UNIVERSITÀ DEGLI STUDI DI BOLOGNA

FACOLTÀ DI INGEGNERIA

Corso di Laurea in Ingegneria Elettronica

Sistemi Informativi I

ANALISI E CONFRONTO DI SISTEMI EBMT PER LA TRADUZIONE ASSISTITA

Tesi di Laurea di:

BRUNO UMBERTO SPAGNA

Relatore:

Chiar.mo Prof. **PAOLO CIACCIA**

Correlatori:

Dott.ssa **FEDERICA MANDREOLI**

Dott. Ing. **RICCARDO MARTOGLIA**

Anno Accademico 2001-2002

Sessione Autunnale

Parole chiave:

EBMT - Example Based Machine Translation
CAT - Computer Aided Translation
Traduzione assistita
Trados
IBM TranslationManager

Ai miei genitori

Sommario

Indice figure

Introduzione

La traduzione è stata per lungo tempo considerata come quell'attività, eseguita nel modo migliore, di trasporto di un testo da una lingua in un'altra. Tradurre nel modo migliore, oggi, significa non solo conoscere le parole, le regole grammaticali e le espressioni tipiche di un paese e del suo idioma, ma anche produrre un documento che sia identico per aspetto all'originale, ne rispetti cioè le caratteristiche formali, e soprattutto sia caratterizzato da un'uniformità terminologica e di stile.

Per sopperire a tale necessità, in relazione anche alle sempre più elevate esigenze del mercato mondiale (quantità maggiori in minore tempo e alla medesima elevata qualità), si è giunti negli ultimi decenni a ricercare tecniche di *automazione* del processo stesso di traduzione. Riuscire ad automatizzare il processo per dare maggiore qualità e quantità in minor tempo è lo scopo che si sono prefissi i più recenti studi sull'uso del calcolatore in appoggio alla traduzione. Tali studi hanno portato alla realizzazione di *sistemi di traduzione assistita*

denominati CAT, Computer Aided Translation, dei quali i sistemi EBMT, Example Based Machine Translation, sono uno dei più promettenti sviluppi di tale ricerca.

Un sistema EBMT traduce per analogia, utilizzando traduzioni precedentemente realizzate per fornire nuovi documenti pretradotti. Il suo funzionamento quindi è riconducibile ad un processo di creazione di una memoria di traduzione, *Translation Memory*, a cui sottoporre, successivamente, un nuovo testo da tradurre. Il risultato che si ottiene è un documento *pretradotto*. Completata poi la traduzione è possibile arricchire la memoria con le nuove parti tradotte in un ciclo continuo di accrescimento della memoria stessa.

I più recenti sviluppi in ambito di Information Retrieval hanno permesso di giungere a tale risultato, di avere cioè sistemi che si basano su database corredati da algoritmi per la determinazione di similarità tra frasi..

Modello a tre livelli

La necessità di tradurre testi in quantità e con qualità sempre maggiori è divenuta oggi tale da richiedere l'utilizzo di sistemi basati su calcolatore come supporto al processo di traduzione vero e proprio. Se in un primo tempo era infatti sufficiente disporre di una schiera di traduttori professionisti, a cui affiancare revisori linguistici, cioè traduttori con elevata esperienza in grado di verificare e quindi validare la traduzione fatta da altri, oggi ciò non è più vero; esigenze come la corretta scelta terminologica, il rispetto delle scelte stilistiche/linguistiche del documento originale, l'esatta corrispondenza al formato dei documenti originali hanno portato alla necessità di introdurre un *nuovo livello organizzativo* che si inserisce tra cliente e traduttore ed agisce sia sul testo originale da tradurre (testo *source* fornito dal cliente) che sul testo tradotto (testo *target* ottenuto dal traduttore). Questo nuovo *modello a tre livelli* nasce nella seconda metà degli anni '70 e vede l'introduzione di una struttura organizzata, intermedia nel processo di traduzione, il cui compito è quello di sovrintendere al lavoro di traduzione, fornendo, a supporto della

traduzione stessa, tutti gli strumenti tecnologici disponibili e le conoscenze specifiche in ambito informatico-linguistico.

Questa struttura è oggi rappresentata dalla cosiddetta *società di servizi linguistici* che, raggruppando conoscenze informatiche, linguistiche e tecnologiche, può affiancare nel proprio compito il traduttore. Essa è costituita da tutte quelle attività collaterali al lavoro di traduzione, quali la corretta ricerca della terminologia da fornire al professionista prima dell'inizio dell'attività di traduzione (glossari e testi di riferimento), l'analisi delle problematiche tecnico-informatiche dei documenti (il problema dei formati dei file in continua evoluzione e sempre più diversificati), il mantenimento dell'uniformità terminologica all'interno di un medesimo testo, oltre all'utilizzo di programmi di supporto alla traduzione.

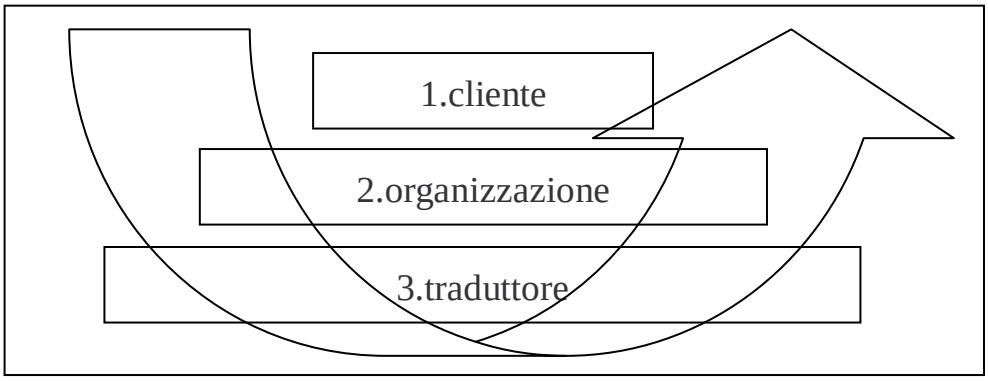


Figura 1 – Struttura a tre livelli

La traduzione diviene quindi un'attività non più individuale ma un processo collettivo di un'organizzazione specializzata.

Attività di traduzione: organizzazione specializzata

L'attività di traduzione di oggi può essere descritta tenendo conto dei vari aspetti che la caratterizzano.

Se si caratterizza semplicemente la traduzione come il processo di traduzione di un testo, si individuano parametri come la *quantità di testo da tradurre*, il *tipo di testo* (contesto), le *lingue coinvolte*, la

qualità della traduzione (correttezza grammaticale, sintattica e terminologica).

Se invece ci si focalizza sull'organizzazione del lavoro di traduzione (secondo livello) si ottengono nuovi parametri come il *tipo di lavoro* (analisi formato, contesto, ricerca terminologica, uniformità terminologica), la *dimensione del lavoro* (analisi dei volumi) e la *programmazione delle risorse* (traduttori coinvolti nel progetto).

Il modello a tre livelli è l'unico in grado ad oggi di garantire tempi di traduzione e costi da sostenere, sufficientemente bassi e in linea con le richieste di qualità del mercato. I volumi attuali di traduzione (normalmente calcolati in righe o cartelle normalizzate) e i relativi costi sarebbero infatti inaccettabili con un approccio che potremmo chiamare *lineare* (un certo numero righe di traduzione è affidato a un certo numero di professionisti entro un complessivo numero di giorni). L'unica soluzione attualmente individuabile è quella in cui un'organizzazione si premura di analizzare un testo da tradurre fornendo al traduttore tutte quelle informazioni e mezzi con cui portare a termine il proprio lavoro nel migliore modo e nel minor tempo possibile.

La traduzione assistita dal calcolatore è quindi approdata nelle società di servizi linguistici come supporto al lavoro di organizzazione e miglioramento della qualità della traduzione.

Tali strumenti permettono infatti sia l'analisi del materiale fornito da tradurre al fine di determinare in breve i tempi e i costi (*fase di preventivazione* del lavoro), che la creazione di memorie di traduzione a partire da materiale inerente l'argomento della traduzione e precedentemente tradotto, e infine la pretraduzione del nuovo materiale sfruttando tali memorie. Altri aspetti della traduzione assistita come la creazione di glossari specifici e l’aggiornamento delle memorie sono ulteriori compiti della struttura organizzativa.

Scopo di questa tesi è quello di documentare la realtà degli strumenti di traduzione assistita EBMT attualmente in commercio e diffusi nelle società di servizi linguistici, le loro caratteristiche e prestazioni,

individuando eventualmente quale tipo di approccio tecnologico-informatico sia alla base di essi.

In particolare verranno analizzati i programmi Trados 5.0.1 e IBM TranslationManager 2.0.6. Il primo è tra i principali software in uso nel mercato della traduzione commerciale, mentre il secondo è stato tra i primi a essere sviluppato partendo da alcuni studi della divisione tedesca del noto produttore in collaborazione con il partner IBM Synthema [19].

La maggior parte del materiale analizzato e sintetizzato nella presente tesi proviene da ricerche eseguite direttamente con i programmi in questione presso la società Intradoc S.r.l. e da ricerche condotte in collaborazione con la Dott.ssa Federica Mandreoli e il Dott. Ing. Riccardo Martoglia dell'Università di Modena e Reggio Emilia.

Per quanto riguarda la struttura della tesi, essa è organizzata nel seguente modo:

Nel **Capitolo 1** vengono descritte quelle che devono essere le caratteristiche peculiari di un buon sistema CAT. Nel **Capitolo 2** sono analizzati i componenti dei due prodotti presi in esame. Nel **Capitolo 3** vengono descritte le principali nozioni di Information Retrieval, mentre nel **Capitolo 4** vengono condotte le esperienze sperimentali al fine di valutare il funzionamento dei due prodotti. Quindi nel **Capitolo 5** si giunge a illustrare i risultati di tali prove in relazione all'efficacia e all'efficienza degli stessi. Nel **Capitolo 6** si evidenziano i limiti di tali sistemi, arrivando nel **Capitolo 7** a fornire una panoramica sugli sviluppi futuri della traduzione assistita.

Capitolo 1 Computer Aided Translation

Scopo della traduzione assistita da calcolatore (Computer Aided Translation - CAT) nel mondo aziendale odierno è quello della pretraduzione automatica al fine di ridurre i costi e i tempi della traduzione stessa mantenendo un elevato standard qualitativo sia dal punto di vista linguistico che di formato. Oltre a questo, lo scopo è anche quello di fornire al traduttore testi pretradotti corredati da ulteriori elementi aggiuntivi utili al completamento del proprio compito.

Il modello a tre livelli precedentemente descritto è l'unico ad oggi a fornire una soluzione efficace a questa richiesta, in quanto da un lato fornisce al cliente garanzia di organizzazione del lavoro, tempi certi di realizzazione a costi preventivati con qualità elevata, e dall'altro è l'unico in grado di consentire un'approfondita ricerca di conoscenze specifiche e in particolare informatiche che solitamente non sono alla portata e competenza di un singolo traduttore.

Un ulteriore motivo all'utilizzo della traduzione assistita è anche quello di generare glossari bilingue a partire da testi completi allo scopo di disporre di un riferimento terminologico preciso da utilizzare anche in occasioni non strettamente legate alla traduzione.

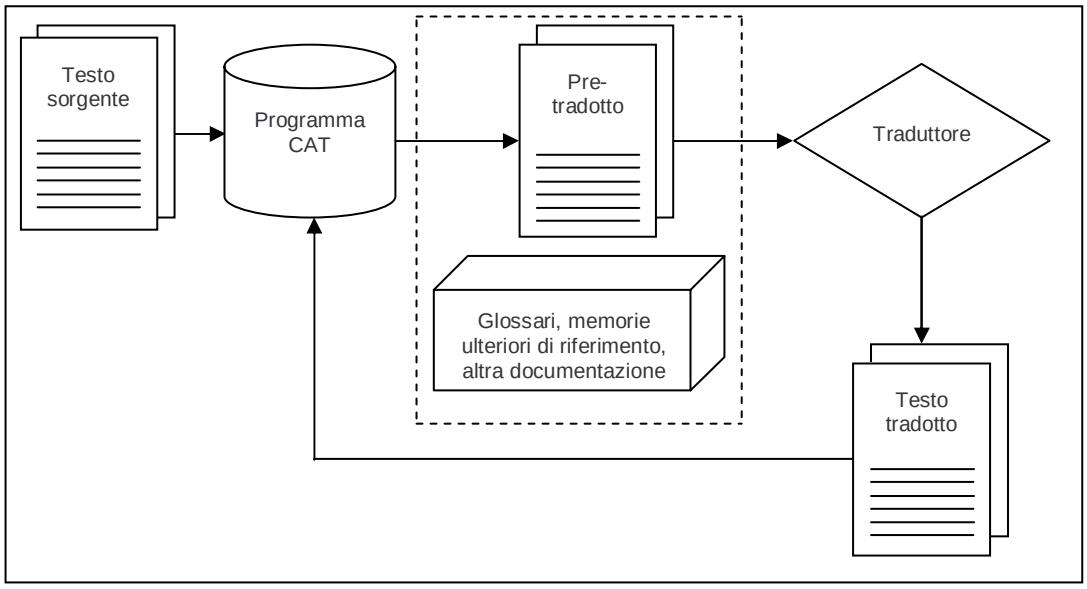


Figura 1.1 - Computer Aided Translation - Programma di traduzione assistita da calcolatore

1.1 Componenti di base e strumenti aggiuntivi

I programmi di traduzione assistita sono normalmente composti da strumenti atti all'analisi e alla pretraduzione di testi da tradurre da una lingua sorgente ad una di destinazione. Questi strumenti sono normalmente racchiusi dentro un'interfaccia grafica principale del programma da cui è possibile impostare un *progetto di traduzione* e accedere ai principali *strumenti* necessari alla gestione dello stesso. L'interfaccia principale del programma consente quindi specificatamente di indicare la posizione dei file da tradurre, la posizione della memoria da utilizzare, le lingue coinvolte nel progetto, lo stato di avanzamento del lavoro, il materiale di riferimento inerente e altre informazioni.

Oltre all'interfaccia principale sono solitamente forniti anche altri componenti utili alla realizzazione di un progetto di traduzione tra cui si possono individuare ad esempio: dizionari multilingua, dizionari dei

sinonimi, tool di conversione tra formati, programmi per la creazione di memorie a partire da testi già tradotti in precedenza.

Oggetto principale di un programma di traduzione assistita è la cosiddetta *Translation Memory* (TM) attorno alla quale ruotano tutti i componenti del programma.

Translation Memory

Il principale componente di un programma di traduzione assistita (CAT) è la cosiddetta *Translation Memory* (TM) o banca dati in cui sono memorizzate le associazioni tra un segmento (o frase) in una lingua e la sua corrispondente in un'altra lingua. Ogni associazione rappresenta un record della banca dati comporto da più campi tra cui anche data di creazione del record, autore, note, ecc.

Tale database costituisce il nucleo attorno a cui lavora il programma vero e proprio che avrà quindi il compito di analizzare un testo sottoposto a traduzione ricercando eventuali somiglianze tra una frase nuova e quelle presenti nel database stesso. Tale approccio alla traduzione assistita basato su esempi viene denominato EBMT - Example Based Machine Translation – intendendo con ciò un sistema per la realizzazione di una traduzione basandosi su precedenti traduzioni memorizzate in una struttura organizzata [2] [16].

Una più precisa definizione è stata fornita dall'Expert Advisory Group on Language Engineering Standards (EAGLES) [6]: *"a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions"* (un archivio testuale multilingua di testi segmentati, allineati, scomposti e classificati che consenta l'archiviazione e la ricerca di segmenti di testo multilingua da tradurre tramite un'interrogazione parametrizzabile.)

Tool di pretraduzione automatica (batch processing)

Il tool di pretraduzione automatica è quello strumento in grado di generare un file bilingue sostituendo al testo originale quei segmenti in lingua trovati in memoria. Il file così ottenuto è poi successivamente sottoposto a traduzione per completarne il risultato.

La caratteristica di questo strumento di operare su testi anche senza conoscenze linguistiche specifiche lo rendono utilizzabile da qualsiasi operatore.

Questo strumento è calibrabile in base ad alcuni parametri come ad esempio il livello di matching con cui cercare nella Translation Memory, il grado di segmentazione del testo, le conversioni automatiche di date, valute e formati numerici, ottenendo così la possibilità di differenti risultati personalizzabili.

Tool di suggerimento

Il tool di suggerimento è quello strumento di interfaccia utente tramite il quale il traduttore traducendo un testo riceve parallelamente alla traduzione individuata nella TM anche alcuni input aggiuntivi (suggerimenti) che sono contestuali, ad esempio suggerimenti di termini attinti da glossari settoriali.

Dizionari Multilingua

Un dizionario multilingua è composto tipicamente da uno o più dizionari bilingua e da un applicativo di ricerca. Questi applicativi sono normalmente concepiti per lavorare assieme ad un programma di videoscrittura e qualche volta permettono anche la gestione delle inflessioni terminologiche (permettono cioè di non dover introdurre la radice di un termine per cercarne la traduzione ma automaticamente cercano e suggeriscono i termini più vicini). Questi applicativi vengono forniti con dizionari settoriali preimpostati, ma modificabili, e sono caratterizzati dalla possibilità di importazione/esportazione dei termini stessi.

Si tratta quindi di uno strumento utile sia per il traduttore in fase di traduzione sia per l'organizzazione che deve fornire al traduttore in fase di inizio lavoro un dizionario contestuale al documento da tradurre.

In taluni casi se la struttura organizzativa e il traduttore utilizzano il medesimo sistema di traduzione assistita, è possibile interscambiare dizionari in un formato originale senza necessità di esportarli al fine di mantenere un'uniformità di strumentazione adoperata e consentendo di arricchire direttamente il dizionario originale.

Dizionari dei Sinonimi

I dizionari dei sinonimi solitamente consistono di due o più dizionari collegati (cross-reference multilingual thesauri) che permettono ricerche di sinonimi (e/o contrari) tramite ricerche di concetto e non solamente di tipo alfabetico. Il fruitore di tale dizionario (corredato da applicativo di ricerca) può quindi scorrere e ricercare la terminologia attraverso più lingue individuando le sotto-aree di pertinenza. Non sono infrequenti informazioni anche di tipo grammaticale, genere, inflessioni e altro.

I traduttori professionisti potrebbero altresì trovare questi dizionari troppo limitati e quindi è prevista la possibilità di modifica, cancellazione, introduzione dei termini stessi. Questi applicativi inoltre forniscono la possibilità di riordinare i dati stessi in base a criteri che vanno al di là della semplice relazione di sinonimia, ad esempio: per associazioni, per contrari (antinomia), per sottoparti, per declinazione. A questo proposito si veda §2.1.1.5 "*MultiTerm*" a pagina 7.

Tool di conversione formati

Oggi i documenti da tradurre sono nei formati più disparati (Quark Xpress, Adobe PageMaker, Adobe FrameMaker, Word XP, Word 2000, Word 97, ecc.) e devono essere tradotti da professionisti che non sempre dispongono di tali software; sono quindi necessari strumenti che consentano la trasformazione di un documento da un

formato ad un altro e viceversa al fine di garantire un processo di traduzione semplificato e il rispetto/mantenimento dell'aspetto del documento originale.

I tool di conversione permettono quindi di convertire in formati di interscambio (Interchange Format File) come l'RTF (Rich Text File) o il MIF (Meta Interchange Format) i documenti originali e di mantenere con opportuni codici (tags) le formattazioni del testo stesso (grassetti, dimensioni caratteri, posizione del testo e delle figure nella pagina) al fine di permettere una corretta contro-conversione nel formato originale.

Solitamente questo tool è di ausilio a chi si occupa di organizzare il lavoro di traduzione e permetterà al traduttore di manipolare documenti senza preoccuparsi del formato originale e senza necessità di disporre di applicativi dedicati. Inoltre il traduttore potrà continuare ad utilizzare il programma di videoscrittura a cui è maggiormente abituato senza dover apprendere nuovi strumenti.

Tale processo di conversione è quindi una normalizzazione del testo che lo rende di fatto indipendente dal formato originale. Tale strumento permette inoltre un mantenimento dell'investimento economico effettuato per l'acquisto del programma, grazie alla possibilità di gestire nuovi formati tramite moduli aggiuntivi *plug-in* da implementare nel programma stesso.

Nella figura seguente si vede un testo esportato dal formato FrameMaker in formato RTF. Esso è corredato da codice di controllo di colore grigio e rosso contenente quelle informazioni aggiuntive non propriamente attinenti il lavoro vero e proprio del traduttore ma necessarie per la sua corretta contro-conversione nel formato originale.

```
<ps·"OperationText0"·10>[<:fc·1>Audio·Tuning·Wizard<:/fc>]画面が表示されます。¶
<ps·"OperationText4"·11><pn·"□<:t>"·2>他の言語の<:fc·
1>Windows<:/fc>を使用している人と通話する場合は、アルファベットまたは数字で設定することをおす
すめします。¶
<ps·"BodyText"·3>¶
<ps·"Operation1"·12><pn·"2<:t>"·
3>デジタルビデオカメラを接続している場合、WAVEデバイスの選択時に、[<:fc·
1>Recording<:/fc>]デバイスを[Panasonic·DV·VideoCamera]にする¶
<ps·"OperationText0"·10>[Panasonic·DV·VideoCamera]を選択できない場合は、<:cs·
"refer"·3><:xr·"5"·1><:/cs>ページの[<:fc·1>Sounds·and·Multimedia·
Properties<:/fc>]画面で設定してください。¶
<ps·"BodyText"·3>¶
<ps·"Operation1"·12><pn·"3<:t>"·4>この後はなにも設定せず、[<:fc·
1>Next<:/fc>]をクリックして先へ進み、設定を完了させる¶
<ps·"BodyText"·3>¶
<ps·"BodyText"·3>¶
<ps·"Operation·T-T"·13><pn·"1<:t>"·
1>設定の詳細についてはそれぞれのソフトのヘルプファイルをお読みください。¶
```

Figura 1.2 - Esempio di file in formato FrameMaker trasformato in formato RTF (traduzione manualistica Panasonic dal giapponese)

Tool di creazione di nuove memorie

Nell'ottica di recuperare materiale precedentemente tradotto viene fornito uno strumento per l'allineamento di testi sorgenti e testi tradotti allo scopo di creare nuove memorie. Queste memorie saranno alla base della pretraduzione di nuovi testi. E' pertanto di fondamentale importanza che questo strumento sia in grado di realizzare velocemente ed efficacemente tali nuove memorie.

Questo strumento, normalmente dotato di interfaccia grafica complessa, si compone di un primo modulo in grado di scomporre in segmenti il testo sorgente e quello tradotto per poi, in una seconda fase, consentire all'utente di verificare, ed eventualmente correggere, l'allineamento segmento-segmento suggerito.

Al termine di questa operazione di allineamento, viene creata una memoria catalogabile secondo vari parametri come il nome del creatore, la data di creazione, il materiale su cui si basa e le lingue presenti (vedi *Figura 1.3*).

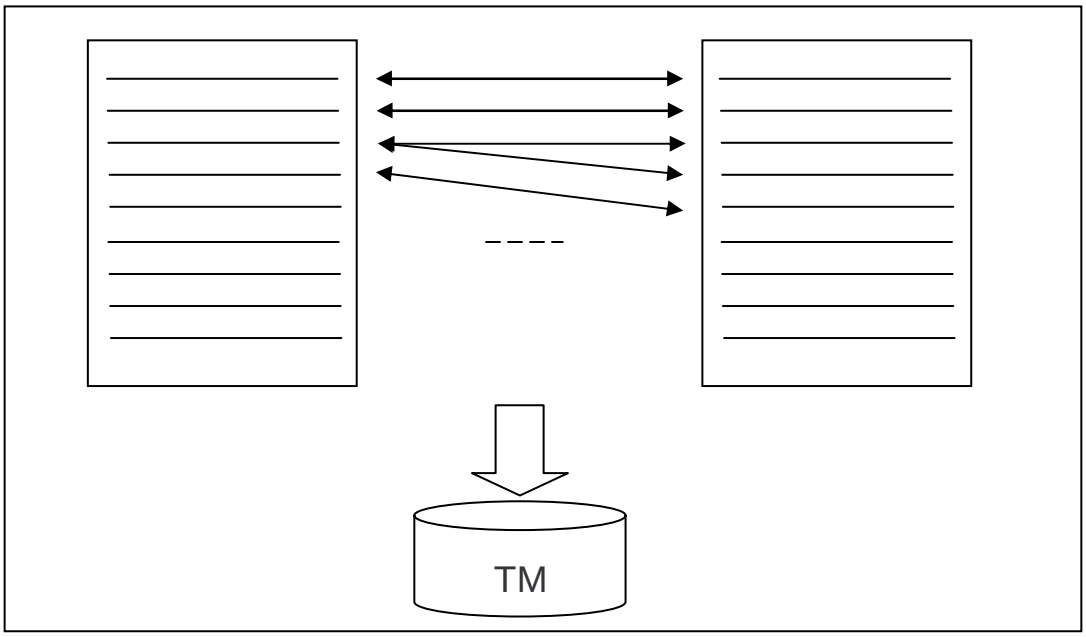


Figura 1.3 - Computer Aided Translation - Allineamento segmenti e creazione di una nuova memoria

1.2 Caratteristiche ideali di un sistema di traduzione assistita

Al fine di poter analizzare un sistema di traduzione assistita presente sul mercato è necessario compiere prima una analisi dei requisiti che un applicativo del genere dovrebbe possedere. Sulla base di queste caratteristiche sarà possibile comprendere meglio le caratteristiche dei prodotti in commercio e i parametri con cui valutarne l'efficacia.

Ricordando che una sistema basato su Translation Memory è una sorta di archivio multilingua contenente del testo segmentato, allineato e classificato, sarà quindi possibile sottoporre al sistema materiale testuale al fine di ottenere in output un dato qualificativo indicante la presenza o meno nell'archivio dello stesso testo o di testo simile ad esso. Tale valutazione di similitudine ovviamente dovrà essere limitata al solo testo e sarà corredato da un valore compreso tra 0 e 1 che ne indica il *grado di similitudine*.

Alcuni software di traduzione assistita cercano e segnalano solo stringhe perfettamente identiche a quelle presenti in memoria, altri invece compiono una ricerca di tipo approssimata (*fuzzy*), suggerendo frasi simili a meno di un valore percentuale che faccia da

discriminante per il traduttore. In questo secondo caso l'applicativo che compone il sistema di traduzione assistita, segnala puntualmente con opportuni parametri, flag o colori le eventuali differenze.

Le segnalazioni sono tanto più accurate quanto più l'applicativo è basato su un algoritmo complesso per l'individuazione di queste somiglianze.

I migliori applicativi segnalano inoltre anche le eventuali ripetizioni presenti nel testo al fine di minimizzare task ripetitivi di traduzione.

Considerando quindi le sopracitate necessità di ricerca ed indicazione del livello di attinenza con i dati presenti in una memoria si evince che un applicativo per la traduzione assistita è un sistema basato su un sistema di tipo query / database.

Infine le caratteristiche di un sistema di traduzione assistita suddividono lo stesso in differenti componenti: *componenti Off-line*, intendendo la capacità del sistema di pre-analizzare un testo ancora da sottoporre a traduzione; *componenti On-line* in grado di accompagnare l'opera di traduzione vera e propria.

1.2.1 Componenti Off-line

1.2.1.1 Analisi del testo da tradurre

L'analisi delle caratteristiche Off-line parte dalla necessità di segmentazione sia del testo sorgente che di quello di destinazione. Tale scomposizione è necessaria da un lato per creare memorie di testi già tradotti associando segmenti nelle due lingue, dall'altro per consentire la ricerca di un segmento nuovo in una memoria esistente.

Segmentazione

Lo scopo della segmentazione è quello di individuare le unità di testo da tradurre che abbiano una specificità, indipendenza e completezza a prescindere dal testo che precede e segue. Intuitivamente la scomposizione di un testo in segmenti può essere effettuata considerando i segni di punteggiatura e gli eventuali codici aggiuntivi o caratteri speciali, come elementi separatori.

Il riconoscimento della punteggiatura richiede ovviamente una conoscenza da parte del sistema dell'ortografia e delle convenzioni grammaticali della lingua, ad esempio per distinguere il normale punto alla fine di una frase da un punto presente alla fine di una abbreviazione o sigla.

Inoltre eventuali presenze di codici aggiuntivi propri del testo devono essere individuati dal sistema e debitamente ignorati e/o utilizzati per la corretta scomposizione del testo.

Si veda a questo proposito l'esempio seguente: una lista di ricambi composta da codici alfanumerici di matricola seguiti dopo un segno distintivo da brevi testi descrittivi, in tale caso il discriminante è proprio il segno distintivo.

```
…
ALFA001 *** Transistor¶
BETA002 *** Triodo¶
GAMMA003 *** Diodo¶
…
```

Figura 1.4 - Esempio di segmentazione del testo

Nell'esempio sopra riportato il sistema deve essere in grado di individuare il simbolo "***" come elemento separatore tra i codici sulla sinistra (da non tradurre) e i termini sulla destra da tradurre. Ovviamente oltre a tale segno, il sistema deve riconoscere anche l'andata a capo (¶) come ulteriore elemento distintivo. Anche altri elementi come ad esempio nomi propri, numeri, date, valute, figure possono essere considerati codici aggiuntivi utili alla scomposizione del testo.

La segmentazione porta quindi alla scomposizione di un testo in elementi di base che successivamente verranno o analizzati ricercandoli nella memoria o allineati con altri segmenti in un'altra lingua al fine di creare una nuova memoria.

I segmenti individuati da un processo di scomposizione non dovranno essere troppo piccoli (necessità di avere un senso compiuto) al fine di non arrivare all'assurda traduzione “parola per parola”; a questo proposito si ricorda la presenza in commercio di programmi di traduzione elementare i cui suggerimenti sono solo applicazione di “filtri” di traduzione parola per parola con risultati spesso scarsi o addirittura errati. La dimensione ottimale varia ovviamente in funzione del tipo di testo che si traduce. Da un lato la traduzione di un nuovo testo utilizzando una memoria basata su una segmentazione troppo *stretta* potrebbe fornire una larga percentuale di uguaglianze non riutilizzabili in un contesto che deve essere completo e omogeneo, dall’altro la traduzione di un testo segmentato strettamente confrontato con memorie dai segmenti più grandi potrebbe avere poche segnalazioni di similitudine.

Altri tipi di segmentazione

Un altro tipo di segmentazione è quella linguistica:

- la *lemmatizzazione (lemming) o stemming* è la riduzione di un termine al suo elemento primitivo, alla radice; ed è utilizzata per la preparazione di una lista di parole chiave ad esempio per successive ricerche in banche dati di tipo differente.
- la *scomposizione sintattica* è utile per individuare quelle sottoparti di un segmento che possono essere considerate come indipendenti e significative. Lo scopo è quello di permettere il riutilizzo della loro traduzione in altre parti del documento stesso (concetto di “molecola-poche parole” rispetto al concetto “atomo-parola”)

Allineamento

Per allineamento si intende l'individuazione delle corrispondenze tra testo (sorgente/traduzione). La segmentazione fornisce all’allineamento quella scomposizione in unità la cui corrispondenza permette di ottenere i migliori risultati nella fase successiva di ricerca nella Translation Memory.

