

How to prepare your data for automated translation?

Andrejs Vasiljevs, ELRC/Tilde

Credits:

Khalid Choukri, ELRA

Josef van Genabith, DFKI

An iceberg floating in the ocean. The small tip above the water represents the public web, while the much larger submerged part represents the deep web. The background is a deep blue gradient.

4%

The Public
Web

96%

The Deep
Web

Multilingual databases
Public sector resources
Organization specific repositories
Legal documents
Scientific reports
Medical records
Etc.

*Information stored inside
institutions or online with
password protection*

- Visible data e.g. [Web](#)
 - Public websites
 - leaflets, brochures, news etc.
- [Invisible Data](#): archives , hidden web, internal repositories
 - translated documents, reports, speeches, meeting minutes etc.
- Data from outsourced translations:
 - From [Language Service Providers](#) and freelance translators
 - Translation Memory as part of contract delivery
 - Respective provisions in the contracts for outsourced translations

The Language Resource Life Cycle

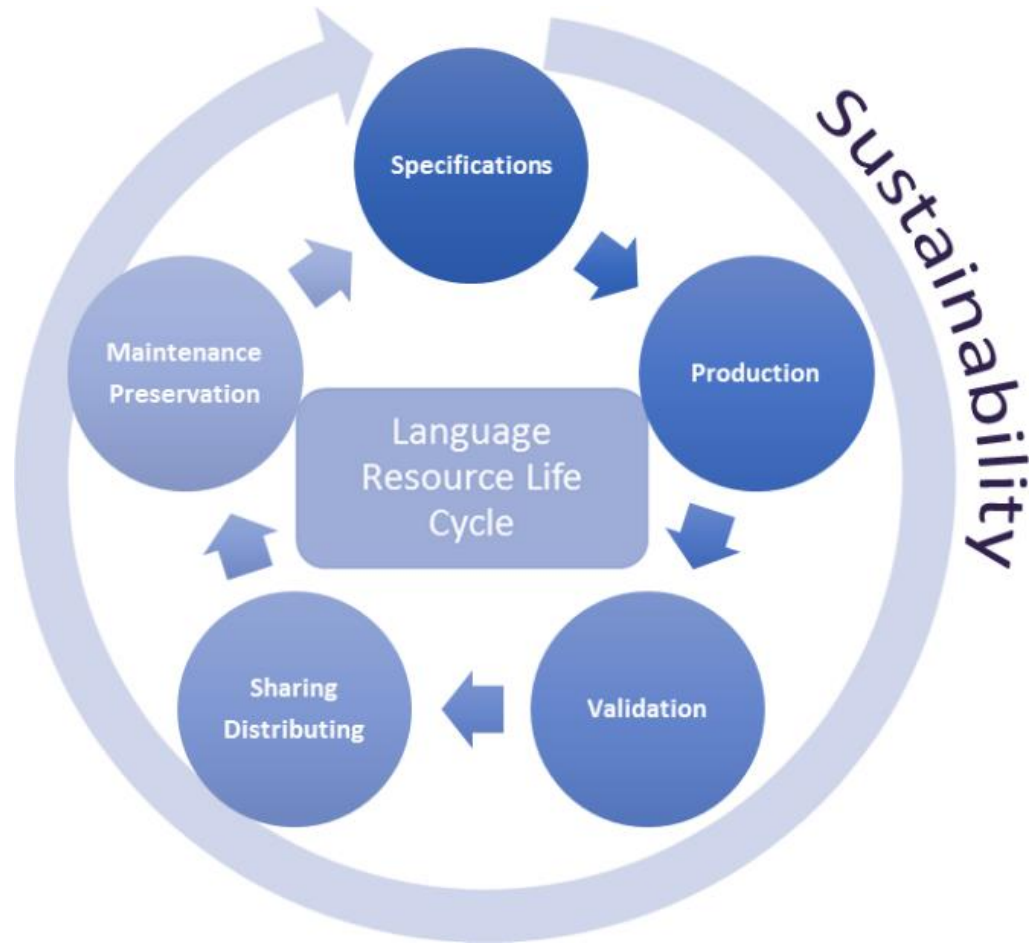
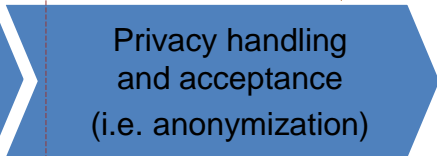
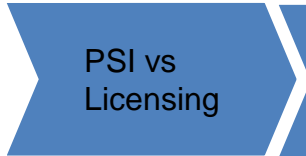


Illustration of Data Packaging Workflow

Data → LR (Language Resources)



