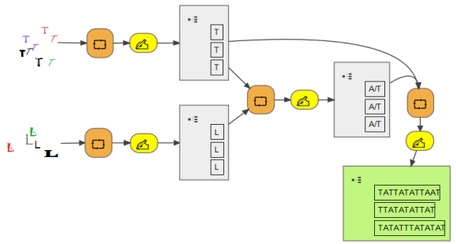
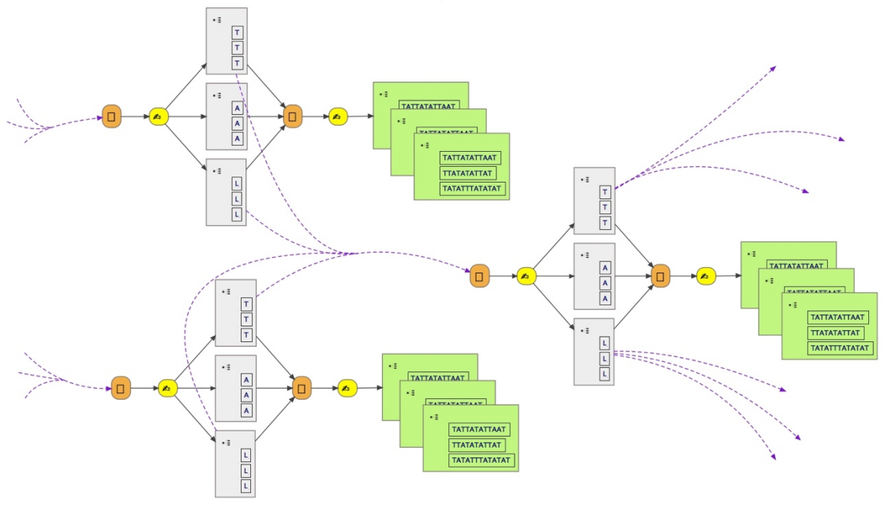
Digital projects today



Cross-project ideal



Five Translations of Aristotle's Categories, or, How to Get Beyond the Siloes of Translation Studies

Joel Kalvesmaki, Editor in Byzantine Studies, Dumbarton Oaks

Follow along:

http://textalign.net/doha

About me

email: kalvesmaki@gmail.com
web: http://kalvesmaki.com
Twitter: @kalvesmaki
Github: https://github.com/Arithmeticus

About the Text Alignment Network

email: kalvesmaki@gmail.com
web: http://textalign.net
Twitter: @textalign
Github: https://github.com/textalign

