



Designing and executing machine translation workflows through the Kepler framework



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Reginald L. Hobbs and Clare R. Voss
Multilingual Computing Branch
Information Sciences Division
hobbs@arl.army.mil , voss@arl.army.mil

Approved For Public Release; Distribution Unlimited



Background



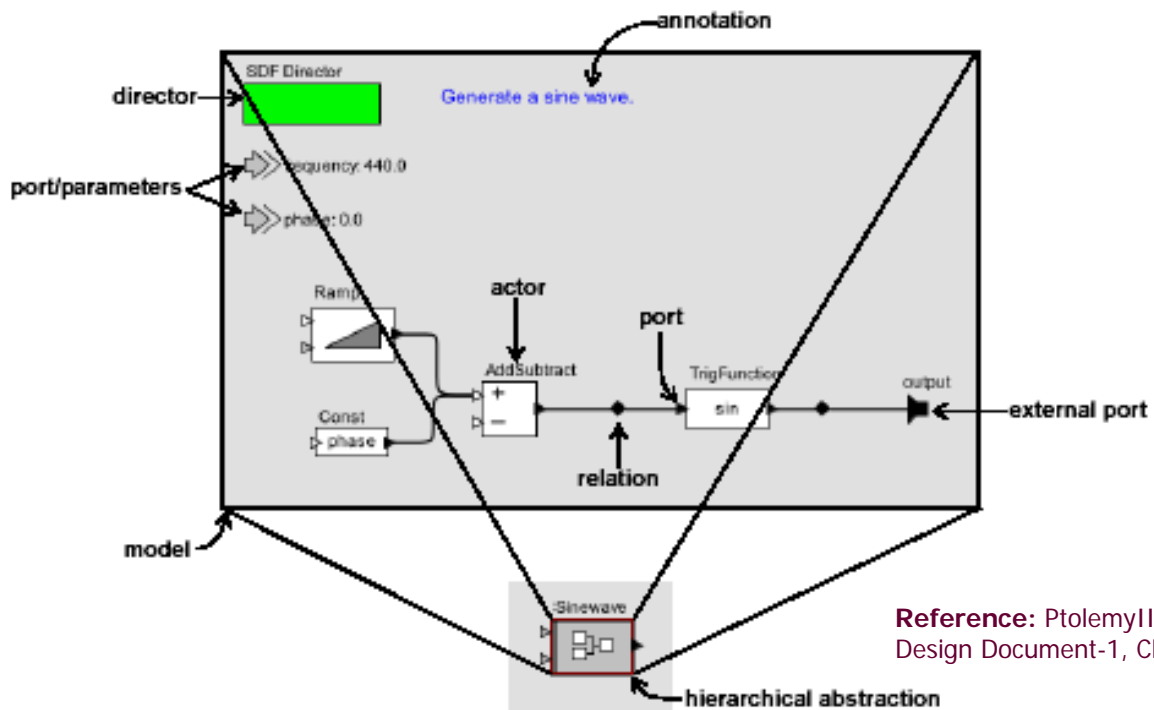
- Who we are in the Multilingual Computing Branch
 - Linguists, Computer Scientists, Translators, Software Developers
- What we do
 - Basic and Applied Research in HLT
 - Research in MT for low density languages customized for military applications.
 - Engineering Lead for Sequoyah Program
- What resources do we have available
 - Configurable, distributed MT testbed with COTS and GOTS systems
 - Linguistic data annotated and archived in Pashto, Dari, Iraqi Arabic and other LCTLs (Less Commonly Taught Languages)

- MLC participated in the NIST Open MT 2008 Workshop
- Urdu-to-English (U2E) track established to challenge participants to build MT for a low-resource language
- Conducted an empirical study of automated post-editing (APE) for augmenting U2E statistical MT
 - Used MOSES tools for building stat MT and APE engines
 - Developed a bitext alignment algorithm for enhancing training data
 - Used automated metrics to evaluate MT output

- Lessons Learned from NIST effort
 - Complicated MT Workflow
 - “Stove-piped” expertise
- Needed a method for documenting the workflow and for configuration management of MT components
- Required support for three different views of the MT problem space
 - Managerial
 - Application
 - Developer