Value chain activity

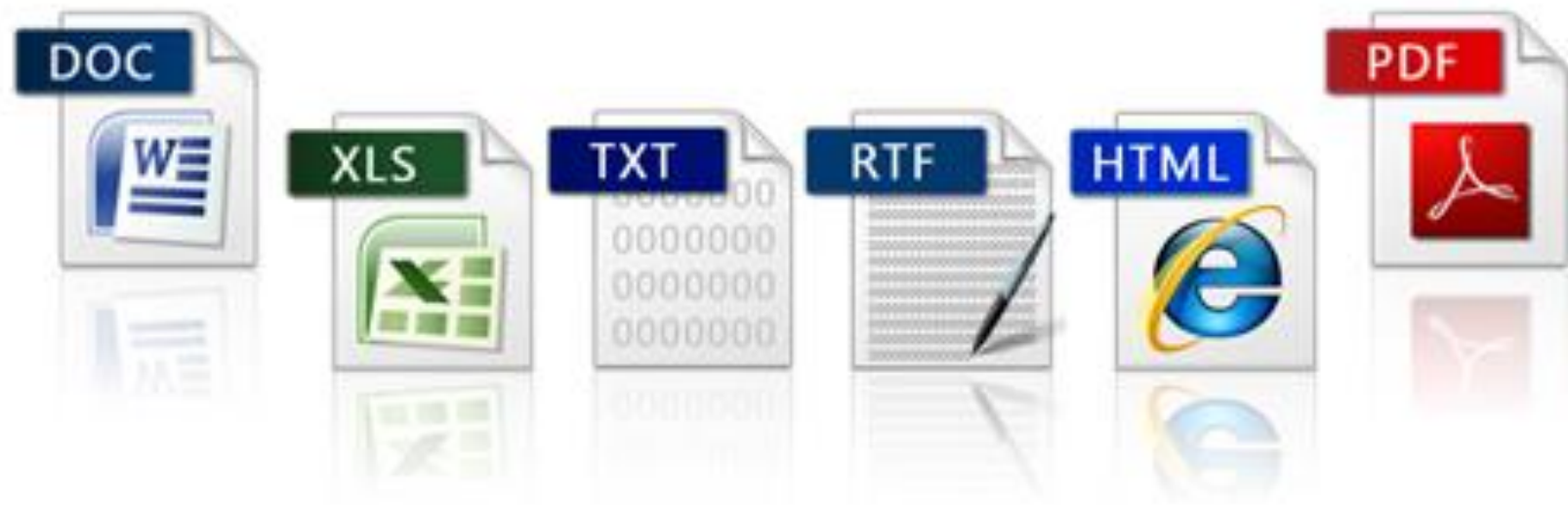


Partnership

ELRC

Public Partner

ELRC / EC

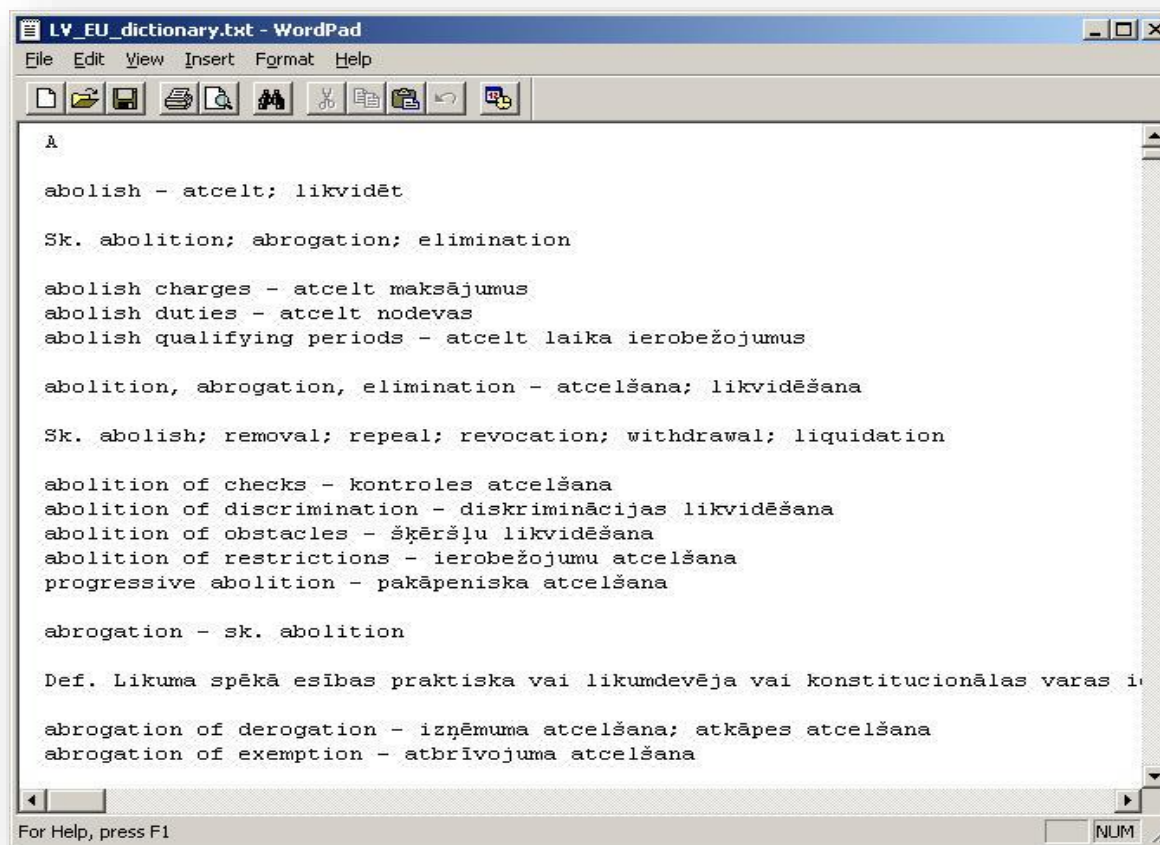


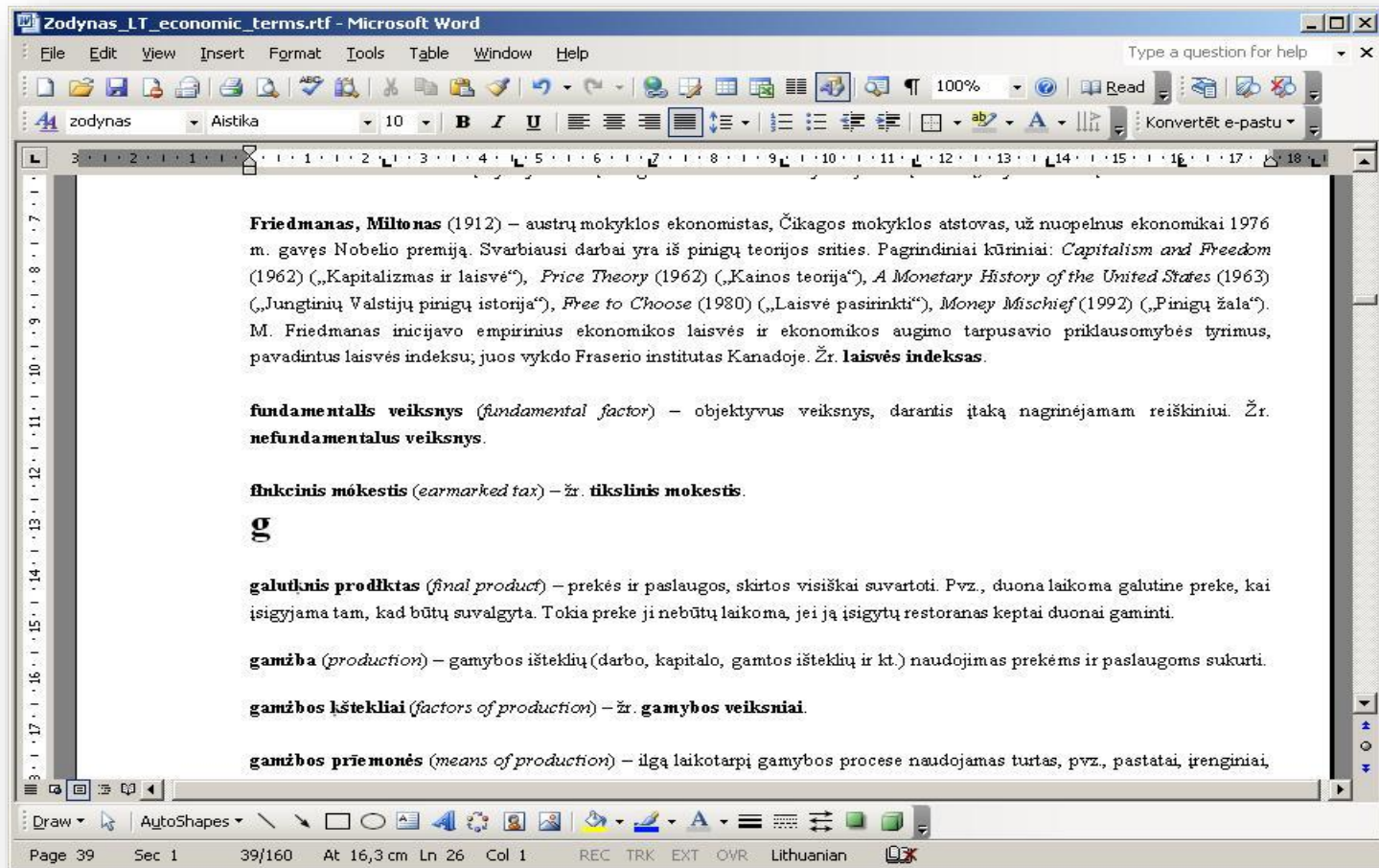
Digital textual data in various formats – valuable language resources