Project	URL	Type	Field	Encoding type	Encoding subtype	Langs	What's being aligned?	Alignment mechanism	Stand-off?	Smallest alignment	Alignments tethered to semantics?	Cross-project alignment?	Sample alignment	Comments
Arabic English Parallel News Text	https://secure.ldc.upenn.edu/	project	CL	SGML	custom	multi	work-version <--> work-version	DOC/p/seg/@id	y	sentence	no	no	<SentPair EnglishSegId= "2,3,4" ArabicSegId= "2" />	
English Translation Treebank: An Nahar Newswire	https://secure.ldc.upenn.edu/	project	CL	SGML	custom	multi	work-version <--> work-version	DOC/p/seg/@id	y	sentence	no	no		
Multiple-Translation Chinese (MTC)	https://secure.ldc.upenn.edu/	project	CL	SGML	custom	multi	work-version <--> work-version	DOC/p/seg/@id	y	sentence	no	no		
PLUG Word Aligner	http://stp.lingfil.uu.se/corpora/plug/pwa/	project	CL	TXT	Linköping	multi	work-version <--> work-version	separate files, aligned by numbered lines	y	sentence	no	no	1X:8:8:8:8:(1) Den kommer att fullföljas med kraft och konsekvens.(2) It will be pursued with firmness and consistency.	
GALE Phase 1 Chinese Newsgroup Parallel Text	https://secure.ldc.upenn.edu/	project	CL	TXT	tab-delimited	multi	work-version <--> work-version	separate files, aligned by numbered newlines	y	sentence	no	no		
PLUG Word Aligner	http://stp.lingfil.uu.se/corpora/plug/pwa/	project	CL	TXT	Uppsala	multi	work-version <--> work-version	single file, lines grouped in numbered sets	n	sentence	no	no	# fields: (id,source,target) svenprf2 ledamöter members	
ACL 2005 project	http://www.cse.unt.edu/~rada/wpt05/	project	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered newlines; companion concordance for token-to-token alignment	y	token	no	no	18 2 2 P 0.7	
BOLT Phase 1 Egyptian Arabic Parallel Word Alignment DF	https://secure.ldc.upenn.edu/	project	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered lines; separate concordance for word tokens	y	token	no	no	22-23(COR) 19[GLU],20-20,22(COR) 6-9(COR) 16-26(COR) 4-7(COR) 9[TOK]-14(COR) 14[MET]-(MTA)	
BOLT Phase 2 Chinese Parallel Word Alignment and Tagging SMS	https://secure.ldc.upenn.edu/	project	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered lines; separate concordance for word tokens	y	token	no	no	22-25[OMN],26,27,28[POS](GIS)	
European Parliament Proceedings	http://www.statmt.org/europarl/	project	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered newlines	y	sentence	no	no		
Hong Kong Laws Parallel Text	https://secure.ldc.upenn.edu/	project	CL	TXT		multi	work-version <--> work-version		y	sentence	no	no		uses <s id="#.#">...</s> but not in XML-compliant manner
Kyoto Free Translation Task	http://www.phontron.com/kftt/	project	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered newlines	y	sentence	no	no		
Moses	http://www.statmt.org/moses/	tool	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered lines; separate concordance for word tokens	y	token	no	no	0-0 0-1 1-2 2-3	
News Commentary	http://www.statmt.org/wmt13/translation-task.html#download	project	CL	TXT		multi	work-version <--> work-version	separate files, aligned by unnumbered newlines	y	sentence	no	no		
Wiki Headlines	http://www.statmt.org/wmt13/translation-task.html#download	project	CL	TXT		multi	work-version <--> work-version	single file, bitexts separated by three pipes	n	token	no	no	Аарон Хант ||| Aaron Hunt	
PLUG Word Aligner	http://stp.lingfil.uu.se/corpora/plug/pwa/	project	CL	XML	custom	multi	work-version <--> work-version	TEI-like align/seg values	n	token	no	no	<align id='svenprf3' link='1-1'><seg lang='sv'><s>Sveriges neutralitetspolitik är av avgörande betydelse för vårt lands fred och oberoende.</s></seg><seg lang='en'><s>Sweden's policy of neutrality is of decisiveimportance for our peace and independence.</s></seg></align>	
InterText	http://wanthalf.saga.cz/intertext	tool	CL	XML	TEI	multi	work-version <--> work-version	Multiple @xml:id identified by stand-off file of link elements	y	token	part	no	<link type='2-1' xtargets='6:1 6:2;7:1' status='man'/>	Permits multiple export formats
MULTEXT-east	http://nl.ijs.si/ME/	project	CL	XML	TEI	multi	work-version <--> work-version	link/@xtargets	y	token	part	no	<link xtargets="Oen.1.1.1.1 Oen.1.1.1.2 ; Oro.1.2.2.1"/>, <link n="1:1" targets="oana-en.xml#Oen.1.1.2.9 oana-ro.xml#Oro.1.2.3.7"/>, <w xml:id="Oen.1.1.2.2.6" lemma="a" ana="#Di">a</w>	Aligns up to ten at a time
PELCRA	http://pelcra.pl/	project	CL	XML	TEI	multi	work-version <--> work-version		y	token	part	no	<w xml:id="w-1"><fs type="morph"><f name="orth">Historia</f></fs>...</w> plus <w xml:id="w-8"><fs type="morph"><f name="orth">History</f></fs>...</w> plus <linkGrp><link target="#w-1 #w-8" type="simple" pelcra:score="0.621777"/></linkGrp>	
Collatex	http://collatex.net	tool	DH	multiple		single	work-version <--> work-version	various: see http://collatex.net/doc/#output	n	token	no	no	various: see http://collatex.net/doc/#output	Five different output methods supported: JSON, 3 XML formats (TEI, GraphML, customized), GraphViz DOT
