

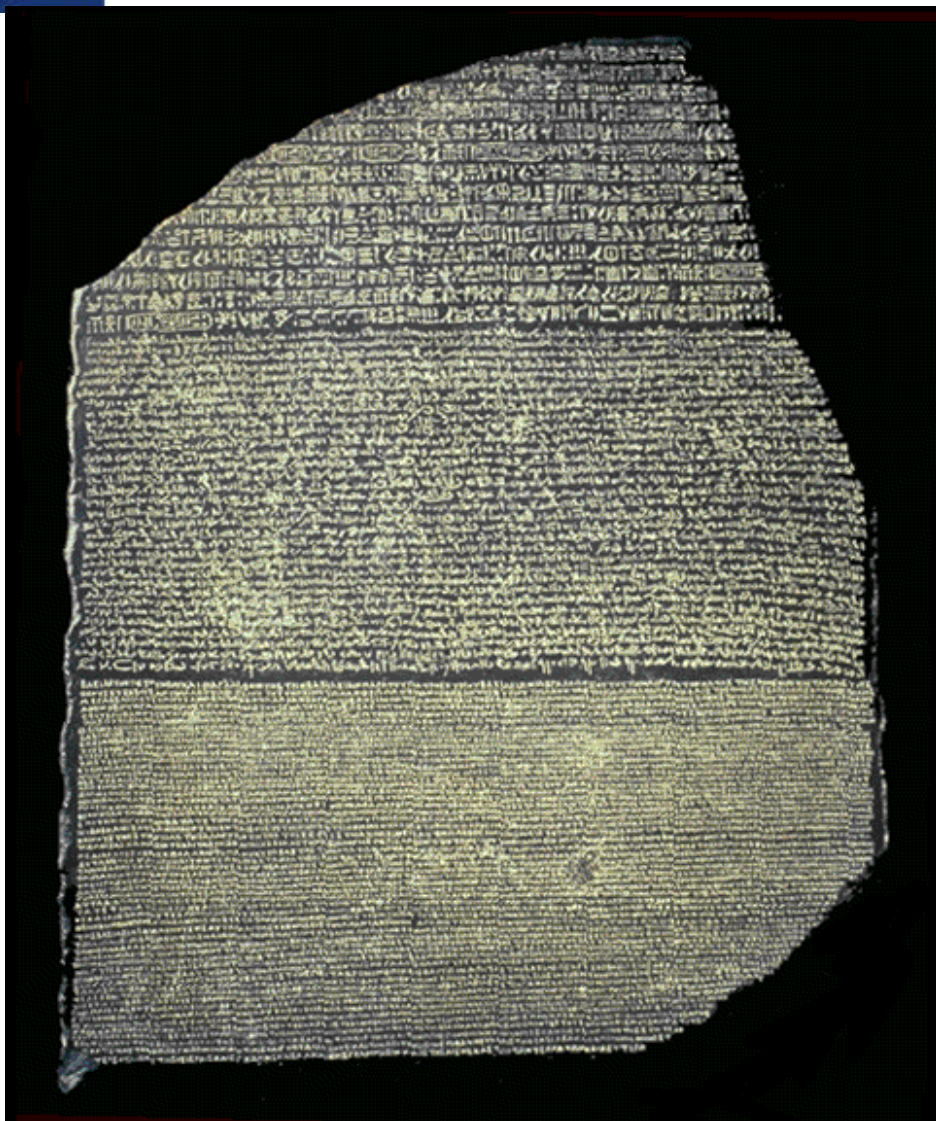


Bevezetés a nyelvtechnológiába

8.

Fordítástámogatás: fordítómemóriák

A Rosetta kő





Párhuzamos korpuszok

A párhuzamos korpuszok

- ☐ Egy könyvben átlagosan néhány százezer szó van
- ☐ Egy művelt ember átlagosan napi tízezer szót olvas
- ☐ Ez összesen évi 3,5 millió és egy életben hozzávetőlegesen néhány százmillió
- ☐ Sok lefordított szöveg akár több százmillió szava elektronikus úton elérhető
- ☐ A gép tehát több szóval tud találkozni, mint egy ember egész életében...
- ☐ Egy lehetséges konklúzió: a fordítási probléma a minták alapján statisztikai alapon megoldható, de legalábbis a fordító munkája másokéval jól támogatható, ha megtaláljuk az épp fordítandóhoz leghasonlóbb mondatot



Kereshető párhuzamos korpusz

A sziklák azonban, amelyekből az Egyetem fölépült, több ezer éven át szívták magukba a mágiát, s annak a rengeteg rendezetlen erőnek kerülnie kellett valahova.	The rocks from which Unseen University was built, however, have been absorbing magic for several thousand years and all that random power has had to go somewhere.
Sajnálattal közlöm, hogy csupáncsak egy ajtó vezet a varázslósághoz, és az a Láthatatlan Egyetem főbejárata, és azon még nő soha nem lépett keresztül.	I am sorry to say there is only one door into wizardry and that is the main gate at Unseen University and no woman has ever passed through it."
Valamikor a nem túl távoli múltban valaki elhatározta valami bizonytalan elképzelés alapján, mely szerint a Tanulás Élvezet Kéne Legyen, hogy az Egyetem ősi folyosóit falfestéssel kell fölvidíteni.	At some time in the recent past someone had decided to brighten the ancient corridors of the University by painting them, having some vague notion that Learning Should Be Fun.
Vastag sugárban dőlt az eső a Láthatatlan Egyetem esővízcsatornáiba s a csurgókba folyt, ahol a nyár óta elhagyatott varjúfészkek úgy lebegtek, mint nagyon rosszul megépített csónakok.	Rain gushed across the leads of Unseen University and poured into the gutters where ravens' nests, abandoned since the summer, floated like very badly-built boats.
Mint az egyetem kincstárnoka azt kell mondjam, hogy mi mindig nagy hangsúlyt fektettünk a közösség érdekeit figyelembe vevő jószomszédi politikára motyogta a varázsló, megpróbálva kikerülni Albert szúrós tekintetét.	'As the bursar of this university I must say that we've always encouraged a good neighbour policy with respect to the community,' mumbled the wizard, trying to avoid Albert's gimlet stare.
A másik kérdésről - amely véleményem szerint ugyancsak nehézségeket okozhatott, s az előbbivel együtt nyilván hozzájárult a Párizsi Egyetem válaszának késlekedéséhez - csak futólag beszélnek.	As to the other point which I have thought may have proved a difficulty and cost the University delay, I will touch but a moment upon that, and pass on.