In generale una differente punteggiatura nel testo tradotto può ingenerare una difficoltà aggiuntiva in fase di allineamento (i migliori applicativi permettono l'individuazione e correzione di tali regole di scomposizione).

Estrazione delle parole chiave

Un ulteriore strumento è quello dell'analisi statistica delle parole presenti nel testo di un documento da tradurre al fine di individuare quei termini che si ripetono molte volte. Questo allo scopo di individuare quelle parole chiave da sottoporre ad ulteriore (e separata) analisi. Si veda §2.1.1.4 "Trados-ExtraTerm" a pagina 7.

Statistiche del testo

Alla necessità di valutare il volume di un nuovo lavoro di traduzione, risponde quell'applicativo capace di fornire numericamente indicazioni sulla quantità di parole, segmenti, ripetizioni presenti nel testo sorgente al fine, ad esempio, di permettere una valutazione tempi/costi di una traduzione che ci si appresta ad iniziare.

1.2.1.2 Importazione dati nella TM

L'importazione dei dati allineati in una TM consiste nell'introduzione automatica di segmenti nella lingua sorgente e i corrispondenti segmenti nella lingua di destinazione.

L'importazione può avvenire da un formato grezzo o da uno nativo.

Il cosiddetto formato grezzo è qualsiasi formato in cui il testo sorgente e quello di destinazione sono formattati eccetto il formato nativo della TM stessa.

In altre parole testi in formato ASCII, ANSI, o di un qualsiasi programma di videoscrittura sono considerati come testo grezzo da segmentare ed allineare.

Il formato nativo invece consiste nell'avere il testo sorgente e quello di destinazione nel formato proprio della Translation Memory. Tale formato conterrà già informazioni come segmentazione e allineamento.

1.2.1.3 Esportazione dati dalla TM

Per esportazione si intende il processo di trasferimento di testo dalla TM in un file esterno di testo. Lo scopo è quello di permettere la successiva importazione da parte del traduttore in un eventualmente differente sistema di traduzione assistita.

1.2.1.4 Altre funzioni Off-line

Altre funzioni off-line possono essere:

- Merging (Join)
- Filtering
- Inversione
- Composizione

Le TM sono dei database in cui il *merging* altro non è che il join tra tabelle. Un record in una tabella corrisponde ad un segmento assieme alla sua traduzione e ad altre informazioni aggiuntive. Un join è basata sulla uguaglianza perfetta o di tipo approssimato tra uno o più campi.

Il merging può essere usato in congiunzione con il filtering. Per *filtering* infatti si intende la possibilità di esportare (filtrare) da un testo sottoposto a traduzione solo quei segmenti che non sono presenti in memoria, al fine ad esempio di sottoporre tali segmenti non tradotti ad altre analisi con altre TM.

Per *inversione* si intende invece la capacità di rovesciare il “verso” della traduzione (scambio tra lingua sorgente e lingua destinazione nel database).

Per *composizione* (sorta di proprietà transitiva delle memorie) si intende la capacità di avere una traduzione in una terza lingua a partire da memorie tra una prima lingua e una seconda e tra questa seconda e la terza in questione.

1.2.2 Componenti On-line

In fase di traduzione un sistema di traduzione assistita deve fornire, a partire dai dati presenti nella TM, suggerimenti su come tradurre un determinato segmento di testo. Tale segnalazione deve essere corredata da varie informazioni aggiuntive:

- diverse alternative di traduzione se in memoria sono presenti varie frasi simili o uguali;
- indicazione del livello di somiglianza tra la frase da tradurre e quella presente in memoria;
- indicazione delle parole differenti (colorandole ad esempio).

La stessa interfaccia utente deve supportare alcune funzionalità specifiche:

- esportazione del nuovo segmento;
- possibilità di modificare la traduzione presente in memoria (eventuali correzioni in memoria);
- possibilità di variare i parametri al fine di ottenere differenti risposte dalla TM (ad esempio ignorare le date, i numeri, i codici, i nomi propri, ecc.).

1.2.2.1 Ricerca

Si consideri la TM come un database e la frase/segmento sorgente come quell'unità da cercare tramite query; si possono così ottenere due tipi di risultati che sono alla base della ricerca.

Exact match

Per Exact match si intende la perfetta uguaglianza carattere per carattere tra il testo sorgente da tradurre e il testo sorgente presente in memoria.

Approximate match

Una uguaglianza di tipo approximate (*fuzzy*) è la somiglianza tra segmenti a meno di alcune differenze. Solitamente gli applicativi

assegnano valori di verosimiglianza. La misurazione della distanza (*Edit distance* [8]) tra le parole è uno dei metodi usati per valutare tale verosimiglianza.

Si tenga conto che il peso assegnato alle differenze varia da programma a programma, ma deve essere in parte parametrizzabile dall'utente. Ad esempio la presenza di codici, date, ecc. e dei rispettivi pesi di traduzione devono essere calibrabili.

1.2.2.2 Aggiornamento e Networking

In fase di traduzione una frase tradotta deve essere immediatamente inserita nella TM al fine di fornirla come suggerimento all'utente qualora si ripresentasse all'interno del testo.

Inoltre in caso di memorie centralizzate deve essere possibile accedere ad esse da parte di più utenti contemporaneamente. Si noti come in caso di memoria condivisa l'inserimento di una traduzione errata può ripercuotersi su altre traduzioni, nasce quindi la necessità della definizione del livello degli utenti (account) in relazione alla memoria, potendo così valutare il suggerimento in relazione al livello degli utenti quali "traduttori", "revisori di lingua", "correttori di bozze", ecc.

Capitolo 2 Programmi in commercio

Nel presente capitolo si vogliono individuare i principali e maggiormente diffusi programmi commerciali.

I programmi sottoposti ad analisi sono Trados nella versione 5.0.1 e IBM TranslationManager nella versione 2.6.0.

2.1 TRADOS

Figura 2.1- Trados 5

Trados è una suite composta da più programmi atti a consentire la traduzione di un testo. L'approccio al sistema è estremamente curato nell'interfaccia grafica e nelle possibilità di forte personalizzazione consentite all'utente.

2.1.1 Componenti e loro caratteristiche

I componenti principali di Trados sono WorkSpace, WorkBench, WinAlign, ExtraTerm, MultiTerm. In particolare tali componenti servono per eseguire differenti *task* legati alla traduzione assistita in relazione anche a differenti scenari possibili di traduzione. L'operatività del sistema potrebbe essere quindi così descritta:

- creazione di memorie a partire da traduzioni precedentemente svolte: usando WinAlign per l'allineamento delle stringhe e WorkBench per la creazione della TM vera e propria si realizzano memorie da utilizzare per successive traduzioni;
- creazione di nuovi progetti di traduzione in ambiente integrato tramite WorkSpace, riunendo file da tradurre, memorie da utilizzare, glossari consultabili;

- creazione di glossari di traduzione esportando le parole chiave da testi da tradurre tramite ExtraTerm, consultabili tramite MultiTerm.

2.1.1.1 WorkSpace

WorkSpace è una shell integrata tramite la quale impostare interi progetti di traduzione in varie lingue indicando ad esempio le cartelle interessate al progetto, i dizionari associati, le memorie di traduzione da utilizzare e aggiornare, oltre a permettere la centralizzazione dei più comuni compiti di traduzione quali analizzare preventivamente il materiale da tradurre (al fine di un preventivo) e arricchire la memoria di traduzione al rientro di un file tradotto (*Figura 2.2*).

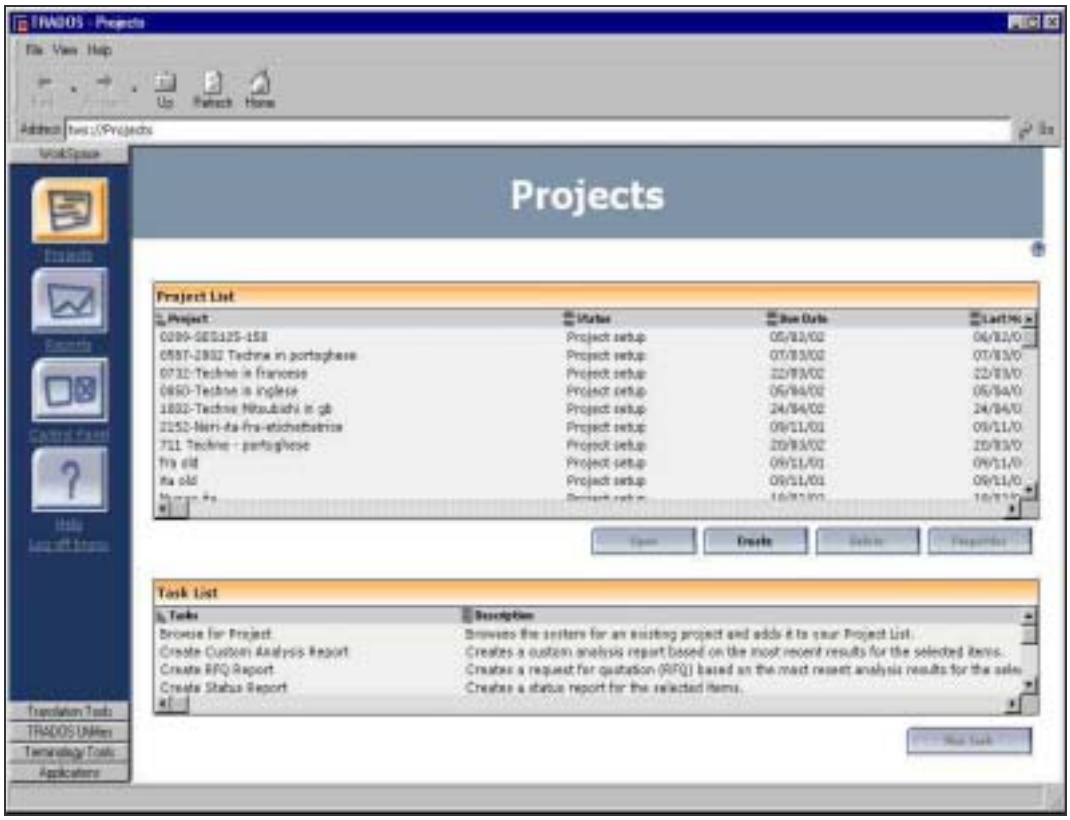


Figura 2.2 - Trados WorkSpace

2.1.1.2 WorkBench

WorkBench è l'interfaccia che si integra con il programma di Word Processing adottato per l'editing del testo da tradurre. Prima della creazione della versione 5, Trados non disponeva della scrivania

virtuale WorkSpace ed era WorkBench stesso a fungere da strumento principale per il traduttore. Ancora oggi si tende ad identificare Trados con lo strumento WorkBench, in quanto risultava l'interfaccia tipica per l'utente, negando di fatto l'esistenza di altri tool.

WorkBench consente il suggerimento on-line in fase di traduzione di segmenti disponibili in memoria a meno di un valore percentuale di similitudine (100% o approximate match); consente inoltre la manutenzione e creazione di memorie a partire da file allineati (realizzati precedentemente con WinAlign).

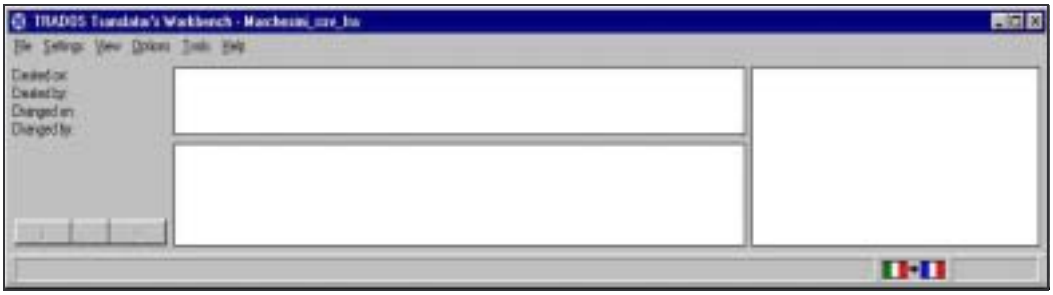


Figura 2.3 - Trados WorkBench

Qui di seguito (*Figura 2.4*) l'interfaccia di importazione e creazione memoria a partire da file allineati in precedenza con WinAlign:

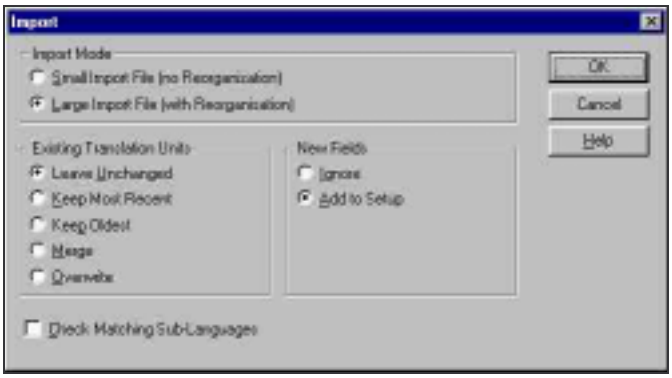


Figura 2.4 - Trados WorkBench - Importazione dati per creazione memoria

Tramite tale schermata è possibile indicare se si tratta di importazioni minori (senza rigenerazione dell'indice della banca dati) o di importazioni massive (*Import Mode*), oltre a permettere l'eventuale aggiornamento di testo già presente in memoria fornendo nuove possibili traduzioni dello stesso (*Existing Translation Units*).

In fase di traduzione il programma è integrato direttamente all'ambiente scelto per la traduzione del testo, tramite pulsanti aggiunti alla toolbar del Word Processor. Tali pulsanti consentono di interrogare il database, di sostituire il testo sorgente con quello tradotto e di aggiornare in tempo reale la memoria (*Figura 2.5*).

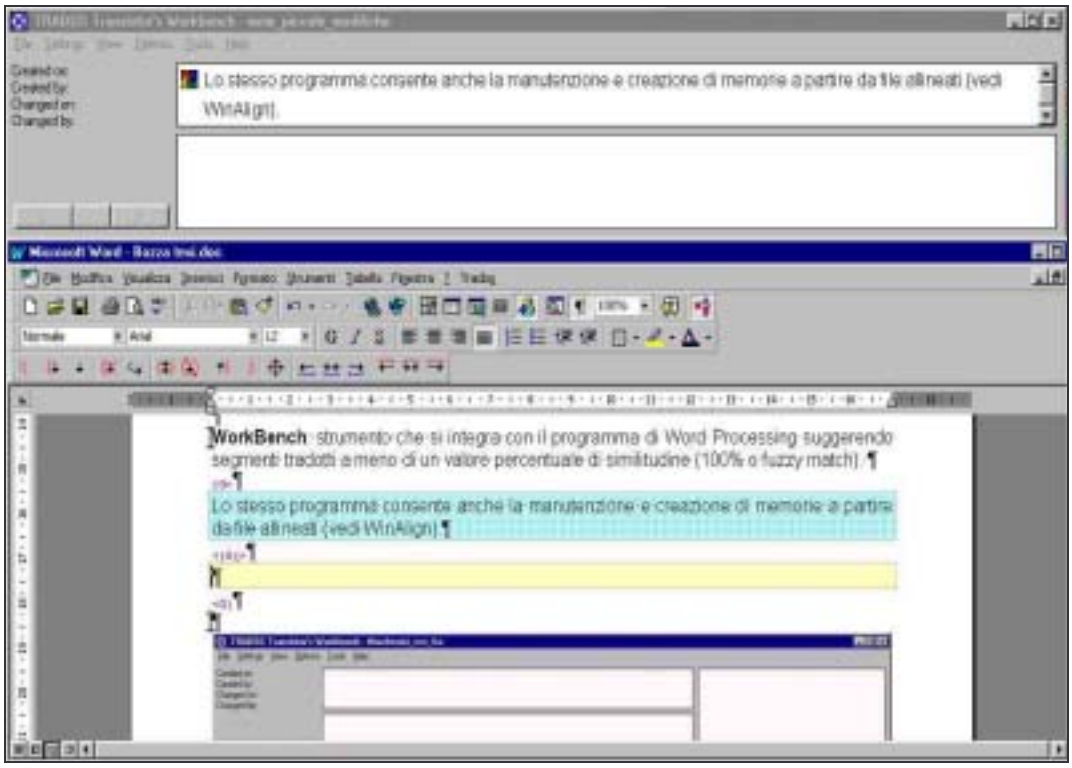


Figura 2.5 - Integrazione Trados - WordProcessor

2.1.1.3 WinAlign

WinAlign è lo strumento per l'allineamento di coppie di file in lingue differenti. Tale allineamento ha lo scopo di generare un file di interscambio successivamente importabile in WorkBench.

Il programma permette l'individuazione di quali file allineare, il formato degli stessi e le lingue di origine e destinazione (*Figura 2.6*).

Figura 2.6 - Impostazione del progetto

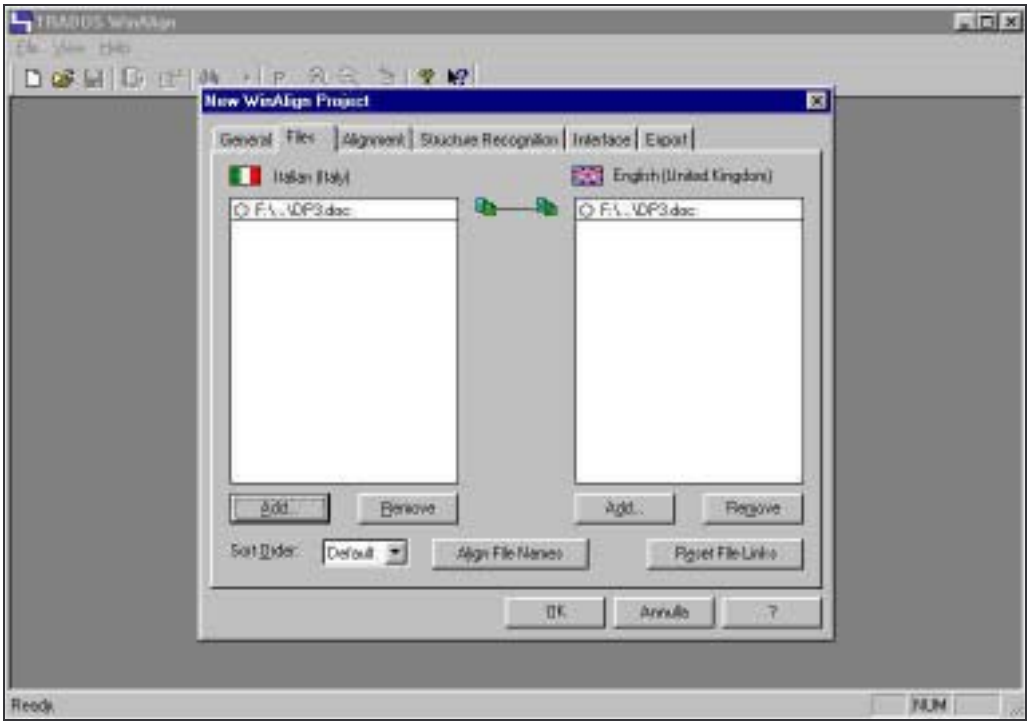


Figura 2.7 - Associazione file sorgente, file tradotto

E' ovviamente possibile impostare interi gruppi di file che compongono progetti complessi.

Dopo questa prima fase di impostazione dei parametri, il programma compie una fase di pre-analisi/segmentazione (*Figura 2.8*). Durante questa fase il programma scompone i file in lingua sorgente e in lingua destinazione in segmenti che nella fase successiva saranno proposti per l'allineamento vero e proprio.

Figura 2.8 - Pre-allineamento batch

Al termine si procede alla verifica dell'allineamento suggerito dal programma tramite un'interfaccia grafica a due colonne in cui sono affiancate le due lingue e i segmenti in cui sono state scomposte. L'operatore/utente può quindi intervenire a vari livelli correggendo eventuali segmenti allineati in maniera errata (*Figura 2.9*).

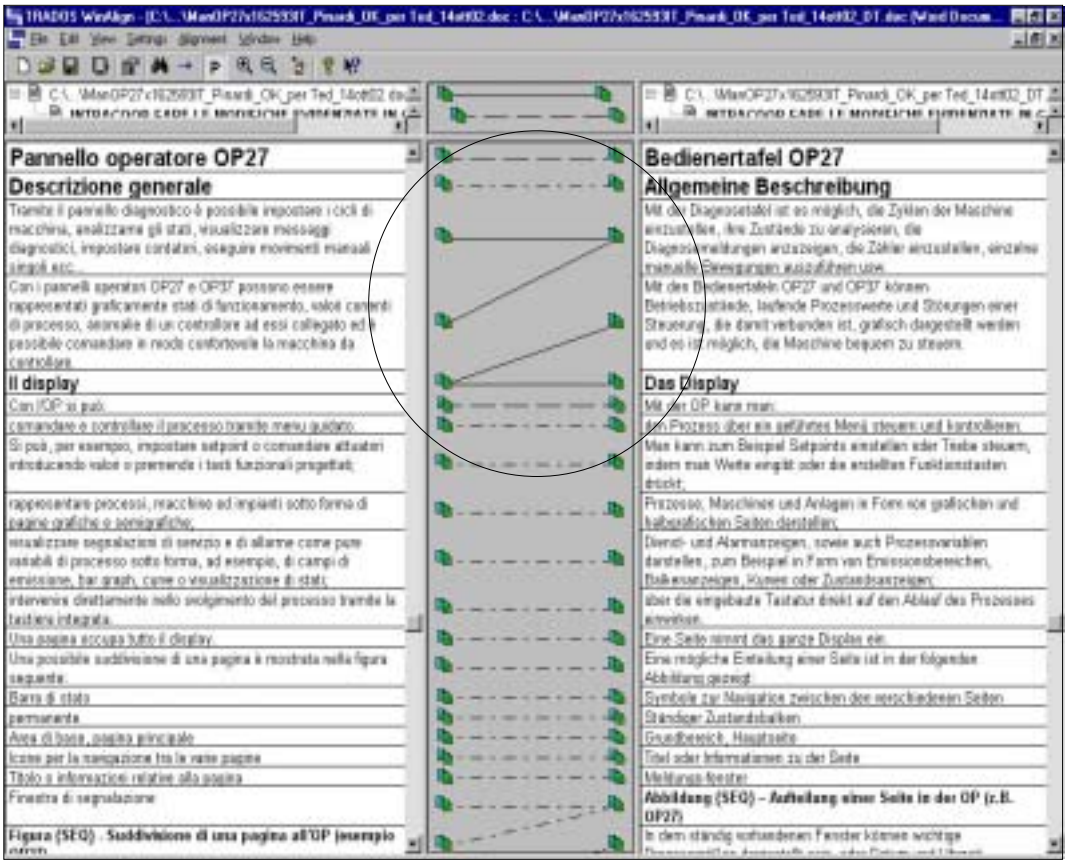


Figura 2.9 - Allineamento segmento-segmento

Durante questa fase è consentito di disallineare interi blocchi di segmenti e di sospendere il lavoro per recuperarlo in un secondo momento. Particolarmente interessante è la funzione che consente di poter modificare il testo di singoli segmenti per correggere eventuali errori di ortografia.

Al termine della fase di verifica dell'allineamento si procede con l'esportazione in un file .TXT per permettere il successivo inserimento nella TM.

Data la particolare importanza della fase di segmentazione del testo è stata inserita la possibilità di impostare alcuni parametri del programma stesso per consentire una calibrazione in base alle proprie esigenze specifiche. In un pannello del programma stesso è infatti possibile impostare, prima di eseguire la pre-analisi/segmentazione del testo, parametri come le regole di segmentazione da adottare (*Segmentation Rules - Figura 2.10*).

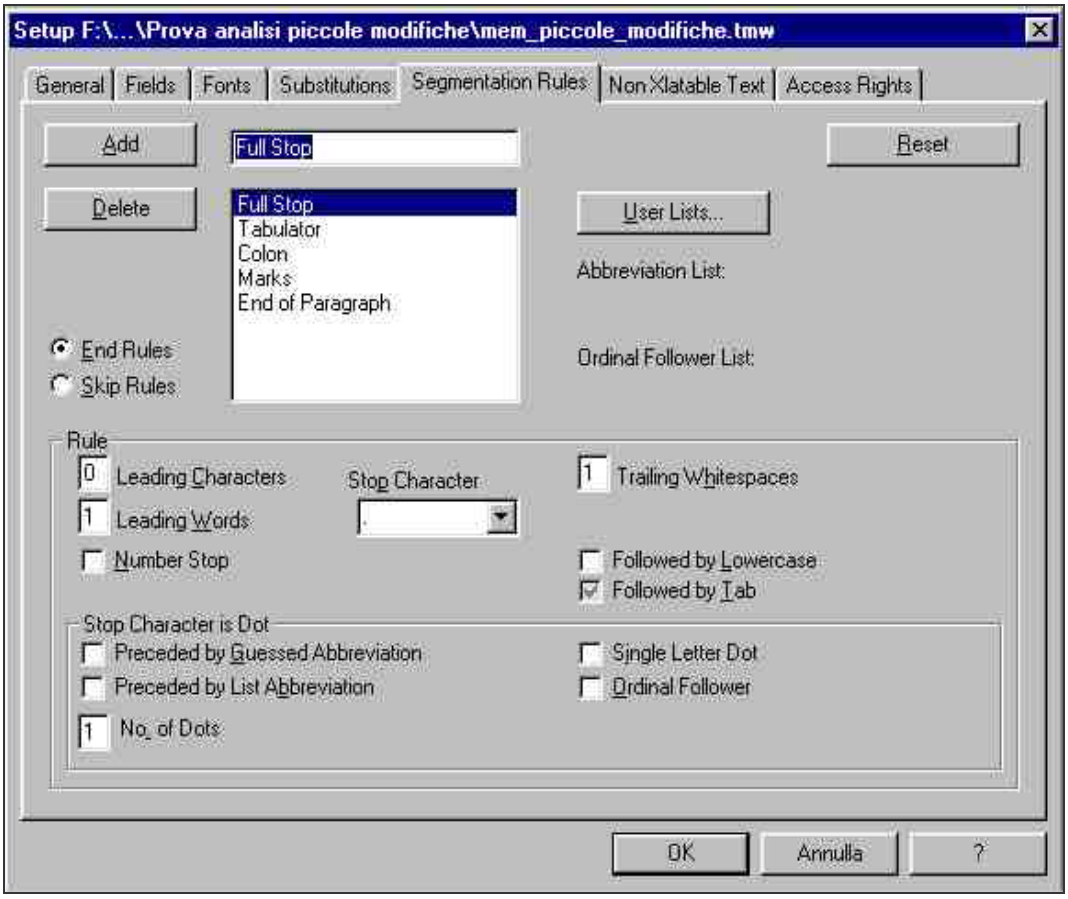


Figura 2.10 - Parametrizzazioni di WinAlign

Si noti come il programma abbia preimpostate alcune regole di segmentazione basate sul punto (Full Stop), sui segni di tabulazione (Tabulator), sui punto e virgola (Colon), su marcatori personalizzabili (Marks) e su segni di fine paragrafo (End of Paragraph), mentre non prevede una segmentazione basata sulla regola della scomposizione in base alla presenza di virgole nel testo ("comma"). Questa opzione è però facilmente implementabile in quanto il programma consente l'introduzione di nuove regole di segmentazione.

2.1.1.4 ExtraTerm

ExtraTerm è lo strumento di estrazione di parole chiave per la realizzazione di glossari (*Figura 2.11*). Il programma è parametrizzabile in funzione del tipo di glossario che si desidera ottenere. Ad esempio è possibile stabilire che il glossario deve contenere singole parole o insiemi di parole composte da un certo numero di termini (*Figura 2.12*).

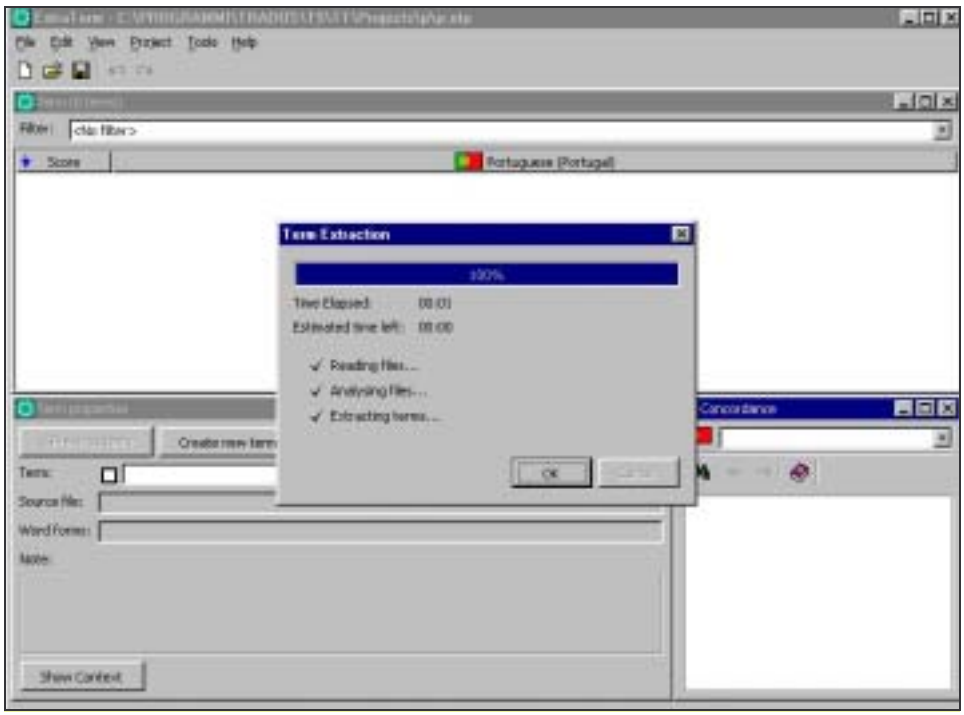


Figura 2.11 - Analisi in ExtraTerm

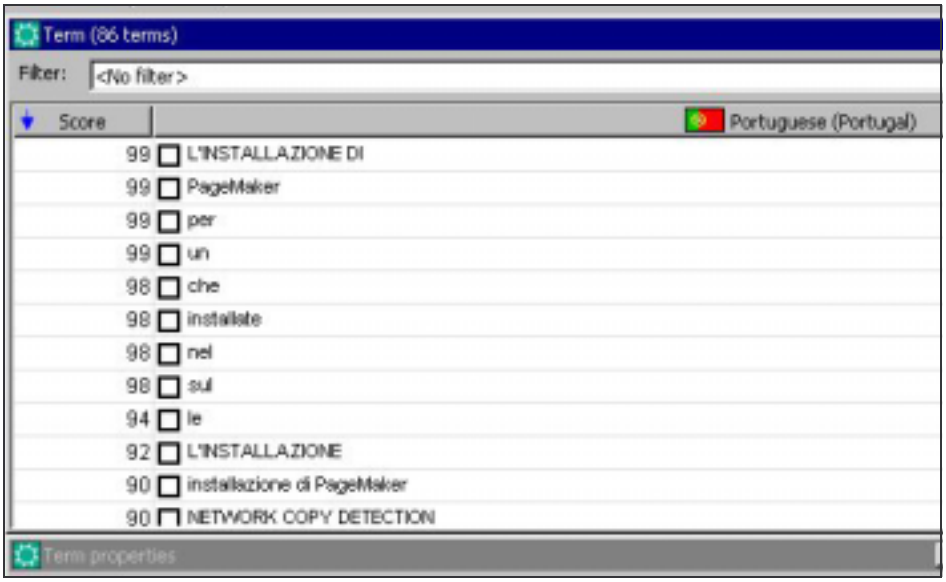


Figura 2.12 - Lista termini esportabili

2.1.1.5 MultiTerm

MultiTerm è l'interfaccia integrata nel programma di Word Processing per la ricerca in tempo reale di singoli termini durante la fase di traduzione.