EE-EN-RU Mathematical terms in printed form

L 361 lähenema	106	L 388 lülitü	M1 maagiline	107	M 33 majoreerimine
L 361 lähenema piirväärtusele	approach a limit	приближаться к пределу	M		
L 362 lähenemine	approaching	приближение	M 1 maagiline ruut	magic square	магический квадрат
L 363 lähim	closest, nearest	ближайший	M 2 maastikuformaat	landscape size	горизонтальный формат
L 364 lähis-	adjacent	прилежащий	M 3 maatriks	matrix	матрица
L 365 lähis- (= lähend-)	approximate	приблизженный	M 4 maatriks-	matrix	матричный
L 366 lähiskaadet	adjacent side	прилежащая катет	M 5 maatriksalgebra	matrix algebra	матричная алгебра
L 367 lähiskülg	adjacent side	прилежащая сторона	M 6 maatriksesitus	matrix representation	матричное представление
L 368 lähismurd	convergent	подходящая дробь	M 7 maatriksi astak	rank of a matrix	ранг матрицы
L 369 lähte-	initial, original, starting	исходный, начальный	M 8 maatriksi diagonaal	diagonal of a matrix	диагональ матрицы
L 370 lähteandmed (⇒ algandmed)	initial data, source data	исходные данные, начальные данные	M 9 maatriksi elementaaritehtused	elementary operations on matrices	элементарные операции над матрицами
L 371 lähtehulk	domain, initial set	множество отправления	M 10 maatriksi jälg	spur of a matrix, trace of a matrix	след матрицы
L 372 lähtepunkt (= alguspunkt)	initial point, starting point	исходная точка, начальная точка, отправная точка	M 11 maatriksi pööramine	inversion of a matrix, matrix inversion	обращение матрицы
L 373 lähtesümbol	initial symbol	начальный символ	M 12 maatriksite liitmine	addition of matrices	сложение матриц
L 374 lähtevõrrand	original equation	исходное уравнение, первоначальное уравнение	M 13 maatriksite ring	ring of matrices	кольцо матриц
L 375 lähtuma	start	исходить	M 14 maatriksmäng	matrix game	матричная игра
L 376 lävi	threshold	порог	M 15 maatriksitähistus	matrix notation	матричная символика
L 377 lühem telg (⇒ väiketelg)	minor axis	малая ось	M 16 maatriksväärtustega funktsioon	matrix-valued function	матрица-функция, матричнозначная функция
L 378 lühend	abbreviation	аббревиатура, сокращение	M 17 Mackey topoloogia	Mackey topology	топология Макки
L 379 lühendama	abbreviate, shorten	сокращать, укорачивать	M 18 Maclaurini rida	Maclaurin's series	ряд Маклорена
L 380 lühendamine	abbreviation, shortening	сокращение, укорачивание	M 19 madalaim numbrikoht	least significant digit	младший разряд
L 381 lühendatud	abbreviated, abridged, contracted	сокращённый	M 20 madalamat järku	of lower order	нижнего порядка
L 382 lühendatud korrutamine	abbreviated multiplication	сокращённое умножение	M 21 magasin (= pinu)	stack	магазин, стек
L 383 lühendatud tähis	abridged notation, contracted notation	сокращённое обозначение	M 22 magasinialgoritm	stack algorithm	магазинный алгоритм
L 384 lühike	short	короткий	M 23 magasiniautomaat	push-down automaton	магазинный автомат
L 385 lühim	shortest	кратчайший	M 24 magasinii tipp	top of stack	вершина стека
L 386 lüke (⇒ translatsioon)	shift, slide, translation	перемещение, перенос, сдвиг	M 25 magistraal	turnpike	магистраль
L 387 lüli	link	звено	M 26 maha tõmbamine	cross out	вычёркивать
L 388 lüliti	switch	переключатель	M 27 mahutõmbamine	crossing out	вычёркивание
			M 28 mahut (= mahukas)	capacity	ёмкость
			M 29 mahukas	capacious	степенное среднее
			M 30 mahukeskmine	power mean	мажоранта
			M 31 majorant	majorant	мажорировать
			M 32 majoreerima	majorize	мажорирование
			M 33 majoreerimine	majorizing	

EN-LV Term collection in TXT format





Friedmanas, Miltonas (1912) – austrų mokyklos ekonomistas, Čikagos mokyklos atstovas, už nuopelnus ekonomikai 1976 m. gavęs Nobelio premiją. Svarbiausi darbai yra iš pinigų teorijos srities. Pagrindiniai kūriniai: *Capitalism and Freedom* (1962) („Kapitalizmas ir laisvė“), *Price Theory* (1962) („Kainos teorija“), *A Monetary History of the United States* (1963) („Jungtinių Valstijų pinigų istorija“), *Free to Choose* (1980) („Laisvė pasirinkti“), *Money Mischief* (1992) („Pinigų žala“). M. Friedmanas inicijavo empirinius ekonomikos laisvės ir ekonomikos augimo tarpusavio priklausomybės tyrimus, pavadintus laisvės indeksu; juos vykdo Fraserio institutas Kanadoje. Žr. **laisvės indeksas**.

fundamentalls veiksnys (*fundamental factor*) – objektyvus veiksnys, darantis įtaką nagrinėjamam reiškiniui. Žr. **nefundamentalus veiksnys**.

flukcinis mokeskis (*earmarked tax*) – žr. **tikslinis mokeskis**.