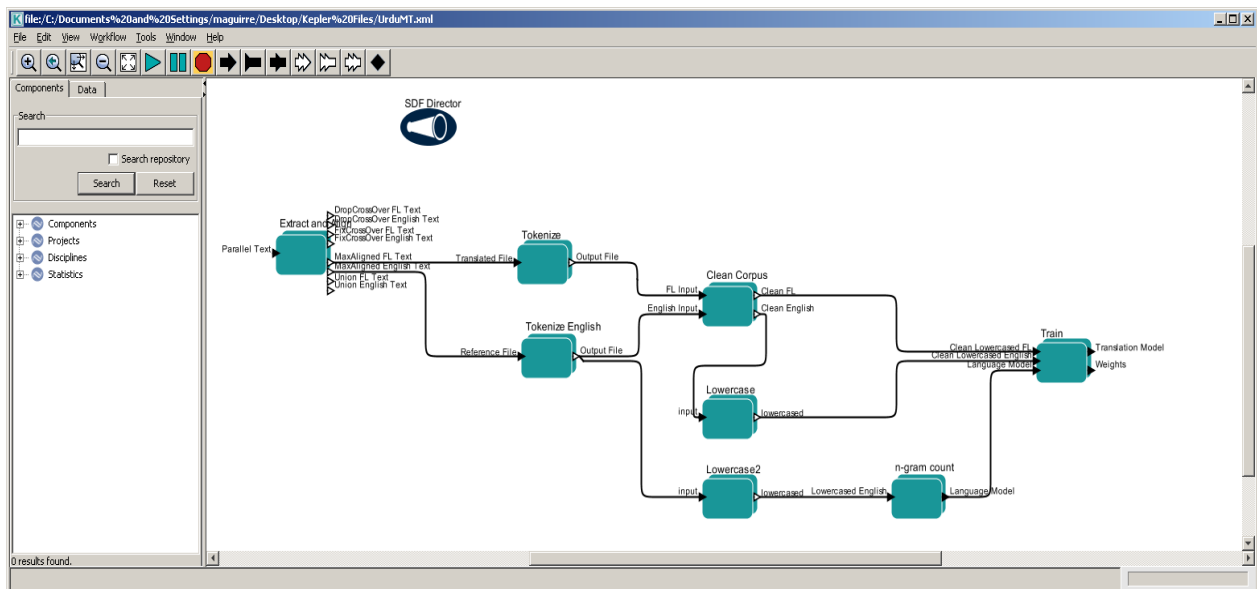
- Scientific workflows (SWFs) are directed graphs with nodes representing computational or process elements and edges representing data channels
- SWFs differ from business workflows (which are process-oriented) in that the focus is on problem-solving through data analysis and transformation.
- SWFs model the flow of data from one step to another in a series of computations that achieve some scientific goal.

- The Kepler project is an open-source scientific workflow system
- Based on Ptolemy II (developed at University of California-Berkeley): a set of Java packages for heterogeneous, concurrent modeling, design and execution.
- Separates the structure of the workflow model from its model of computation
- Workflows represented using an XML-based formal language – MoML (Modeling Markup Language)
- Kepler has been used in other disciplines, particularly bioinformatics, to create repositories for scientific collaboration and data sharing