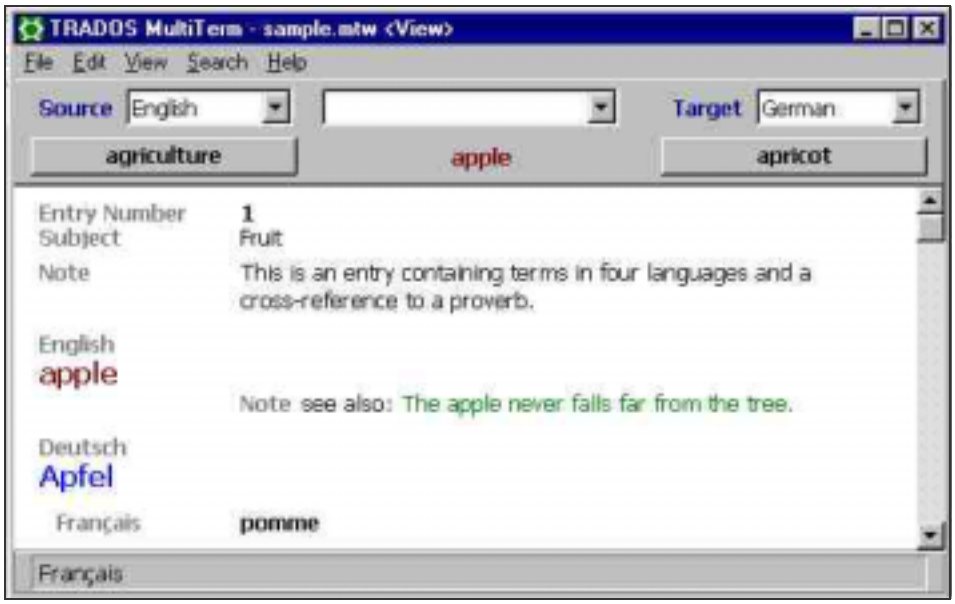


Figura 2.13 - Interfaccia MultiTerm

2.1.2 Ulteriori caratteristiche

Lo strumento WorkBench consente la pretraduzione automatica di tipo batch di un documento indicando un livello minimo di verosimiglianza da applicare. E' cioè consentito di ottenere in maniera semiautomatica un documento pretradotto in cui le frasi uguali a quelle in memoria sono sostituite con quelle nella lingua di destinazione. Il testo così ottenuto sarà composto da frasi ancora nella lingua sorgente e frasi nella lingua destinazione.

La particolarità di questo strumento è la possibilità di discriminare la soglia di verosimiglianza. In altre parole è consentito impostare a livelli inferiori il 100% il parametro con cui il programma sostituisce le frasi quasi uguali (Approximate match).

Ad esempio è possibile pretradurre un documento indicando che segmenti nella lingua sorgente uguali solo al 97% a segmenti presenti in memoria siano trattati alla stregua di Exact match (*Figura 2.14*). Questo genere di approccio è ovviamente estremamente delicato in quanto si presuppone la certezza da parte dell'operatore di essere di fronte a casi in cui la verosimiglianza al 97% sia effettivamente una verosimiglianza assoluta.

Questo genere di azione è consigliabile solo dopo alcune prove e in casi particolari in cui, ad esempio, i codici di controllo, e non il testo vero e proprio, differiscono da quelli in memoria ed è comunque prevista una successiva verifica degli stessi.

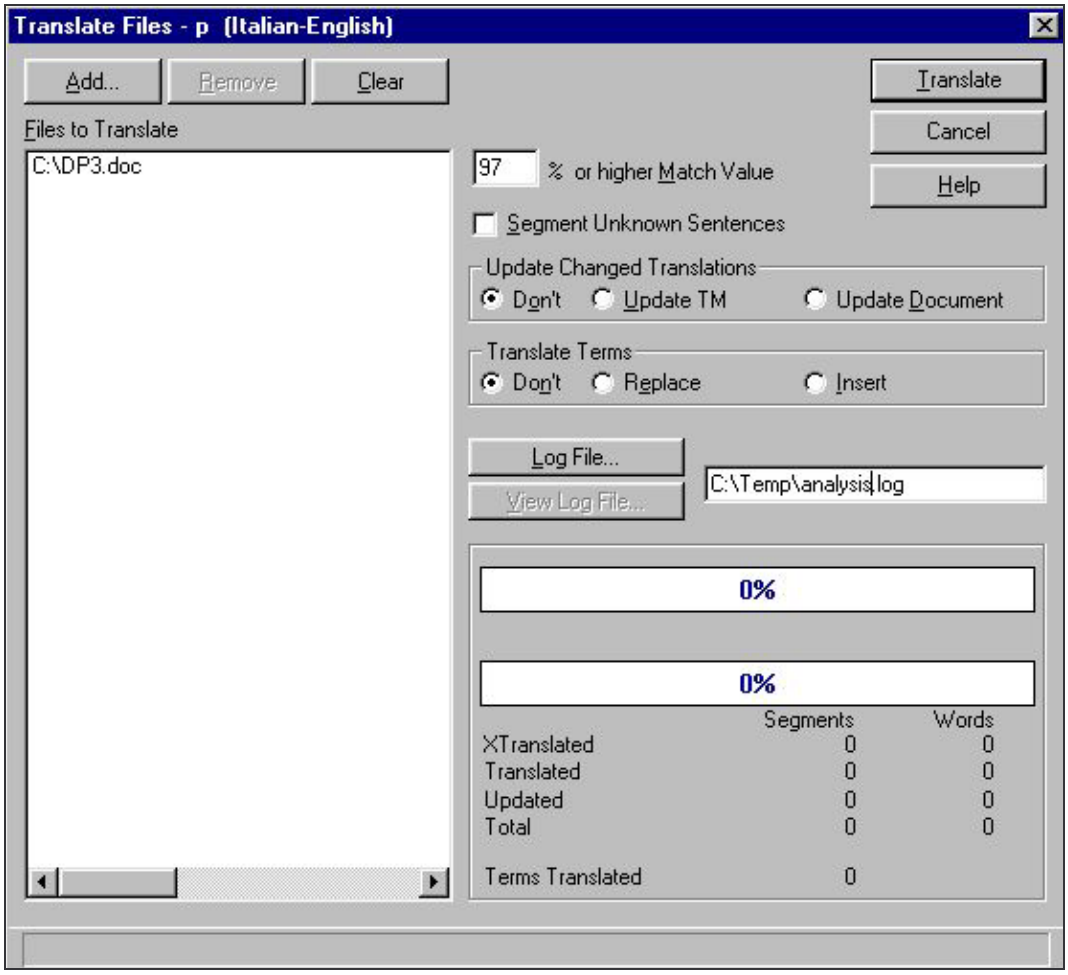


Figura 2.14 – Trados WorkBench - Pretraduzione di tipo batch

2.2 IBM TRANSLATION MANAGER

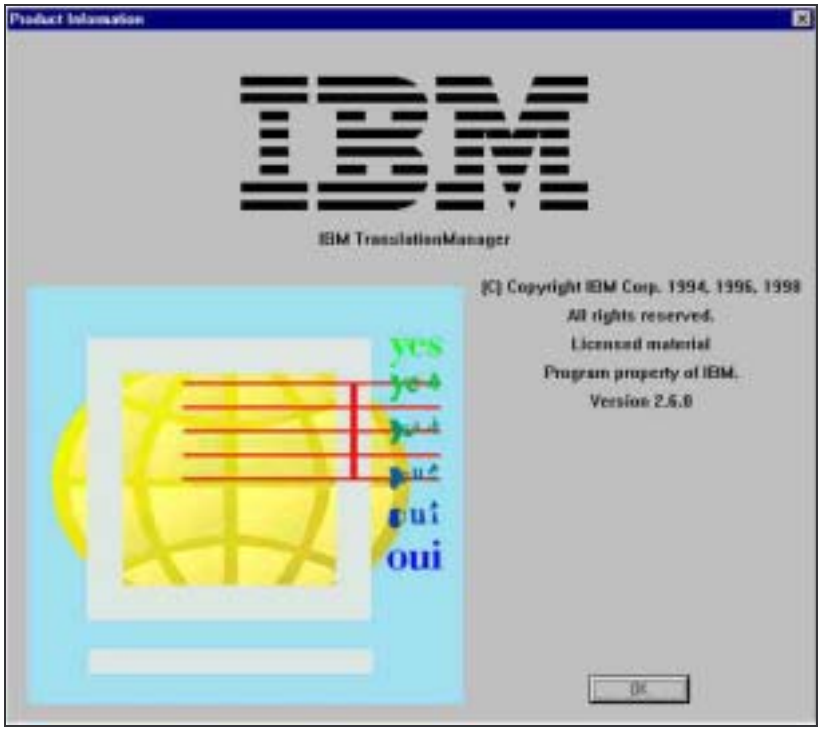


Figura 2.15 - IBM TranslationManager

IBM TranslationManager è un programma incentrato su una scrivania virtuale di lavoro, unica e centralizzata, e da uno strumento aggiuntivo per la creazione di memorie.

2.2.1 Componenti

Il programma è costituito da due applicativi ciascuno specifico per un determinato compito. La shell principale di sistema TranslationManager consente di cui creare progetti, importare file, associare memorie a progetti, consultare memorie, ecc. Il secondo applicativo si chiama Initial Translation Memory Tool e serve per la creazione di memorie permettendo l'allineamento di file scomponendoli in singole *sentence*.

2.2.2 Caratteristiche

La scrivania virtuale TranslationManager contiene una serie di finestre ciascuna atta ad uno scopo diverso. Una prima finestra contiene ad

esempio la lista delle cartelle progetto in cui è possibile importare e impostare i formati dei file o aprire i file da pretradurre e completare. Altri strumenti/finestre presenti sono la cartella delle memorie consultabili e la cartella degli eventuali dizionari settoriali (*Figura 2.16*).

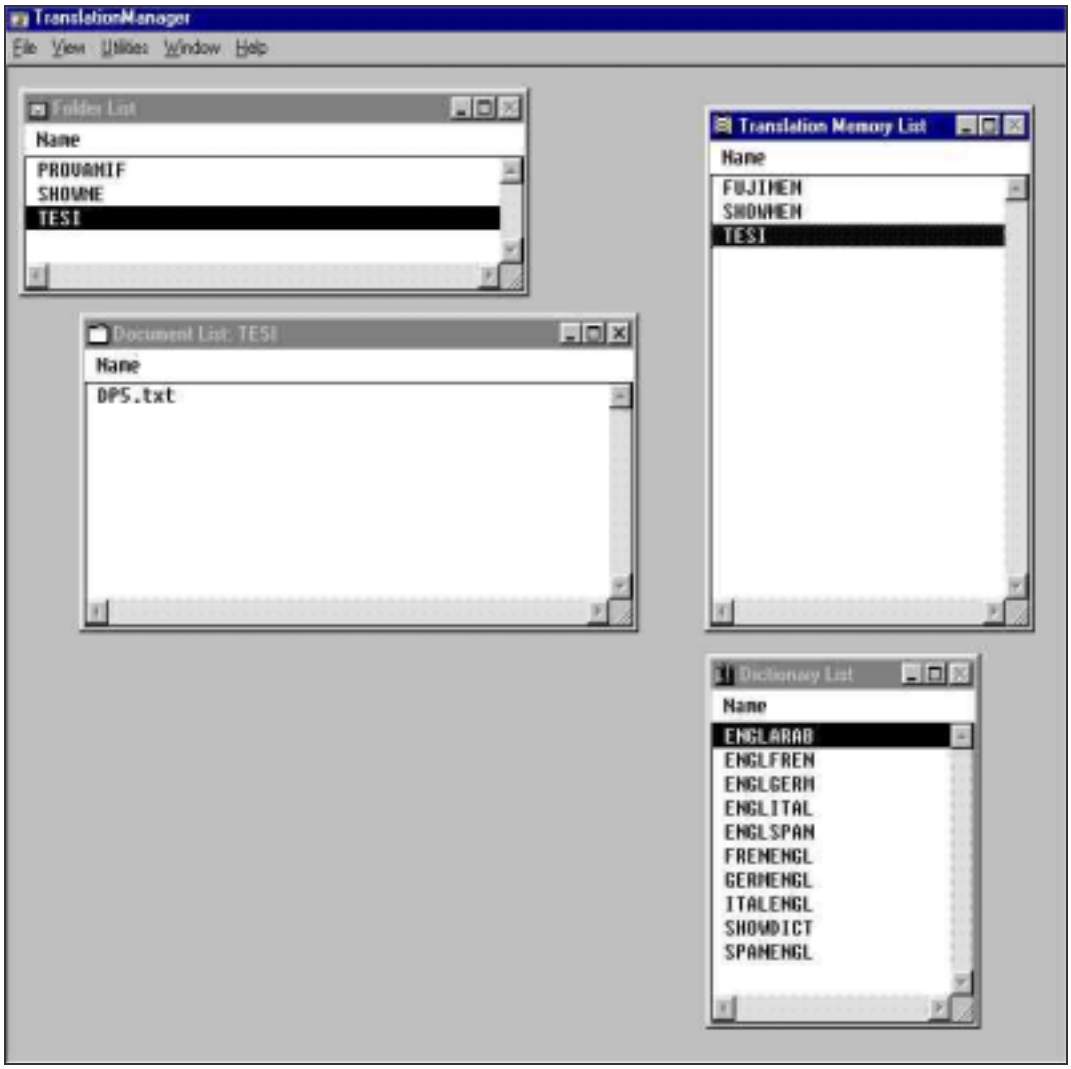


Figura 2.16 - Interfaccia principale IBM TranslationManager

L'Initial Translation Memory Tool (*Figura 2.17*) si presenta invece con una interfaccia articolata in cui individuare i file da allineare allo scopo di permetterne la scomposizione e successiva realizzazione di una nuova memoria.

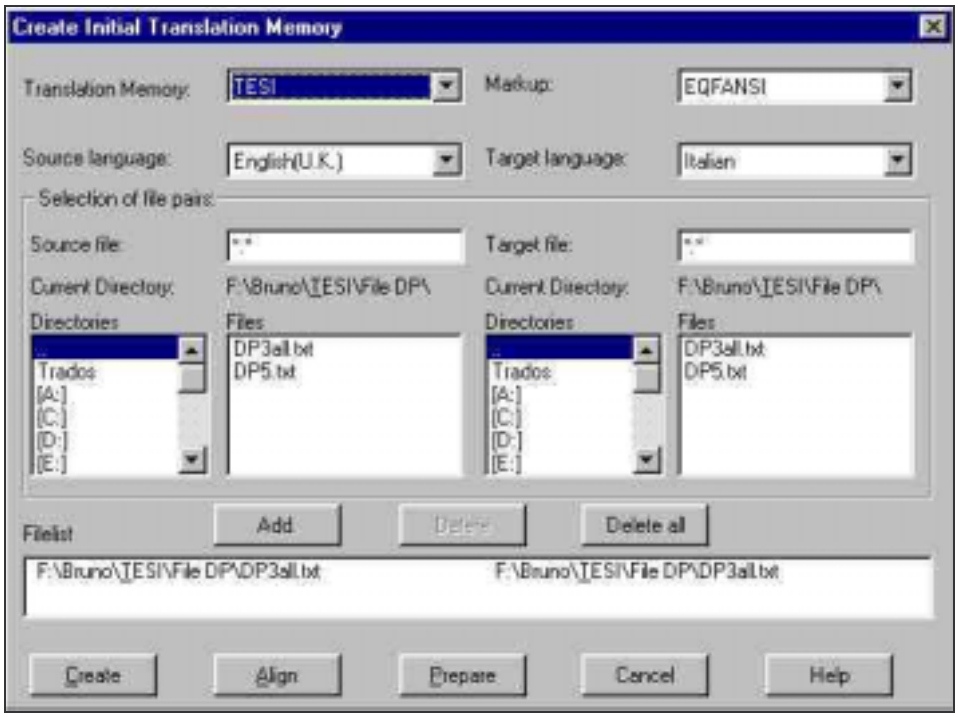


Figura 2.17 - Initial Translation Memory Tool

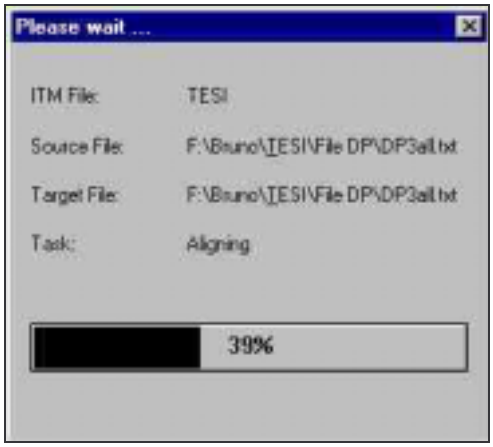


Figura 2.18 – Allineamento con l'Initial Translation Memory Tool

Dopo la prima fase di segmentazione e allineamento dei due testi (*Figura 2.18*), il programma presenta un'interfaccia a due colonne in cui vengono visualizzati i due testi di cui verificarne l'allineamento (*Figura 2.19*).

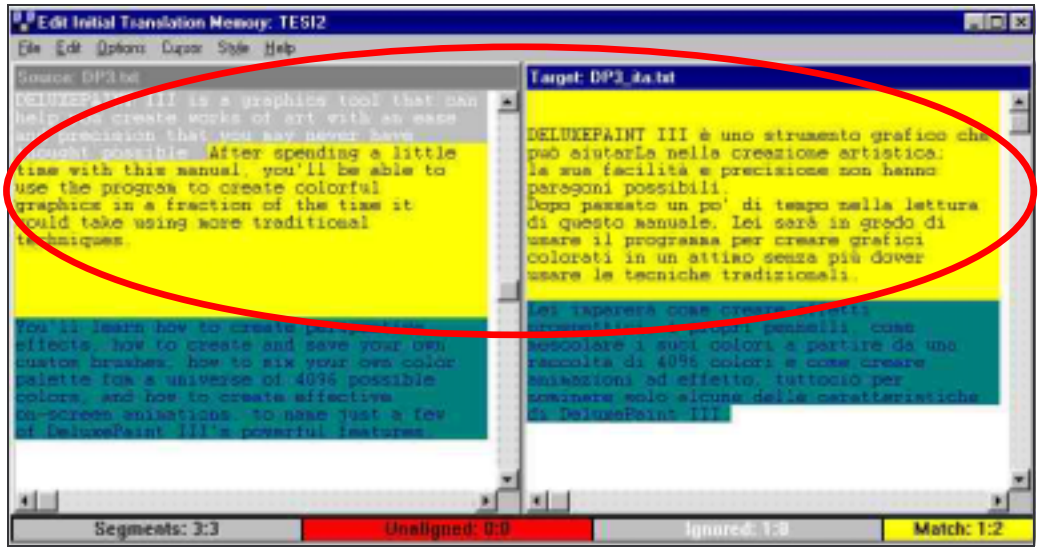


Figura 2.19 – Initial Translation Memory Tool - Allineamento segmenti

Particolarità di questo approccio è l'arricchimento della memoria in tempo reale; mentre infatti si procede con la conferma delle singole frasi, la memoria viene aggiornata ed è subito utilizzabile da un altro eventuale utente. Ovviamente il lavoro di conferma di allineamento può essere sospeso, consentendo di rimandare particolare decisioni e non introdurre in memoria dati inesatti.

Al termine della fase di allineamento dei due testi si conferma la creazione della memoria completando il processo (*Figura 2.20*).

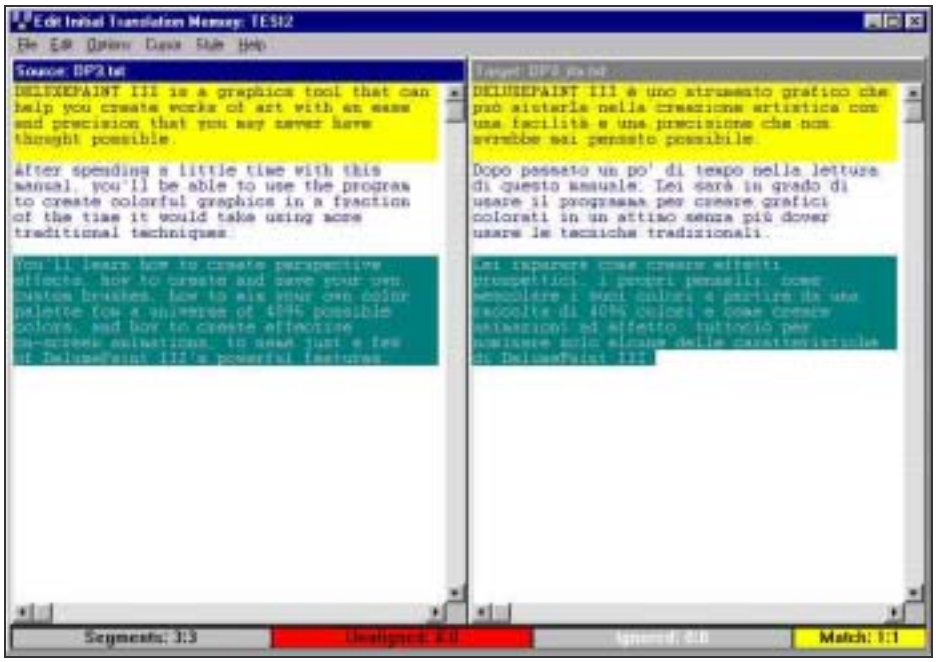


Figura 2.20 - Conferma allineamento

Questa interfaccia a differenza del precedente Trados si presenta estremamente scarna. In particolare risulta scomoda e poco intuitiva per il neofita a causa dell'assenza di icone e menu personalizzabili. E' quindi possibile, solo dopo molta pratica, apprendere i numerosi comandi abbreviati da tastiera. Ciò è prevalentemente dovuto all'origine del programma in ambiente Unix testuale in cui era stato sviluppato e a cui i progettisti non hanno rinunciato nel *porting* in ambiente Windows.

2.2.3 Ulteriori caratteristiche

Come nel caso del precedente programma analizzato, si vuole ora valutare la differente tecnica di approccio alla traduzione del sistema IBM. I tipi di analisi dei file da tradurre è anche in questo caso distinguibile in analisi di tipo batch e analisi di tipo On-line.

Analisi di tipo batch

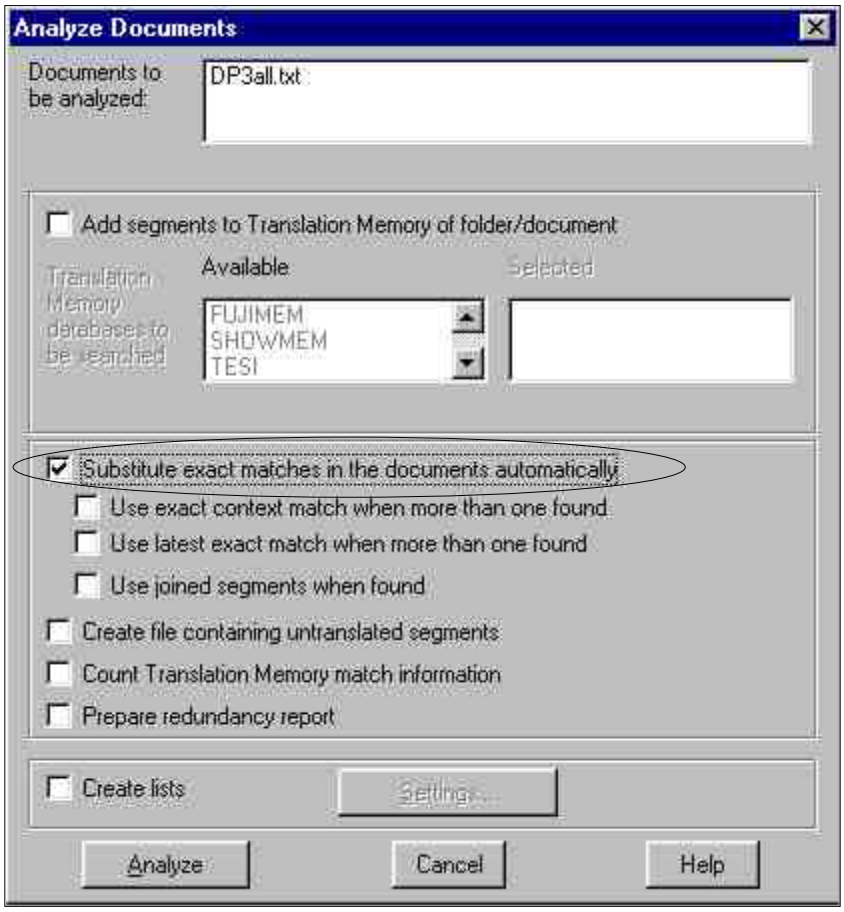


Figura 2.21 - Analisi batch in IBM TranslationManager

Partendo dalla scrivania del sistema è possibile aprire e analizzare un file singolo di un progetto impostato. Tale tipo di analisi prevede solo pochi settaggi possibili per la pretraduzione automatica di un file.

Una caratteristica principale è che questo tipo di approccio può essere definito di tipo assoluto, in quanto le frasi eventualmente uguali (*Exact match*) vengono sostituite senza lasciare traccia nel documento originale, perdendo quindi il vantaggio di un eventuale suggerimento al traduttore. Non è inoltre possibile indicare un eventuale livello minimo di verosimiglianza da utilizzare come soglia per ottenere un documento pretradotto in cui anche le frasi simili a meno di un certo valore percentuale siano automaticamente sostituite come fossero esattamente uguali. Per questo motivo è preferibile fornire al traduttore sia il file originale che la memoria, al fine di permettergli una traduzione on-line che consenta in tempo reale la consultazione delle frasi suggerite dalla memoria. Questo processo però genera nel traduttore la necessità di tradurre o meglio confermare anche quelle frasi che sono in realtà già rilevate dal sistema come uguali al solo fine di permettere al traduttore di venirne a conoscenza

In conclusione questo approccio è sicuramente meno funzionale del cosiddetto approccio *a due vie* consentito da altri programmi, tra cui Trados, in cui è possibile ottenere un file di testo pretradotto in maniera batch in cui sono affiancati sia il testo sorgente che la traduzione corrispondente a meno un valore di verosimiglianza indicato. L'approccio consentito da IBM TranslationManager inoltre rende necessario disporre del programma stesso da parte del traduttore.

Analisi On-line

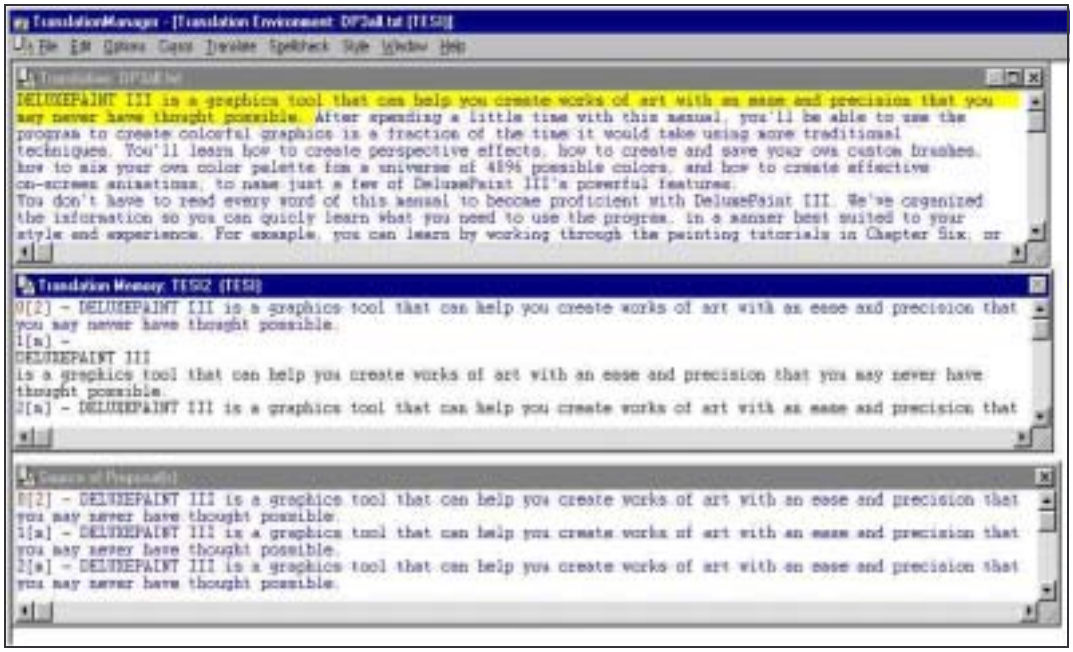


Figura 2.22 - Analisi batch in IBM TranslationManager

L'analisi on-line in questo caso avviene in un ambiente grafico che non è il classico wordprocessor a cui il traduttore è abituato, bensì di un ambiente a finestre in cui è possibile vedere: il testo nuovo da tradurre, i segmenti trovati simili in memoria e i segmenti originali da cui queste traduzioni sono scaturite.

Tale ambiente si occupa quindi delle eventuali conversioni da un formato nativo (RTF, ANSI, altri) in questo documento di tipo testuale, perdendo l'immediatezza d'uso. Per questo motivo il traduttore dovrà quindi utilizzare una serie di comandi direttamente dai menu a tendina o le combinazioni di tasti di questa particolare interfaccia.

Il programma prevede la possibilità di nascondere eventuali *tags* al fine di permettere una traduzione più agevole al traduttore che però può in qualsiasi momento vedere (senza modificare) questi codici di controllo. L'esempio riportato in figura essendo basato su file di testo ANSI non presenta questi codici di controllo.

Capitolo 3 Valutazione di un programma CAT

In questo capitolo si intende fornire una panoramica degli approcci possibili per la valutazione di un programma CAT.

3.1 Efficacia - Concetto di Precision e Recall

La determinazione di uno strumento di misura valido per la determinazione dell'efficacia di un programma CAT deve ovviamente rifarsi agli stessi strumenti forniti dalla letteratura sull'Information Retrieval per la misurazione di una query SQL e dei risultati forniti.

Information Retrieval

La nascita del problema del recupero di informazioni da una raccolta di testi è databile per lo meno al terzo secolo avanti Cristo, quando iniziarono ad apparire biblioteche con centinaia di migliaia di documenti; tuttavia è solo recentemente, con l'avvento degli strumenti

informatici, che l'*Information Retrieval* ha subito una "spinta evolutiva" ed ha assunto consistenza dal punto di vista teorico.

Per *Information Retrieval* si intende "classicamente" quell'insieme di tecniche che consentono un accesso mirato ed efficiente a grandi raccolte di oggetti contenenti principalmente testo (ad esempio, il recupero, in una biblioteca, di tutti i libri inerenti l'argomento "*Information Retrieval*"). Attualmente l'*Information Retrieval* riguarda tecniche applicabili in modo algoritmico da "macchine" (*calcolatori*) in grado di accedere agli oggetti di una raccolta. Ogni oggetto è riassunto nei suoi attributi da "descrittori", che tipicamente consistono in un testo. Queste descrizioni possono anche includere descrittori assegnati dal creatore (autore) dell'oggetto o da un qualche indicizzatore (un uomo o una "macchina") o utilizzati per descrivere eventuali relazioni con altri oggetti nella raccolta.