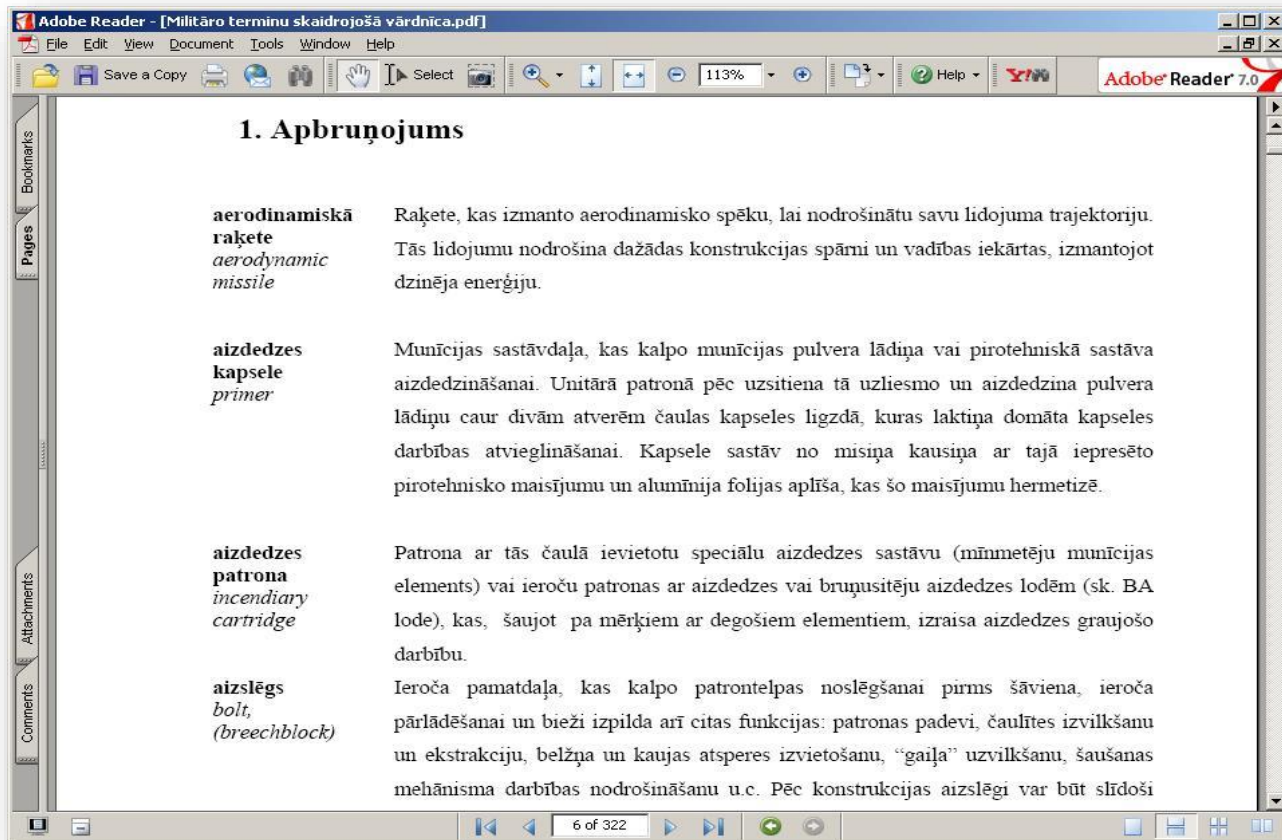
Ū

galutknis prodiktas (*final product*) – prekės ir paslaugos, skirtos visiškai suvartoti. Pvz., duona laikoma galutine preke, kai išsigyjama tam, kad būtų suvalgyta. Tokia preke ji nebūtų laikoma, jei ją išsigytų restoranas keptai duonai gaminti.

gamžba (*production*) – gamybos išteklių (darbo, kapitalo, gamtos išteklių ir kt.) naudojimas prekėms ir paslaugoms sukurti.

gamžbos kštekliai (*factors of production*) – žr. **gamybos veiksniai**.

gamžbos priemonės (*means of production*) – ilgą laikotarpį gamybos procese naudojamas turtas, pvz., pastatai, įrenginiai.



Adobe Reader - [Militāro terminu skaidrojošā vārdnīca.pdf]

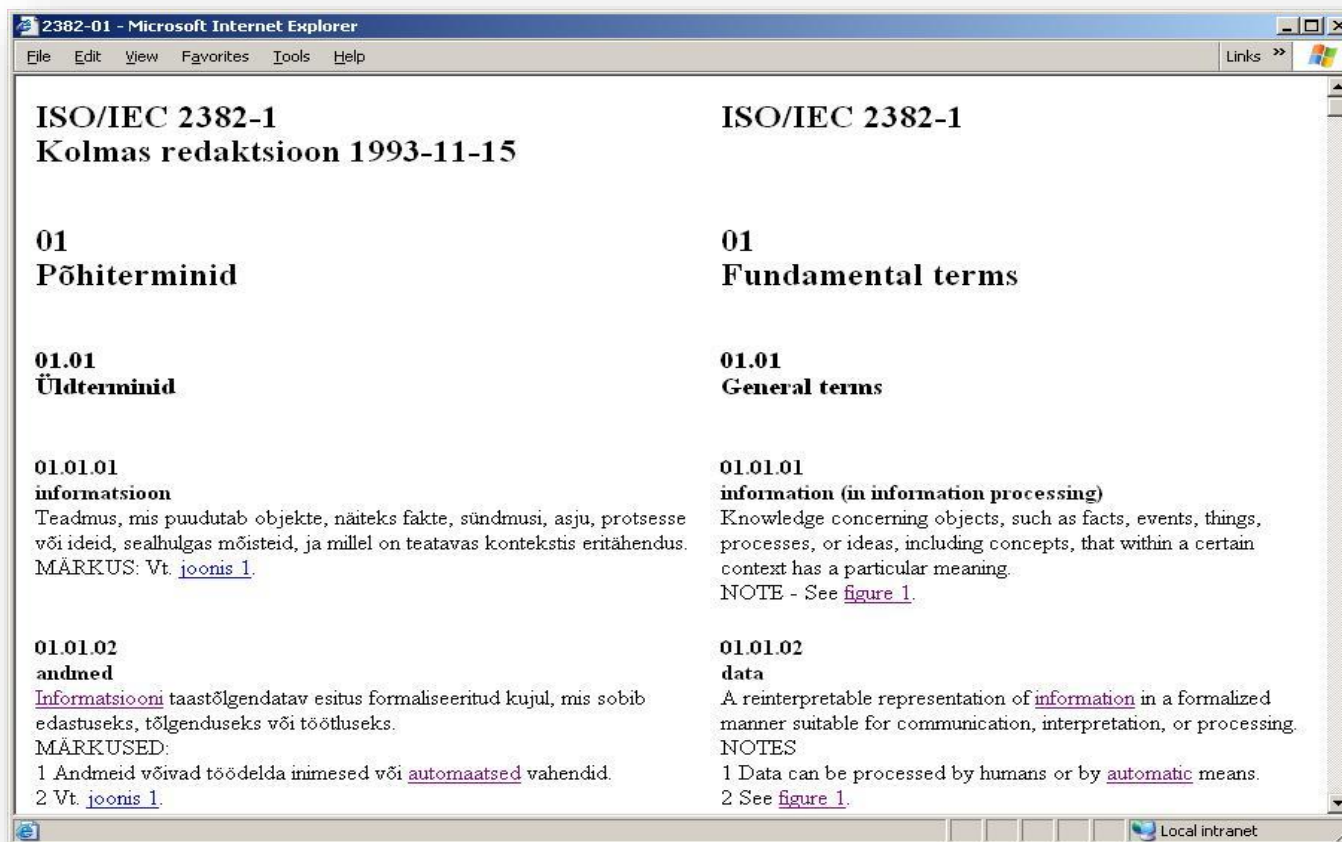
File Edit View Document Tools Window Help

Save a Copy Select 113% Help Adobe Reader 7.0

1. Apbruņojums

aerodinamiskā raķete <i>aerodynamic missile</i>	Raķete, kas izmanto aerodinamisko spēku, lai nodrošinātu savu lidojuma trajektoriju. Tās lidojumu nodrošina dažādas konstrukcijas spāri un vadības iekārtas, izmantojot dzinēja enerģiju.
aizdedzes kapsēle <i>primer</i>	Munīcijas sastāvdaļa, kas kalpo munīcijas pulvera lādiņa vai pirotehniskā sastāva aizdedzināšanai. Unitārā patronā pēc uzsitiena tā uzliesmo un aizdedzina pulvera lādiņu caur divām atverēm čaulas kapsēles ligzdā, kuras laktiņa domāta kapsēles darbības atvieglināšanai. Kapsēle sastāv no misiņa kausiņa ar tajā iepresēto pirotehnisko maisījumu un alumīnija folijas aplīša, kas šo maisījumu hermetizē.
aizdedzes patrona <i>incendiary cartridge</i>	Patrona ar tās čaulā ievietotu speciālu aizdedzes sastāvu (mūmmetēju munīcijas elements) vai ieroču patronas ar aizdedzes vai bruņusitēju aizdedzes lodēm (sk. BA lode), kas, šaujot pa mērķiem ar degošiem elementiem, izraisa aizdedzes graujošo darbību.
aizslēgs <i>bolt, (breachblock)</i>	Ieroča pamatdaļa, kas kalpo patronatelpas noslēgšanai pirms šāviena, ieroča pārlādēšanai un bieži izpilda arī citas funkcijas: patronas padevi, čaulītes izvilkšanu un ekstrakciju, belzņa un kaujas atsperes izvietošānu, "gaiļa" uzvilkšanu, šaušanas mehānisma darbības nodrošināšanu u.c. Pēc konstrukcijas aizslēgi var būt slidoši

