one of the most outstanding parliamentarians and speakers this country has witnessed.
6.	Je suis heureux d'avoir pu quitter le cabinet de mon dentiste qui vient tout juste de m'extraire une dent.	I am pleased to be able to **rise** from the dentist's chair where a few moments ago I had a tooth jerked out.
7.	Madame la Présidente, on a présenté les hommages et dit tout ce qui s'imposait, mais je tenais à me lever pour faire l'éloge de notre Président.	Madam Speaker, everything has pretty well been said, but I felt I wanted to **rise** to pay tribute to our Speaker.
8.	Monsieur le Président, j'invoque le Règlement.	Mr. Speaker, I **rise** on a point of order.

Source: *TransSearch*,
Laboratoire de Recherche Appliquée en Linguistique Informatique,
Université de Montréal

http://www-rali.iro.umontreal.ca

Parallel scrolling screens


MicroTac French Assistant
File Edit Tools Search Format Translate Options Window Help
Translation Project [C:\MYDOCU~1\BOOKS\BENJAM~1\ARMS-ISA.MTP]

ainsi qu'un fonction capable d'identifier les expressions idiomatiques du texte.

De son côté, la fonction de correspondance C n'est nullement limitée à priori. Le couplage des segments de T1 et T2 n'est pas tenu de respecter des contraintes d'ordre ou des contraintes de biunovicité (c'est-à-dire de couplage un-pour-un).

3.3 Critères d'évaluation

Notre définition de la notion de correspondance de segments de textes place donc assez peu de contraintes à priori sur les modèles de segmentation et de correspondance admissibles. Il s'agit tout au plus d'un cadre général à l'intérieur duquel on pourra proposer des modèles spécifiques.

Trois critères au moins pourront nous servir à évaluer les modèles proposés.

Le premier critère concerne la complétion: dans quelle mesure le modèle permet-il de rendre explicites toutes les correspondances traductionnelles observables dans un couple de textes donnés ? Au stade actuel, on devra en général se contenter de modèles qui ne sont capables de

every word of text the "lemma" correspondent as well as a capable function of identifying the idiomatic expressions of text.

Of its side, the function of C correspondence not is not at all limited a priori. The coupling of the segments of T1 and T2 are not held of respecting some constraints of command or some constraints of biuniqeness (that is of coupling one for one).