Quello che rende un sistema del genere un *Retrieval System* è la capacità di descrivere e tentare di soddisfare un **fabbisogno informativo** (f.i.), cioè un interesse specifico, dell'utente. Le caratteristiche del f.i. sono descritte in query. Una query è espressa tipicamente in linguaggio naturale, ma sono possibili anche altre forme, come l'uso di espressioni booleane o di esempi di documenti.

Dando per scontato che l'utente sia in grado di formalizzare delle *query* corrette e coerenti con il proprio f.i., egli dovrebbe essere in grado di constatare, effettuata una ricerca, se gli oggetti recuperati dalla "macchina" rientrano o meno nel suo interesse. Questa centralità dell'utente nella valutazione del *processo di retrieval* è una caratteristica importante dell'*Information Retrieval* tradizionale e svincola in qualche modo il problema da quello attinente l'interpretazione del testo.

Sintetizzando, il processo di *retrieval* può essere visto come una serie di passi:

1. si genera una rappresentazione del significato o del contenuto di ogni oggetto, basata sulla sua descrizione,

2. si genera una rappresentazione del significato del fabbisogno informativo,
3. si confrontano le due rappresentazioni e si scelgono quegli oggetti che sembrano attenersi maggiormente al f.i.

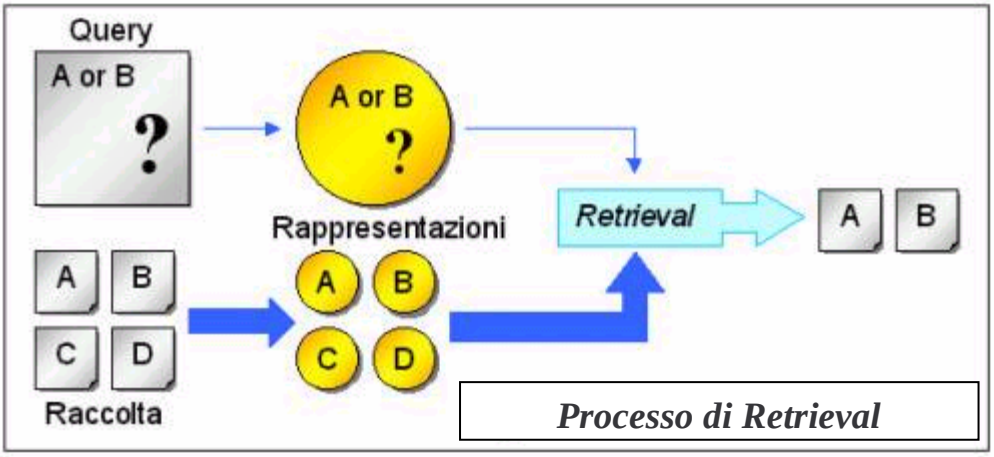


Figura 3.1 – Processo di Retrieval

Risulta evidente allora come il problema sia quello di cercare delle ***buone rappresentazioni*** del documento e del f.i., e di come poter paragonare queste rappresentazioni.

Come valutare un meccanismo di ricerca

Compito dell'*Information Retrieval* (IR) è quello di memorizzare, rappresentare ed estrarre da una raccolta di documenti la più grande quantità di informazioni su un dato argomento e null'altro [5][18]. E' necessario specificare che per documenti si intende ovviamente un insieme omogeneo di dati che può quindi essere diversamente costituito rispetto agli storici insiemi di testi quali si è solito fare riferimento pensando a banche dati.

Se il reperimento di informazioni è l'aspetto su cui si incentra maggiormente questa Tesi, allora l'aspetto da approfondire di un Sistema IR è l'efficacia con cui esso è in grado reperire in una collezione di documenti proprio i documenti ricercati, cioè la sua capacità di fornire tutte le informazioni rilevanti presenti in una collezione. La determinazione della rilevanza di un risultato fornito dal Sistema IR è ovviamente soggettiva in quanto a medesimi risultati

forniti da una stessa ricerca, utenti diversi potrebbero assegnare giudizi diversi, cioè interpretare diversamente il risultato fornito. Tutto dipende dalle conoscenze a priori che si hanno sull'argomento ricercato. In particolare è necessario osservare come, in relazione ad una data ricerca, vi sono dati restituiti rilevanti e dati restituiti non rilevanti. I risultati ottenuti dalla ricerca potranno quindi essere suddivisi nei seguenti quattro casi:

- rilevanti e reperiti, cioè corretti;
- rilevanti e non reperiti, cioè omessi;
- non rilevanti e reperiti, cioè inesatti;
- non rilevanti e non reperiti, cioè da omettere;

	Rilevanti REL	**Non rilevanti NREL**
Documenti reperiti RET	Corretti	Inesatti
Documenti non reperiti NRET	Omessi	Da omettere

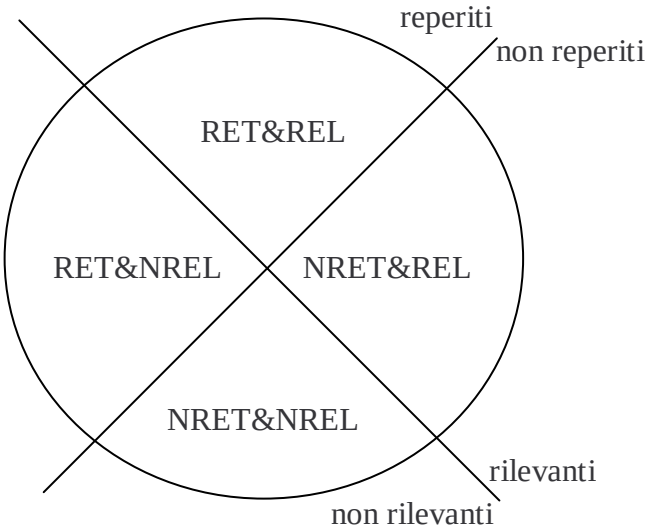


Figura 3.2 – Possibili risultati di una ricerca in una collezione

Scopo di un Sistema IR è quello di massimizzare i documenti rilevanti e reperiti (RET&REL) e minimizzare il "rumore" costituito dai

documenti reperiti e non rilevanti. I documenti rilevanti e non reperiti (NRET&REL) sono invece i risultati omessi dal sistema e quindi sono da minimizzare ugualmente.

I parametri adottati a livello internazionale per misurare l'efficacia di un Sistema IR sono i fattori **recall** (richiamo, ricordo) e **precision** (precisione).

Il fattore *recall* viene calcolato sui documenti rilevanti reperiti, e ne misura la percentuale rispetto al totale contenuto nella collezione, mentre il fattore *precision* invece riguarda i documenti reperiti dalla ricerca e rappresenta la percentuale di documenti rilevanti.

$$\text{recall} = \frac{RET \,\&\, REL}{REL} \qquad \text{precision} = \frac{RET \,\&\, REL}{RET}$$

Ovviamente il fattore recall presuppone di conoscere quanti sono i documenti rilevanti in tutta la collezione (deus ex-machina), mentre il fattore precision è calcolabile a partire dal risultato ottenuto.

In genere aumentando il fattore recall, diminuisce il fattore precision e viceversa.

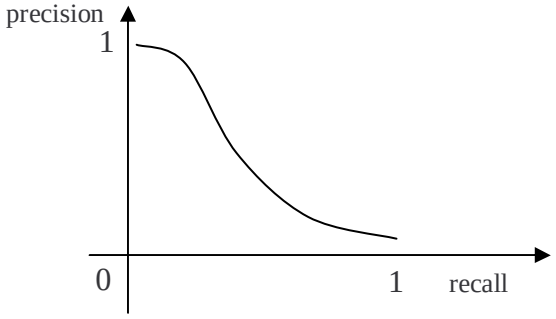


Figura 3.3 – Curva precision e recall

Un buon modello di ricerca sarà allora quello in grado di massimizzare i fattori di recall e precision, quindi di minimizzare rispettivamente il silenzio cioè l'assenza di informazione (informazione omessa) e il rumore cioè il deterioramento dell'informazione (informazione inesatta).

Esempio

Ad esempio se disponiamo di una raccolta di 10 documenti, di cui 4 hanno per argomento le torte di mele, mentre gli altri 6 contengono ricette per cucinare il pesce. Effettuando una ricerca sulle torte, il processo di retrieval restituisce 5 documenti, di cui 3 riguardano le torte di mele mentre gli altri 2 la cottura della trota. Si avrà una precision pari a 3/5 cioè 60%, mentre la recall sarà pari a 3/4 cioè 75%.

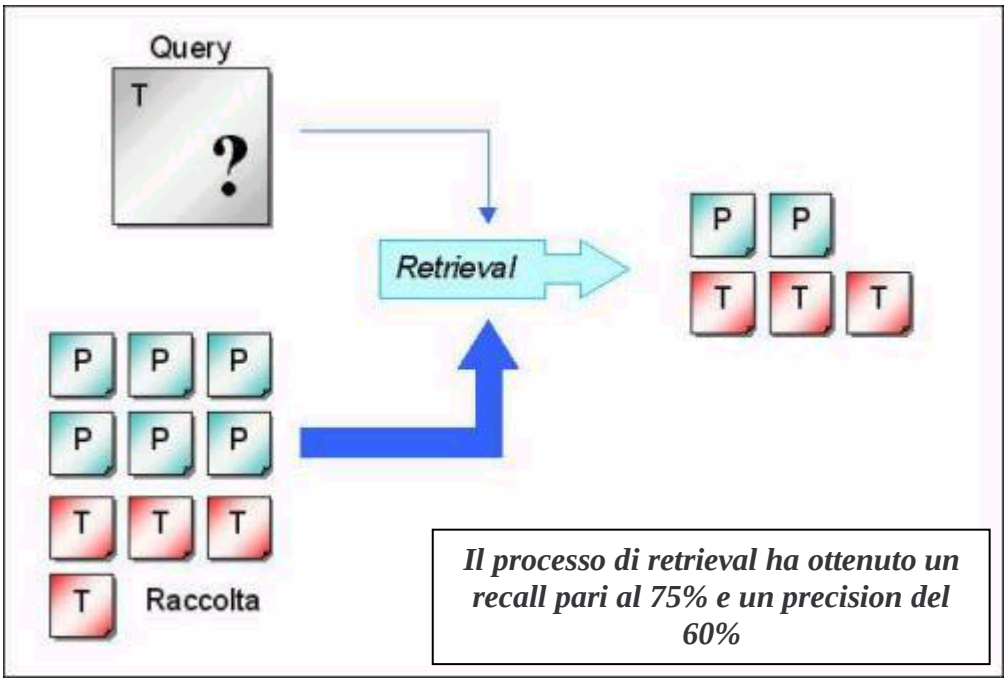


Figura 3.4 – Precision e recall - Esempio

3.2 I principali modelli di ricerca

Come si è visto, per attuare il processo di retrieval bisogna avere a disposizione delle descrizioni di documenti da confrontare con la descrizione del f.i. Quello che serve è allora un modello di retrieval [1], che stabilisca i dettagli delle rappresentazioni usate e il criterio per paragonare i due tipi di descrizioni.

Modello booleano

Nel modello booleano i documenti e le query sono rappresentati come insiemi (*set*) di parole chiave; pertanto si dirà che il modello è di tipo *set theoretic* [15]. Esso è il primo modello utilizzato storicamente e si

contraddistingue in quanto i documenti sono rappresentati da insiemi di termini chiave (*keywords*) estratti manualmente o automaticamente dal testo e le ricerche sono condotte tramite parole chiave connesse da operatori logici. La sua diffusione è dovuta prevalentemente alla sua semplicità ed efficienza, ma a volte fornisce limitata efficacia.

In particolare quindi dato un insieme finito di feature R = {r1, r2…, rk}, il documento è rappresentato come un assegnamento di qualche feature. Tale assegnamento viene generalmente rappresentato da un vettore a valori booleani V di lunghezza k: assegnare la feature r_i ad un documento significa impostare a true l'i-mo elemento di V; se una data feature non è presente nel documento, il corrispondente elemento nel vettore viene posto a false.

Ad esempio se l'insieme delle feature è:

R = {…, informatica, information retrieval, intelligenza artificiale, …}

il presente documento potrebbe essere rappresentato dal vettore:

[…, TRUE, TRUE, FALSE, …]

Un fabbisogno informativo è descritto da una espressione booleana, in cui gli operandi rappresentano dei concetti, mentre vengono ritenuti rilevanti tutti e solo quei documenti che soddisfano l'espressione suddetta.

La valutazione della query partiziona allora l'insieme dei documenti in due sottoinsiemi, quelli "rilevanti" e quelli "non rilevanti", ma non fornisce alcuna informazione sulla relativa probabilità che documenti ritenuti "rilevanti" soddisfino il f.i.

Il modello booleano è alla base della maggior parte dei servizi commerciali di retrieval, ma generalmente è considerato difficile da utilizzare. Inoltre, non classificando i documenti nella raccolta, ottiene bassi risultati di recall e precision.

Modello dello spazio vettoriale

Nel modello vettoriale, i documenti e le query sono rappresentati come vettori in uno spazio k-dimensionale; pertanto si definisce questo modello come *algebrico* [15].

Tale modello utilizza nuovamente il vettore di feature dove i documenti V_D e le query Q sono rappresentati da vettori di lunghezza k, questa volta a valori reali, in cui ogni elemento costituisce un peso. Possono essere utilizzate diverse tecniche per determinare i vari pesi, ma comunemente ci si basa sulla frequenza di un termine in un singolo documento e nell'intera raccolta. Il confronto fra documenti e query avviene utilizzando una funzione di somiglianza: ad esempio il coseno dell'angolo tra i vettori.

Questo modello è storicamente importante, avendo avviato un filone di ricerca fin dagli anni '60, ma viene criticato per il fatto di essere un modello ad hoc, peraltro povero di giustificazioni teoriche.

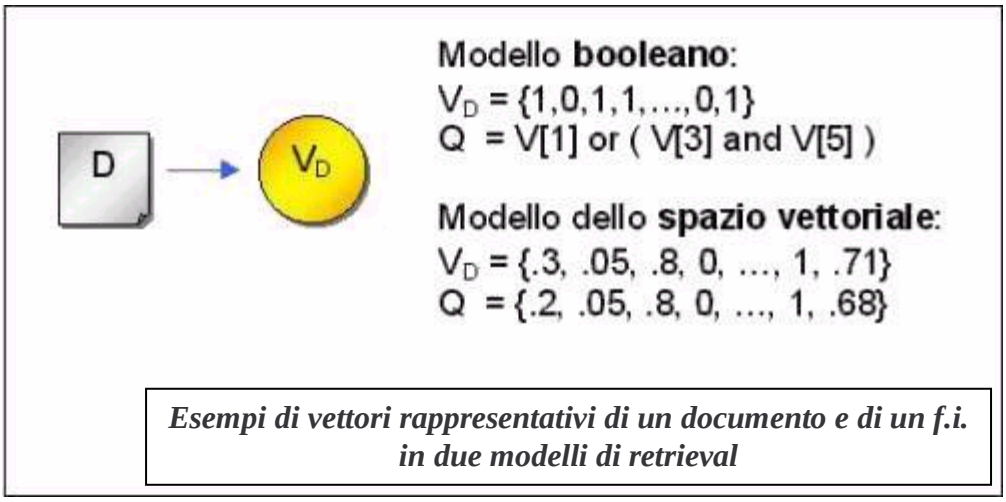


Figura 3.5 – Modello booleano

Un'evoluzione di questo modello in grado di tenere conto della somiglianza fra i documenti è quello basato sui cluster (grappoli). A fondamento del modello è la Cluster Hypothesis, che afferma che documenti simili sono conformi ad uno stesso f.i.

Invece che paragonare le rappresentazioni dei singoli documenti alla rappresentazione del f.i., viene effettuata una prima catalogazione dei documenti suddividendoli in cluster (in base ad una qualche misura di somiglianza) e per ogni cluster si crea un documento "medio" che rappresenta i documenti corrispondenti. Il processo di retrieval restituisce tutti quei documenti appartenenti a cluster che soddisfano la query.

E' da notare come questo modello non fornisca una definizione standard della rilevanza, dipendendo quest'ultima dai cluster utilizzati e dai criteri adottati per la riduzione di documenti simili ad uno stesso cluster. Infine, esistono diversi criteri per identificare i cluster da recuperare, specie se si ricorre a tecniche di strutturazione che permettono la navigazione nella gerarchia dei cluster.

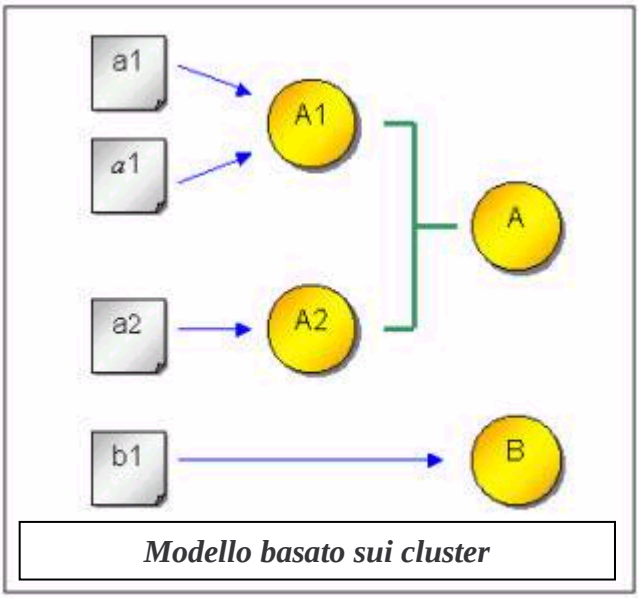


Figura 3.6 – Modello basato su cluster

Modello probabilistico

Un modello simile al *vector-space model*, ma con una più solida base teorica, è il modello probabilistico. Il principio alla base del modello è detto Probabilistic Ranking Principal e afferma che la miglior efficacia di retrieval si raggiunge quando i documenti sono classificati in ordine decrescente della probabilità di rilevanza.

Il documento d_i e il f.i. f_j vengono rappresentati da un vettore booleano a k dimensioni. Detto F l'insieme delle rappresentazioni per i fabbisogni informativi e D l'insieme delle rappresentazioni dei documenti, si definisce uno spazio degli eventi F x D. L'obiettivo è quello di determinare le coppie (d_i, f_j) rilevanti, cioè stimare la probabilità $P(R \mid d_i, f_j)$.

Di fatto si fanno delle supposizioni sulla indipendenza della distribuzione delle feature nei documenti e nelle query.

Successivamente si effettua una stima delle probabilità che le singole feature siano assegnate a documenti "rilevanti" o "non rilevanti", e si utilizza il teorema di Bayes per derivare una funzione di classificazione che calcoli $P(R \mid d_i, f_j)$ in termini di queste probabilità. Nella pratica, esistono differenti modelli probabilistici, ciascuno caratterizzato da un particolare insieme di supposizioni di indipendenza, e quindi da differenti funzioni di classificazione.

Si noti, infine, come i vari modelli probabilistici e quello dello spazio vettoriale, possano essere considerati delle estensioni del modello booleano, in quanto fanno ricorso al vettore di feature e consentono ricerche in termini di corrispondenza parziale (partial matching) delle feature tra query e documenti.

Capitolo 4 Prove sperimentali

Scopo di questa sperimentazione è di individuare quale approccio i programmi di traduzione assistita di tipo commerciale adottano per individuare similitudini tra la frase sottoposta e quelle presenti in memoria.

In particolare si adotta il seguente schema sperimentale: si realizza inizialmente una memoria a partire da un testo precedentemente tradotto realizzando così una Translation Memory; utilizzando questa TM viene sottoposto un nuovo testo da tradurre tramite programma CAT. I risultati ottenuti sono indicati tramite alcuni valori numerici di "distanza" dal tesato presente in memoria. In generale infatti un programma CAT indica che la traduzione suggerita è valida o meno fornendo un valore numerico di somiglianza che va dal 50% al 100% (approximate match o exact match).

Si è scelto di procedere utilizzando un testo descrittivo di un software (DeLuxe Paint III) e crearne così una TM. Successivamente si è

proceduto a effettuare alcune modifiche al testo al fine di valutare quali siano le penalità applicate dai programmi CAT.

```
DELUXEPAINT III is a graphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

Figura 4.1 - Testo originale inserito in memoria

4.1 Valutazione delle penalizzazioni

Per penalizzazione si intende il grado di distanza/somiglianza che il programma assegna ad un segmento sottoposto a traduzione.

4.1.1 Trados

4.1.1.1 Eliminazione di una parola di 8 lettere

Nel brano sottoposto (originariamente composto da 109 parole per 605 caratteri - spazi compresi) si è proceduto ad eliminare una parola di 8 lettere (108 parole-596 caratteri).

```
DELUXEPAINT III is a graphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

Figura 4.2 - Eliminazione di una parola di 8 lettere

In Trados l'analisi di tipo batch (off-line) indica i seguenti valori:

```
Start Analyse: Sat May 11 14:06:50 2002
Translation Memory: Prova analisi piccole
modifiche\mem_piccole_modifiche.tmw
Prova analisi piccole modifiche\DP3_modificato.doc

 Chars/Word       4.55
 Chars Total       492

Analyse Total (1 file):

 Match Types  Segments        Words Percent Placeables
 XTranslated         0            0       0          0
 Repetitions         0            0       0          0
 100%                0            0       0          0
```

```
 95% - 99%          2          83      77          0
 85% - 94%          1          25      23          0
 75% - 84%          0           0       0          0
 50% - 74%          0           0       0          0
 No Match           0           0       0          0
 Total              3         108     100          0

 Chars/Word       4.55
 Chars Total       492

Analyse finished successfully without errors!

Sat May 11 14:06:51 2002
=================================================================
```

Figura 4.3 - Report analisi batch Trados - Eliminazione di una parola di 8 lettere

Volendo valutare l'impatto che l'eliminazione della parola comporta nella fase di analisi da parte di Trados, si procede con la pretraduzione del testo che permetterà di analizzare la segmentazione e la penalizzazione valutata dal programma CAT.

La parola "graphics" è stata eliminata

```
<0>DELUXEPAINT·III·is·a·tool·that·can·help·you·create·works·of·art·with·an·ease·
and·precision·that·you·may·never·have·thought·possible.<193>DELUXEPAINT·III·is·a·
graphics·tool·that·can·help·you·create·works·of·art·with·an·ease·and·precision·
that·you·may·never·have·thought·possible.<0>·<0>After·spending·a·little·time·with·
this·manual,·you'll·be·able·to·use·the·program·to·create·colorful·graphics·in·a·
fraction·of·the·time·it·would·take·using·more·traditional·techniques.<197>After·
spending·a·little·time·with·this·manual,·you'll·be·able·to·use·the·program·to·
create·colorful·graphics·in·a·fraction·of·the·time·it·would·take·using·more·
traditional·techniques.<0>·<0>You'll·learn·how·to·create·perspective·effects,·how·
to·create·and·save·your·own·custom·brushes,·how·to·mix·your·own·color·palette·
fom·a·universe·of·4096·possible·colors,·and·how·to·create·effective·on-screen·
animations,·to·name·just·a·few·of·DeluxePaint·III's·powerful·
features.<197>You'll·learn·how·to·create·perspective·effects,·how·to·create·and·
save·your·own·custom·brushes,·how·to·mix·your·own·color·palette·fom·a·universe·
of·4096·possible·colors,·and·how·to·create·effective·on-screen·animations,·to·
name·just·a·few·of·DeluxePaint·III's·powerful·features.<0>¶
```

Figura 4.4 - Pretradotto - Eliminazione di una parola di 8 lettere

Dalla schermata qui sopra si comprende che il programma ha penalizzato le frasi identiche a quelle in memoria con un fattore del

3% a causa del recupero della frase da un progetto di allineamento fatto con WinAlign (questo dato è chiaramente dichiarato nella manualistica di Trados e modificabile dall'utente ad esempio impostando questa penalizzazione di default a 0%). La prima frase, che è stata oggetto dell'eliminazione della parola, è stata penalizzata di un ulteriore **4 %** arrivando ad un fuzzy match del 93%.

La segmentazione ha spezzato il testo in corrispondenza del punto come da parametrizzazione di default del programma.

Se si procede con l'eliminazione di una parola di 3 lettere, risulta:

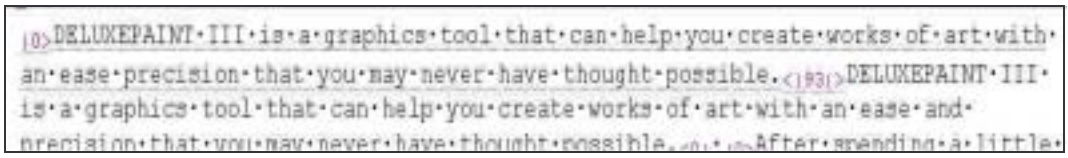


Figura 4.5 - Pretradotto - Eliminazione di una parola di 3 lettere

Penalizzazione del **4%**.

Sembrerebbe quindi che la penalizzazione non dipenda dalla lunghezza della parola eliminata.

Si procede ora ad eliminare 2 parole dal testo campione ("and precision") su un totale di 127 caratteri vengono eliminati quindi 13 caratteri (riduzione del 10% circa):

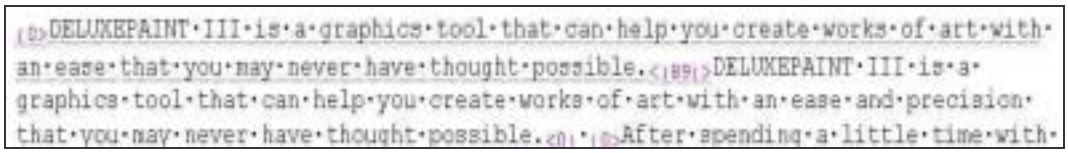


Figura 4.6 - Pretradotto - Eliminazione di due parole

La penalizzazione passa al **8%**. Una parola di 9 lettere e uno spazio hanno quindi incrementato la penalizzazione di un ulteriore 4%.

Si procede ora ad eliminare 3 parole dal testo campione ("graphics", "and precision"):

DELUXEPAINT·III·is·a·tool·that·can·help·you·create·works·of·art·with·an·ease·
that·you·may·never·have·thought·possible.DELUXEPAINT·III·is·a·graphics·tool·
that·can·help·you·create·works·of·art·with·an·ease·and·precision·that·you·may·
never·have·thought·possible.·After·spending·a·little·time·with·this·manual,·

Figura 4.7 - Pretradotto - Eliminazione di tre parole

La penalizzazione passa al **13%** (sono state eliminate tre parole per un totale di 20 lettere e tre spazi). Ulteriore incremento quindi di 5 punti percentuali (da sola la parola "`graphics`" aveva dato una penalizzazione del 4%).

Da quest'ultimo dato si comprende come la penalizzazione applicata dal programma Trados non dipenda dalla sola quantità di parole eliminate ma anche da altri parametri quali la lunghezza totale della frase, che in questo ultimo caso è sostanzialmente più corta di quella della prima prova in cui si è eliminata la sola parola "`graphics`".

L'algoritmo di penalizzazione di Trados tiene pertanto conto non solo della quantità di parole eliminate ma anche della lunghezza totale della frase nuova rispetto a quella nella TM.

4.1.1.2 Aggiunta di una parola di n lettere

Si procede ad aggiungere una parola di 8 lettere al testo originale al fine di valutare la distanza assegnata dal programma Trados a tale modifica:

```
DELUXEPAINT III is a standard graphics tool that can help you
create works of art with an ease and precision that you may never
have thought possible. After spending a little time with this
manual, you'll be able to use the program to create colorful
graphics in a fraction of the time it would take using more
traditional techniques. You'll learn how to create perspective
effects, how to create and save your own custom brushes, how to
mix your own color palette from a universe of 4096 possible
colors, and how to create effective on-screen animations, to name
just a few of DeluxePaint III's powerful features.
```

Figura 4.8 - Aggiunta di una parola di 8 lettere

```
{0}DELUXEPAINT·III·is·a·standard·graphics·tool·that·can·help·you·create·works·of·
art·with·an·ease·and·precision·that·you·may·never·have·thought·
possible.<}94{>DELUXEPAINT·III·is·a·graphics·tool·that·can·help·you·create·works·
of·art·with·an·ease·and·precision·that·you·may·never·have·thought·possible.<0}·
{0}After·spending·a·little·time·with·this·manual,·you'll·be·able·to·use·the·
program·to·create·colorful·graphics·in·a·fraction·of·the·time·it·would·take·
using·more·traditional·techniques.<}97{>After·spending·a·little·time·with·this·
manual,·you'll·be·able·to·use·the·program·to·create·colorful·graphics·in·a·
fraction·of·the·time·it·would·take·using·more·traditional·techniques.<0}·
{0}You'll·learn·how·to·create·perspective·effects,·how·to·create·and·save·your·
own·custom·brushes,·how·to·mix·your·own·color·palette·fom·a·universe·of·4096·
possible·colors,·and·how·to·create·effective·on-screen·animations,·to·name·just·
a·few·of·DeluxePaint·III's·powerful·features.<}97{>You'll·learn·how·to·create·
perspective·effects,·how·to·create·and·save·your·own·custom·brushes,·how·to·mix·
your·own·color·palette·fom·a·universe·of·4096·possible·colors,·and·how·to·create·
effective·on-screen·animations,·to·name·just·a·few·of·DeluxePaint·III's·powerful·
features.<0}¶
```

Figura 4.9 - Pretradotto - Aggiunta di una parola di 8 lettere

Il risultato è una penalizzazione del **3%** (97%-3%=94%)

Se si aggiunge invece una parola composta da 12 lettere il valore rimane lo stesso (la penalizzazione resta del 3%). L'algoritmo nuovamente non tiene in considerazione la lunghezza della parola aggiunta.

Prova ulteriore con due parole aggiunte (per un totale di 8+3+9=20 caratteri + 3 spazi=23):

```
DELUXEPAINT III is a standard and efficient graphics tool that can
help you create works of art with an ease and precision that you
may never have thought possible. After spending a little time with
this manual, you'll be able to use the program to create colorful
graphics in a fraction of the time it would take using more
traditional techniques. You'll learn how to create perspective
effects, how to create and save your own custom brushes, how to
mix your own color palette from a universe of 4096 possible
colors, and how to create effective on-screen animations, to name
just a few of DeluxePaint III's powerful features.
```

Figura 4.10 - Aggiunta di due parole

```
DELUXEPAINT·III·is·a·standard·and·efficient·graphics·tool·that·can·help·you·
create·works·of·art·with·an·ease·and·precision·that·you·may·never·have·thought·
possible.DELUXEPAINT·III·is·a·graphics·tool·that·can·help·you·create·works·
of·art·with·an·ease·and·precision·that·you·may·never·have·thought·possible.
```

Figura 4.11 - Pretradotto - Aggiunta di due parole

La penalizzazione passa al **8%**. Anche in questo caso quindi l'aggiunta di x caratteri non influenza in maniera direttamente proporzionale il valore di penalizzazione indicato da Trados, il quale considera anche la lunghezza della frase nuova sottoposta a traduzione.