6 of 322



2382-01 - Microsoft Internet Explorer

File Edit View Favorites Tools Help Links »

ISO/IEC 2382-1 Kolmas redaktsioon 1993-11-15	ISO/IEC 2382-1
01 Põhiterminid	01 Fundamental terms
01.01 Üldterminid	01.01 General terms
01.01.01 informatsioon Teadmus, mis puudutab objekte, näiteks fakte, sündmusi, asju, protsesse või ideid, sealhulgas mõisteid, ja millel on teatavas kontekstis eritähendus. MÄRKUS: Vt. joonis 1 .	01.01.01 information (in information processing) Knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning. NOTE - See figure 1 .
01.01.02 andmed Informatsioon taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötluks. MÄRKUSED: 1 Andmeid võivad töödelda inimesed või automaatsed vahendid. 2 Vt. joonis 1 .	01.01.02 data A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. NOTES 1 Data can be processed by humans or by automatic means. 2 See figure 1 .

Local intranet

LV-RU Multi-Domain term list in Excel



Microsoft Excel - Dainu skapis 2006 Jan.xls

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial 8 B I U

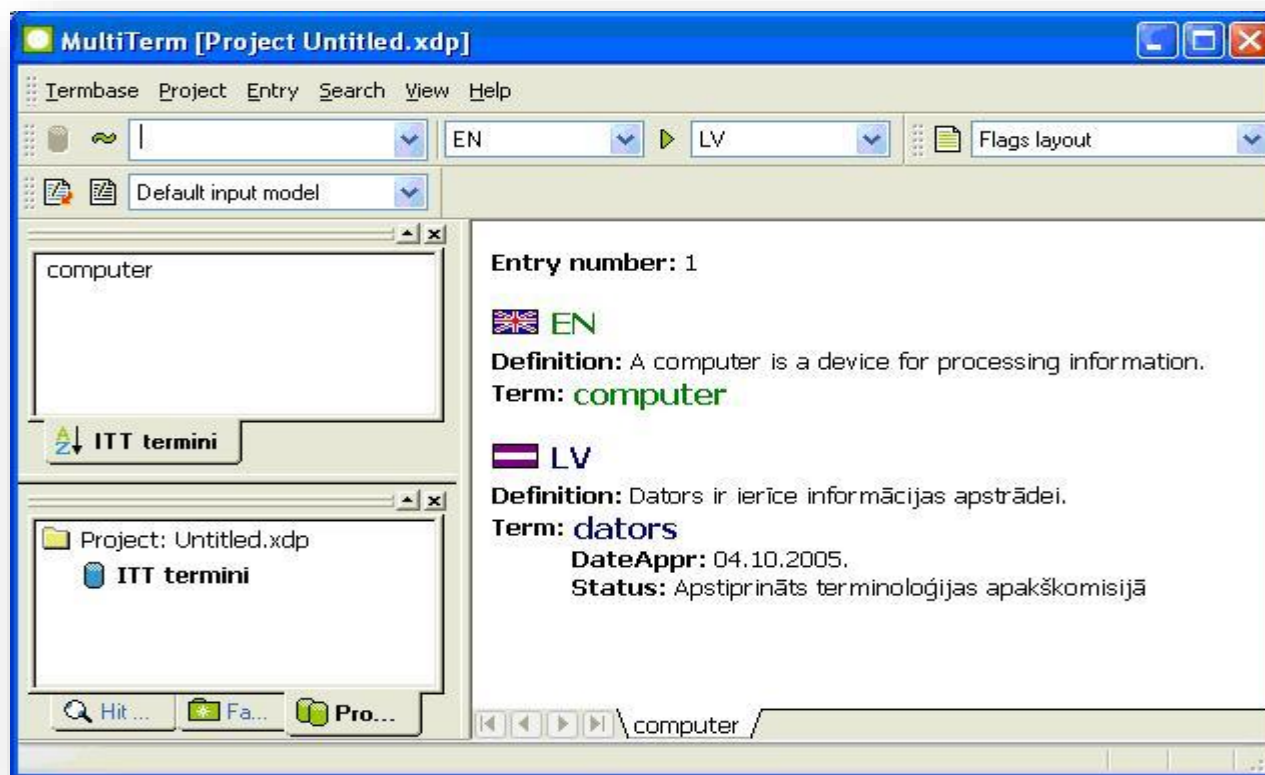
A8 diagnosticēšanas noteikumi

	A	B	C	D	E	F	G
977	elektronoptiskais pārveidotājs	электронно-оптический преобразователь	krimināl.	1979			10
978	hidropārveidotājs	гидропреобразователь	lauks. tehn.	1974			33
979	induktīvais pārveidotājs	электродинамический преобразователь	elektron.	1978			26
980	induktīvais pārveidotājs	индуктивный преобразователь	elektron.	1980			40
981	kapacitīvais pārveidotājs	емкостной гидропреобразователь	elektron.	1980			40
982	kapacitīvais pārveidotājs	электростатический преобразователь	elektron.	1978			26
983	magnetrostriktīvais pārveidotājs	магнитострикционный гидропреобразователь	elektron.	1978			14
984	neapgriežamais pārveidotājs	необратимный преобразователь	elektron.	1978			26
985	pjezoelektriskais pārveidotājs	пьезоэлектрический преобразователь	elektron.	1978			14
986	pneimohidropārveidotājs	пневмогидропреобразователь	lauks. tehn.	1974			33
987	signālu pārveidotājs	устройство преобразования сигналов; УПС	elektron.	1978			18
988	pārvēlums	перекат	šūš.	1980			20
989	kursa pārvešana	перевод румбов	jūrn.	1981			8
990	meridonālais siltuma pārvietojums	межширотный перенос	meteor.		38	120	
991	ārpustreses pārvietošanās trīsdimensiju telpā	внетрассовое перемещение в трехмерном пространстве	ek. ģ.	1975			14
992	atpakaljā pārvietošanās	"поплатное" перемещение	ek. ģ.	1975			14
993	ciklona anomālā pārvietošanās	анормальное перемещение циклона	meteor.		38	120	
994	kravas pārvietošanās	смещение груза	jūrn.	1981			9
995	pierobežas pārvietošanās	пограничное перемещение	ek. ģ.	1975			14
996	sadalošā pārvietošanās	распределительное перемещение	ek. ģ.	1975			14
997	laikapstākļu pārvietošanās; laikapstākļu pārnese	перенос погоды	meteor.		38	120	
998	mitruma pārvietošanās atmosfērā	перенос влаги в атмосфере	meteor.		38	120	