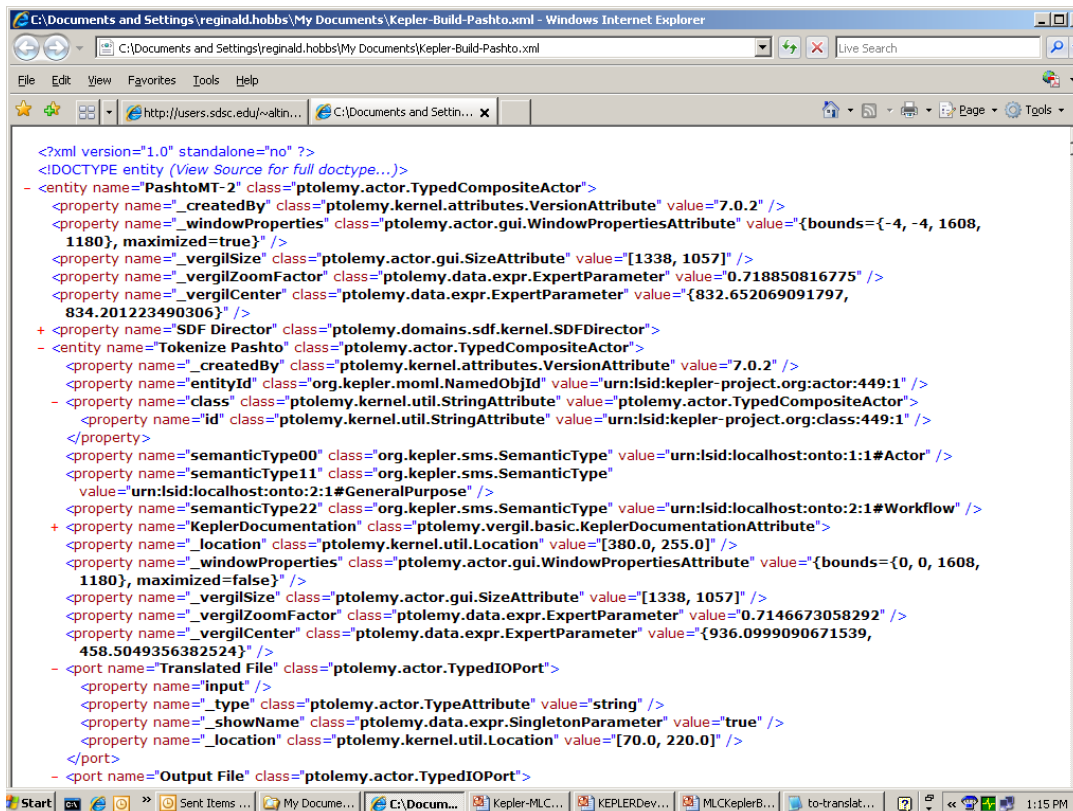


- Actor prototyping tool
- Generic Web and Grid Service
- RDBMS Connection and Querying
- Generic User Interface and Transformation
- Biological Service and Data Access
- Rock Classification
- EML Data Ingestion
- GARP Native Species Pipeline (Using JNI to utilize C++ code)