4.1.1.3 Modifica di una parola

Si vuole ora procedere alla modifica di una sola lettera di una parola in un segmento già presente in memoria per individuare il valore di penalizzazione individuato da Trados con tale modifica. Si procede ad eliminare la "s" finale della parola "graphics"

```
DELUXEPAINT III is a graphic tool that can help you create works
of art with an ease and precision that you may never have thought
possible.
```

Figura 4.12 - Modifica di una parola

```
{0>DELUXEPAINT·III·is·a·graphic·tool·that·can·help·you·create·works·of·art·with·
an·ease·and·precision·that·you·may·never·have·thought·possible.<{96{>DELUXEPAINT·
III·is·a·graphics·tool·that·can·help·you·create·works·of·art·with·an·ease·and·
precision·that·you·may·never·have·thought·possible.<{0{·{0>After·spending·a·little·
```

Figura 4.13 - Pretradotto - Modifica di una parola

La penalizzazione è di un solo punto percentuale (**1%**).

L'eliminazione di una seconda lettera nella stessa parola comporta una penalizzazione sempre del 1%.

L'eliminazione di due lettere in due parole diverse comporta invece una penalizzazione del **2%**.

L'aggiunta di una lettera ad una parola comporta una penalizzazione del **1%**.

L'aggiunta di due lettere in due parole diverse comporta una penalizzazione del **4%**.

Il sistema Trados dunque penalizza maggiormente eventuali lettere aggiunte piuttosto che quelle eliminate.

4.1.1.4 Unione di due frasi

Si procede ora a congiungere due intere frasi al fine di determinare come il programma CAT valuti queste due frasi già presenti singolarmente in TM. La congiunzione tra le due è in una prima fase effettuata con l'"and". In una seconda fase verranno congiunte con una semplice virgola. In una terza fase le due frasi verranno messe una all'interno dell'altra.

- Prima prova: "A and B"
- Seconda prova: "A, B"
- Terza prova: "A1, B, A2"

Prima prova:

```
DELUXEPAINT III is a graphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible and after spending a little time with this manual, you'll
be able to use the program to create colorful graphics in a
fraction of the time it would take using more traditional
techniques. You'll learn how to create perspective effects, how to
create and save your own custom brushes, how to mix your own color
palette from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

Figura 4.14 - Prova "A and B"

```
DELUXEPAINT·III·is·a·graphics·tool·that·can·help·you·create·works·of·art·with·
an·ease·and·precision·that·you·may·never·have·thought·possible·and·after·
spending·a·little·time·with·this·manual,·you'll·be·able·to·use·the·program·to·
create·colorful·graphics·in·a·fraction·of·the·time·it·would·take·using·more·
traditional·techniques.DELUXEPAINT·III·is·a·graphics·tool·that·can·help·you·
create·works·of·art·with·an·ease·and·precision·that·you·may·never·have·thought·
possible·and·after·spending·a·little·time·with·this·manual,·you'll·be·able·to·
use·the·program·to·create·colorful·graphics·in·a·fraction·of·the·time·it·would·
take·using·more·traditional·techniques.'You'll·learn·how·to·create·
```

Figura 4.15 - Pretradotto - Prova "A and B"

Il risultato dell'analisi batch è incredibilmente del 0%.

Lo stesso risultato si ottiene usando la virgola come separatrice tra la frase A e la frase B (*seconda prova*).

Il problema è che nelle regole di separazione dei paragrafi (segmentazione) la virgola non è considerata. Si tenga conto però dell'alto grado di personalizzazione possibile in Trados:

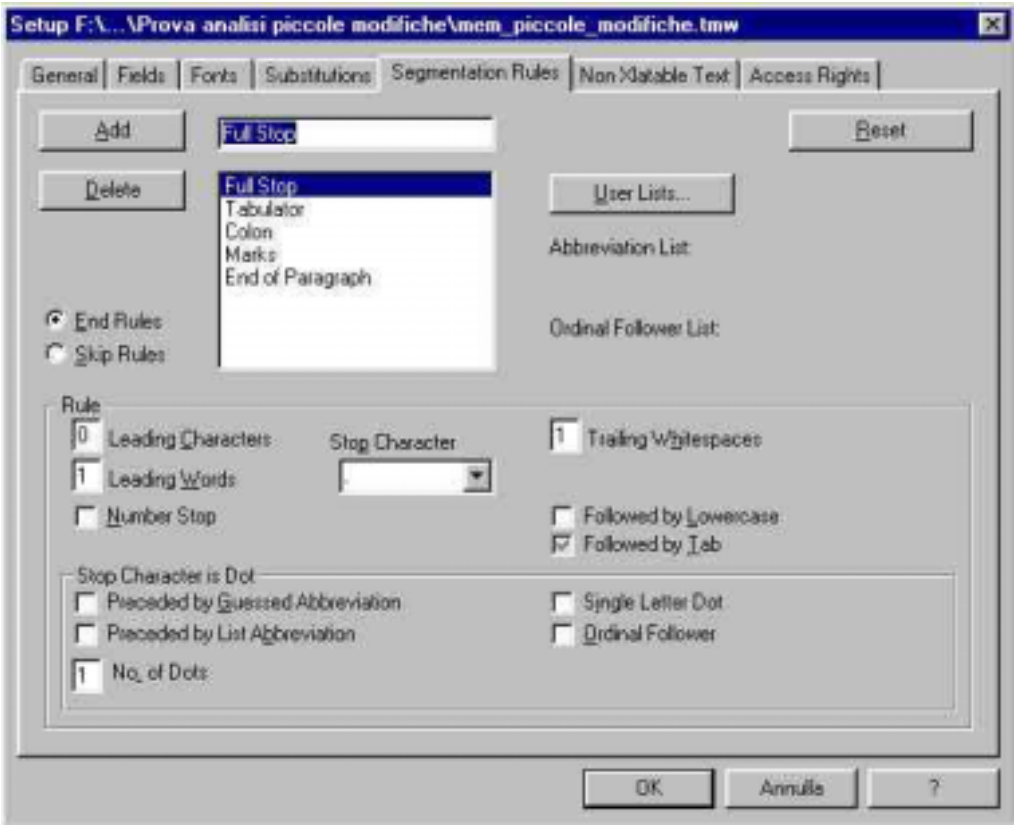


Figura 4.16 - Trados - Impostazione di nuove regole di segmentazione

Dopo l'introduzione di una nuova regola di segmentazione basata sulla virgola ("comma" in inglese), si ottiene (*Figura 4.17*):

DELUXEPAINT·III·is·a·graphics·tool·that·can·help·you·create·works·of·art·with·an·
ease·and·precision·that·you·may·never·have·thought·possible,·<10>After·spending·a·
little·time·with·this·manual,·you'll·be·able·to·use·the·program·to·create·
colorful·graphics·in·a·fraction·of·the·time·it·would·take·using·more·traditional·
techniques.<197>After·spending·a·little·time·with·this·manual,·you'll·be·able·to·
use·the·program·to·create·colorful·graphics·in·a·fraction·of·the·time·it·would·
take·using·more·traditional·techniques.<0>·<10>You'll·learn·how·to·create·
perspective·effects,·how·to·create·and·save·your·own·custom·brushes,·how·to·mix·
your·own·color·palette·fom·a·universe·of·4096·possible·colors,·and·how·to·create·
effective·on-screen·animations,·to·name·just·a·few·of·DeluxePaint·III's·powerful·
features.<197>You'll·learn·how·to·create·perspective·effects,·how·to·create·and·
save·your·own·custom·brushes,·how·to·mix·your·own·color·palette·fom·a·universe·
of·4096·possible·colors,·and·how·to·create·effective·on-screen·animations,·to·
name·just·a·few·of·DeluxePaint·III's·powerful·features.<0>¶

Figura 4.17 - Pretradotto dopo l'introduzione di una nuova regola di segmentazione

Il programma individua correttamente la frase B come presente in memoria ma non trova la frase A (ora leggermente diversa da quella in TM in quanto termina con una virgola e non con un punto).

In conclusione i simboli di segmentazione sono determinanti per la corretta individuazione in TM della frase ricercata. Inoltre, anche se altamente personalizzabile, il programma non considera tali simboli di segmentazione come secondari, ma al contrario assegna loro una grande importanza. Per fare una controprova, si vuole valutare cosa accade nel togliere semplicemente il punto di separazione tra le due frasi (*Figura 4.18*):

DELUXEPAINT·III·is·a·graphics·tool·that·can·help·you·create·works·of·art·with·an·
ease·and·precision·that·you·may·never·have·thought·possible·After·spending·a·
little·time·with·this·manual,·you'll·be·able·to·use·the·program·to·create·
colorful·graphics·in·a·fraction·of·the·time·it·would·take·using·more·traditional·
techniques.·<10>You'll·learn·how·to·create·perspective·effects,·how·to·create·and·
save·your·own·custom·brushes,·how·to·mix·your·own·color·palette·fom·a·universe·
of·4096·possible·colors,·and·how·to·create·effective·on-screen·animations,·to·
name·just·a·few·of·DeluxePaint·III's·powerful·features.<197>You'll·learn·how·to·
create·perspective·effects,·how·to·create·and·save·your·own·custom·brushes,·how·
to·mix·your·own·color·palette·fom·a·universe·of·4096·possible·colors,·and·how·to·
create·effective·on-screen·animations,·to·name·just·a·few·of·DeluxePaint·III's·
powerful·features.<0>¶

Figura 4.18 - Pretradotto dopo l'eliminazione del punto di separazione

Il programma in questo caso infatti non distingue più le due frasi e non trova alcuna corrispondenza in TM (indicatore a 0).

Terza prova (A1, B, A2):

DELUXEPAINT III, after spending a little time with this manual, you'll be able to use the program to create colorful graphics in a fraction of the time it would take using more traditional techniques, **is a graphics tool that can help you create works of art with an ease and precision that you may never have thought possible**. You'll learn how to create perspective effects, how to create and save your own custom brushes, how to mix your own color palette from a universe of 4096 possible colors, and how to create effective on-screen animations, to name just a few of DeluxePaint III's powerful features.

Figura 4.19 - Trados - "A1, B, A2"

{0>DELUXEPAINT·III,·after·spending·a·little·time·with·this·manual,·you'll·be·able·to·use·the·program·to·create·colorful·graphics·in·a·fraction·of·the·time·it·would·take·using·more·traditional·techniques,·is·a·graphics·tool·that·can·help·you·create·works·of·art·with·an·ease·and·precision·that·you·may·never·have·thought·possible.<}93}>After·spending·a·little·time·with·this·manual,·you'll·be·able·to·use·the·program·to·create·colorful·graphics·in·a·fraction·of·the·time·it·would·take·using·more·traditional·techniques.<0}·{0>You'll·learn·how·to·create·perspective·effects,·how·to·create·and·save·your·own·custom·brushes,·how·to·mix·your·own·color·palette·fom·a·universe·of·4096·possible·colors,·and·how·to·create·effective·on-screen·animations,·to·name·just·a·few·of·DeluxePaint·III's·powerful·features.<}97}>You'll·learn·how·to·create·perspective·effects,·how·to·create·and·save·your·own·custom·brushes,·how·to·mix·your·own·color·palette·fom·a·universe·of·4096·possible·colors,·and·how·to·create·effective·on-screen·animations,·to·name·just·a·few·of·DeluxePaint·III's·powerful·features.<0}¶

Figura 4.20 - Pretradotto - Trados - "A1, B, A2"

L'analisi di tipo batch suggerisce una somiglianza del **50% circa**, indicando come possibile traduzione la sola parte B e ignorando completamente la parte A che è stata spezzata in due parti.

Analisi di tipo on-line

Se si procede con l'utilizzo dell'interfaccia utilizzata dal traduttore professionista l'interazione con il programma WorkBench permette il suggerimento di frasi presenti in memoria sia per la parte A che per la parte B, previa evidenziazione e ricerca di tipo "Concordance":

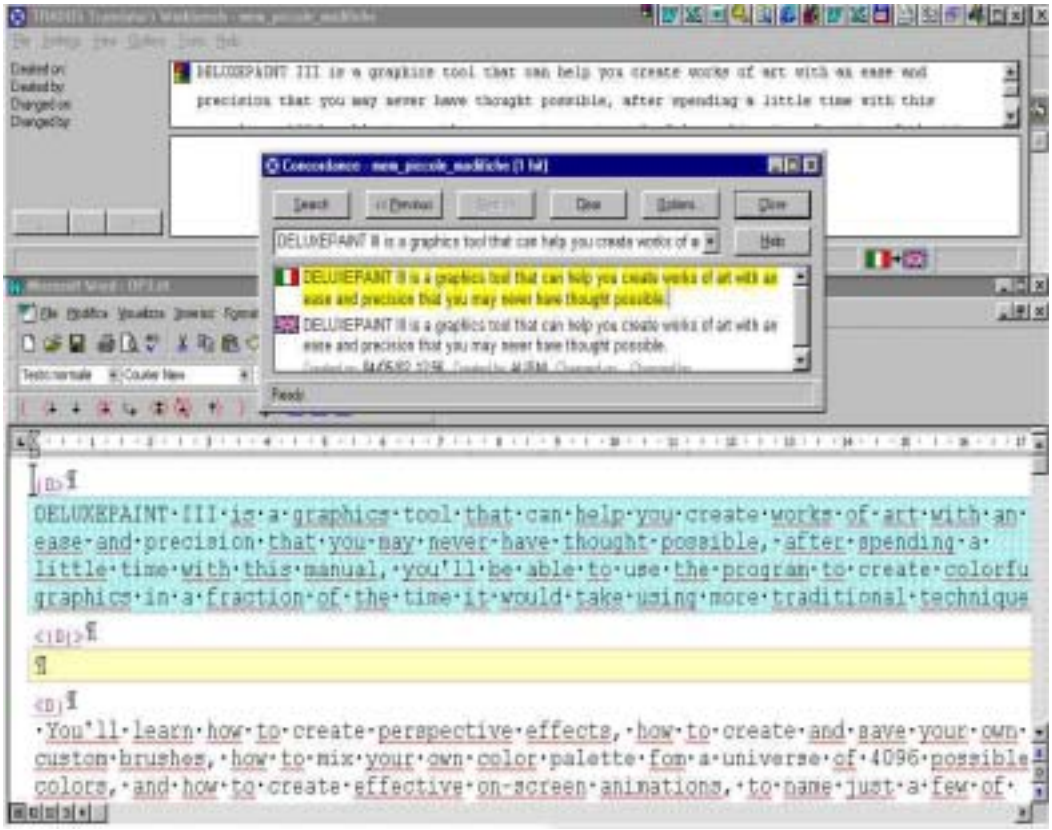


Figura 4.21 - Trados - Analisi on-line

Conclusioni sull'analisi di tipo batch

Il sistema Trados non è in grado di riconoscere parti di frasi come presenti in memoria, ma tratta solo segmenti strutturalmente simili a quelli in memoria. La memoria di Trados, o meglio l'algoritmo che Trados utilizza per "leggere" la propria memoria ed analizzare un testo da tradurre, non possiede la capacità di analizzare una frase/segmento suddividendola in sottoparti e cercando queste ultime nella TM.

Si noti però come il prodotto nel suo aspetto on-line risulti essere invece buon suggeritore di dove e come sono presenti differenze tra memoria e testo da tradurre, arrivando addirittura assieme ad altri programmi della suite (leggi MultiTerm) a fornire ulteriori aiuti al traduttore.

4.1.1.5 Stemming in Trados

In questo paragrafo ci si prefigge il compito di analizzare il sistema Trados in relazione alle cosiddette "radici" di una parola. Si vuole infatti determinare se il sistema è in grado di individuare le varianti di un termine, singolari/plurali, coniugazioni di verbi, ecc. e se quindi riesce a trovare similitudine tra un segmento nuovo e uno in memoria simile.

Per stemming si intende appunto la capacità di estrarre da un termine la sua radice eliminando prefissi e suffissi:

connected, connecting, connection, connections ➔ *connect*

Si vuole quindi determinare se il sistema è in grado di suggerire varianti di un termine.

Prima prova - approccio batch

La prova consisterà quindi nella modifica di un paio di termini nel testo originale per sottoporlo poi ad un'analisi di tipo batch:

```
DELUXEPAINT III is a graph tool that could help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

Figura 4.22 - Trados - Stemming

Si procede modificando la parola "graphics" in "graph" e "can" in "could".

Il risultato ottenuto è il seguente:

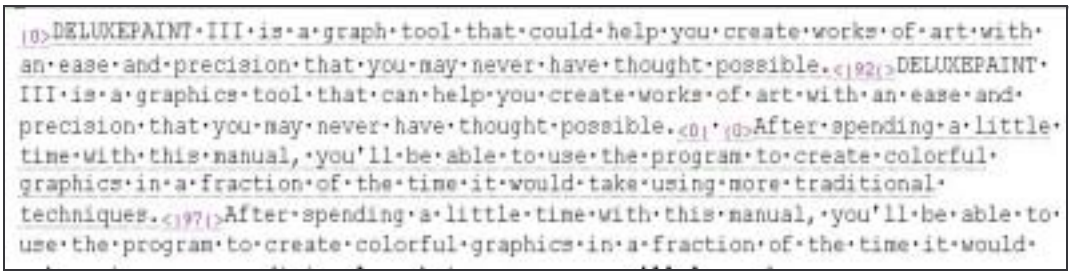


Figura 4.23 - Pretradotto - Trados - Stemming

La pretraduzione di Trados individua una similitudine tra la prima frase e l'originale in memoria pari al **95%** (penalizzazione pari a 5%) senza però segnalare quali termini siano "lontani" dalla memoria.

Prima prova - approccio on-line

Se invece del procedimento batch si utilizza l'aspetto on-line di Trados è interessante vedere come il sistema in effetti individui esattamente i termini modificati, senza peraltro suggerire possibili traduzioni per essi:

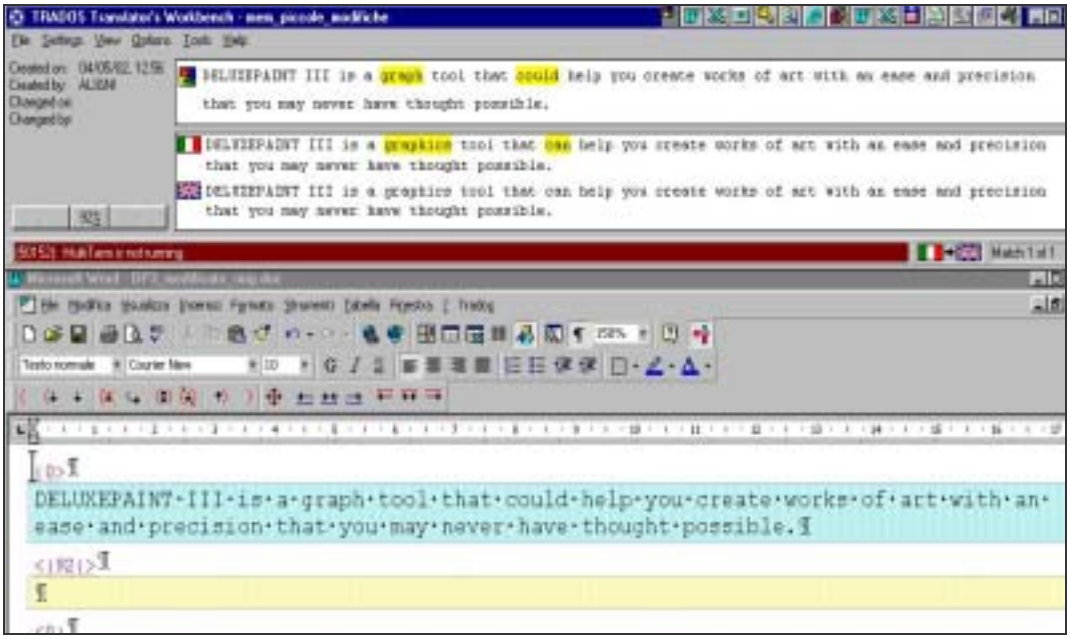


Figura 4.24 - Trados - Stemming - on-line

Si noti come il programma comunque segnali l'assenza del componente "MultiTerm" il quale come glossario potrebbe contenere suggerimenti validi per i termini evidenziati.

Seconda prova - approccio batch

Si vuole ora provare a valutare il comportamento di Trados in relazione alla modifica della parola "graphics" in "raphics" e di "graphics" in "graphic":

```
DELUXEPAINT III is a raphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.

DELUXEPAINT III is a graphic tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

Figura 4.25 - Trados - Stemming

Lo scopo è quello di comprendere quanto il sistema penalizzi questa differenza di un solo carattere. Il sistema infatti dovrebbe penalizzare maggiormente il primo caso in cui i due termini non hanno linguisticamente niente in comune, e penalizzare meno il secondo caso.

Il risultato ottenuto è il seguente:

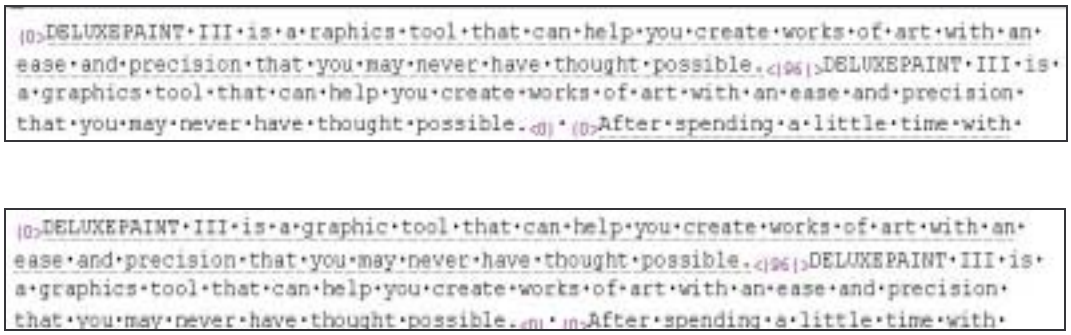
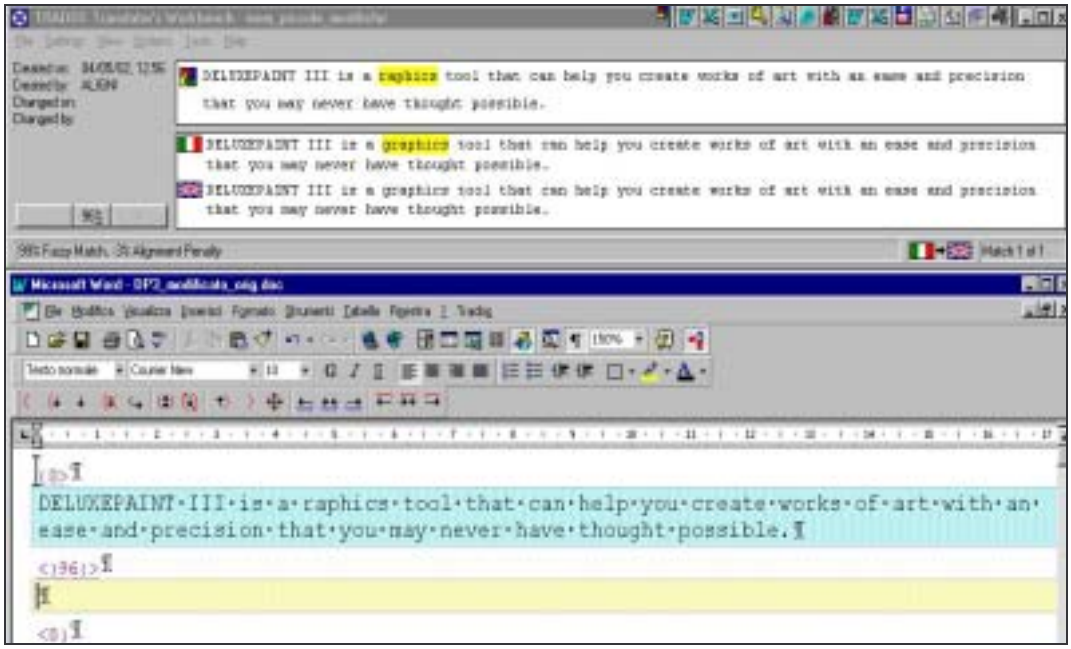


Figura 4.26 - Pretradotto - Trados - Stemming

Il sistema sia nel primo caso ("graphics" in "raphics") che nel secondo ("graphics" in "graphic") penalizza con il medesimo **1%** non distinguendo i due casi.

Seconda prova - approccio on-line

Se si utilizza la procedura on-line del programma si ottengono le seguenti segnalazioni:



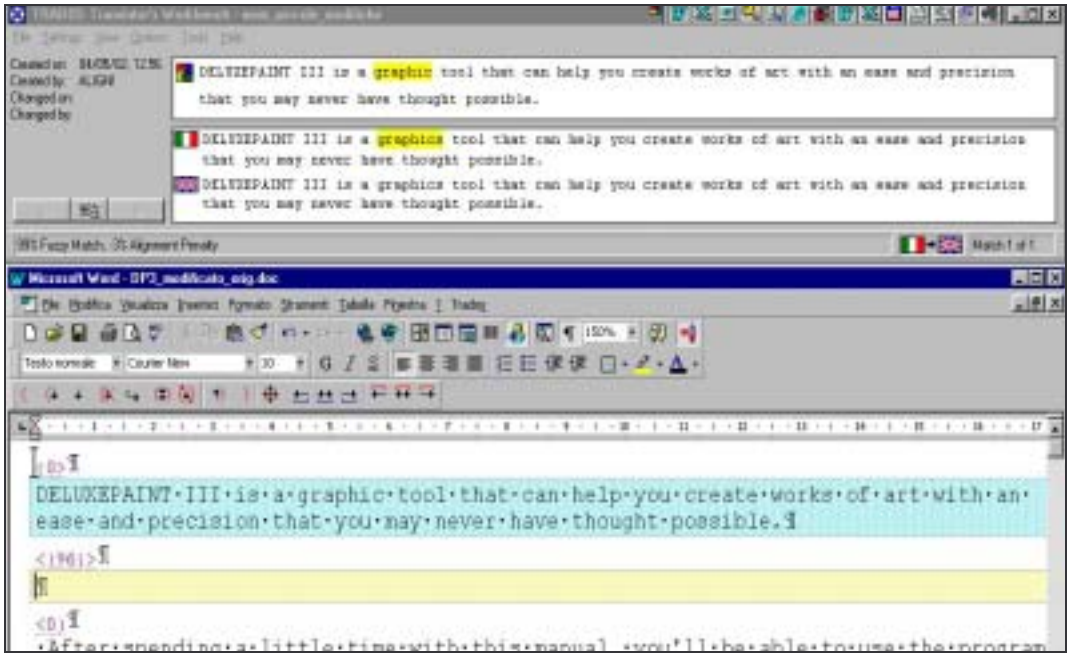


Figura 4.27 - Trados - Stemming - on-line

In cui si vede che il sistema semplicemente evidenzia la parola diversa senza valutare in cosa lo sia, confermando così che si tratta di un sistema che non valuta le singole parole ma interi segmenti (si veda a questo proposito il paragrafo §6.1.1 "Approccio grammaticale" a pagina 7).

4.1.1.6 Ordine delle parole

Si vuole ora analizzare il comportamento in relazione all'inversione di ordine delle parole all'interno di uno stesso segmento. Sono state quindi invertiti i termini "*graphics tool*" in "*tool graphics*" e "*ease and precision*" in "*precision and ease*":

```
DELUXEPAINT III is a tool graphics that can help you create works
of art with an precision and ease that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques
```

Figura 4.28 - Trados - Ordinamento delle parole

Il risultato:

```
{0>DELUXEPAINT·III·is·a·tool·graphics·that·can·help·you·create·works·of·art·with·
an·precision·and·ease·that·you·may·never·have·thought·possible.<{33{>DELUXEPAINT·
III·is·a·graphics·tool·that·can·help·you·create·works·of·art·with·an·ease·and·
precision·that·you·may·never·have·thought·possible.<0{·{0>After·spending·a·little·
```

Figura 4.29 - Pretradotto - Trados - Ordinamento delle parole

Anche in questo caso l'analisi batch non segnala le singole differenze ed indica una distanza pari al **4%**.

L'analisi on-line fornisce però anche in questo caso maggiori dettagli:

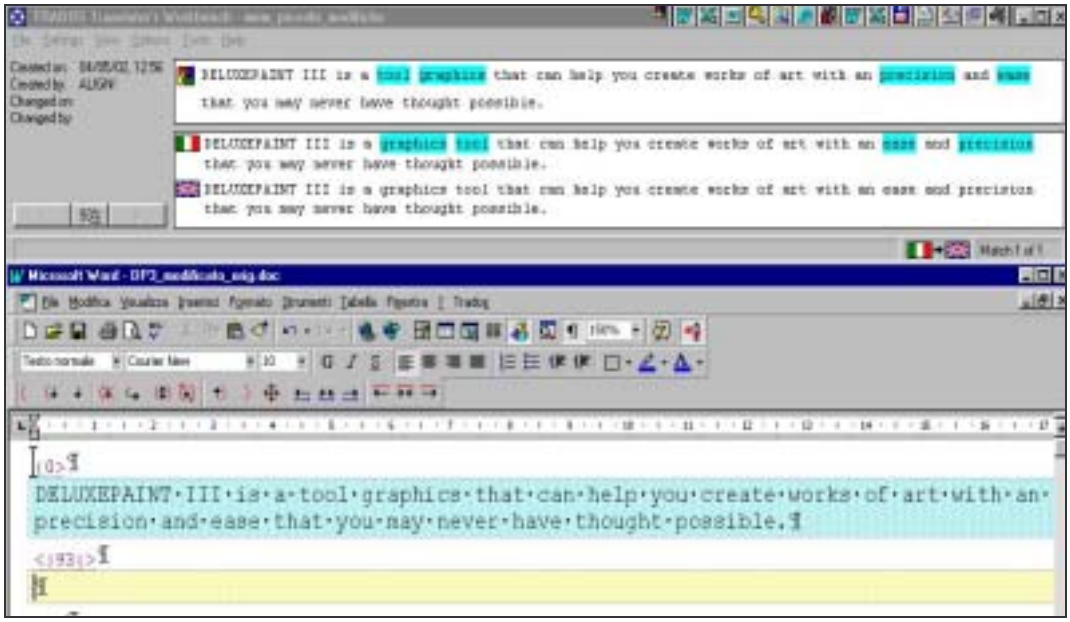


Figura 4.30 - Trados - Ordinamento delle parole - on-line

La segnalazione è precisa e arricchita da colori che indicano un semplice cambiamento di ordine delle parole. In particolare l'Help del programma dichiara:

Blue	Indicates that a part of the segment has moved. A clause like *for instance* can occur in several different places in the segment without changing its meaning. This means that the suggested translation does not always need further adaptation.

In altre parole il sistema segnala in azzurro (Blue) quelle parti di segmento che sono in realtà state solo spostate e che la traduzione del segmento non dovrebbe quindi essere molto diversa da quella proposta dal sistema stesso.

4.1.1.7 Conclusioni su stemming e ordine delle parole

Per quanto riguarda l'analisi stemming ed ordine delle parole è naturale ottenere dal programma suggerimenti non completi, perché Trados è un sistema basato da un lato su memorie di associazione tra segmenti e dall'altro da algoritmi che "pesano" le frasi da tradurre in relazione a quelle in memoria (*String based translation memory* - STM, chiamato anche TMS-*Translation Memory System*); non si tratta quindi di una Machine Translation (MT) dotata di algoritmi e soprattutto regole di sintassi linguistica proprie di ogni idioma e in grado quindi di distinguere semplici modifiche di singolari/plurali, coniugazione dei verbi, ecc. Tali ultimi dispositivi (*Lexeme-based translation memory* - LTM) sono descrivibili come insiemi di regole di inflessione e derivazione, mentre i sistemi STM, come Trados, semplicemente memorizzano la corrispondenza "esteriore" delle stringhe senza compiere nessun tipo di analisi lessicale.

In particolare Trados si basa su FindLink (database retrieval) creato da CONNEX nel 1996, il quale nella fase di "apprendimento" memorizza nel proprio database un insieme di segmenti in lingua sorgente e destinazione, codificando le stringhe in *n-grammi*, mentre in fase di "traduzione" la stringa ricercata è analizzata per individuare il livello di similarità con quelle in memoria, risultato fornito tramite il già visto indice tra 0% e 100%.

A proposito di questo si veda il Capitolo 6 "Limiti attuali dei programmi commerciali" ed in particolare il paragrafo §6.1.1 "Approccio grammaticale" a pagina 7.

4.1.2 IBM Translation Manager

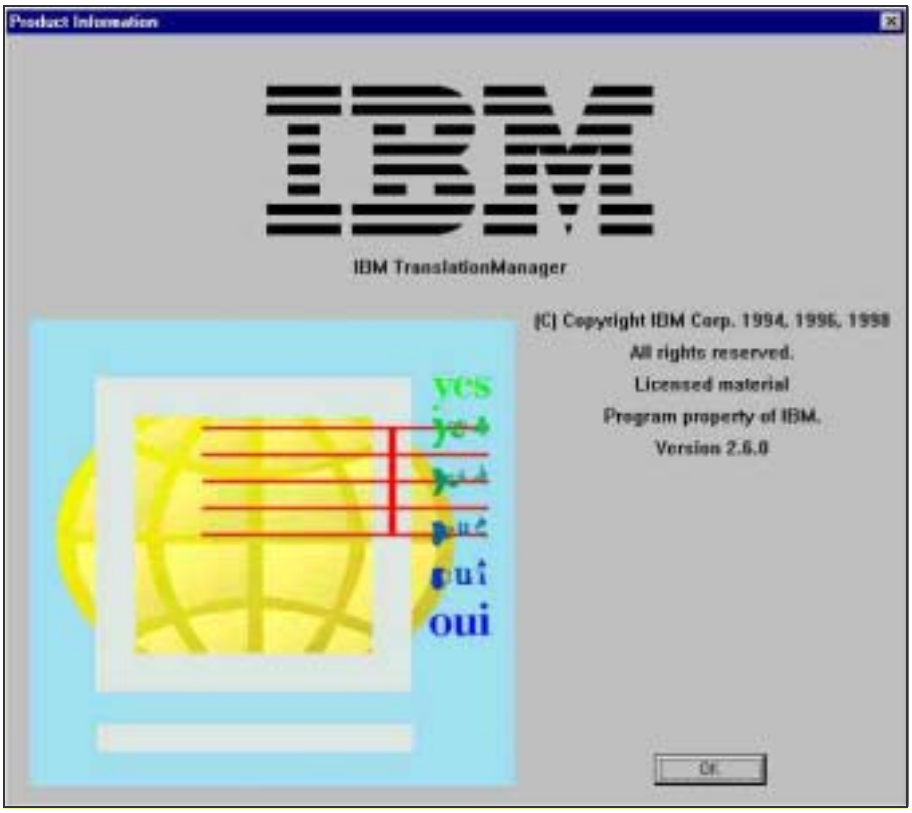


Figura 4.31 – IBM TranslationManager

4.1.2.1 Eliminazione di una parola di 8 lettere

Nel brano sottoposto si è proceduto ad eliminare una parola di 8 lettere.