Párhuzamosan is kereshető korpusz

Járt egyetemre , nem?	You went to college , right?
Ezt követően a Bizottság a Cranfield-i Repülőműszaki Egyetem Légi közlekedési Csoportjának a vezetőjét, Doganis professzort bízta meg független szakértői feladatok ellátásával.	Thereafter, the Commission appointed Professor Doganis, Head of the Air Transport Group of the College of Aeronautics at Cranfield University as a new independent expert.
A legtöbben a sebészeti és az általános tanfolyamot párhuzamosan hallgatták, de a törekvőbbek vagy a szorgalmasabbak ezenfelül még további tanulmányokat folytattak, amelyek végeztével elnyerhették a londoni egyetem doktorátusát.	Most of the students took the curriculum of the Conjoint Board of the College of Surgeons and the College of Physicians; but the more ambitious or the more industrious added to this the longer studies which led to a degree from the University of London.
Végig csaltam az egyetemen	I have cheated my way through college
Azt hiszem, a portói az okosabb választás, 1934-ben került az egyetem pincéjébe, ha nem tévedek, ez az egyik legjobb évjárat, ezenfelül brandy nincsen is.	The port I think is the better bet, laid down by the college in 1934, one of the finest vintages I think you'll find, and on the other hand I don't actually have any brandy.
Az egyetemen , addig, amíg végzek.	In college before I dropped out.
Két kölyök egy puccos egyetemen	Two kids in a fancy college



Alkorpuszokra bontott párhuzamos korpusz

Dictionary **egyetem**

college

Subtitles - Ez egy **egyetem**!

- This is **college**!

Subtitles Az **egyetemen** volt.

She was at **college**.

Literature Még **egyetemre** járt.

She was in **college**."

Subtitles Máris mész az **egyetemre**!

Already off to **college**.

Literature Hogy Brett **egyetemre** járhatson.

Toward's Brett's **college**.

Subtitles Járt **egyetemre**, nem?

You went to **college**, right?

Literature Hol járt Claire **egyetemre**?

"Where did Claire go to **college**?"

Subtitles Két kölyök egy puccos **egyetemen**.

Two kids in a fancy **college**.

Subtitles Végig csaltam az **egyetemen**.

I have cheated myway through **college**.

Literature Én akkor már **egyetemre** jártam.

I was in **college** by then.

Literature Ó, **egyetem**... **egyetem**... zihálta Kinderman bánatosan, kimeresztette a szemeit, és bólogatott.

"Ah, **college**, **college**..." breathed Kinderman ruefully, nodding as he stared.

Subtitles Ilyen lányok miatt jöttem az **egyetemre**.

That's the reason I went to **college**.

Subtitles Az **egyetemen** meg hívogatsz állandóan.

Now in **college**, you call me a good time.

Subtitles Istenem, az **egyetem** óta nem.

I haven't had one of these since **college**.

A Hunglish korpusz

- ☐ Magyar–angol párhuzamos korpusz
- ☐ Szoftverdokumentáció,
jogi szövegek,
szépirodalom,
film,
újságok
- ☐ 54,2m token
- ☐ 2,07m mondat
- ☐ BME MOKK, MTA NYTI



Egy tipikus párhuzamos korpusz: Acquis Communautaire

Language ISO code	N° of texts	Text body			Signatures	Annexes	Total No words (text + signatures + annexes):
		Total No words	Total No characters	Average No words	Total No words	Total No words	
bg	11384	16140819	104522671	1417.85	2170075	14114612	32425506
cs	21438	22843279	148972981	1065.55	7225300	16763733	46832312
da	23624	31459627	213468135	1331.68	2629786	16855213	50944626
de	23541	32059892	232748675	1361.87	2542149	16327611	50929652
el	23184	36453749	239583543	1572.37	2973574	16459680	55887003
en	23545	34588383	210692059	1469.03	3198766	17750761	55537910
es	23573	38926161	238016756	1651.3	3490204	19716243	62132608
et	23541	24621625	192700704	1045.9	1336051	14995748	40953424
fi	23284	24883012	212178964	1068.67	2677798	12547171	40107981
fr	23627	39100499	234758290	1654.91	3021013	19978920	62100432
hu	22801	28602380	213804614	1254.44	2529488	15056496	46188364
it	23472	35764670	230677013	1523.72	3120797	18331535	57217002
lt	23379	26937773	199438258	1152.22	2436585	15018484	44392842
lv	22906	27592514	196452051	1204.6	1673124	15437969	44703607
mt	10545	20926909	128906748	1984.53	1336042	15620611	37883562
nl	23564	35265161	231963539	1496.57	3039580	18467115	56771856
pl	23478	29713003	214464026	1265.57	2513141	17027393	49253537
pt	23505	37221668	227499418	1583.56	3034308	19350227	59606203
ro	6573	9186947	60537301	1397.68	514296	11185842	20887085
ro-19211 (readme)	19211	30832212	182631277	1604.92	---	---	30832212
sk	21943	26792637	179920434	1221.01	3227852	16190546	46211035
sl	20642	27702305	178651767	1342.04	3103193	16837717	47643215
sv	20243	29433037	199004401	1453.99	2575771	14965384	46974192
Total	463792	636216050	4288962348	1387.23	60368893	358999011	1055583954



A Biblia mint párhuzamos korpusz

Latin: Biblia Sacra Vulgata

respondens autem lob dixit

Job 23:1 Spanish: La Biblia de las Américas (©1997)

Entonces respondió Job, y dijo:

Hiob 23:1 German: Luther (1912)

Hiob antwortete und sprach:

Job 23:1 French: Louis Segond (1910)

Job prit la parole et dit:

約伯記 23:1 Chinese Bible: Union (Simplified)

约伯回答说:

King James Bible

Then Job answered and said,

American King James Version

Then Job answered and said,

Job 23:1 Danish

Så tog Job til Orde og svarede:

Job 23:1 Dutch Staten Vertaling

Maar Job antwoordde en zeide:

Jób 23:1 Hungarian: Karoli

Felele pedig Jób, és monda:

ljob 23:1 Esperanto

ljob respondis kaj diris:

JOB 23:1 Finnish: Bible (1776)

Job vastasi ja sanoi:

JOB 23:1 Finnish: Pyhä Raamattu (1933/1938)

Job vastasi ja sanoi:

Job 23:1 Greek OT: Septuagint

υπολαβων δε ιωβ λεγει

Jobo knyga 23:1 Lithuanian

Jobas atsakydamas tarė:

Job 23:1 Maori

Katahi ka whakautua e Hopa, i mea ia,

Jobs 23:1 Norwegian: Det Norsk Bibelselskap (1930)

Da tok Job til orde og sa:

Polish: Biblia Gdanska

A odpowiadając ljob rzekł:

Jó 23:1 Portuguese Bible

Então Jó respondeu:

lov 23:1 Romanian: Cornilescu

lov a luat cuvîntul și a zis:

Иов 23:1 Russian: Synodal Translation (1876)

И отвечав Иов и сказал:



Szövegszinkronizáció

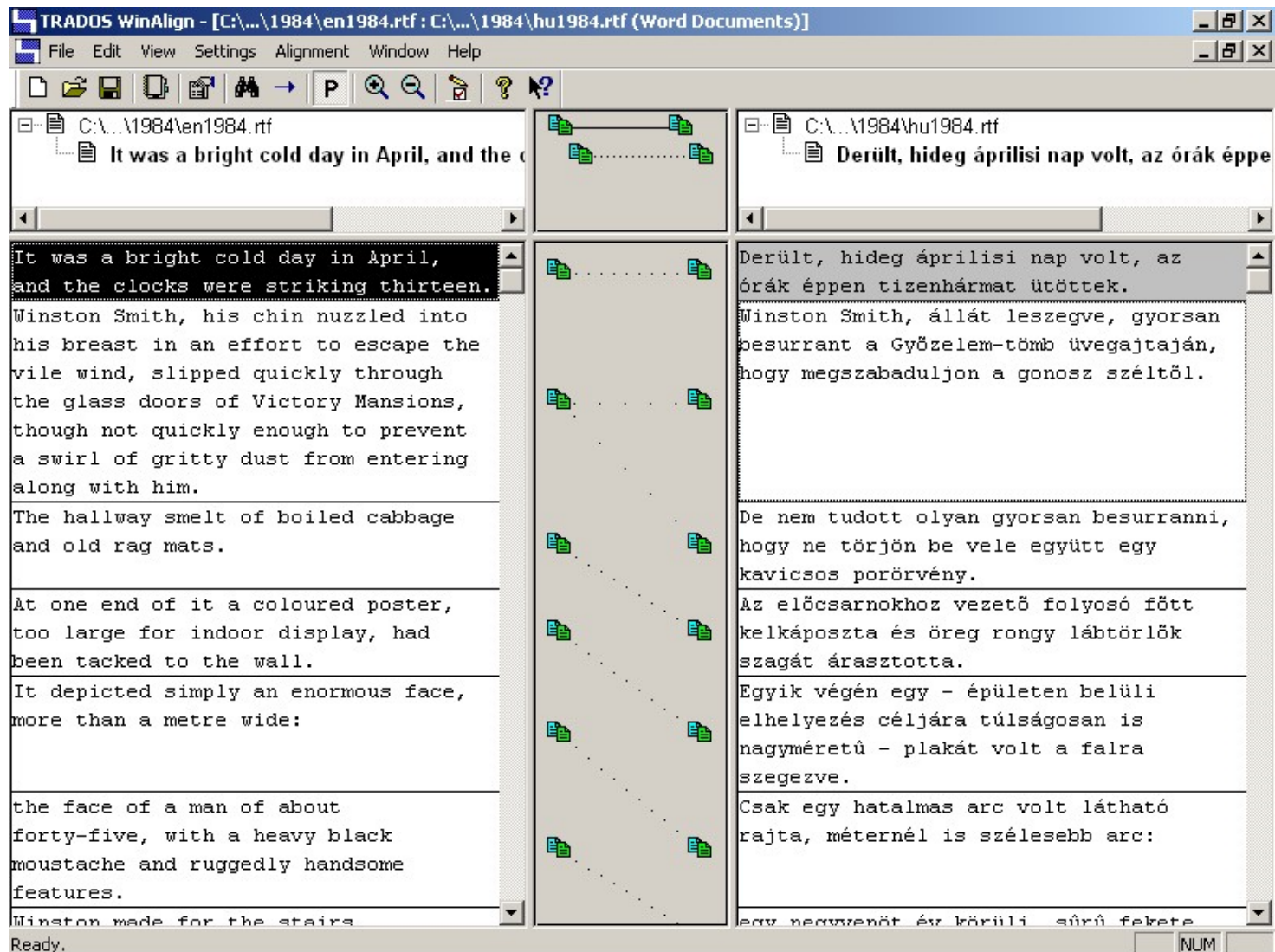
Szövegszinkronizáció-típusok

- ☐ „Alignment”
- ☐ Kétnyelvű szövegpárban az egymás fordításának tekinthető szövegegységek meghatározása
- ☐ Nyelvi egységek szerinti szövegszinkronizáció-típusok:
 - Bekezdésszintű
 - Mondatszintű
 - Kifejezésszintű
 - Szószintű
- ☐ Pontosság/fedés szerinti szövegszinkronizáció-típusok:
 - teljes
 - részleges