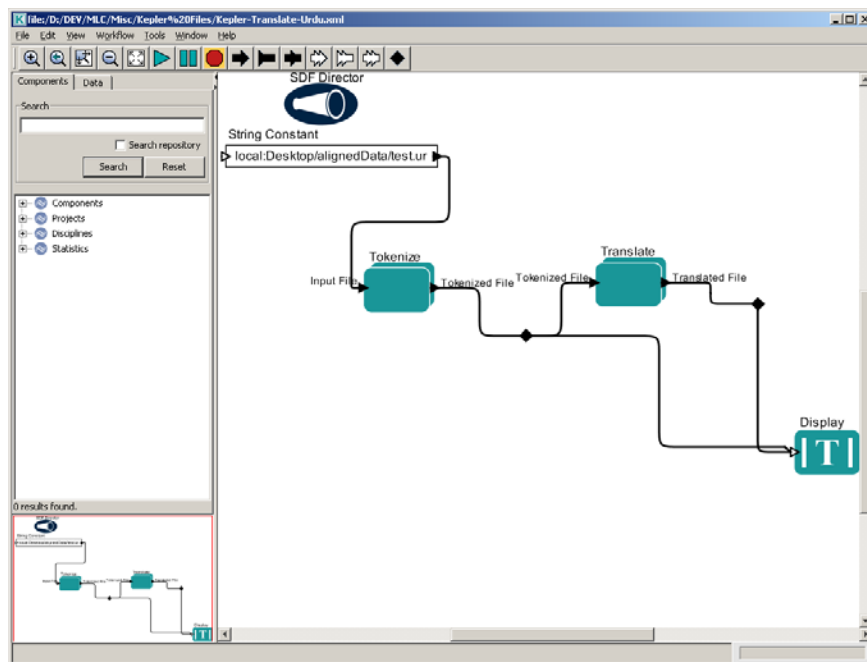
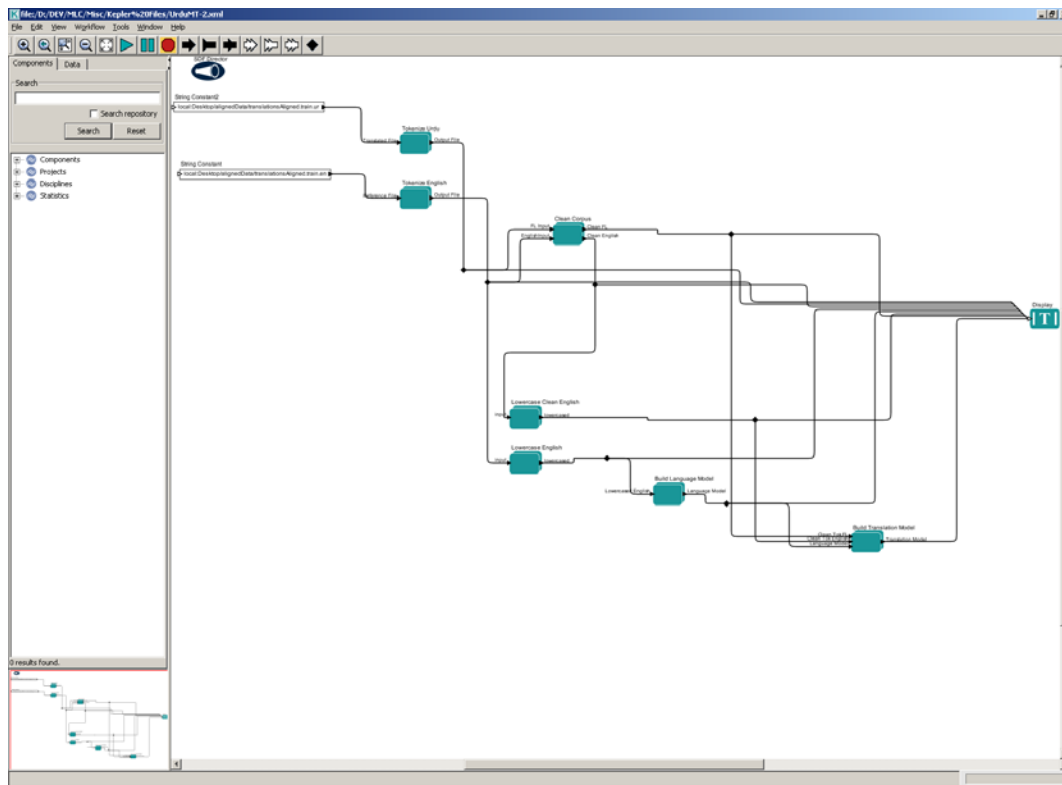
- Software installation and configuration
- Modeling the Urdu-to-English workflow in MoML
- Successfully built an U2E statistical MT engine using Kepler to automate the entire workflow