```
DELUXEPAINT III is a graphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

L'analisi del nuovo testo privo della parola "`graphics`" è individuata dalla funzione *Analyze Documents* del programma IBM:

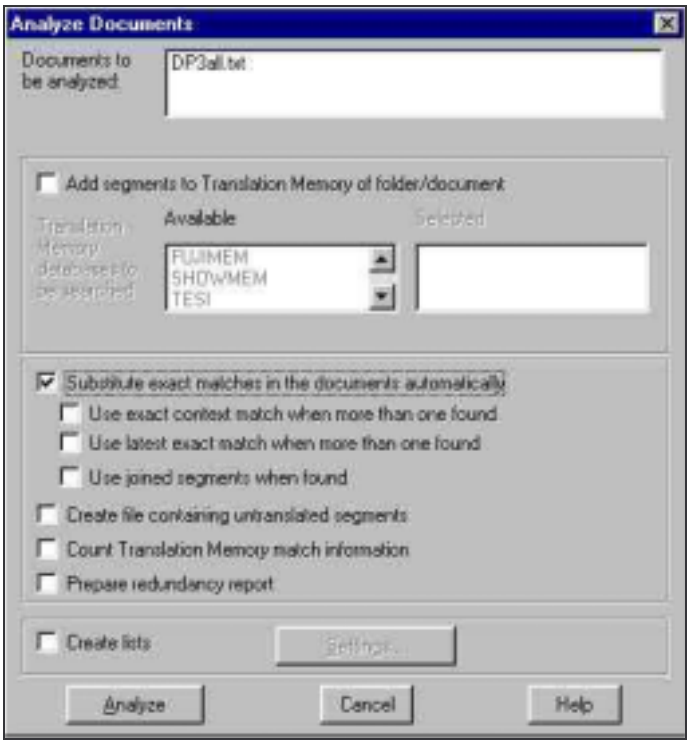


Figura 4.32 – IBM TranslationManager – Batch processing

I cui risultati visibili nella seguente schermata:

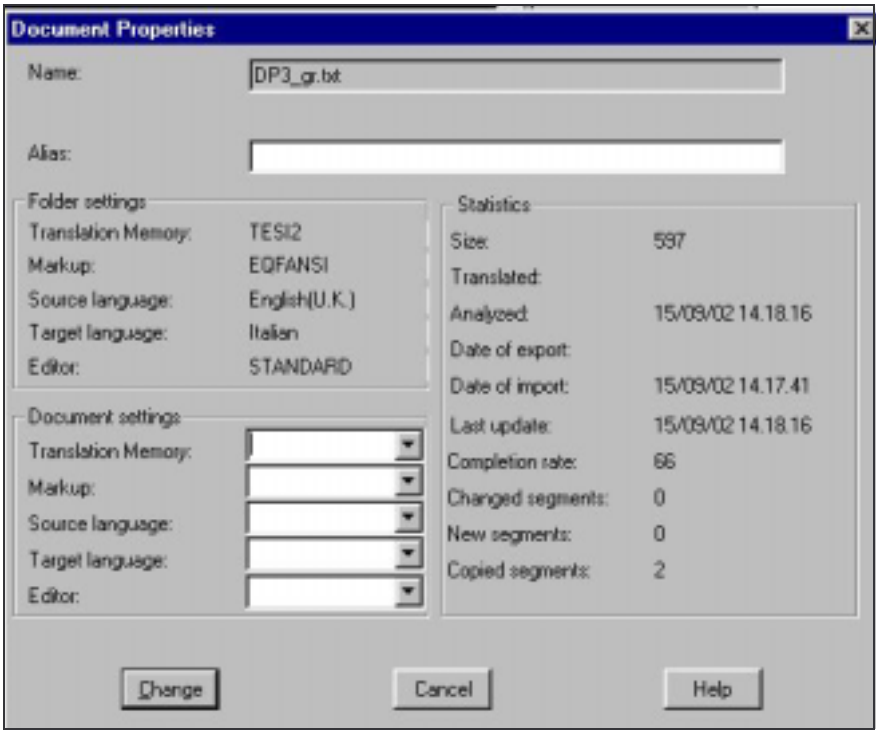


Figura 4.33 – IBM TranslationManager – Impostazione parametri del progetto

In altre parole, dei tre segmenti che costituiscono il campione preso ad esempio solo due sono stati riconosciuti come uguali al 100% (*copied segments*), mentre uno non è stato riconosciuto come tale (*Completion rate* = 66 indica che il 66% del testo originale è stato tradotto) (*Figura 4.33*).

La funzione conteggio presente nella schermata principale del programma evidenzia infatti che solo 81 parole su 106 sono tradotte, mentre le restanti 25 non sono traducibili.

Word Count Results - Translation

Documents	Translated	Untranslated	Target Words	Markup
DP3_gr.txt	81	25	70	0
=========	=========	=========	=========	=========
Total	81	25	70	0

Figura 4.34 – IBM TranslationManager – Analisi risultati

Passando ora alla funzionalità on-line vera e propria si possono valutare i risultati sopra indicati:

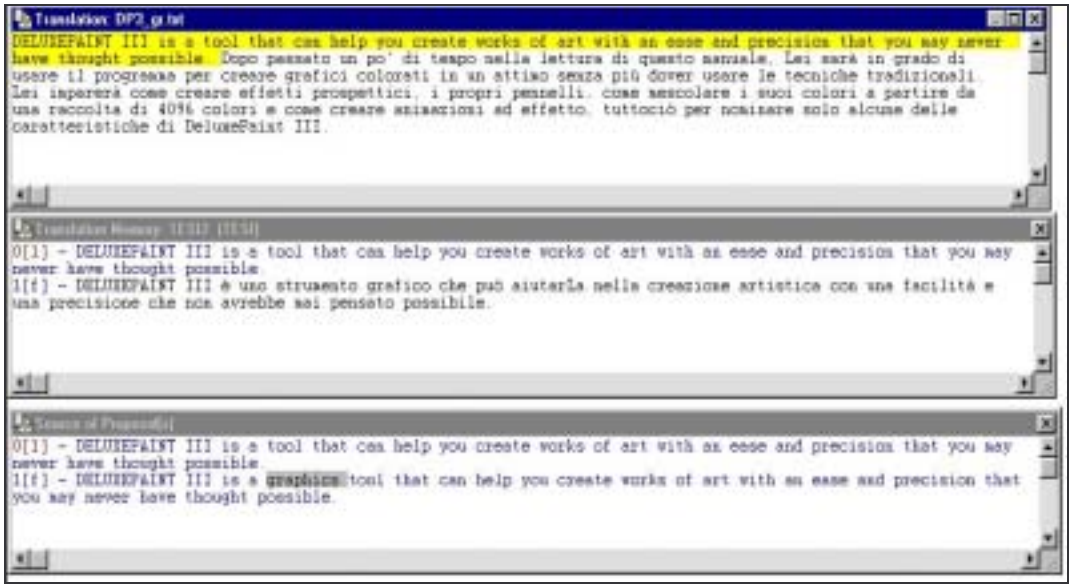


Figura 4.35 – IBM TranslationManager – Interfaccia operativa

Il programma ha infatti già sostituito i due segmenti identici (il secondo e il terzo nell'esempio), mentre per quanto riguarda il primo segmento viene evidenziato in grigio il termine differente.

Si vuole ora valutare il risultato che si ottiene dal programma con l'eliminazione di due o più parole, al fine di valutare quale sia la penalizzazione *fuzzy*:

```
DELUXEPAINT III is a graphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

L'analisi evidenzia, in maniera simile al caso precedente, una sostituzione di 81 parole, contro 24 non ancora tradotte:

Word Count Results - Translation

Documents	Translated	Untranslated
DP3_elim2.txt	81	24
=========	=========	=========
Total	81	24

Figura 4.36 – IBM TranslationManager –Risultato dopo eliminazione di due parole

L'analisi on-line infatti rivela quanto segue:

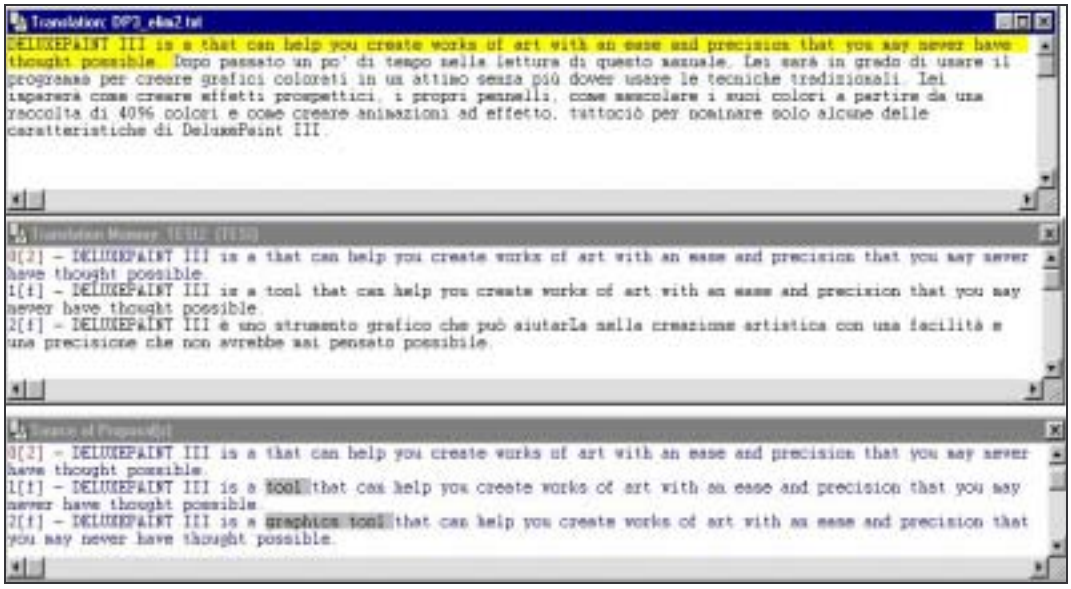


Figura 4.37 – IBM TranslationManager – Eliminazione di due parole

Entrambi i termini "`graphics`" e "`tool`" sono evidenziati in grigio ad indicare dove sia la differenza con il testo originale.

Si noti come la precedente analisi abbia creato/aggiunto un segmento nella Translation Memory che il sistema indica al punto 1, fornendo quindi al traduttore un'informazione aggiuntiva su come poter tradurre quel determinato segmento. In altre parole (come in Trados) il processo di autoapprendimento è continuo e i risultati validati in precedenza sono ora rimessi a disposizione del traduttore stesso.

In generale si può quindi concludere che la quantità di parole eliminate (una o due) non è valutata dal programma in maniera differente.

Sorge quindi spontanea la domanda su quale sia la soglia minima di distanza tollerata in una penalizzazione *fuzzy*. A questo proposito si procede con un'ulteriore prova (*Figura 4.38*) in cui viene eliminata una prima e seconda parte sostanziale della frase originale.

```
DELUXEPAINT III is a graphics tool that can help you create works
of art with an ease and precision that you may never have thought
possible. After spending a little time with this manual, you'll be
able to use the program to create colorful graphics in a fraction
of the time it would take using more traditional techniques.
You'll learn how to create perspective effects, how to create and
save your own custom brushes, how to mix your own color palette
from a universe of 4096 possible colors, and how to create
effective on-screen animations, to name just a few of DeluxePaint
III's powerful features.
```

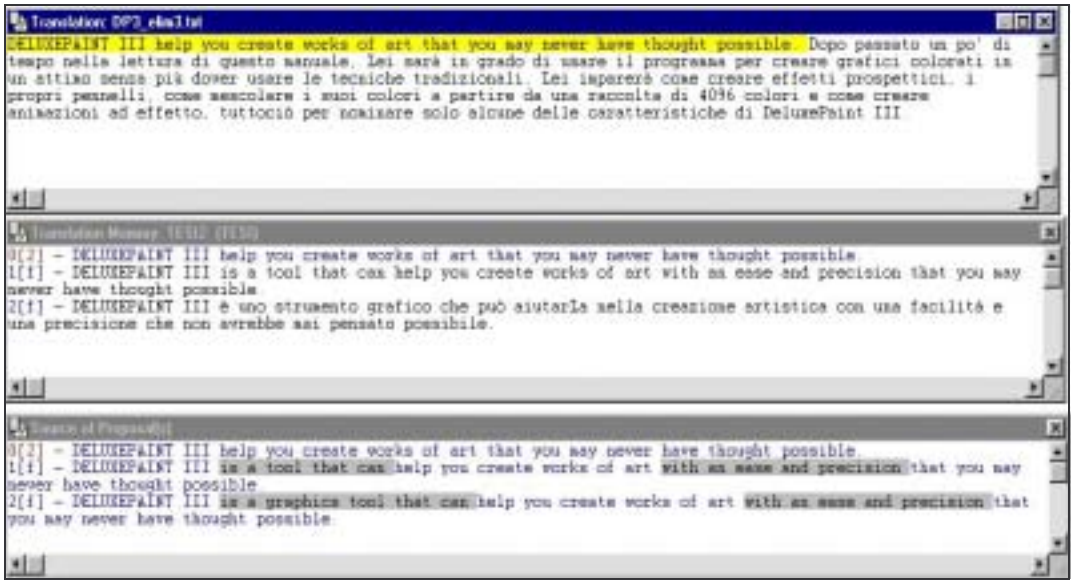


Figura 4.38 – IBM TranslationManager – Eliminazione di parti consistenti

Anche in questo caso il programma ha riconosciuto la similarità tra il segmento sottoposto a traduzione e i segmenti presenti in memoria.

Una successiva eliminazione di una ulteriore parte del testo porta infine al non riconoscimento da parte del programma (*Figura 4.39*):

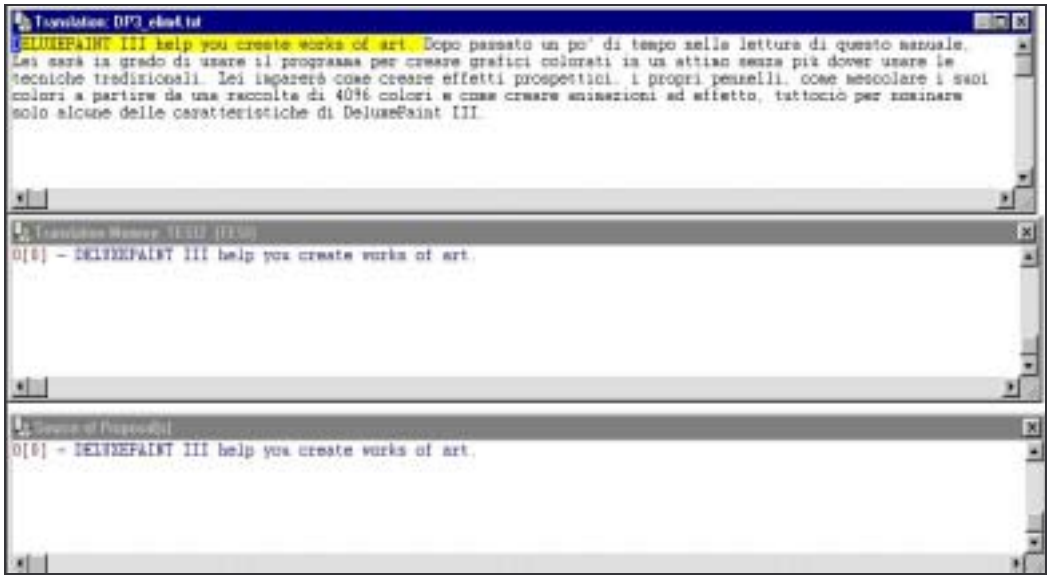


Figura 4.39 – IBM TranslationManager – Eccesso di distanza

4.1.2.2 Aggiunta di una parola di n lettere

Si procede ad aggiungere una parola di 8 lettere al testo originale al fine di valutare la distanza assegnata dal programma IBM a tale modifica:

```
DELUXEPAINT III is a standard graphics tool that can help you
create works of art with an ease and precision that you may never
have thought possible. After spending a little time with this
manual, you'll be able to use the program to create colorful
graphics in a fraction of the time it would take using more
traditional techniques. You'll learn how to create perspective
effects, how to create and save your own custom brushes, how to
mix your own color palette from a universe of 4096 possible
colors, and how to create effective on-screen animations, to name
just a few of DeluxePaint III's powerful features.
```

Il risultato che si ottiene (*Figura 4.40*) evidenzia la corretta individuazione da parte del programma della nuova parola:

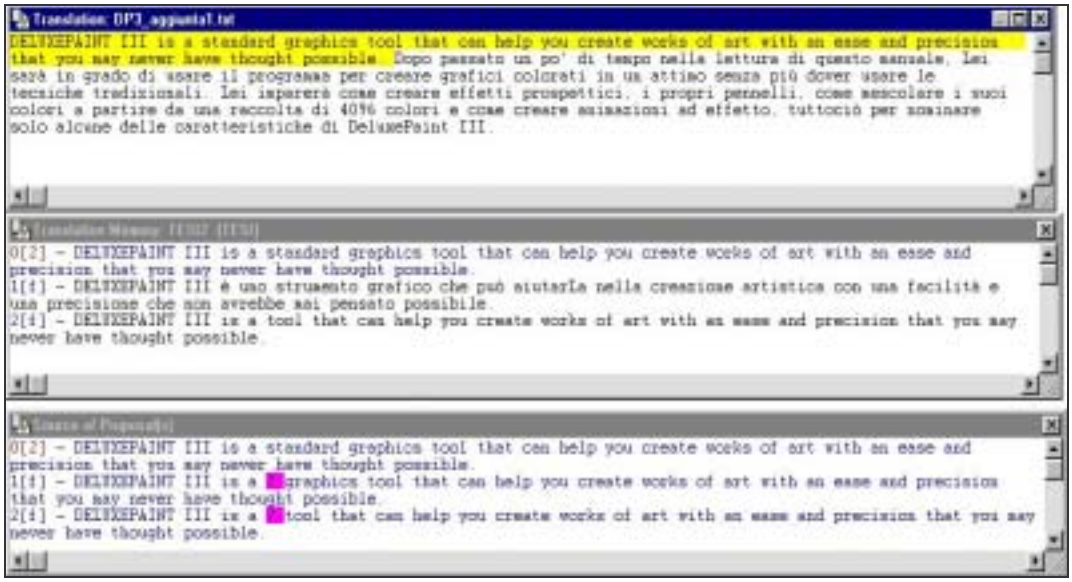


Figura 4.40 – IBM TranslationManager – Aggiunta di una parola

La nuova parola viene indicata nelle possibili traduzioni con un punto interrogativo evidenziato.

4.1.2.3 Modifica di una parola

Si vuole ora procedere alla modifica di una sola lettera di una parola presente in memoria. Si procede ad eliminare la "s" finale della parola "graphics"

```
DELUXEPAINT III is a graphic tool that can help you create works
of art with an ease and precision that you may never have thought
possible.
```

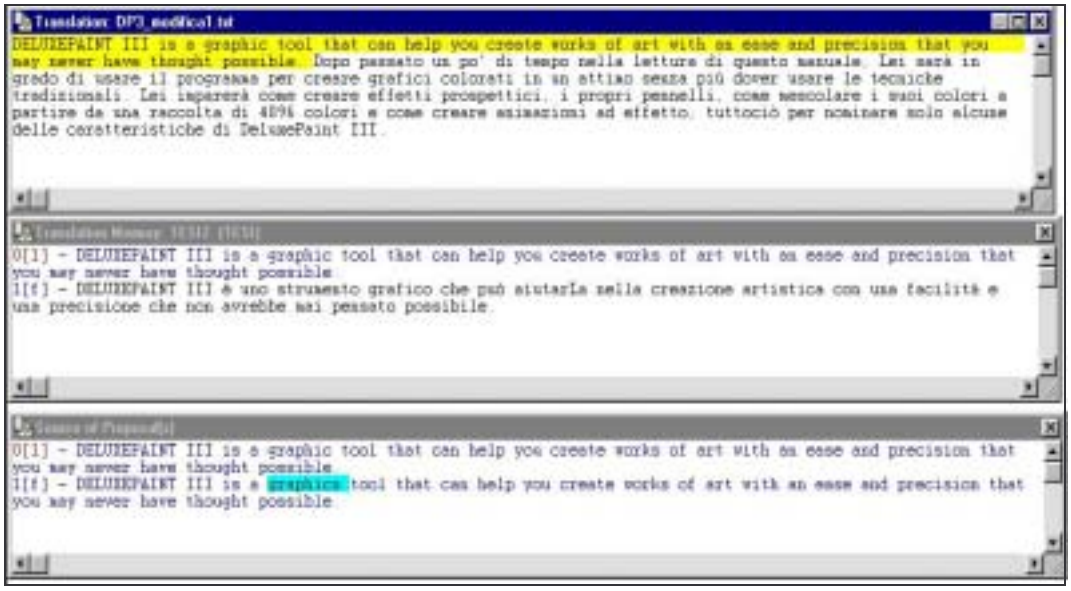


Figura 4.41 – IBM TranslationManager – Modifica di una parola

Anche in questo caso il programma segnala correttamente la differenza evidenziando la parola.

Si noti come tale esempio di eliminazione della “s” finale corrisponda anche ad una prova sui singolari e plurali; permette infatti di concludere che il sistema non è in grado di distinguere i casi singolari e plurali di un termine nel testo sottoposto ad analisi.

4.1.2.4 Unione di due frasi

Si procede ora a congiungere due intere frasi al fine di determinare come il programma valuti queste due frasi già presenti singolarmente in TM. La congiunzione tra le due è in una prima fase effettuata con "and". In una seconda fase le due frasi vengono composte mettendole una all'interno dell'altra.

- Prima prova: "A and B"
- Seconda prova: "A1 B A2"
- Terza prova: "A1 parte del segmento B A2"
- Quarta prova: "“A1 B modificato A2”

Prima prova: "A and B"

DELUXEPAINT III is a graphics tool that can help you create works of art with an ease and precision that you may never have thought possible and after spending a little time with this manual, you'll be able to use the program to create colorful graphics in a fraction of the time it would take using more traditional techniques. You'll learn how to create perspective effects, how to create and save your own custom brushes, how to mix your own color palette from a universe of 4096 possible colors, and how to create effective on-screen animations, to name just a few of DeluxePaint III's powerful features.

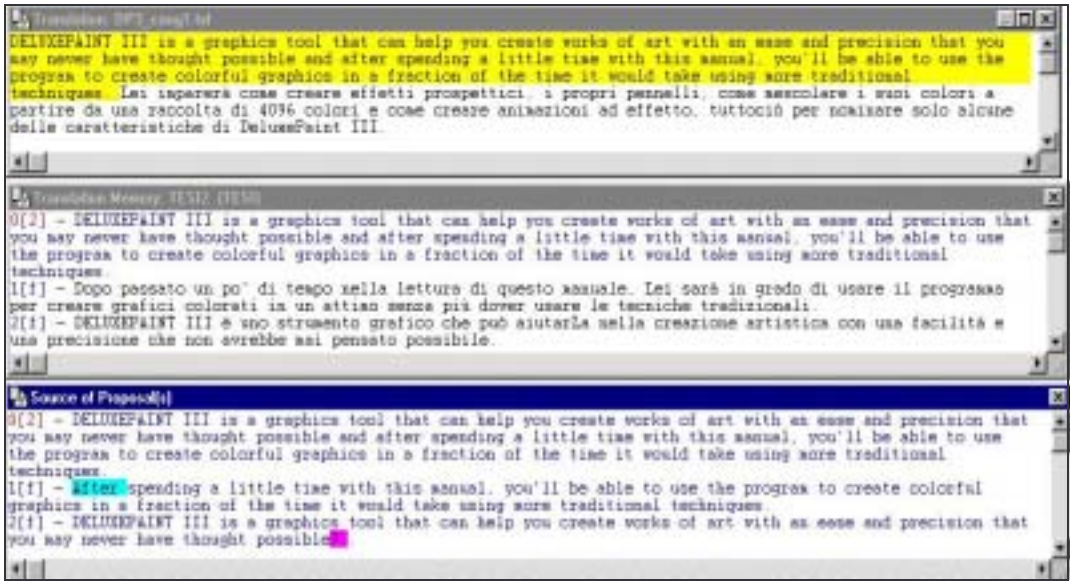


Figura 4.42 – IBM TranslationManager – “A and B”

Il programma non propone direttamente una frase composta come ci si aspetterebbe, bensì indica correttamente i due segmenti che possono concorrere alla traduzione del segmento "A and B" completo. Si noti in particolare l'individuazione della differenza tra la parola "`After`" (in memoria) e la parola "`after`" sottoposta, e come il programma indichi con il punto interrogativo alla fine del segmento in memoria la presenza di una parola aggiuntiva ("`and`").

Non viene invece indicato il punto ("`.`") al termine del primo segmento in memoria come oggetto di necessaria modifica.

Seconda prova: "A1 B A2"

DELUXEPAINT III is a graphics tool that After spending a little time with this manual, you'll be able to use the program to create colorful graphics in a fraction of the time it would take using more traditional techniques **can help you create works of art with an ease and precision that you may never have thought possible.** You'll learn how to create perspective effects, how to create and save your own custom brushes, how to mix your own color palette from a universe of 4096 possible colors, and how to create effective on-screen animations, to name just a few of DeluxePaint III's powerful features.

Anche in questo caso il programma riesce ad individuare il testo sottoposto come simile a due segmenti presenti in memoria, indicando con il punto interrogativo evidenziato le parti di "ingresso" ed "uscita" del segmento:

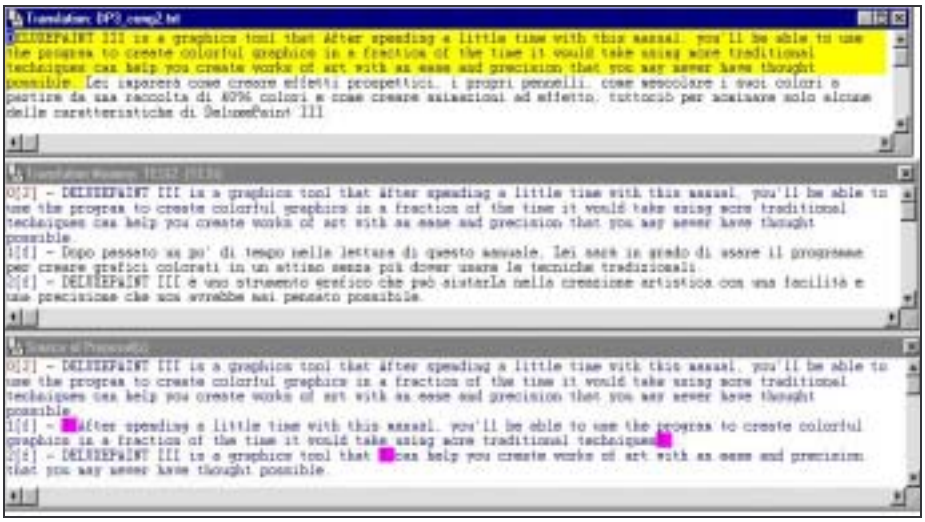


Figura 4.43 – IBM TranslationManager – "A1 B A2"

Terza prova: "A1 parte del segmento B A2"

DELUXEPAINT III is a graphics tool that you'll be able to use the program **can help you create works of art with an ease and precision that you may never have thought possible**. You'll learn how to create perspective effects, how to create and save your own custom brushes, how to mix your own color palette from a universe of 4096 possible colors, and how to create effective on-screen animations, to name just a few of DeluxePaint III's powerful features.

Si ottiene il seguente risultato:

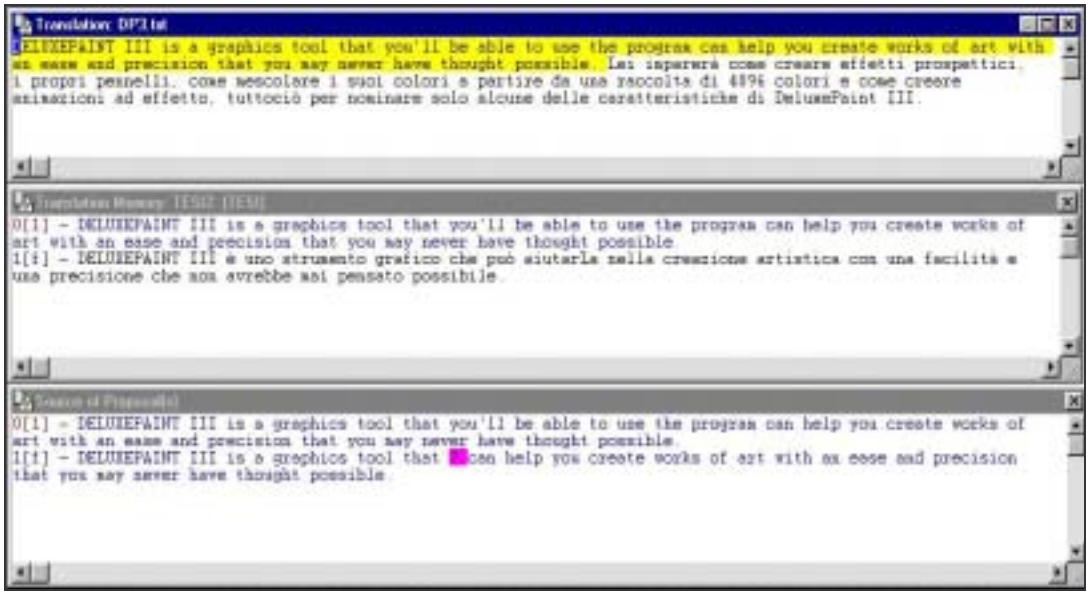


Figura 4.44 – IBM TranslationManager - "A1 parte di B A2"

In questo caso il programma evidenzia la presenza di una parte di segmento sconosciuta, ma non individua correttamente la eventuale presenza in altri segmenti in memoria.

Se il testo sottoposto fosse composto invece da una parte del segmento B più significativa si otterrebbe il seguente risultato:

```
DELUXEPAINT III is a graphics tool that you'll be able to use the
program to create colorful graphics in a fraction of the time can
help you create works of art with an ease and precision that you
may never have thought possible. You'll learn how to create
perspective effects, how to create and save your own custom
brushes, how to mix your own color palette from a universe of 4096
possible colors, and how to create effective on-screen animations,
to name just a few of DeluxePaint III's powerful features.
```

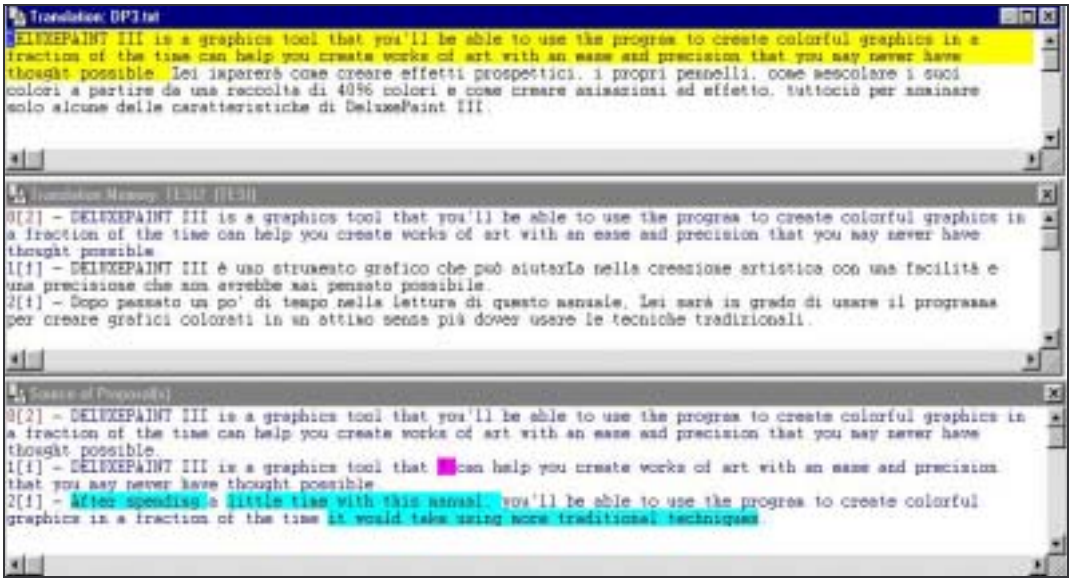


Figura 4.45 – IBM TranslationManager - "A1 parte di B significativa A2"

In questo caso il programma è riuscito ad individuare correttamente la presenza della parte significativa del segmento B come presente in un altro segmento in memoria. Si noti come in questo caso il suggerimento della corrispondente traduzione in italiano sia fornita indicando globalmente il segmento in italiano, senza individuare la parte del testo corrispondente.

A questo punto si introduce un'ulteriore modifica al testo a partire dai risultati di quest'ultima prova.

Quarta prova: piccole modifiche al segmento B inserito

Lo scopo di questa prova è di valutare se il programma è in grado di riconoscere ancora la parte di testo B, inserita all'interno del segmento A, se ad essa vengono applicate piccole modifiche (indicate con una sottolineatura).

DELUXEPAINT III is a graphics tool that you'll be able to use the program to **modify** colorful graphics in a fraction of the time **can help you create works of art with an ease and precision that you may never have thought possible**. You'll learn how to create perspective effects, how to create and save your own custom brushes, how to mix your own color palette from a universe of 4096 possible colors, and how to create effective on-screen animations, to name just a few of DeluxePaint III's powerful features.

Si ottiene:

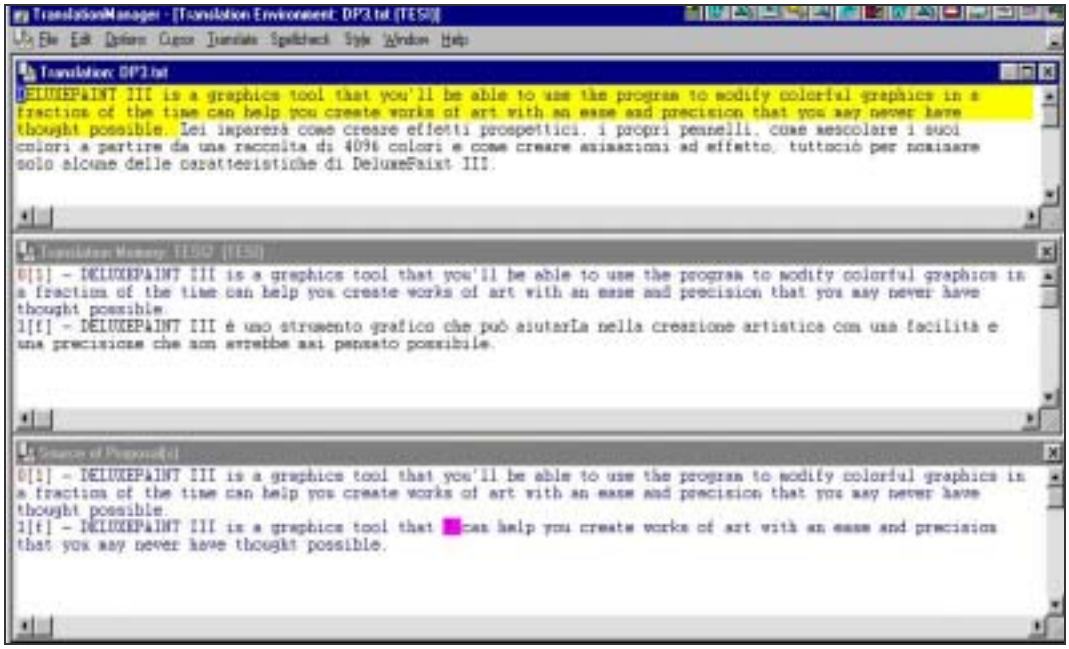


Figura 4.46 – IBM TranslationManager - "A1 B modificato A2"

Ovviamente se si prova a modificare maggiormente il testo B inserito come qui sotto illustrato:

DELUXEPAINT III is a graphics tool that you **are** able to use the programs to **modify** colorful graphics in a fraction of the time **can help you create works of art with an ease and precision that you may never have thought possible**. You'll learn how to create perspective effects, how to create and save your own custom brushes, how to mix your own color palette from a universe of 4096 possible colors, and how to create effective on-screen animations, to name just a few of DeluxePaint III's powerful features.

allora, il programma fornisce il seguente risultato:

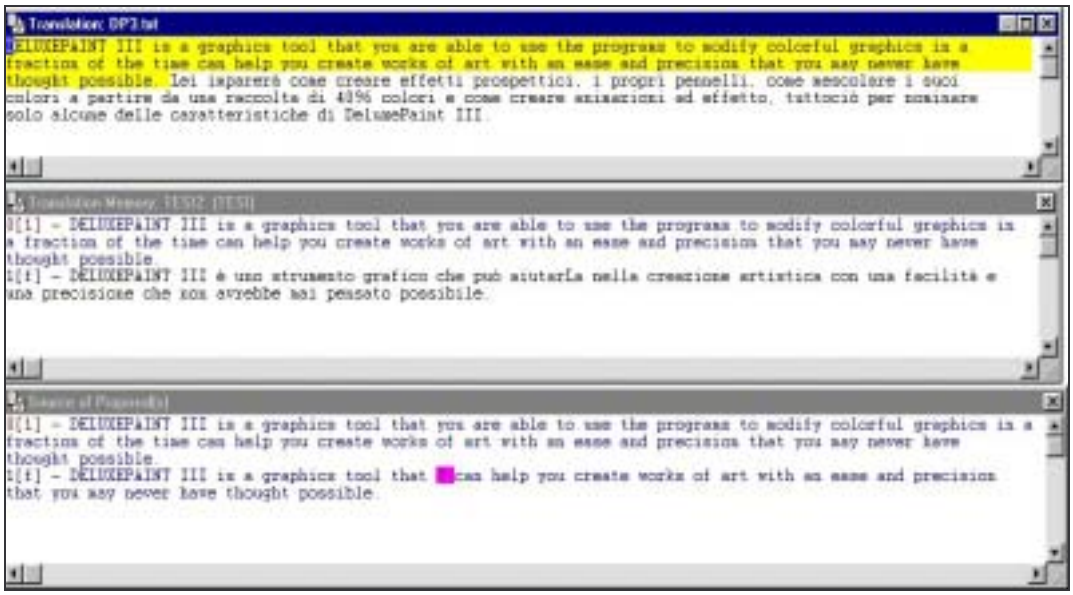


Figura 4.47 – IBM TranslationManager - "A1 B ulteriormente modificato A2"

Il programma non è quindi nuovamente in grado di fornire suggerimenti utili sulla parte inserita e modificata.

4.1.2.5 Markup Table

Si è già valutato come il sistema IBM TranslationManager sia dotato di un'interfaccia poco *user friendly*; molti dei comandi infatti sono poco intuitivi e a volte presenti solo sotto forma di comandi a menu. Dal punto di vista batch è inoltre assente la possibilità di tradurre automaticamente segmenti simili a meno di un impostato valore di verosimiglianza come invece era presente nel programma Trados. L'unica possibilità disponibile è quello della traduzione automatica dei segmenti uguali al 100% (exact match).

Un ulteriore limite è la limitata disponibilità di Markup Table fornite. Tali Markup Table sono i filtri con cui il programma importa il testo al fine di sottoporlo ad analisi e traduzione. Inoltre anche disponendo del filtro giusto tale procedimento di lavoro costringe l'utente finale a lavorare/tradurre nell'ambiente IBM piuttosto che direttamente nell'applicativo originale; questo significa che l'utente finale non potrà utilizzare gli strumenti nativi a cui potrebbe essere necessario ricorrere (ad esempio correzione automatica, spostamento del testo sopra/sotto altro testo, ecc.). Si tenga infine conto che l'importazione da formati proprietari (ad esempio Word di Microsoft) comporta la presenza di codici *tag* di controllo che ulteriormente distolgono l'attenzione dell'utente dalla traduzione.

Esempio di TAGS nell'IBM TranslationManager

A questo proposito si vuole sottoporre ad analisi il seguente segmento in formato Microsoft Word per Windows 6.0:

DELUXEPAINT III is a **graphics tool** that can help you create works of art with an ease and precision that you may never have thought possible.

Tale testo si presenta con font Arial dimensione 12, alcune parole sottolineate, altre in grassetto e altre ancora colorate.

L'analisi on-line fornisce i seguenti risultati:

Figura 4.48 – IBM TranslationManager – Word per Windows

Il programma ha nuovamente individuato correttamente il segmento sottoposto a traduzione come presente in memoria, ma ne ha anche evidenziato tutti gli aspetti differenti. In particolare si veda come sono chiaramente indicati i tags di formattazione ([bold], [underline], [color:06], ecc.) e i rispettivi punti di differenza presenti nella frase presente in memoria (priva di tali tags) ben evidenziati.

Se ora l'utente decide di accettare la traduzione suggerita dalla memoria si ritroverà la seguente sostituzione:

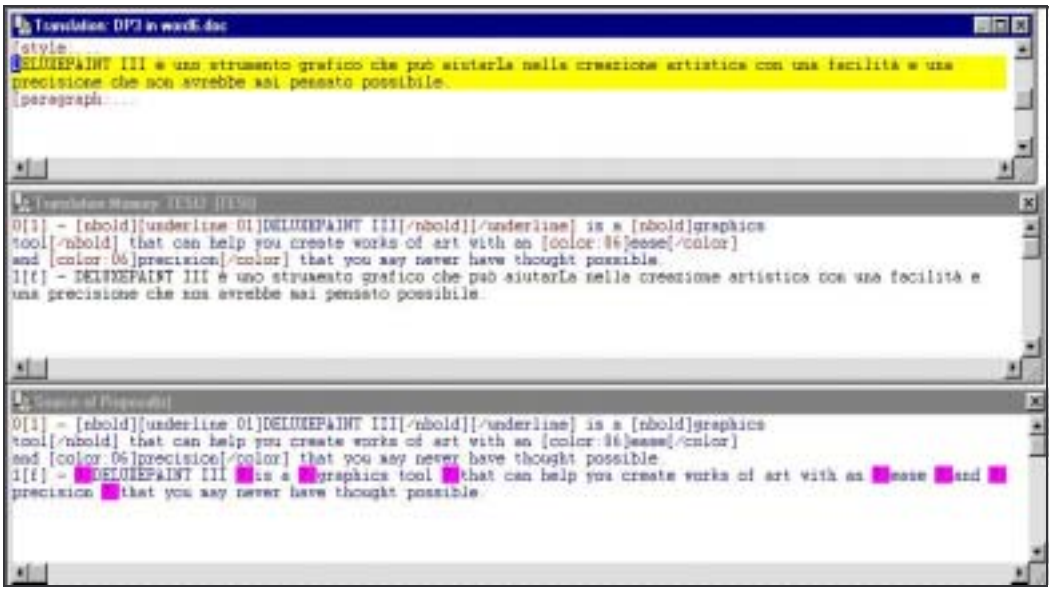


Figura 4.49 – IBM TranslationManager – Word per Windows

A questo punto quindi l'utente dovrà a mano inserire i codici di controllo della formattazione.

Il programma a questo proposito fornisce con il tasto destro del mouse un menu contestuale con il quale poter inserire nel segmento tradotto (ma incompleto) i tags già presenti nel segmento originale:

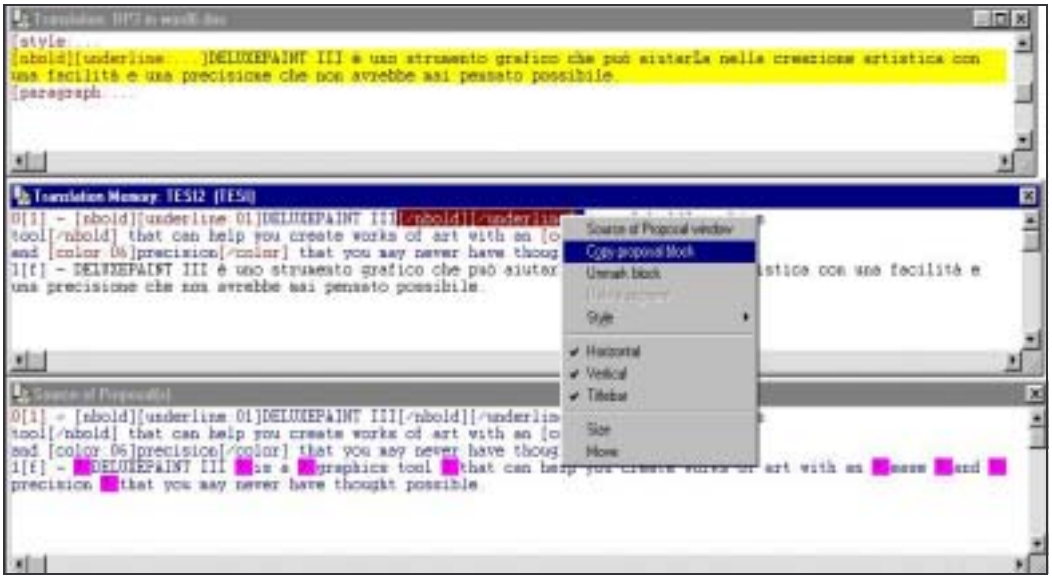


Figura 4.50 – IBM TranslationManager – Word per Windows

Così facendo il segmento verrà correttamente tradotto e validato dal programma al successivo passaggio al segmento seguente da tradurre.

Il risultato finale che si ottiene è:

DELUXEPAINT III è uno strumento **grafico** che può aiutarLa nella creazione artistica con una facilità e una precisione che non avrebbe mai pensato possibile.

Il risultato è corretto ma è costato all'utente un'enorme perdita di tempo.

Si noti infine come tale risultato sia poi inesatto in quanto sono presenti delle imperfezioni quali l'inserzione di doppi spazi a separazione di alcune parole. Ciò è dovuto allo strumento di copia dei tags presente nel programma che non offre nessuna forma di automatismo ma è affidata totalmente alla "mano" dell'utente finale.

4.2 Efficacia in relazione ad un testo campione (nuovo)

Nel presente capitolo si analizzano i valori forniti dai due programmi CAT nella traduzione di un testo più corposo, utilizzando nello specifico un file DeLuxePaint V composto da circa 6000 parole come source language input, evoluzione del testo presente nella TM.

4.2.1 Risultati ottenuti

I due programmi, utilizzando le opzioni standard di configurazione degli stessi, forniscono i seguenti risultati indicati più sotto.

Trados WorkBench

Si ricordi che la realizzazione della TM è stata automaticamente penalizzata dal sistema con una distanza del 3% per cui un fuzzy-match del 100% non è possibile. L'exact match, la piena corrispondenza tra testo sottoposto a traduzione e segmento presente in memoria, è il 97% e a tale valore ci si dovrà riferire come massimo consentito in questa analisi.

L'analisi fornita è la seguente:

Match Types	Segments	Words	Percent	Placeables
XTranslated	0	0	0	0
Repetitions	12	116	2	0
100%	0	0	0	0
95% - 99%	20	311	5	0
85% - 94%	22	423	7	0
75% - 84%	16	300	5	0
50% - 74%	8	201	3	0
No Match	313	4,921	78	0
Total	391	6,272	100	0

Chars/Word	4.74
Chars Total	29,763

Il log fornito è dettagliato in relazione alla distanza dei segmenti analizzati rispetto alla TM. Su un totale di 391 segmenti, pari a 6272 parole, il programma ha individuato 20 segmenti (311 parole) come uguali al 100% o poco meno (il range illustrato in realtà va dal 95% al 99% cioè, riportandoci a valori in centesimi, dal 98% al 100% riferendosi alla TM utilizzata). Un'altra consistente parte di testo analizzato (22 segmenti pari a 423 parole) risulta presente in memoria

a meno di una distanza dal 85% al 94% (cioè normalizzando, dal 87% al 97%).

Per quanto riguarda invece le parti non trovate in memoria, cioè con distanza superiore al 50%, risultano 313 segmenti pari a 4921 parole, corrispondenti al 78% del documento.

Si noti come sia caratteristico di questo programma fornire alcuni valori di valutazione aggiuntivi tra cui il valore medio di caratteri per parola, il numero totale di caratteri e soprattutto il numero delle *repetitions* intendendo con questo il numero di segmenti presenti nel file analizzato, e non presenti in memoria, che si ripetono all'interno dello stesso testo e per i quali è sufficiente tradurli una sola volta per averli poi in memoria al successivo incontro. Nel nostro esempio ben 12 segmenti diversi sono presenti nel testo varie volte ciascuno e risultano quindi segnalati come ripetizioni di traduzioni. Ovviamente essi devono essere considerati a tutti gli effetti come segmenti non corrispondenti ad alcun segmento della TM e per questo motivo il numero totale di no match diventa di 325 segmenti pari a 5037 parole.

IBM Translation Manager

L'analisi del nuovo file sottoposto a traduzione automatica fornisce i seguenti risultati:

Word Count Results - Translation Memory Matches

Documents	Total	Exact-Exact	Exact (1)	Exact (2+)	Fuzzy	No match
DP5.txt	6227	0	190	0	1544	4493
Total	6227	0	190	0	1544	4493

In questo caso i risultati vengono illustrati in maniera molto più sintetica ed in particolare il numero totale di parole è di 6227. L'*Exact match* è presente per un totale di 190 parole coinvolte, mentre i casi di *Fuzzy match* sono pari a 1544 parole.

Si noti come innanzitutto il risultato illustrato non differenzi i casi di *Fuzzy match* per scaglioni di distanza ma genericamente li raggruppi assieme non consentendo di valutare distanze effettivamente elevate

da altre minori cioè con intervento minimo richiesto al traduttore. Inoltre il risultato non tiene assolutamente conto del numero dei segmenti fornendo un risultato solo in parole.

La quantità di no match è pari a 4993 parole.

Confronto di risultati ottenuti

La valutazione dei risultati deve basarsi sul confronto dei valori numerici suddividendo i campi di analisi in tre casi: la quantità di *Exact match* che quindi non prevedono intervento del traduttore ma sono a tutti gli effetti segmenti presenti identici a segmenti in memoria; la quantità di *Fuzzy match* che richiedono un intervento parziale; la quantità di no match.

La quantità di exact match è a favore del sistema Trados con 20 segmenti pari a 311 parole contro le 119 parole del sistema IBM.

Per quanto riguarda i casi di *Fuzzy match* il sistema IBM non fornisce dettagli quali il sistema Trados e quindi il confronto dei risultati vede la somma dei valori *No match* di Trados (46 segmenti pari a 924 parole) a confronto con il valore genericamente indicato dal programma IBM (1544 parole).

Il risultato finale sulla quantità di testo ancora da tradurre (*No match*) riporta 325 segmenti pari a 5037 parole per Trados contro 4493 parole per IBM.

	Trados	IBM
Exact match	311	119
Fuzzy match	924	1544
No match	5037	4493
Totale	6272	6227

Il valore differente di totale numero parole porta alla seguente tabella di valori normalizzati:

	Trados	IBM
Exact match	4,96	1,91
Fuzzy match	14,74	24,79
No match	80,30	73,30
Totale	100	100

Tale risultato non è da intendersi a favore del sistema Trados rispetto a quello IBM in quanto se ad una approssimativa analisi si potrebbe concludere che il sistema Trados individua una quantità maggiore di *Exact match* e quindi richiede da parte del traduttore una quantità minore di interventi per completare la traduzione, in realtà è la quantità di *Fuzzy match* individuata a fornire un valido metro di paragone. Il sistema IBM è superiore in questo caso di un 10% circa (24,79 contro 14,74) rispetto al concorrente. Questo significa che il prodotto IBM è più efficace proprio in quel terreno meno definito dei *Fuzzy match* utili al professionista come suggerimenti per completare il proprio compito.

Si noti infine come la quantità di *No match* sia paragonabile a meno di un 7% sempre a favore del sistema IBM.

4.3 Efficienza in relazione ad un testo campione (nuovo)

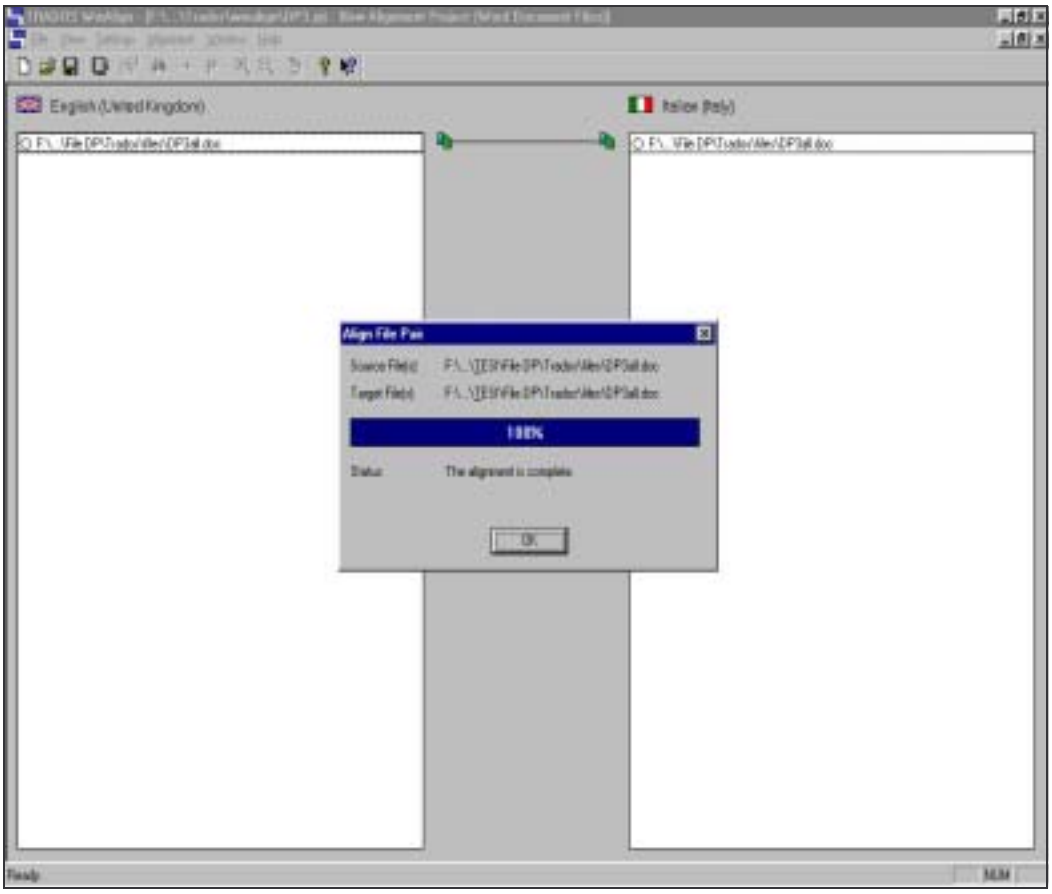
4.3.1 Risultati ottenuti

Si vuole determinare il tempo impiegato dai due sistemi per:

- allineare due file per creare una nuova memoria;
- analizzare un file nuovo sottoposto a traduzione;
- esportare il file così pretradotto.

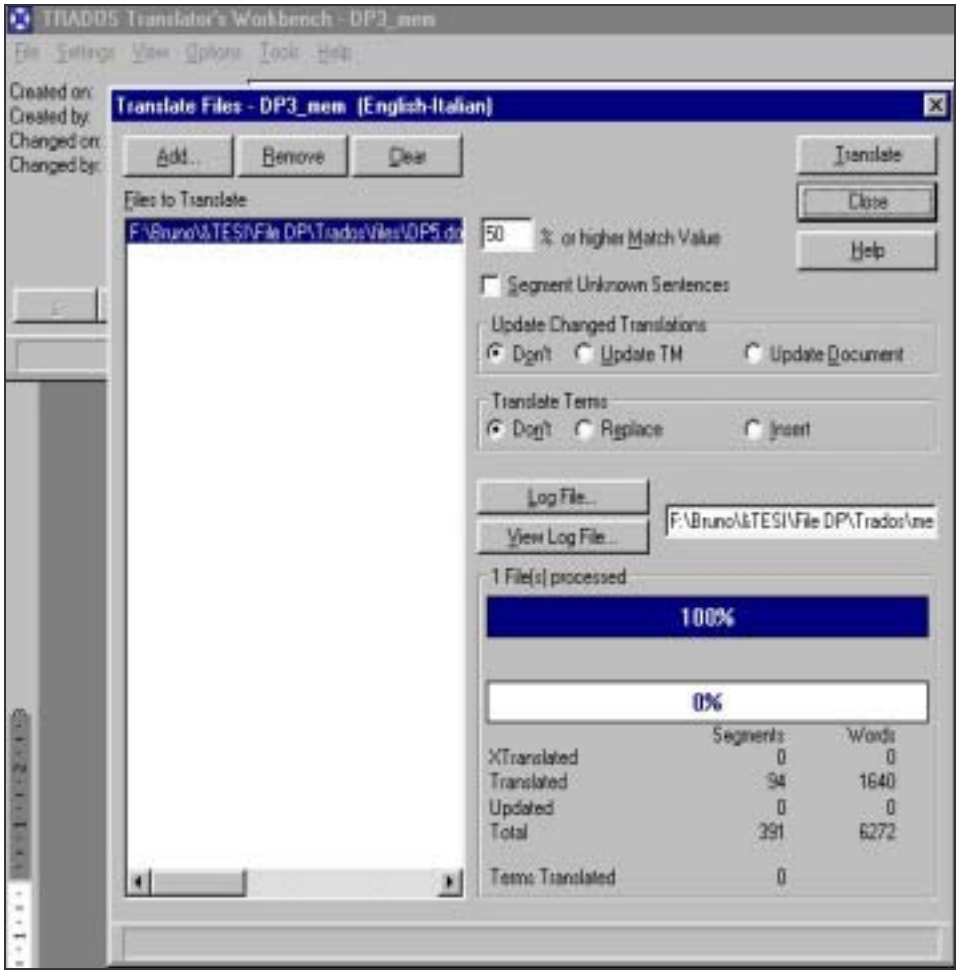
Trados WorkBench

Creazione memoria:



Tempo impiegato per l'allineamento: 18 secondi

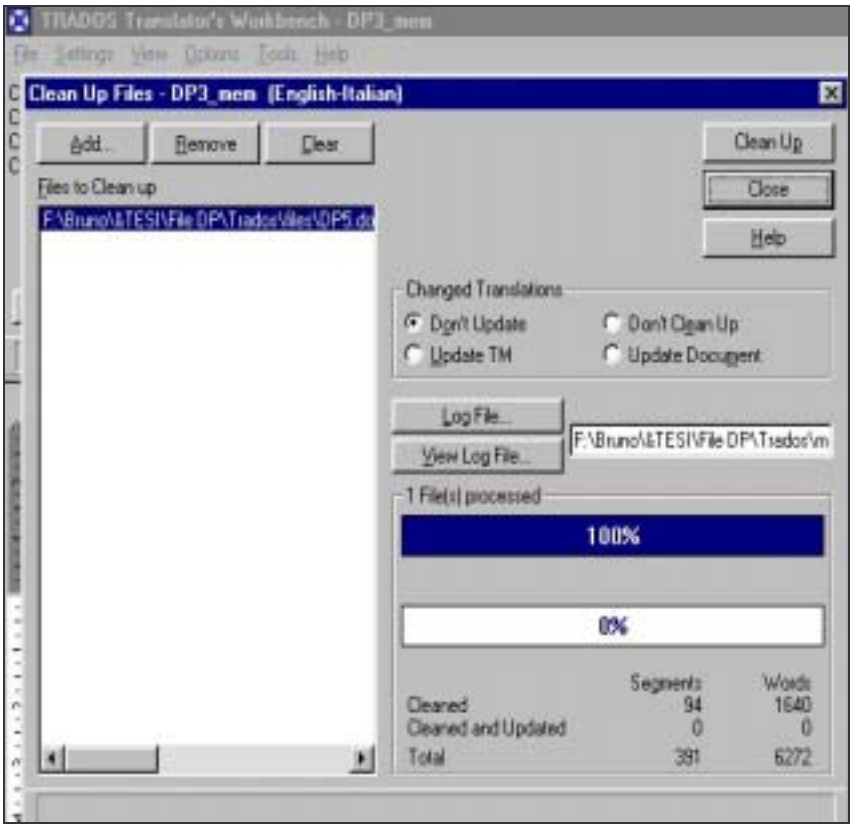
Analisi del file da tradurre:



Traduzione automatica: 23 secondi

Esportazione del file:

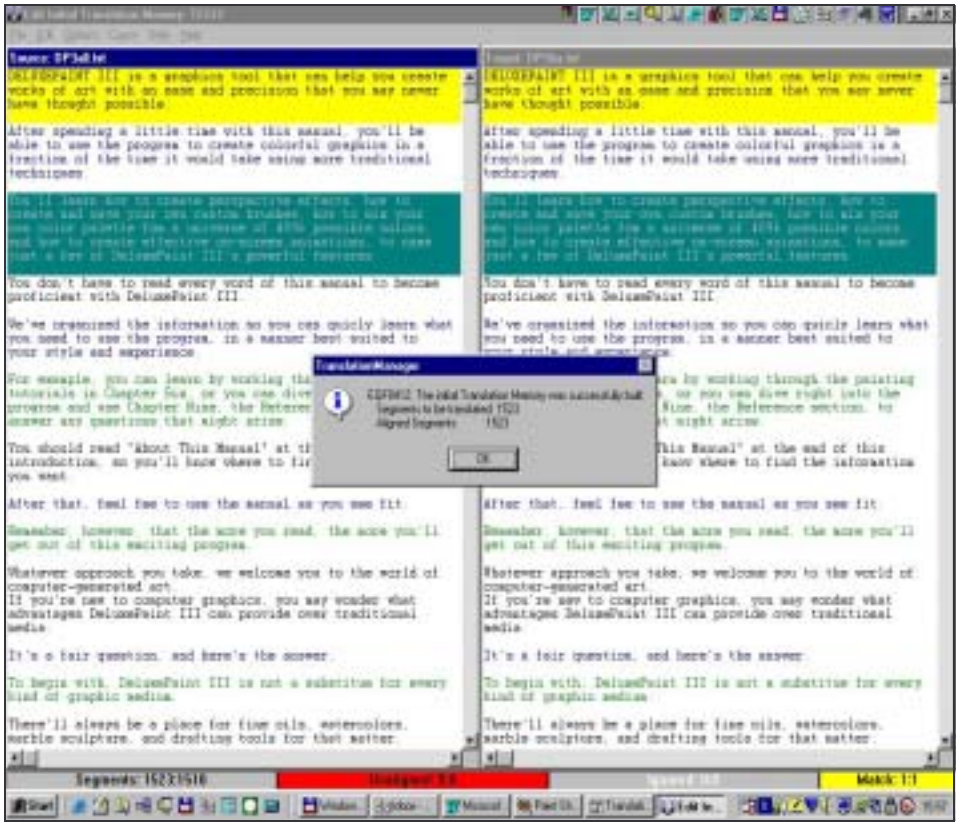
Il programma Trados prevede un successivo passaggio di pulizia dai TAGS che si può far corrispondere alla fase di esportazione del file pretradotto.



Tempo impiegato: 6 secondi

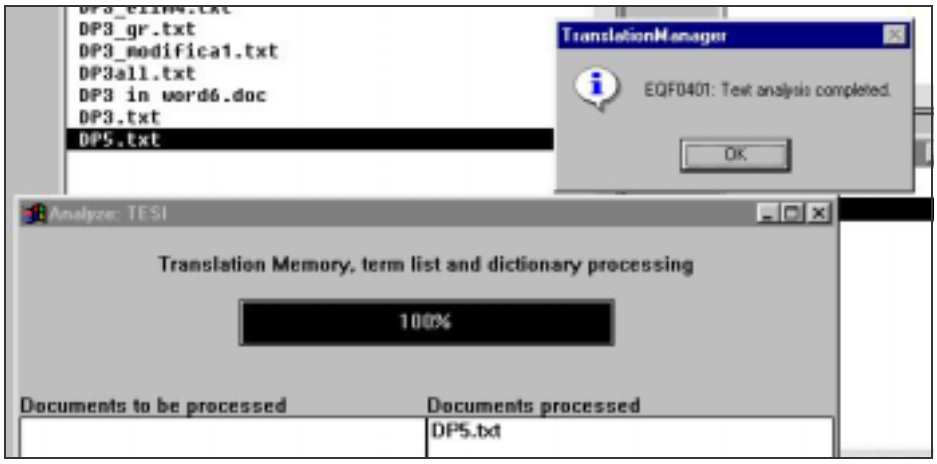
IBM Translation Manager

Creazione memoria:



Tempo impiegato per l'allineamento: 5 secondi

Analisi del file da tradurre:



Traduzione automatica (analisi): 3,5 secondi

Esportazione del file:

Tempo impiegato: 0,5 secondi

Conclusioni:

	Trados	IBM
Allineamento	18 s	5 s
Traduzione	23 s	3,5 s
Esportazione	6 s	0,5 s
Totale	47 s	9 s

La tabella mette in risalto i tempi complessivi ottenuti nelle prove sperimentali per ottenere il testo pretradotto.

4.3.2 Scalabilità

L'analisi della scalabilità consiste nella valutazione dei tempi di risposta del sistema al variare della dimensione dei dati in input.

Lo scopo è quindi quello di verificare se, creata una memoria, la pretraduzione di testi grandi il doppio o il quadruplo, generi prestazioni lineari alla quantità di parole analizzate. A questo scopo si è creata una nuova memoria allineando segmenti di testo pari a circa 7000 parole corrispondenti a 553 segmenti allineati, per poi sottoporre ad analisi un testo nuovo composto da 20000 parole (2007 segmenti).

Successivamente si è analizzato il comportamento dei due sistemi con testi nuovi corrispondenti al 25%, 50% e 75% del testo integrale di 20000 parole.

Trados WorkBench

	25%	50%	75%	100% (20000 parole)
Traduzione	4,6 s	8,5 s	12,2 s	15,8 s
Esportazione	2,6 s	4,1 s	5,3 s	6,9 s
Totale	8,2 s	12,6 s	17,5 s	22,7 s

IBM TranslationManager

	25%	50%	75%	100% (20000 parole)
Traduzione	4,8 s	8,2 s	11,4 s	14,8 s
Esportazione	1,1 s	1,1 s	1,1 s	1,2 s
Totale	6,2 s	9,3 s	12,5 s	16,3 s

Da questo test si verifica come i sistemi forniscono risultati quasi lineari all'aumentare della quantità di testo da pretradurre.

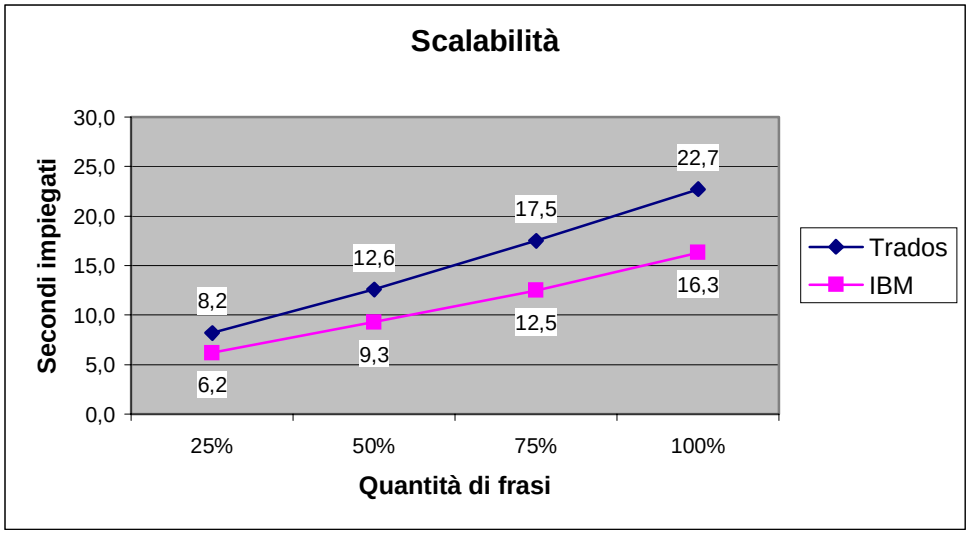


Figura 4.51 – Scalabilità dei sistemi

4.4 Caratteristiche dell'esperimento

E' stato utilizzato il seguente Personal Computer:

Pentium III - 700 MHz (microprocessore Intel) con 512 MB RAM.

Il testo sottoposto a traduzione automatica è il file descrittivo delle caratteristiche della versione "V" del programma DeLuxe Paint, mentre la memoria utilizzata è stata ottenuta allineando con se stesso il file della versione "III".

Note sul calcolo dell'efficienza

Per quanto riguarda il programma Trados si osserva che la traduzione in automatico avviene in due fasi distinte, la prima di pretraduzione con aggiunta del testo trovato in memoria al testo originale (distinti tramite marcatori nascosti), la seconda con l'eliminazione del testo originale indicato come identico a quello proposto dalla memoria. Si tratta quindi di sommare due tempi.

Si tenga inoltre in considerazione che l'analisi con il programma IBM Translation Manager non fornisce in output direttamente il file pretradotto vero e proprio in quanto si lavora nell'ambiente

proprietario del programma. L'output del file si ottiene solo con una successivo comando di esportazione che genera il file vero e proprio nel formato originale. Anche in questo caso quindi si devono considerare i tempi delle due fasi.

Capitolo 5 Analisi dei risultati

Scopo di questo capitolo è quello di analizzare i risultati ottenuti dai due sistemi di traduzione assistita sia in relazione a quelli che sono i risultati ideali sia confrontando le caratteristiche proprie di ciascuno.

5.1 Commento ai risultati

5.1.1 Caratteristiche comuni

Entrambi i programmi consentono innanzitutto di creare una memoria di analisi mettendo in *allineamento* due testi precedentemente tradotti (sorgente/traduzione) grazie ad un'interfaccia grafica a due finestre verticali. L'allineamento risulta essere più difficoltoso nell'ambiente IBM TranslationManager in quanto si basa sull'uso o di comandi posti menu a tendina o di combinazioni di tasti da apprendere mnemonicamente. L'interfaccia del prodotto Trados risulta invece estremamente ben realizzata e intuitiva per l'utente.

Entrambi i prodotti permettono l'importazione di vari *formati dei file*, sia per la realizzazione della memoria sia per quanto riguarda la fase di traduzione vera e propria. Il prodotto IBM TranslationManager risulta però essere meno aggiornato e addirittura non predisposto nei riguardi di programmi di impaginazione avanzata. Entrambi si prestano bene invece alla gestione dei più comuni formati di videoscrittura.

Entrambi possiedono un *formato proprietario della memoria* ma comunque dotato di possibilità di esportazione in comuni file di testo.

5.1.2 Differenze

L'interfaccia utente è decisamente un elemento di differenza tra i due prodotti. Il prodotto Trados infatti si interfaccia direttamente o con il programma di videoscrittura o comunque consente la gestione dei file direttamente nel loro formato nativo senza imporre alcuna conversione (salvo casi specifici). Il prodotto IBM invece costringe l'utente ad importare nel proprio ambiente di lavoro i file trasformandone con opportuni filtri il formato. Se da un lato questo porta ad una uniformità di interfaccia indipendente dal formato specifico dei file sottoposti a traduzione, d'altro canto limita la libertà e l'intuitività delle operazioni eseguibili sui file.

La principale differenza tra i risultati ottenuti dai due applicativi è la capacità del prodotto IBM TranslationManager di individuare e quindi suggerire anche sottoparti di segmento. Questa sua caratteristica ovviamente è legata, come si è visto, alla lunghezza della parte suggerita. Se essa risultasse troppo corta in relazione alla lunghezza del segmento presente in memoria, il suggerimento verrebbe a mancare. Il prodotto Trados fornisce invece una serie di risultati più rigidi anche se affiancati da tool secondari importanti (vedi §5.1.4 "Strumenti aggiuntivi" a pagina 7).