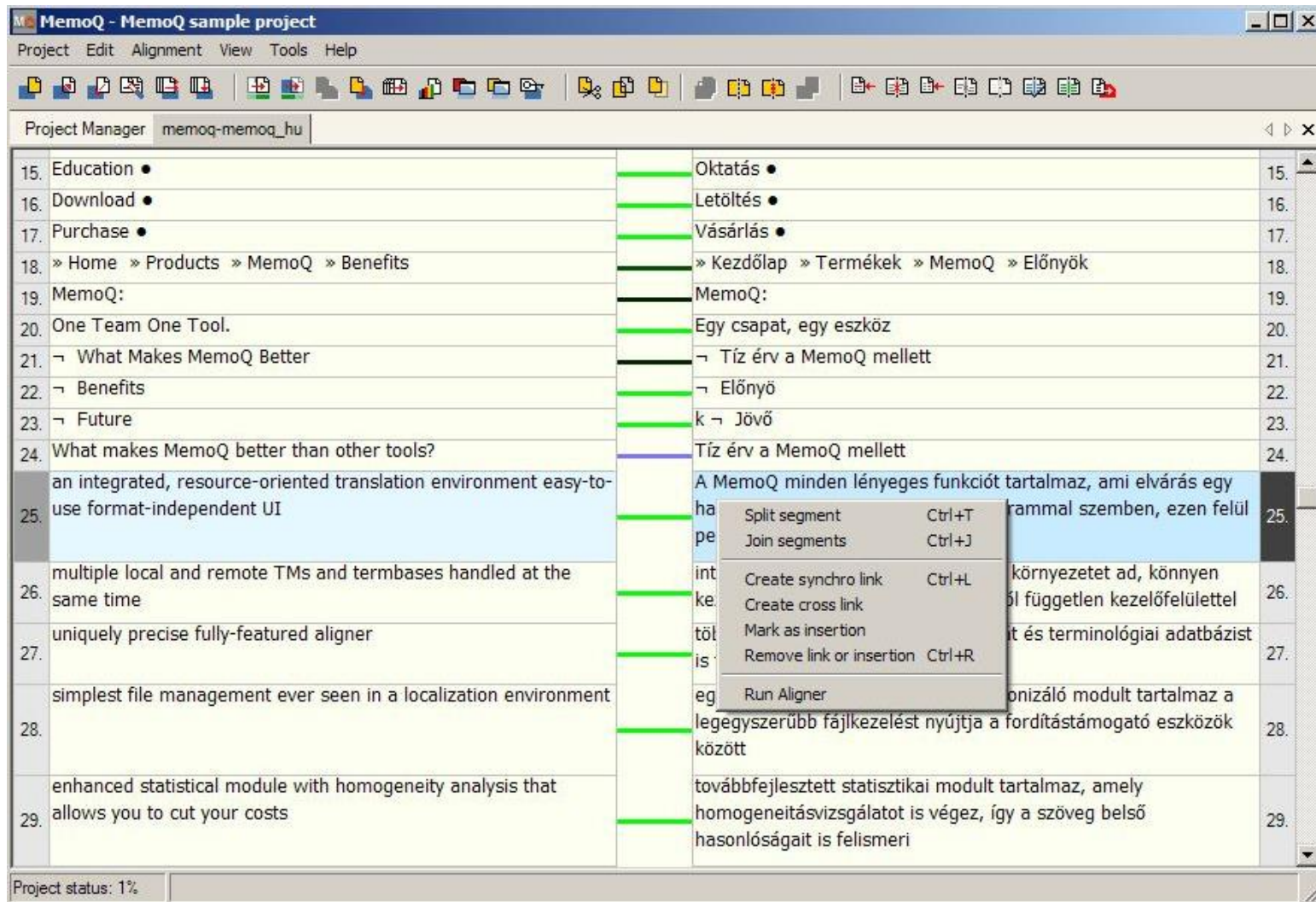
A szövegszinkronizáció felhasználása

- ☐ Párhuzamos korpuszok építése fordítástudományi felhasználásra
 - két- vagy többnyelvű korpuszok szinkronizálása fordítási megfelelések vizsgálatára
- ☐ Fordítások terminológiai konzisztenciájának vizsgálata
 - a forrásszövegben megtalált szakkifejezéseket a fordításban megkeressük
- ☐ Kétnyelvű terminológia gyűjtése
- ☐ Fordítómemória építése
 - a párok kész fordításként felhasználhatók a későbbiekben
- ☐ Statisztikai gépi fordítás

Szövegszinkronizáló eszközök (TRADOS)



Szövegszinkronizáló eszközök (MemoQ)



The screenshot shows the MemoQ software interface. The top menu bar includes Project, Edit, Alignment, View, Tools, and Help. Below the menu is a toolbar with various icons. The Project Manager pane on the left shows a project named 'memoq-memoq_hu'. The main workspace displays a list of segments with their IDs and descriptions. A context menu is open over the segment 'an integrated, resource-oriented translation environment easy-to-use format-independent UI', showing options like Split segment, Join segments, Create synchro link, Create cross link, Mark as insertion, Remove link or insertion, and Run Aligner. The Project status at the bottom left is 1%.

ID	Segment Description	Alignment	Target Description	ID
15.	Education •	Green	Oktatás •	15.
16.	Download •	Green	Letöltés •	16.
17.	Purchase •	Green	Vásárlás •	17.
18.	» Home » Products » MemoQ » Benefits	Green	» Kezdőlap » Termékek » MemoQ » Előnyök	18.
19.	MemoQ:	Green	MemoQ:	19.
20.	One Team One Tool.	Green	Egy csapat, egy eszköz	20.
21.	↪ What Makes MemoQ Better	Green	↪ Tíz érv a MemoQ mellett	21.
22.	↪ Benefits	Green	↪ Előnyö	22.
23.	↪ Future	Green	k ↪ Jövő	23.
24.	What makes MemoQ better than other tools?	Green	Tíz érv a MemoQ mellett	24.
25.	an integrated, resource-oriented translation environment easy-to-use format-independent UI	Green	A MemoQ minden lényeges funkciót tartalmaz, ami elvárás egy	25.
26.	multiple local and remote TMs and termbases handled at the same time	Green	ha	26.
27.	uniquely precise fully-featured aligner	Green	int	27.
28.	simplest file management ever seen in a localization environment	Green	ke	28.
29.	enhanced statistical module with homogeneity analysis that allows you to cut your costs	Green	töl	29.



Mondatszegmentálás



A mondatszegmentálás bonyolult feladat

He can be reached at john.smith@sun.com. The class was written to be part of the java.util.Collections framework. He watched "E!" channel. There is a Jos. A. Banks clothes store on Newbury St. He worked at Smith Corp. "U.S. News reports that he was a model employee," said the anchor. She prefers the title MS. and don't forget it! He saw a 600-lb. gorilla. He flew at 500 m./h. and really high. He had a Ph.D. from UCLA. His name was R. J. Smith, Esq., and don't forget it. I bought fish, fruits, etc. Do you want some? Use a period ("\.") to indicate end of sentence. With \$15.7 billion in annual revenues, Sun can be found in more than 170 countries and on the World Wide Web at <http://sun.com>.

Egy egyszerűbb (?) mondatszegmentálási példa

A Mersey-parti vörösök a hazai vezetés után jobban odatették magukat, aminek eredményeként először Craig Bellamy a 38. percben – először még csak – az oldalhálóba fejtelt, a játékrész lefújása előtt nem sokkal viszont már a gólvonal mögé tette a labdát: Steven Gerrard álompasszát a walesi csatár csukafejessel küldte kapura, de Victor Valdes kapus már csak bentről tudta kiemelte a labdát (Dirk Kuyt a biztonság kedvéért azért még a hálóba bombázott).