Ready NUM

```
DH-KOZGE.SGH - Notepad
File Edit Format View Help
<dh-kozge>
<entry>
<orth>Abbuchungsauftrag</orth>
<sense>Lehkvíjsi megbkzjs. A számlavezető banknak adott visszavonható megbkzjs, meghatározott összegnek a folyószámlájáról való lehkvíjsjra.</sense>
</entry>
<entry>
<orth>Aberdepot</orth>
<orthvar>
<orth>Summenverwahrung</orth>
</orthvar>
<sense>Értékpapírok bank általi megőrzésének a gyakorlatban kevésbé alkalmazott módja, amikor az értékpapírt betevő ügyfél nem ugyanannak az értékpapírnak, hanem csak ugyanolyan minőségű és mennyiségű értékpapírnak visszavételére jogosult. Nincs szó tehát igazi letétről, a betett értékpapír tulajdonjoga a megőrző banké lesz.</sense>
</entry>
<entry>
<orth>Abfallbeseitigung</orth>
<sense>Hulladék éltívolktíjsa. Híztartíjsi, ipari és mezőgazdasíjsi hulladék begyűjtése, kezelése és tírolíjsa. Az NSZK-ban a tartomínyok kötelesek terveket készítení a hulladék éltívolktíjsjra, feltüntetve a hulladéktírolók helyét is.</sense>
</entry>
<entry>
<orth>Abfertigung</orth>
<sense>Írukezelés. A száíllíktíjsban az íru felvétele, jogilag a fuvarozíjsi szerződés előkészítése és megkötése.</sense>
</entry>
<entry>
<orth>Abfertigungsgebühr</orth>
<sense>Kezelési költség. A fuvarozó íltal a feladíjsi és a rendeltetési íllomíjson végzett szolgáltatíjsok ellenértéke. A kezelés és a száíllíktíjs költsége képezi a fuvardíjsjat.</sense>
</entry>
<entry>
```