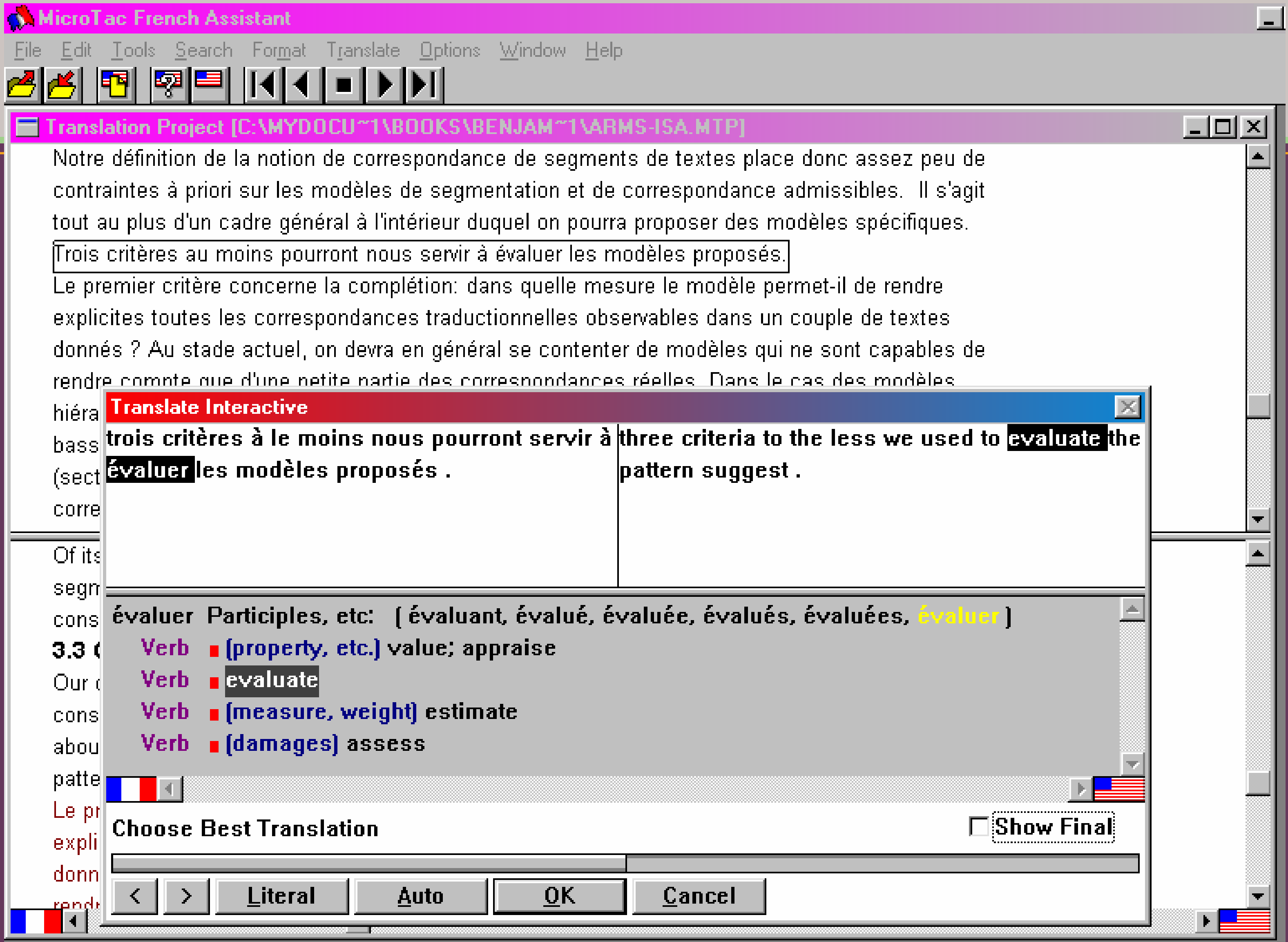
3.3 Criteria of assessment

Our definition of the notion of correspondence of segments of texts position therefore enough few constraints to [priori] on the patterns of segmentation and of admissible correspondence. Il s'agit tout au plus d'un cadre général à l'intérieur duquel on pourra proposer des modèles spécifiques.

Trois critères au moins pourront nous servir à évaluer les modèles proposés.

Le premier critère concerne la complétion: dans quelle mesure le modèle permet-il de rendre explicites toutes les correspondances traductionnelles observables dans un couple de textes

Interactive translation



MicroTac French Assistant
File Edit Tools Search Format Translate Options Window Help
Translation Project [C:\MYDOCU~1\BOOKS\BENJAM~1\ARMS-ISA.MTP]

Notre définition de la notion de correspondance de segments de textes place donc assez peu de contraintes à priori sur les modèles de segmentation et de correspondance admissibles. Il s'agit tout au plus d'un cadre général à l'intérieur duquel on pourra proposer des modèles spécifiques. Trois critères au moins pourront nous servir à évaluer les modèles proposés. Le premier critère concerne la complétion: dans quelle mesure le modèle permet-il de rendre explicites toutes les correspondances traductionnelles observables dans un couple de textes donnés ? Au stade actuel, on devra en général se contenter de modèles qui ne sont capables de rendre compte que d'une petite partie des correspondances réelles. Dans le cas des modèles

hiéra
bass
(sect
corre

Of its
segm
cons
3.3 (
Our c
cons
abou
patte
Le pr
expli
donn
rend

trois critères à le moins nous pourront servir à évaluer les modèles proposés .

three criteria to the less we used to evaluate the pattern suggest .

évaluer Participles, etc: (évaluant, évalué, évaluée, évalués, évaluées, évaluer)

Verb	■	(property, etc.) value; appraise
Verb	■	evaluate
Verb	■	(measure, weight) estimate
Verb	■	(damages) assess

Choose Best Translation

3.3 Use of low-quality output

- To get a rough idea of content, and to identify which parts need to be translated "properly"
- … especially with "exotic" languages
- Widely used on the Internet for browsing, chat-rooms and email:
- Despite low quality, users seem satisfied

MT on the web

- Despite limitations, it is now the widest use of MT
- Pioneered by CompuServe in 1992, now AltaVista’s use of Systran in babelfish is most well known
- Users at first “amazed”, then disappointed, then pragmatic
- Task is especially difficult due to odd grammar, spelling, punctuation (GIGO), and wide variety of subject matter, often mixed
- Some MT products now customized for web-page translation, e.g. take HTML mark-up into account

3.4 Sublanguage and controlled language

- Restrictions may be natural or imposed
- Related terms: special language, jargon, register, LSP
- For human: (usually) more readable, less ambiguous, more “focussed”
- For MT:
- – fewer syntactic constructions
- – closed vocabulary with fewer homonyms
- – greater certainty about interpretation

Features of sublanguage

- Lexicon
  - smaller size: less concepts to cover
  - finite/closed: innovation is controlled
  - nature: less homonymy, some synonyms (dis)favoured
  - grammatical use: fewer category ambiguities
- Syntax
  - reduced range of structures
  - some structures (dis)favoured
  - less flexibility in choice of structure
  - some deviance from “standard” grammar

Controlled languages

- Widely used in technical authoring
- Similar features to sublanguage
- Can be coupled with grammar checker
- Permits multilingual authoring