Szinkronizációs módszerek

A szinkronizációs módszerek csoportosítása

- ☐ Hossz-alapú módszerek
 - teljes szinkronizáció
- ☐ Lexikai információn alapuló teljes szinkronizációt megvalósító módszerek
 - csak részleges szinkronizáció
 - külső adatbázis kell
- ☐ Horgonykereső módszerek
 - csak részleges szinkronizáció
- ☐ Hibrid módszerek
 - több módszert kombinálnak

Hossz-alapú szinkronizáció

- ☐ „Length-based alignment”
- ☐ Gale & Church (1991, 1993)
- ☐ Csak a szövegegységek hosszait vizsgálja
- ☐ Valószínűségi modell
- ☐ Teljes szinkronizáció
- ☐ Dinamikus programozás →
globálisan optimális megoldást keres

A hossz-alapú szinkronizáció előnyei/hátrányai

☐ Pozitívumok

- Globálisan optimális megoldást keres
- Teljes szinkronizációra képes
- Robusztus: minden szövegegységnek van hossza

☐ Negatívumok

- Nem elég a szövegegységek hosszait vizsgálni, ezért könnyen „eltéved”
- A beszúrásokat és az elhagyásokat rosszul kezeli
- A dinamikus programozás költséges ($O(n^2)$), de ha a mátrixnak csak az átló körüli elemeit számítjuk ki, akkor jóval hatékonyabb módszert kapunk

Horgony-alapú szinkronizáció

- ☐ A *horgonyok* a szövegegységek egymásnak nagy pontossággal megfeleltethető pontjai (pl. nevek, számok)
- ☐ Kezdetben két horgony a szövegpár végpontjait köti össze
- ☐ Ezek után olyan szavakat keres, amelyek gyakran fordulnak elő együtt a szövegpár két oldalán nagyjából azonos pozícióban (heurisztika), majd ezekből horgonyokat határoznak meg (de pl. az egymást keresztező horgonyokat elvetik)
- ☐ A horgonyok által kisebb szinkronizálendő szegmensekre bontható a szöveg, ezeken belül újra horgonyokat keresnek
- ☐ Egyes horgonyokkal csak részleges szinkronizáció valósítható meg
- ☐ Az eredmények nem elég jók

A horgonyválasztásról

Latin írást használó nyelvek esetén (kivétel: német)

- ☐ azonos alakú (nagybetűs) szavak
- ☐ azonos szótövű (nagybetűs) szavak
(morfológia kell hozzá, ráadásul olyan,
amelyik az ismeretlen szavakkal is elboldogul)
- ☐ azonosan kezdődő, hasonló szavak („cognates”)
pl. *Parliament, Parlament; Europa, Europe, Európa*
- ☐ görög / cirill átírás??

Nem-latin írást használó nyelvek esetén

- ☐ horgonyjelöltek szótár alapján
- ☐ sok a nem egy-egy típusú megfeleltetés a szótárban

Statisztikai horgonyszűrés

- ❑ A horgonyjelöltek pozíciói (forrásbeli pozíció és a fordításbeli pozíció) alapján kiszámítják a lineáris regressziós egyenest
- ❑ Az egyenestől mért távolságok alapján egy hisztogramot készítenek, és ezzel szűrik a horgonyokat
- ❑ Kiszámítják a lineáris regressziós egyenes egy elfogadható konfidenciasávját, és a sávon kívülre került horgonyjelölteket eldobják



Fordítómemóriák

A fordítómemória

- ☐ A fordítómemória (fordítási memória, translation memory, TM) szegmenspárokat (általában mondatpárokat) tárol
- ☐ A TM forrásnyelvi szegmensek és célnyelvi fordítások párjait tárolja, némi adalékinformációval (rögzítés dátuma, fordító neve, fordítás státusza stb.)
- ☐ Ha később egy hasonló forrásnyelvi szegmenst kell fordítani, a TM program megtalálja és felkínálja a szegmens korábbi fordítását
- ☐ A javasolt mondat
 - (1) teljes egészében megfelelhet: egy az egyben el lehet fogadni
 - (2) jó kiindulópont lehet: némi módosítással fel lehet használni
 - (3) teljesen rossz: elvethető

A fordítómemóriák alapfunkciói

Export/Import

Szövegelemzés

- ☐ központosítás: egyfajta előfeldolgozás
- ☐ markup: lehetséges segítség
- ☐ „kódolt” szövegrészek (dátum, ábra, stb.) azonosítása

Szegmentálás

- ☐ a fordítási egységek megállapítása
- ☐ egynyelvű felszíni elemzéssel
- ☐ a szinkronizálás alapvető feltétele

Szövegszinkronizálás

- ☐ forrásnyelvi–célnyelvi párok létrehozása

Terminológiai kivonatolás

- ☐ induló szótárral, szövegstatistikák segítségével
- ☐ a fordítási munkával kapcsolatos becslések alapjául is szolgál

Szövegelőfeldolgozás (MemoQ)

Statistics

Scope

- ☒ Project
- ☐ Document
- ☐ Selection
- ☐ From cursor

☐ Show results for each file

Counts

☐ Include target counts

☒ Status report

Word counts

- ☒ MemoQ
- ☐ TRADOS-like

Analysis

- ☒ Project TM's
- ☐ Details by TM
- ☒ Homogeneity
- ☒ Fragment assembling

Calculate

Save as HTML

Save as CSV

Close

Counts

Scope: Project

Type	Segments	Source words	Source chars	Percent
All	61	470	2690	100
Repetition	0	0	0	0
Not started	47	426	2431	91
Pre-translated	12	23	121	5
Fragments	0	0	0	0
Edited	1	19	128	4
Confirmed	1	2	10	0

Analysis

Scope: Project

Resources: Every TM, Homogeneity

Type	Segments	Source words	Source chars	Percent
100%	2	3	18	1
95%-99%	0	0	0	0
85%-94%	1	19	128	4
75%-84%	0	0	0	0
50%-74%	17	39	197	8
Fragments	0	0	0	0
No match	41	409	2347	87

A fordítómemóriák további (ritkább) funkciói

Visszakeresés

- ☐ pontos találat
- ☐ környezet szerint is pontos találat (ICE: In Context Exact)
- ☐ fuzzy találat
- ☐ nyelvileg helyes találat

Módosítás, javítás

- ☐ update-folyamat

Automatikus fordítás

- ☐ automatikus visszakeresés
- ☐ automatikus csere
- ☐ igazi gépi fordítás
- ☐ „intelligens” fordítómemória-funkciók

A fordítómemóriák előnyei és hátrányai

+

- ☐ Konzisztens dokumentumok előállítása: nagy projekteknél elengedhetetlen
- ☐ Gyorsítja a folyamatot: egyszer kell fordítani
- ☐ Pénzt takarít meg: a fordítóiroda tudja, mit tartalmaz az induló memória, illetve meg tudja számolni, hogy mit kellett „igazán” fordítani
- ☐ Egyetlen projekten belül is gazdaságos: menet közben segít, nemcsak a következő projekt indulásakor