EN-LV IT Termbase in Trados MultiTerm

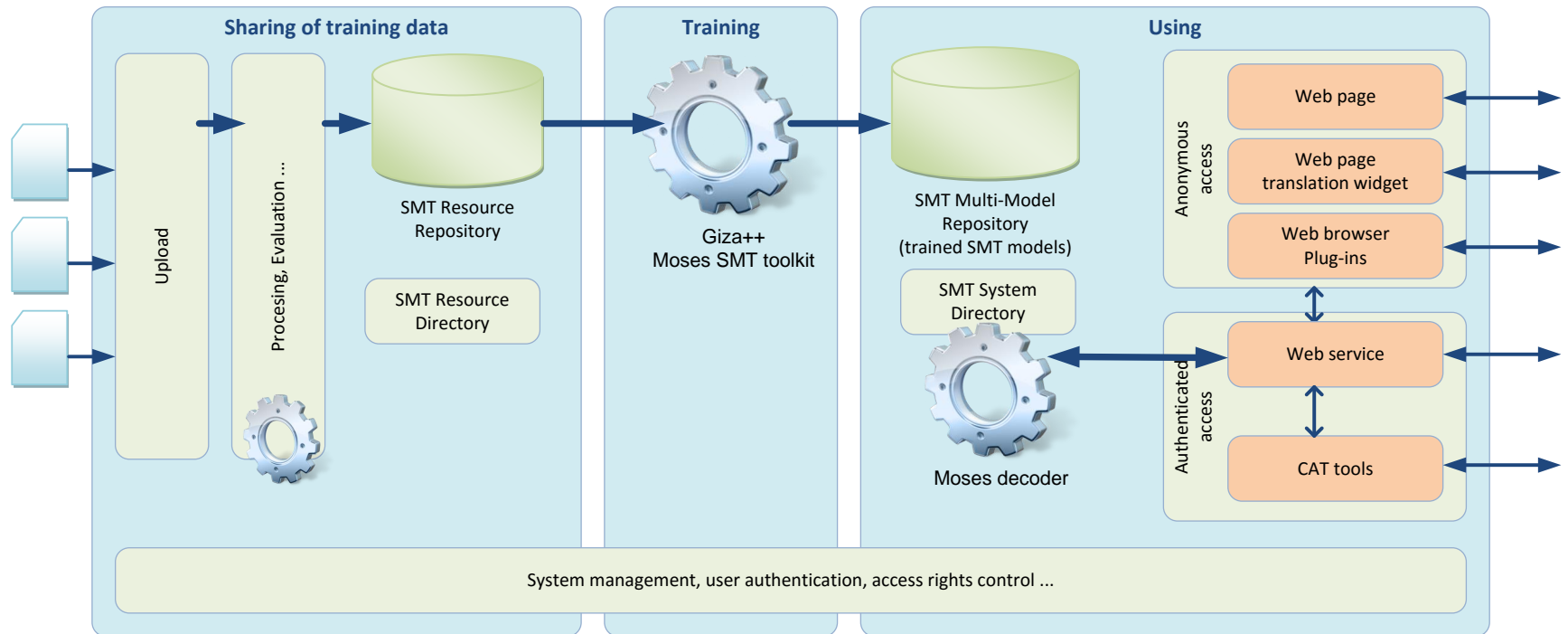


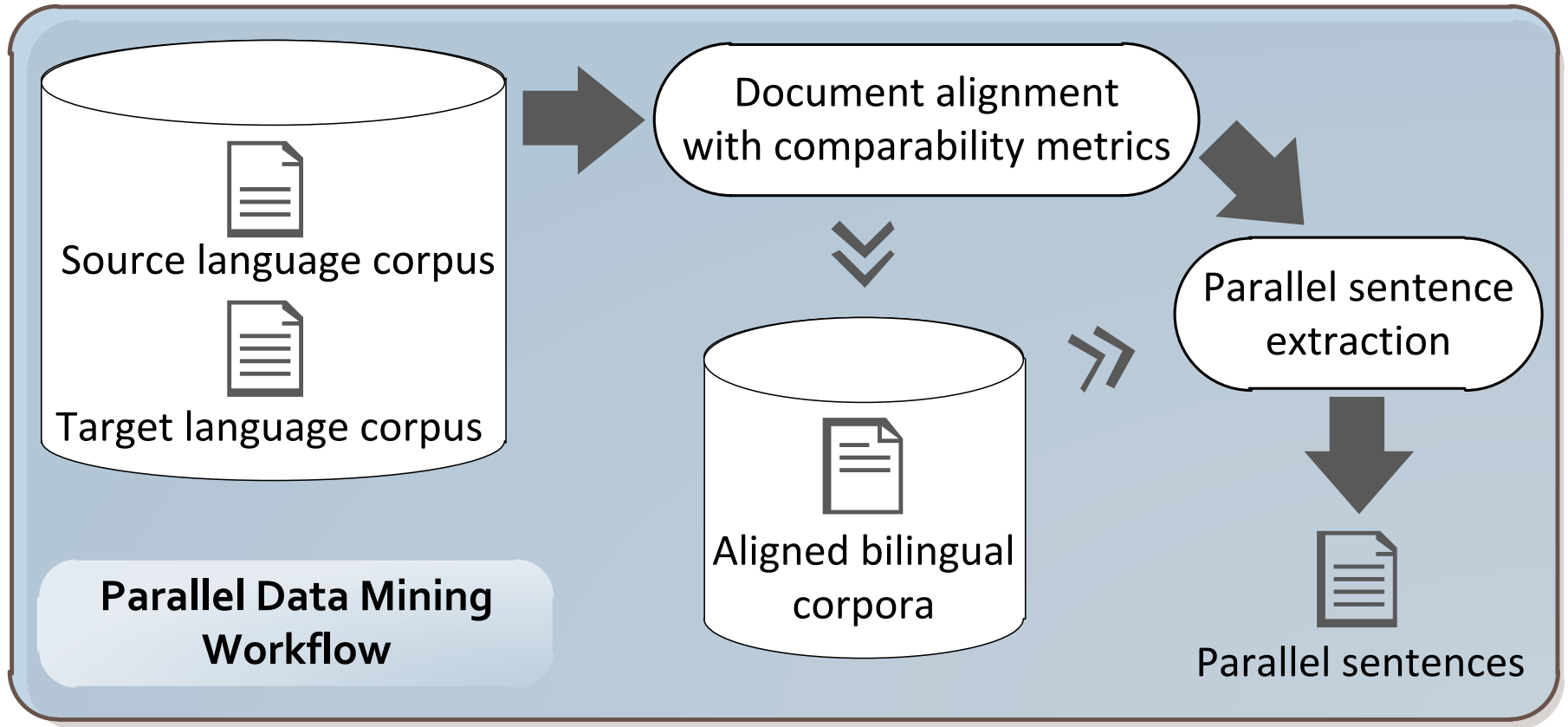


- Identification of sources, identification and selection of data sets (raw data)
 - Data can be obtained from the visible sources
e.g. harvested from web
 - Data can be handed over by the public sector players
 - Public sector players can boost the identification of visible sources
- Can be carried out in cooperation by the ELRC and the data provider

- **Cleaning** of data format
Discarding formatting, encoding character sets e.g. UTF8, formatting features e.g. bold, italic; graphics, ads, tables, html tags, etc.
- File **conversions**
e.g. conversion to XML, XLIFF, etc.
- **Data preparation** for Automated Translation tools
e.g. extraction and alignment
- **Validation** and Quality Control of the output
Language Resource format, content, storage
- **Description** of the Language Resource (meta-data)
- Packaging and **delivery** (data repository with e-sharing) to EC and Owner

Let's MT!





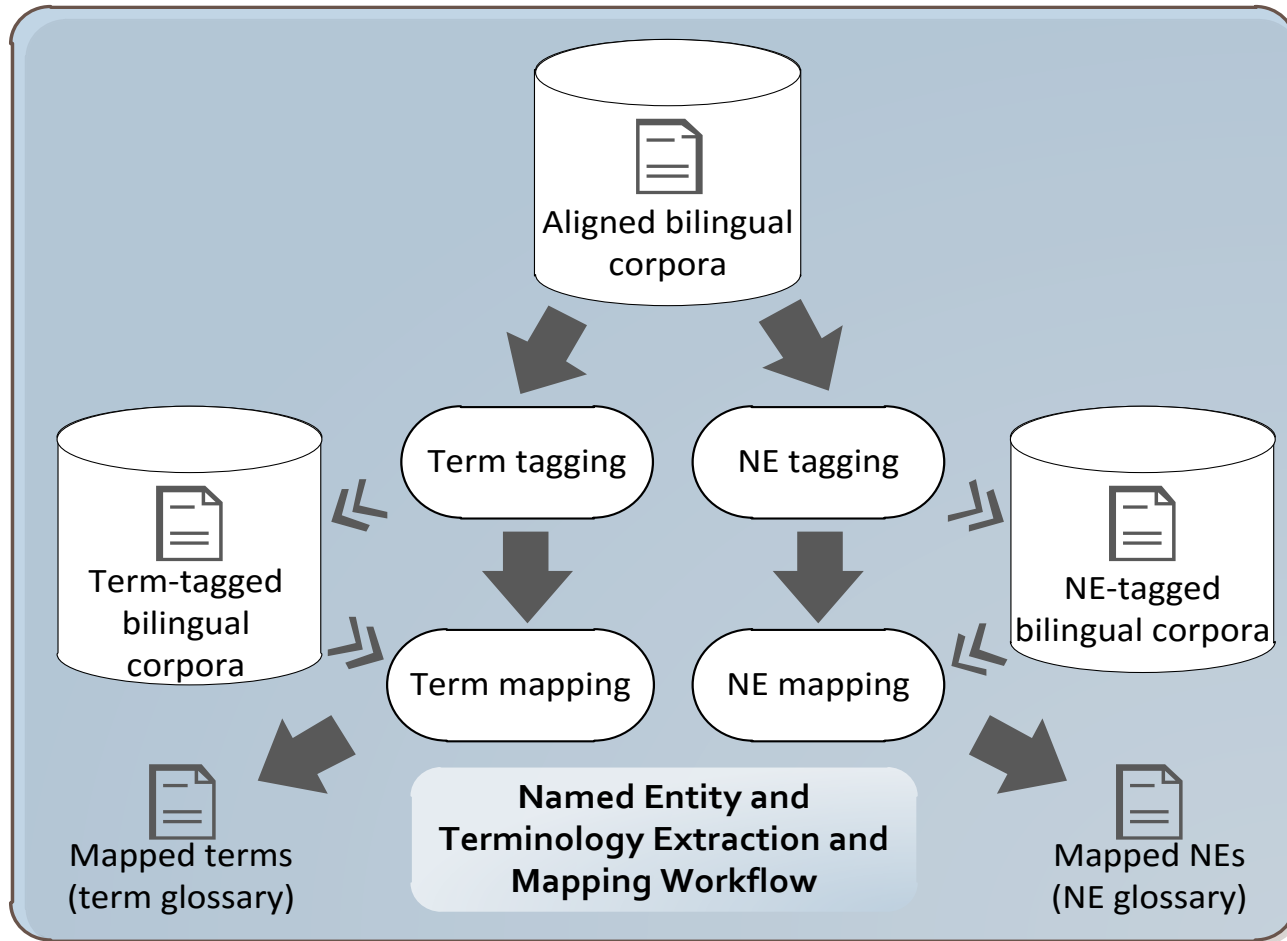
Processing comparable corpora to extract parallel sentence pairs