Per quanto riguarda le prestazioni dei due prodotti (si veda §4.3 "Efficienza in relazione ad un testo campione (nuovo)" a pagina 7) ovviamente i risultati in termini di tempo di esecuzione sono difficilmente confrontabili. Questi risultati a favore del prodotto IBM

sono influenzati dall'approccio di base al problema dell'interfaccia con l'utente che nel caso IBM prevede la trasformazione dei formati nativi in un formato proprio, mentre nel caso di Trados prevede il mantenimento del formato stesso a favore di un utilizzo conservativo direttamente nell'applicativo che ha realizzato il testo da tradurre.

Nel caso di Trados si deve però riconoscere che ciò è vero solo nel caso di file di videoscrittura dal formato comune, in quanto in caso di specifici formati di impaginazione (Quark Xpress, PageMaker, FrameMaker - si veda "Tool di conversione formati" a pagina 7) anche qui si è costretti alla trasformazione dei formati arrivando in casi particolari all'utilizzo di un formato proprietario (Trados Tag Text - TTX) simile al formato XML.

5.1.3 Limiti

Un primo limite dei sistemi di traduzione assistita descritti è la necessità di apprendere l'uso specifico di differenti componenti per ciascuna suite. Ciascuna di queste componenti è da considerarsi come un programma vero e proprio e come tale va appreso nei minimi dettagli per ottenere da esso il massimo delle prestazioni e l'individuazione dei limiti. Trados ha cercato di ovviare a questo problema centralizzando nel programma WorkSpace (vedi §2.1.1.1 "WorkSpace" a pagina 7) l'accesso al progetto traduzione e ai vari programmi della suite. IBM ha ridotto il numero di strumenti totale introducendo il concetto di scrivania di lavoro centralizzata in maniera simile all'altro programma, ma comunque fornendo una serie di finestre che riportano al problema della conoscenza specifica delle singole sottoparti di programma.

Infine entrambi si basano su un approccio *example based* al problema della traduzione assistita che non tiene conto delle eventuali sottoparti presenti in memoria, e quindi non sono in grado di fornire suggerimenti per sottoparti di segmento. Questo limite non è proprio dei sistemi example based, piuttosto sembra essere uno scarso sviluppo ed implementazione di algoritmi di ricerca e di metrica di similarità tra segmenti. A questo proposito infatti esistono efficaci

studi sulla ricerca di similarità su base sintattica tra cui il recente sviluppo dell'ambiente EXTRA presso l'Università di Modena e Reggio Emilia [15].

Dalle prove sperimentali condotte risulta infine che il prodotto IBM è dotato di una capacità di indicizzazione del testo superiore a quella del prodotto Trados, data la sua capacità di suggerire anche traduzioni per sottoparti non troppo brevi di segmento.

5.1.4 Strumenti aggiuntivi

Solo Trados fornisce MultiTerm (glossario), strumento in grado di suggerire termini singoli o successioni di termini come aiuto in fase di traduzione on-line. Il prodotto IBM consente anch'esso la ricerca in dizionari specifici dall'interfaccia però meno intuitiva.

Trados consente inoltre l'estrazione di termini da testi sottoposti a traduzione per la creazione di glossari puntuali o per sottoporre tali liste ad altri strumenti di ricerca o TM (vedi §2.1.1.4"ExtraTerm" a pagina 7).

5.2 Deduzione del modello

Entrambi i prodotti si sono rivelati privi di ogni spiegazione sul modello di analisi del testo adottato. La spiegazione di tale riservatezza va ovviamente ricercata nel segreto industriale. Si noti come addirittura il modello di memoria del prodotto Trados stia per essere sottoposto a crittografia per garantirne il non utilizzo da parte di software di terze parti [4].

Per quanto riguarda il modello adottato da IBM TranslationManager si può affermare che esso effettua una rianalisi di un segmento appena questo si discosta da quello presente in memoria di una certa quantità di parole. Tale rianalisi viene effettuata allo scopo di ricercare ulteriori suggerimenti per sequenze di parole non riconosciute e presenti all'interno di un segmento in fase di traduzione. Questa nuova analisi ovviamente può essere descritta come una reiterazione (nested loop) del primo processo di ricerca di traduzione in memoria effettuata ad

un secondo livello questa volta specificatamente sulla *sequenza contaminante* il segmento.

Il prodotto Trados invece non rimette in analisi parti di frasi non riconosciute e quindi semplicemente indica una distanza maggiore tra il segmento in fase di traduzione e quello presente in memoria.

Capitolo 6 Limiti attuali dei programmi commerciali

6.1 Possibili sviluppi futuri

Gli attuali programmi di traduzione assistita che utilizzano un approccio basato su esempio (EBMT - Example Based Machine Translation), creano e gestiscono memorie (banche dati) di segmenti in lingua sorgente in associazione con altri segmenti in lingua target.

Nel capitolo Capitolo 4 “Prove sperimentali” a pagina 7 sono stati analizzati i prodotti Trados e IBM TranslationManager e si è ottenuto una verifica del fatto che il loro approccio sia valido per cambiamenti di minore entità, mentre per modifiche sostanziali o ricerche di sottoparti di segmento sia inefficace. Tale approccio è forse l’unico possibile? Esistono miglioramenti che possono essere richiesti agli attuali programmi di traduzione assistita per risultare più efficaci? Con quali criteri confrontare i vari sistemi di traduzione assistita?

Si proverà qui di seguito a indicare possibili sviluppi ai limiti degli attuali programmi in commercio di questo genere, ricordando l'ottimo risultato ottenuto dal prodotto EXTRA sviluppato presso l'Università di Modena e Reggio Emilia [15].

6.1.1 Approccio grammaticale

Un ottimo spunto per i futuri programmi di traduzione assistita è l'approccio basato sull'analisi grammaticale di una lingua. Lo stesso vale anche per la una combinazione tra quest'ultimo e il classico approccio *example based* (*sistemi ibridi*).

L'approccio grammaticale è caratterizzato dai seguenti elementi:

- Allineamento delle parti elementari di una frase: soggetto, verbo, complemento oggetto e quanto altro compone una frase; a questo proposito si segnala il programma MORPHIX che fornisce un robusto algoritmo per la scomposizione di frasi [7] (*efficient and robust lexically-driven parser*);
- Traduzione di testo "unrelated": oltre alla traduzione di testo presente in memoria e simile sarebbe interessante vedere un suggerimento (basato su regole grammaticali certe) anche per frasi non presenti in memoria; come dire una traduzione "parola per parola" ma efficiente e corretta [17].

Da qui i due tipi di approccio TMS-Translation Memory System (*"a TMS stores in a computer all translations made by a translator. In case of re-translation, these translations are retrieved automatically"*) e MT-Machine Translation (*"an MT system applies grammatical rules and information from dictionaries to a given source sentence in order to translate it"*) devono trovare punti di contatto: infatti nel primo caso un traduttore che si fida della memoria presente nel TMS procederà con la traduzione delle sole parti nuove, al contrario in un contesto MT il traduttore dovrà verificare che la macchina abbia suggerito correttamente e in molti casi dovrà correggere l'output fornito dall'MT. Sembrerebbe perciò che l'MT, che cerca di sostituirsi al traduttore e per questo motivo è soggetto ad errore, sia peggiore del

TMS. Sappiamo però che l'MT ha anche la capacità di scomporre una frase in sottoparti "atomiche" che potrebbero essere analizzate successivamente da un sistema MT. Per questo motivo sarebbe auspicabile l'integrazione delle caratteristiche di scomposizione dell'MT nei tool del TMS.

Si veda a questo proposito [9] la dichiarazione in cui si suggerisce da parte del responsabile ricerca e sviluppo della società Trados di integrare la cosiddetta "proposal machine" o MT nel sistema TMS (Trados).

Alla stessa conclusione giungono Michael Carl e Silvia Hansen [3] nel loro "Linking Translation Memories with Example-Based Machine Translation" affermando che *"più un sistema MT è abile nel decomporre e generalizzare le frasi da tradurre, per poi ricomporle, maggiore sarà la sua efficacia. Una combinazione di differenti approcci al problema sembrano essere la chiave per incrementare i risultati globali ottenibili"*. Arrivando a concludere che *"un sistema TMS raggiunge precisioni di traduzione superiori quando può individuare nel proprio database match molto prossimi, mentre quando il testo in memoria non contiene alcuna similarità sono migliori gli approcci tramite Lexeme-based translation memory (LTM)"*.

6.1.2 Approccio sintattico

L’approccio sintattico si basa sulla definizione di una nuova metrica di similarità e sulla creazione di una nuova tecnica di ricerca. L’ambiente EXTRA ne è un esempio [15]. Tramite questo approccio è infatti possibile ottenere suggerimenti di traduzione fornendo in particolare anche indicazioni in lingua sulla traduzione di sottoparti. Oltre a questo aspetto innovativo sono inoltre possibili calibrazioni tramite filtri personalizzabili al fine di migliorare l’efficienza del processo di ricerca [13][14].

6.1.3 Criteri di valutazione standardizzati

Si dovrebbe cercare di individuare nuovi criteri di valutazione per i programmi CAT che tengano conto dei differenti approcci che si possono seguire riguardo il problema della traduzione e che forniscano quindi una valutazione complessiva e obiettiva dello strumento CAT [12].

In particolare si possono individuare il "*user-oriented evaluation criteria*". Questa valutazione dovrà tenere conto dell'attinenza della traduzione fornita con il settore linguistico del testo sorgente, il rispetto delle formattazioni originali, le ricerche all'interno del testo e tutte le "classiche" problematiche delle GUI (Graphic User Interface).

Un altro metro di misura sarà il cosiddetto "*linguistic criteria*". Esso dovrà tenere conto del rispetto della grammatica e delle espressioni linguistiche proprie di ciascuna lingua.

Il "*technical criteria*" dovrà invece tenere conto dell'apertura del modello software utilizzato, al fine di non pregiudicare sviluppi futuri, scalabilità, modularità, costo della manutenzione del sistema; a questo proposito ci sarà anche un "*economic criteria*" per individuare il costo per licenza del prodotto, i conseguenti costi di follow-up (architettura hardware da implementare). Il "*strategic criteria*" infine considererà i fattori di espandibilità del prodotto in relazione alle mutevoli esigenze di business delle aziende.

Capitolo 7 Conclusioni e futuro della Traduzione Assistita

L'analisi degli attuali programmi CAT ha indicato quali siano ad oggi i limiti della traduzione assistita da calcolatore, rilevando soprattutto la loro incapacità di riconoscere sottoparti di segmento. L'uso di sistemi di tipo *example based machine translation* basati su Translation Memory (TM) e algoritmi di ricerca si è rivelato non sufficiente a garantire un'elevata qualità della traduzione e il recupero delle informazioni eventualmente presenti in più sorgenti di dati, qualora non siano presenti direttamente in memoria.

Come indicato nel capitolo 6.1 "Possibili sviluppi futuri" una promettente direzione da seguire è, per alcuni studiosi, quella di integrare gli attuali sistemi basati su TM con sistemi basati su *proposal machine* [9]. Lo studio di nuovi sistemi di traduzione assistita è oggi infatti indirizzato verso la *Machine Translation* e il *Natural Language Processing*, intendendo con questo sistemi di traduzione automatici ed evoluti, in quanto in grado di avvicinarsi alla

qualità della traduzione umana [10]. Il recente aumento delle richieste di traduzioni in ambito commerciale da parte di colossi giapponesi come Fujitsu, Toshiba, NTT, Brother, Catena, Matsushita, Mitsubishi, Sharp, Sanyo, Hitachi, NEC, Panasonic, Kodensha, Nova, Oki ma anche da parte di realtà degli Stati Uniti (Motorola) e del nord Europa (Nokia, Ericsson) ha portato ad una intensificazione delle ricerche per lo sviluppo di sistemi automatici di traduzione di questo genere. Anche la realtà di Internet ha incrementato tali ricerche, si veda ad esempio come CompuServe abbia introdotto da alcuni anni un servizio di traduzione automatica basato sul sistema Transcend sviluppato da Intergraph. La stessa Commissione Europea ha adottato rapidamente il sistema Systran per poter tradurre in tutte le lingue della comunità le migliaia di leggi, note, documenti che ogni giorno produce e deve rendere disponibili a tutti i paesi.

Nuove soluzioni si sono quindi affacciate sul mercato con sistemi basati tipicamente su mainframe come AppTek (www.apptek.com), CITAC (www.citac-mt.com), EJ Bilingual (www.onesource.com), LEC (www.lec.com), Neocor, PC-Translator, e Globalink. Tali soluzioni sono basate sui più recenti studi del *Natural Language Processing* (NLP) e sono solitamente specifiche per settore (cliente) e coppia di lingue.

Ovviamente i produttori di software non specifici, quali Trados, IBM e altri, non sono rimasti insensibili a questa crescita di soluzioni specifiche e se da un lato hanno abbassato il costo delle proprie soluzioni basate su Translation Memory (per aumentare il proprio bacino d’utenza), dall'altro hanno tentato un'evoluzione specifica dei propri prodotti. Ad esempio Personal Translator, è un prodotto realizzato con la collaborazione tra IBM e von Rheinbaben & Busch: basato su un sistema LMT (Logic-Programming based Machine Translation) è disponibile come componente aggiuntivo MT per IBM TranslationManager. Altri sistemi basati su PC sono Hypertrans per traduzioni tra italiano e inglese, il sistema Al-Nakil per arabo, francese e inglese, il sistema Winger per danese-inglese, francese-inglese e

inglese-spagnolo e il sistema TranSmart della Kielikone Ltd per finnico-inglese.

In particolare quanto più il *Natural Language Processing* è stato oggetto di studio nei dipartimenti di ricerca e sviluppo, uscendo dall'area della sola ricerca teorica, tanto più le società di Information Technology hanno potuto realizzare prodotti basati su tale metodologia. Ad esempio sia Winger che TranSmart hanno sviluppato prodotti specifici per alcuni loro clienti; nel caso di TranSmart, si è trattato inizialmente dello sviluppo di un sistema per Nokia Telecommunications, per giungere in seguito a nuove versioni installandole anche in altre compagnie di telecomunicazioni finlandesi. Allo stesso modo la GSI-Erli che per le proprie esigenze ha sviluppato un sistema di traduzione combinando un motore MT con altri strumenti su una piattaforma già esistente (AlethTrad), ha reso poi disponibile il sistema stesso ai propri clienti.

Altri ambiti di ricerca

Un altro settore di ricerca è quello indirizzato verso lo sviluppo di sistemi di traduzione della parola che combinano l'uso del calcolatore con i più recenti studi sulle *reti neurali*.

Altri ambiti di ricerca sono il *parallel processing*, l'*approccio sintattico-statistico all'analisi del testo*, la *scomposizione secondo regole grammaticali*, l'uso di *sistemi ibridi* che utilizzano più approcci contemporaneamente.

E' quindi ragionevole concludere che l'attuale stadio di sviluppo dei sistemi commerciali EBMT, con i sopradescritti punti di forza e limiti, è insufficiente a soddisfare le notevoli richieste del mercato. Per questo motivo la ricerca di soluzioni MT innovative, più efficaci ed efficienti, sarà oggetto di ulteriori studi da parte dell’ambiente scientifico e dell'industria [13][14][15].

Fonti e riferimenti bibliografici

[1] R.Baeza-Yates and B.Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999

[2] R.D.Brown, *Example-Based Machine Translation in the Pangloss System*. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), p. 169-174. Copenhagen, Denmark, August 5-9, 1996.

[3] M.Carl, S.Hansen, *Linking Translation Memories with Example-Based Machine Translation.* Institut für Angewandte Informationsforschung, Martin-Luther-Straße 14, 66111 Saarbrücken, Germany - carl@iai.uni-sb.de

[4] Champollion, *WordFast*, interfaccia alternativa e gratuita per utilizzare eventuali memorie Trados o crearne di nuove. www.champollion.net

[5] P.Ciaccia, *Dispense corso di Sistemi Informativi II*, Università degli studi di Bologna, Facoltà di Ingegneria, Dipartimento CSITE-CNR

[6] Expert Advisory Group on Language Engineering Standards (EAGLES). www.ilc.pi.cnr.it/EAGLES/home.html

[7] W. Finkler and G. Neumann, *Morphix: a fast realization of a classification-based approach to morphology*. In H Trost, editor, *Proceedings of 4th ÖFAI*, Berlin, 1988. Springler.

[8] L.Gravano, P.G. Ipeirotis, H.V.Jagadish, N.Koudas, S.Muthukrishnan and D.Srivastava, *Approximate String Joins in a Database (Almost) for Free*. In Proceeding of 27th VLDB Conference, 2001.

[9] M.Heyn, *Integrating Machine Translation into Translation Memory Systems*. In EAMT Workshop, TKE'96, Vienna, Austria, 29 and 30 August 1996; in particolare i paragrafi 3 (*MT as an "add-on"*) e 4 (*Integrating MT into TMS*), www.eamt.org/archive/vienna.pdf

[10] J.Hutchins, *Computer-based translation systems and tools*. ELRA Newsletter vol.1 no.4, www.eamt.org, December 1996. www.eamt.org/archive/hutchins_intro.html

[11] IBM TranslationManager, www-3.ibm.com/software/ad/translat/tm/

[12] E.Maier, A.Clarke, H.-U. Stadler, *Evaluation of machine translation systems at CLS Corporate Language Services AG*. MT Summit VII '99, September 13-17, 1999, Singapore. it.jeita.or.jp/aamt/mtsummit99. CLS Corporate Language Services AG and Canoo Engineering AG. www.sail-labs.com/products/CLS.pdf

[13] F.Mandreoli, R.Martoglia and P.Tiberio, *A Syntactic Apprroach for Searching Similarities within Sentences*, In Proc. Of the 11th Conference of Information and Knowledge Management (CIKM), 2002

[14] F.Mandreoli and R.Martoglia and P.Tiberio, *Searching Similar (Sub)sentences for Example Based Machine Translation*, In Proc. of the 10th convegno su Sistemi evoluti per Basi di Dati (SEBD), 2002

[15] R.Martoglia, *EXTRA: Progetto e Sviluppo di un Ambiente per Traduzioni Multilingua Assistite*. Tesi di Laurea presso l'Università degli studi di Modena e Reggio Emilia. Anno accademico 2000/2001.

[16] S.Nirenburg, S.Beale and C.Domashnev, *A Full-Text Experiment in Example-Based Machine Translation*, School of Computer Science, Carnegie Mellon University

[17] R.Rapp, *Automatic Identification of Word Translations from Unrelated English and German Corpora*. University of Mainz, FASK (Germany) January 1999

[18] G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval*. Mc-Graw-Hill, 1983

[19] Sythema,
http://www.synthema.it

[20] Trados Corporation,
www.trados.com