–

- ☐ Azon alapul, hogy a mondatok „újrafelhasználhatók”
- ☐ A ma szokásos fordítói munkafolyamat átszervezésével jár
- ☐ Vannak dokumentumformátumok, melyeket még mindig nem támogatnak közvetlenül
- ☐ Idő, míg valaki megtanulja és rászokik
- ☐ Szabadúszó fordítóknak drága és sokszor bonyolult
- ☐ Minden bemenetnek elektronikusnak kell lennie
- ☐ A korábbi fordítások is átalakítandók a fordítómemóriának megfelelő alakúra
- ☐ A hibák is jobban szaporodhatnak a segítségével

Keresés a fordítómemóriában

- ☐ Minden TM-beli mondathoz egy numerikus értéket generálnak (fuzzy kulcs) és egy speciális indexfájlban tárolják egy egyértelmű azonosítóval
- ☐ Kereséskor kiszámítják a keresendő mondat fuzzy-indexét és a legkisebb távolságú fuzzy-indexfájlbeli mondat lesz a találat
- ☐ Indexfájl: egyfajta térkép, ahol egy adott ponttól egy bizonyos sugáron (a megengedett maximális „fuzzyságon”) belül keresünk
- ☐ A fuzzykulcs általában valamilyen n -gram (leggyakrabban a trigram-) technológián alapul
- ☐ Normalizálás: a hosszú mondatok kis helyen való tárolását meg kell oldani (kisbetűsítés, nem alfabetikus karakterek, pl. betűközök eltávolítása stb.): erre a minél kisebb karakterhalmazzal való kalkulálás miatt van szükség

Füzér-összehasonlítás

- ❑ A füzér-összehasonlítás általában a Levenshtein-távolsággal történik (minimális számú törlés, beszúrás és helyettesítés)
- ❑ Dinamikusan épül egy mátrix, melynek minden pozícióján az adott részfüzérék Levenshtein-távolsága áll, pl.

	l	o	v	e	s
h	1	2	3	4	5
a	2	2	3	4	5
t	3	3	3	4	5
e	4	4	4	3	4
s	5	5	5	4	3

- ❑ Értelemszerűen mondatokra is alkalmazható, nemcsak szavakra
- ❑ Az algoritmus optimalizálható azzal, hogy csak az első különböző karaktertől készítjük el (mindkét irányban, tehát hátulról is):

	l	o	v
h	1	2	3
a	2	2	3
t	3	3	3

A fűzér-összehasonlítás eredménye

- ☐ Két hasonló mondat:

Directive 77/93/EEC is hereby amended as indicated in the Annex to the Directive.

Directive 83/107/EEC is hereby amended as indicated in the Annex to the Directive.

- ☐ Az eredeti mátrix $82 \times 83 = 6806$ cellát tartalmazna, míg a redukált mátrix csak $5 \times 6 = 30$ cellát (ami kevesebb, mint fél százalék)
- ☐ Súlyozhatjuk a karaktermanipulációkat is:
 - nagybetű–kisbetű-helyettesítés: 1
 - diakritikus jelek cseréje: 2
 - egyéb csere: 4
 - TM-beli karakter törlési súlya < egy beszúrás súlya

TMX

A TMX-et két szinten definiálják:

1. fordítási egység (Translation Unit: <tu>)
2. egy alacsonyabb szintű meta-markup (Segment: <seg>)

A TMX két implementációs szintje:

1.szint (Plain Text Only):

az adatok a <seg> elemeken belül sima szövegek
Content Markup nélkül, ami szoftverüzenetek
fordításakor elég, formázott dokumentumoknál nem

2. szint (Content Markup):

formázott dokumentumokra (ld. SRX)

Egy TMX-példa

```
<?xml version="1.0"?> <!-- Example of TMX document -->
<tmx version="1.4">
<header creationtool="XYZTool" creationtoolversion="1.01-023" datatype="PlainText" segtype="sentence"
adminlang="en-us" srclang="EN" o-tmf="ABCTransMem" creationdate="20020101T163812Z" creationid="ThomasJ"
changedate="20020413T023401Z" changeid="Amity" o-encoding="iso-8859-1" > <note>This is a note at document
level.</note> <prop type="RTFPreamble">{\rtf1\ansi\tag etc...\fonttbl}</prop> <ude name="MacRoman"
base="Macintosh"> <map unicode="#xF8FF" code="#xF0" ent="Apple_logo" subst="[Apple]"/> </ude>
</header>
<body>
<tu tuid="0001" datatype="Text" usagecount="2" lastusedate="19970314T023401Z" >
<prop type="x-Domain">Computing</prop>
<prop type="x-Project">P&#x00E6;gasus</prop>
<tuv xml:lang="EN" creationdate="19970212T153400Z" creationid="BobW" >
<seg>data (with a non-standard character: &#xF8FF;).</seg>
</tuv>
<tuv xml:lang="FR-CA" creationdate="19970309T021145Z" creationid="BobW" changedate="19970314T023401Z"
changeid="ManonD" >
<prop type="Origin">MT</prop>
<seg>donn&#xE9;es (avec un caract&#xE8;re non standard: &#xF8FF;).</seg>
</tuv>
</tu>
<tu tuid="0002" srclang="*all*" >
<prop type="Domain">Cooking</prop>
<tuv xml:lang="EN"> <seg>menu</seg> </tuv>
<tuv xml:lang="FR-CA"> <seg>menu</seg> </tuv>
<tuv xml:lang="FR-FR"> <seg>menu</seg> </tuv>
</tu>
</body>
</tmx>
```

Fordítómemóriákkal kapcsolatos szabványok

TMX [Translation Memory Exchange format](#). This standard enables the interchange of translation memories between translation suppliers. TMX has been adopted by the translation community as the best way of importing and exporting translation memories. The current version is 1.4b - it allows for the recreation of the original source and target documents from the TMX data.

TBX [Termbase Exchange format](#). This standard allows for the interchange of terminology data including detailed lexical information. The framework for TBX is provided by two ISO 12620, ISO 12200 and ISO Committee Draft 16642, known as TMF or Terminological Markup Framework. ISO 12620 provides an inventory of well-defined “data categories” with standardized names that function as data element types or as predefined values. ISO 12200 (also known as MARTIF) provides the basis for the core structure of TBX. TMF includes a structural metamodel for Terminology Markup Languages in general, regardless of which XML style of representation is used.

SRX [Segmentation Rules Exchange format](#). SRX is intended to enhance the TMX standard so that translation memory data that is exchanged between applications can be used more effectively. The ability to specify the segmentation rules that were used in the previous translation increases the leveraging that can be achieved.

GMX [GILT Metrics](#). GILT stands for (Globalization, Internationalization, Localization, and Translation). The GILT Metrics standard comprises three parts: GMX-V for volume metrics, GMX-C for complexity metrics and GMX-Q for quality metrics. The proposed GILT Metrics standard is tasked with quantifying the workload and quality requirements for any given GILT task.