Classical Text Editor	http://cte.oeaw.ac.at	tool	DH	TXT	custom	multi	work-version <--> work-version	anchors in separate files	y	character	no	no	Text 1: ...ἀνεχώρησε. {Q\s:33\Q}Μετὰ... and Text 2: ...departed. {Q\s:33\Q}Coming...	private, customized markup scheme; alignment also supports nine levels of apparatus critici, which is kept in the same file the main text is in
Parallel Aligned Hebrew-Aramaic and Greek texts of Jewish Scripture	http://ccat.sas.upenn.edu/gopher/text/religion/biblical/parallel/	project	DH	TXT		multi	work-version <--> work-version	aligned word tokens on same line, tab delimited	n	token	no	no	--+ " =;XLM <ge41.1^> E)NU/PNION EI)=DEN	Uses betacode, specialized transliteration scheme
Juxta	http://www.juxtasoftware.org/	tool	DH	XML	custom	single	work-version <--> work-version		y	token	no	no		alignment identified in the software, not in the data
Alpheios	http://repos1.alpheios.net/exist/rest/db/app/align-entersentence.xhtml	tool	DH	XML	custom	multi	work-version <--> work-version	@n and @n-ref values cross-pointing	n	token	no	no	<wds lnum="L1"><w n="1-1"><text>I</text><refs nrefs="1-3"/></w></wds><wds lnum="L2"><w n="1-3"><text>mir</text><refs nrefs="1-1"/></w></wds>	embedded in html code
GATE (General Architecture for Text Engineering)	https://gate.ac.uk/sale/tao/splitch20.html	tool	DH	XML	custom	multi	work-version <--> work-version		n	token	no	no	See URL	exports to XML
Open Scripture Information Standard	http://www.bibletechnologies.net/OSISinformation/	project	DH	XML	custom	multi	Bible-version/commentary <--> Bible-version/commentary	separate files, adopting XML schema	y	clause	no	yes		cross-project alignment applies only to Bibles
Unified Standard Format Markup	http://ebible.org/usfx/#schema	project	DH	XML	custom	multi	Bible-version/commentary <--> Bible-version/commentary	separate files, adopting XML schema	y	clause	no	yes		cross-project alignment applies only to Bibles
XCES	http://www.xces.org/	project	DH	XML	custom (based on TEI)	multi	work-version <--> work-version	two layers of links based on element ids	y	token	no	no	Too many to illustrate concisely. See examples at http://www.xces.org/schema/#standoff	
Computational Historical Semantics	http://www.comphistsem.org/	project	DH	XML	TEI	single	Editions <--> lexicon	w/@lemma	y	token	part	no	@lemma="[{'id':1284294,'ls':[{'id':7478,'name':'prologus','wfs':[{'id':1284309,'name':'prologus'}]}],'name':'prologus@NN'}]"	Not a bitext alignment, but the lexicon alignment mechanism is instructive
Digital Comparative Edition and Translation of the Shorter Chinese Samyukta Āgama	http://buddhistinformatics.chibs.edu.tw/BZA/	project	DH	XML	TEI	multi	work-version <--> work-version	standoff file using elements outside TEI namespace	y	text clusters	part	no	<mets:file MIMETYPE="text/xml" CHECKSUM="47aa467c7246ca102511c24d57ad2311" SIZE="6843" ID="b356za1336" CHECKSUMTYPE="MD5"><mets:FLocat LOCTYPE="URL" xlink:href="file:///xml/b356za1336.xml"/></mets:file>	
Digital Thoreau	http://digitalthoreau.org/walden	project	DH	XML	TEI	single	work-version <--> work-version	textcrit module + xml:id	y	phrases, readings	part	no	<app xml:id="walc04-app-0004"><lem>shutter</lem><rdg wit="#wc_0a">crevices</rdg><rdg wit="#wc_0d">~~crevices~~<add rendition="pencil">shutter</add></rdg></app>	
English-Norwegian Parallel Corpus	http://www.hf.uio.no/ilos/english/services/omc/enpc/	project	DH	XML	TEI	multi	work-version <--> work-version	s/@id	y	sentence	part	no	<s id=DL2T.1.s19 corresp=DL2.1.s18>"Ikke glem at du har levd godt i fire år nå."</s>	includes morphology on word tokens; unusual English POS tags
Folger Shakespeare	http://www.folgerdigitaltexts.org/	project	DH	XML	TEI	single	Editions <--> commentary		y	token	part	no	<w xml:id="w0000980" n="1.1.8">relief</w>	alignment possibilities not yet fully realized
SAWS	http://ancientwisdoms.ac.uk	project	DH	XML	TEI	multi	work-version <--> work-version, apparatus	lb/@n plus app/@loc; stand-off app + children	y	lines (scriptum-oriented), phrases, readings	part	no	<lb type="WJ" n="2.30"/>τοῦ ἀνθρώπου εἴτε καὶ ἀπὸ φθόνου, σοφῶς ὑπελθὼν ἔξελε τὸν κατη	
Versioning Machine	http://v-machine.org	tool	DH	XML	TEI	single	work-version <--> work-version	textcrit module + xml:id	y	phrases, readings	part	no	<app> <rdg wit="#a660 #ll227">When </rdg> <rdg wit="#h201 #h72 #p1891 #l1894 #cp32">For </rdg> </app>	
VVV Shakespeare	http://www.delightedbeauty.org	project	DH	XML	TEI	multi	work-version <--> work-version		n		part	no	various	TEI imported into native SQL database
mkAlign	http://www.tal.univ-paris3.fr/mkAlign/	tool	DH	XML	TMX	multi	work-version <--> work-version	single file, alignments as sole children of a common parent element, <vu>	n	token	no	no	<tu><tuv xml:lang="EL"><seg> {καὶ,καί.I+Part} {πάντα,πᾶς.A} {προστίθησιν,προστίθημι.V} {ὑμῖν,ὑμεῖς.PRO+Per2p} {τὰ,ὁ.DET} {ἑαυτοῦ,ἑαυτοῦ.PRO+Ref3s}. </seg></tuv><tuv xml:lang="KA"><seg> {და,და.I+Conj} {ყოველსა,ყოველი.PRO+Det} {თავისა,თავი.PRO+Ref} {თქუისსა,თქუი.PRO+Pos} {მოგცემს,მოცემა.V+Mas{თქუენ,თქუენ.PRO+Pers}. </seg></tuv></tu>	Can include part of speech data, embedded as plain text in leaf elements. For TMX see http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

Notes
No known DH or CL projects make use of XLIFF, which one might think conducive to this type of work.