Sentence alignment from parallel documents



en	el
Greece of art and science	Η Ελλάδα των τεχνών και της επιστήμης
Greece is a place of culture, the arts and sciences.	Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών.
Its tradition of contribution to global cultural and scientific communities, combined with its outstanding natural beauty and excellent infrastructure, has made it an ideal place in which to hold conferences.	Η μακράιωνη συμβολή της στο παγκόσμιο γίνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις άριτες υποδομές, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων.
Over the last few years, Greece has more and more frequently welcomed people of letters, sciences and the arts, who have participated in symposia, conferences and exhibitions.	Τα τελευταία χρόνια, η ελληνική επικράτεια υποδέχεται όλο και συχνότερα ανθρώπους των γραμμάτων, των επιστημών και των τεχνών, οι οποίοι συμμετέχουν σε συμπόσια, συνέδρια και εκθέσεις.
Athens International Airport 'Eleftherios Venizelos', one of the most modern airports in the world in operation since 2001, greatly boosted the organization of international conferences.	Ο Διεθνής Αερολιμένας Αθηνών «Ελευθέριος Βενιζέλος», ένα από τα πλέον σύγχρονα αεροδρόμια παγκοσμίως, ο οποίος λειτουργεί από το 2001, έδωσε μεγάλη ώθηση στη διοργάνωση διεθνών συνεδρίων.
Conference tourism is extremely interdependent: it requires of course a high level of background support from the host country, and at the same time it can actively contribute to improving the overall standard of services in the region.	Ο συνεδριακός τουρισμός είναι άκρως αλληλεπιδραστικός: απαιτεί, βέβαια, ένα υψηλού επιπέδου υπόβαθρο από τη χώρα υποδοχής, ταυτόχρονα όμως συμβάλλει ενεργά στην αναβάθμιση της συνολικής ποιότητας μιας περιοχής.
It is logical that a country chosen as a conference location should be involved in the cultural 'product', giving the public, both residents and visitors, the chance to experience human achievement and innovative thought.	Είναι λογικό, ένας χώρος ο οποίος προτιμάται για τη διεξαγωγή συνεδρίων, να μετέχει προνομιακά στο πολιτιστικό «προϊόν», μιας και δίνει τη δυνατότητα σε κοινό, κατοίκους και επισκέπτες, να έρθουν σε επαφή με τα ανθρώπινα επιτεύγματα και τις καινοτομίες.
The Greece of the pre-Socratic philosophers, of the great poets, of Pheidias the sculptor and Asclepius the physician, extends its hospitality and its warmest welcome, honouring people of intellect and creativity, commerce and scientific progress.	Η Ελλάδα των προσωκρατικών φιλοσόφων, των μεγάλων ποιητών, του Φειδία και του Ασκληπιού υποδέχεται φιλόξενα και τιμά τους ανθρώπους του πνεύματος, του εμπορίου και της προόδου.
Scientific conferences in the land that gave birth to science	Συνέδρια στη χώρα που γέννησε τις επιστήμες
Greece has a large number of esteemed scientists, both here in the country and abroad.	Η Ελλάδα διαθέτει μεγάλο και υψηλής αξίας επιστημονικό δυναμικό, τόσο εντός όσο και εκτός συνόρων.
Greek scientists, with their inventions, innovations and research work, play a leading part in the international scientific community.	Οι Έλληνες επιστήμονες, με τις εφευρέσεις, τις καινοτομίες και το ερευνητικό τους έργο πρωταγωνιστούν στη διεθνή επιστημονική κοινότητα.
Numerous important scientific conferences take place in Greece, reflecting the significance the country places on innovative science.	Τα επιστημονικά συνέδρια που λαμβάνουν χώρα στην Ελλάδα είναι και πολλά και σημαντικά, αντανακλώντας τη σημασία που δίνει η χώρα στις καινοτόμες επιστήμες.
Medical, architectural, natural and humanistic scientific conferences enrich Greece's cultural life, and at the same time give participants the opportunity to experience the	Ιατρικά συνέδρια, αρχιτεκτονικά, φυσικών και ανθρωπιστικών επιστημών, πλουτίζουν την πολιτιστική ζωή της

ACCURAT Tools for Named Entity and Terminology Extraction



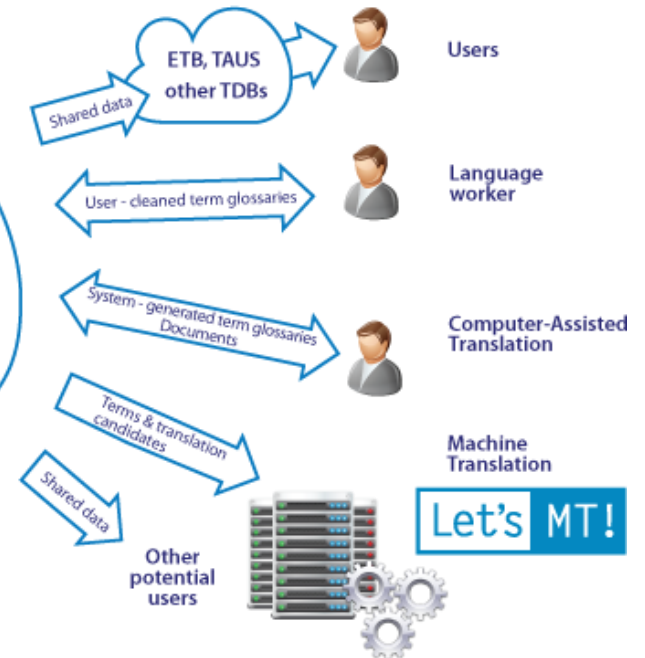
ACQUIRE & PROCESS



PROCESS, STORE & SHARE



REUSE





- Legal Issues
 - Legal status determination (accept/reject decision)
 - Accuracy and acceptance of privacy processing (e.g., anonymization)
 - Application of PSI versus need for a License
- Role of the ELRC Consortium
 - Support on practical issues
 - Technical/legal helpdesk, consultancy
 - Model licensing agreements
 - Government Open Licenses
 - Standard Re-use Licenses
 - License interoperability

- Identify a large source of data on individuals, organizations etc.
- Use a Named Entity Recognizer (NER) to find and remove private bio-data (names, locations, dates, birth information, etc.) and replace with generic placeholders.
- Confirm results meet acceptable requirements
 - Reject data if anonymization not accurate as required

– Sharing/distribution

- Ensure your data falls within the PSI directive as transposed in your country
- If not, foresee an open and permissive licence
- If Privacy is an issue, plan necessary procedures to handle these

– Maintenance/preservation

- The best option is often to partnership with a data center
- See how ELRC can assist you
- There is also the option of national open data portal
- “Putting” data on the web is not an optimal option


We need your involvement



- You know your data
 - visible vs. invisible
- Access to archives, deep web, etc. is often not possible from the outside.
- Not all data is already under PSI or a permissive license
- Access to derived forms (e.g., PDF) is less efficient than access to internal source content repositories.



Languages — the heart of Multilingual Europe







Search

Filter by:

- ▾ Language Name
- ▾ Language Script
- ▾ Resource Type
- ▾ Availability
- ▾ Licence
- ▾ Conditions of Use
- ▾ Linguality Type
- ▾ MIME Type
- ▾ Domain

Resource Type:












- Corpus: 
- Lexical/Conceptual: 
- Language: 
- Description: 

3 Language Resources

Order by: Resource Name A-Z ▾


INTERA Corpus - the Greek part of the EL-EN pair 0 19

English Modern Greek (1453-)

-  ODbL
-  ODC-BY
-  LO-OL-FR
-  DL-DE-BY
-  IODL-IT
-  OGL-UK
-  NLS OD-L-FIN
-  NCGL-UK
-  DL-DE-ZERO
-  NLOD-NO
-  PDDL




my movies 2 8

Danish Modern Greek (1453-)


-  PDDL

press releases of the mkouts Ministry 5 9

Croatian

-  CC BY-NC
-  PUBLIC DOMAIN
-  PDDL

The European Language Resource Coordination - Connecting Europe Facility repository
created on the basis of the META-SHARE repository software



CC-BY-NC-SA 4.0



Helpdesk

Got a question? We're here to help!

We are happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

Please feel free to contact us through one of the following channels:

a Web forum

[Web forum](#)

Telephone*

+33 970 440 522

reach the Secretariat Support at

+49 681-8575 5285

Skype

CEF-AT-Helpdesk

E-mail

help@cef-at-helpdesk.org