OLIF [Open Lexicon Interchange Format](#). OLIF is an open, XML-compliant standard for the exchange of terminological and lexical data. Although originally intended as a means for the exchange of lexical data between proprietary machine translation lexicons, it has evolved into a more general standard for terminology exchange.

XLIFF [XML Localization Interchange File Format](#). It is intended to provide a single interchange file format that can be understood by any localization provider. XLIFF is the preferred way of exchanging data in XML format in the translation industry.

TransWS [Translation Web Services](#). TransWS specifies the calls needed to use Web services for the submission and retrieval of files and messages relating to localization projects. It is intended as a detailed framework for the automation of much of the current localization process by the use of Web Services.

xml:tm [xml:tm](#) This is a radical new approach to translation memory based on the concept of text memory which composes author and translation memory. One of the first significant advances in translation memory technology since its inception.



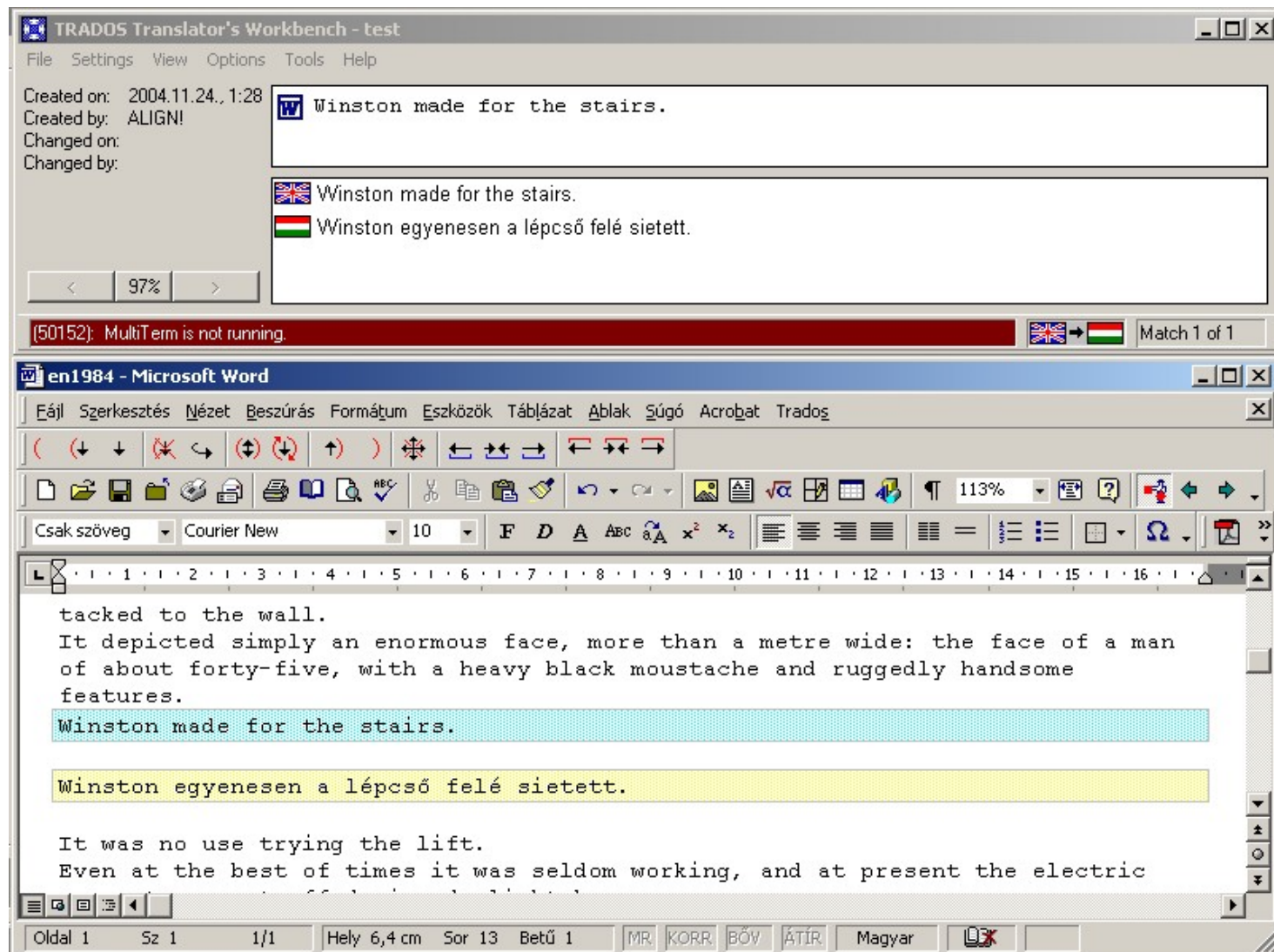
Fordítómemóriák a gyakorlatban



Fontosabb fordítómemóriák

Across Language Server	Windows	http://www.across.net/en/
Déjà Vu	Windows	http://www.atril.com/content/last-update
GlobalSight	Linux, Win	http://www.globalsight.com/
memoQ	Windows	https://www.memoq.com/en/
Memsources	Web, Win/Mac/Linux	https://www.memsources.com/en
MetaTaxis	MS Word add-in	http://www.metataxis.com/
OmegaT	Java	http://omegat.org/
Open Language Tools	Java	https://open-language-tools.java.net/
Poedit	Cross-platform	https://poedit.net/
Pootle	Cross-platform	http://www.translate.org.za/
SDL Trados	Windows	http://www.translationzone.com/
SmartCAT	Cross-platform	https://ru.smartcat.ai/
Star Transit	Windows	https://www.star-group.net/en/products/transit.html
Virtaal	Cross-platform	http://virtaal.translatehouse.org/
Wordfast Classic	MS Word (Win/Mac)	http://www.wordfast.com/
Wordfast Pro	Java	http://www.wordfast.com/

Integrált fordítási környezet (TRADOS)



Önálló fordítási környezet (MemoQ)

MemoQ - MemoQ sample project

Project Edit Translation Format View Tools Help

Project Manager memoq

Line	Source Text	Target Text	Progress	Status
37.	Benefits	Előnyök	100%	✓
38.	MemoQ, the world's most advanced <i>integrated localization environment</i> optimizes the productivity of translation and provides the quickest ROI ever.	A MemoQ, a világ legfejlettebb <i>integrált fordítási környezete</i> , optimalizálja a fordítás és lokalizáció hatékonyságát és grantálja a befektetés lehető legyorsabb megtérülését.	0%	✗
39.	In the MemoQ ILE, all translation resources are only a mouse-click away resulting in an ergonomomy never seen before.		0%	✗
40.	During the design of MemoQ we had only one thing in mind: the optimal environment for translators and project managers.		0%	✗
41.	Kilgray worked hard to avoid all useless keystrokes, file operations, and the result is a one-stop environment with every tool integrated in the most transparent manner.		0%	✗
42.	MemoQ is an easy-to-use concept-oriented system.		0%	✗
43.	An average translator can learn to use 90% of the program's functionality within just two hours.		0%	✗
44.	The remaining 10% is the domain of project managers and clients: you can achieve enormous savings by customizing		0%	✗

Translation Results

Source Text	Target Text
MemoQ, the world's most advanc...	1 A MemoQ, a világ legfejlettebb integrá...
MemoQ	2 MemoQ
integrated localization environment	3 integrált fordítási környezet

MemoQ, the world's most advanced integrated localization environment optimizes the productivity of translation and provides the quickest ROI ever.

MemoQ, the world's most advanced integrated localization environment optimizes the productivity of translation *and localization* and provides the quickest ROI ever.

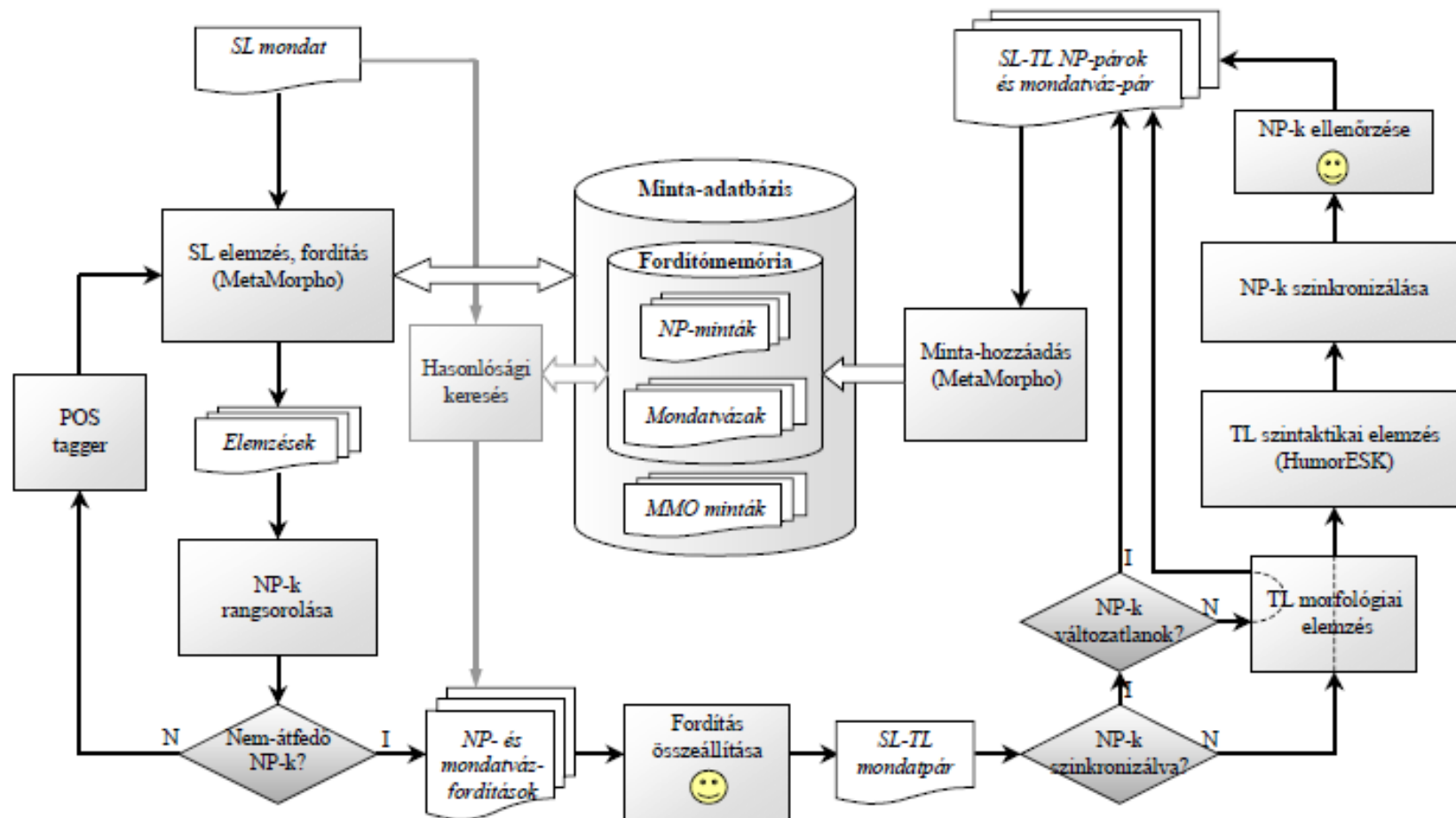
A MemoQ, a világ legfejlettebb integrált fordítási környezete, optimalizálja a fordítás és lokalizáció hatékonyságát és grantálja a befektetés lehető legyorsabb megtérülését.

Pro 101/123 **CLI Famous Client**
Dom IT **Sub CAT**

Aligned Corrected
 MemoQ gabor
 5/3/2006 1:04:18 PM **89%**

Project status: 1% Document status: 1% Conf: 1 Ed: 1 Empty: 47 Pre: 12 Frag: 0 ✗ Errors: 11 Ready

Az intelligens fordítómemória felé





Automatikus kiértékelés

A fordítások hasonlósága: az automatikus kiértékelés alapjai

- ☐ A kézi kiértékelés lassú, és szubjektív
- ☐ Vannak statisztikai módszereink...
- ☐ Nagy tesztkorpuszon közelít az „objektívhez”
- ☐ 2002, IBM: Bilingual Evaluation Understudy (BLEU)
- ☐ A lényeg:
átfedés a referenciafordításokkal
- ☐ Módosítások kellenének:
morfológia, szinonimák, ...

A fordítási minőség mérésének alapjai

Referenciafordítás:

The gunman was shot to death by the police .

BLEU-módszer:

The gunman was shot kill .

Wounded police jaya of

The gunman was shot dead by the police .

The gunman arrested by police kill .

The gunmen were killed .

The gunman was shot to death by the police .

The ringer is killed by the police .

Police killed the gunman .

zöld = 4-gram illeszkedik (jó!)

piros = nem-illeszkedő szó (rossz!)