- Can we do this for another low-resource language that uses Arabic script?
- How much of the U2E workflow can be re-used?
- Overview of Pashto proof-of-concept task
 - Source and characteristics of training data
 - Installation of Kepler client software on a laptop for remote access
 - Design of Pashto build/translate workflows
 - Approach
 - Use Kepler to remotely build a Pashto stat MT engine
 - Use Kepler to translate Pashto source text through generated MT Engine



```
<?xml version="1.0" standalone="no" ?>
<!DOCTYPE entity (View Source for full doctype...)>
- <entity name="PashtoMT-2" class="ptolemy.actor.TypedCompositeActor">
  <property name="_createdBy" class="ptolemy.kernel.attributes.VersionAttribute" value="7.0.2" />
  <property name="_windowProperties" class="ptolemy.actor.gui.WindowPropertiesAttribute" value="{bounds={-4, -4, 1608, 1180}, maximized=true}" />
  <property name="_vergilSize" class="ptolemy.actor.gui.SizeAttribute" value="[1338, 1057]" />
  <property name="_vergilZoomFactor" class="ptolemy.data.expr.ExpertParameter" value="0.718850816775" />
  <property name="_vergilCenter" class="ptolemy.data.expr.ExpertParameter" value="{832.652069091797, 834.201223490306}" />
+ <property name="SDF Director" class="ptolemy.domains.sdf.kernel.SDFDirector">
- <entity name="Tokenize Pashto" class="ptolemy.actor.TypedCompositeActor">
  <property name="_createdBy" class="ptolemy.kernel.attributes.VersionAttribute" value="7.0.2" />
  <property name="entityId" class="org.kepler.moml.NamedObjId" value="urn:lsid:kepler-project.org:actor:449:1" />
  <property name="class" class="ptolemy.kernel.util.StringAttribute" value="ptolemy.actor.TypedCompositeActor">
  <property name="id" class="ptolemy.kernel.util.StringAttribute" value="urn:lsid:kepler-project.org:actor:449:1" />
  </property>
  <property name="semanticType00" class="org.kepler.sms.SemanticType" value="urn:lsid:localhost:onto:1:1#Actor" />
  <property name="semanticType11" class="org.kepler.sms.SemanticType" value="urn:lsid:localhost:onto:2:1#GeneralPurpose" />
  <property name="semanticType22" class="org.kepler.sms.SemanticType" value="urn:lsid:localhost:onto:2:1#Workflow" />
+ <property name="KeplerDocumentation" class="ptolemy.vergil.basic.KeplerDocumentationAttribute">
  <property name="location" class="ptolemy.kernel.util.Location" value="[380.0, 255.0]" />
  <property name="_windowProperties" class="ptolemy.actor.gui.WindowPropertiesAttribute" value="{bounds={0, 0, 1608, 1180}, maximized=false}" />
  <property name="_vergilSize" class="ptolemy.actor.gui.SizeAttribute" value="[1338, 1057]" />
  <property name="_vergilZoomFactor" class="ptolemy.data.expr.ExpertParameter" value="0.7146673058292" />
  <property name="_vergilCenter" class="ptolemy.data.expr.ExpertParameter" value="{936.0999090671539, 458.5049356382524}" />
- <port name="Translated File" class="ptolemy.actor.TypedIOPort">
  <property name="input" />
  <property name="type" class="ptolemy.actor.TypeAttribute" value="string" />
  <property name="showName" class="ptolemy.data.expr.SingletonParameter" value="true" />
  <property name="location" class="ptolemy.kernel.util.Location" value="[70.0, 220.0]" />
  </port>
- <port name="Output File" class="ptolemy.actor.TypedIOPort">
```



- DIALOG BOX 1:
 - Error at copy execution:
 - Exception caught at command: scp -f kepler/pashto/corpus/translationsAligned.train.tok.ur
 - org.kepler.ssh.SshException: Error at acknowledgement: scp: kepler/pashto/corpus/translationsAligned.train.tok.ur: No such file or directory
- /home/kepler/pashto-scripts/tokenize-pashto.sh: line 24: test.ps: No such file or directory
- DIALOG BOX 2:
 - kepler@sandbox-1.arlada.net:kepler/pashto/corpus/translationsAligned.train.tok.en
 - kepler@sandbox-1.arlada.net:kepler/pashto/corpus/translationsAligned.train.tok.ur
 - kepler@sandbox-1.arlada.net:kepler/pashto/corpus/translationsAligned.train.clean.en
 - kepler@sandbox-1.arlada.net:kepler/pashto/lm/translationsAligned.train.lowercased
 - kepler@sandbox-1.arlada.net:kepler/pashto/lm/translationsAligned.train.lm
 - kepler@sandbox-1.arlada.net:kepler/pashto/corpus/translationsAligned.train.clean.lowercased.en
 - kepler@sandbox-1.arlada.net:kepler/pashto/corpus/translationsAligned.train.clean.ur
 - local:Desktop/kepler.tar.gz
- DIALOG BOX 3:
 - Error at copy execution:
 - Exception caught at command: scp -f kepler/pashto/corpus/translationsAligned.train.tok.en
 - org.kepler.ssh.SshException: Error at acknowledgement: scp: kepler/pashto/corpus/translationsAligned.train.tok.en: No such file or directory
- /home/kepler/pashto-scripts/tokenize-english.sh: line 24: /home/kepler/kepler/pashto/corpus/translationsAligned.train.tok.en: No such file or directory
- DIALOG BOX 4:
 - clean-corpus.perl: processing //home/kepler/kepler/pashto/corpus/translationsAligned.train.tok.ps & .en to
 - /home/kepler/kepler/pashto/corpus/translationsAligned.train.clean, cutoff 1-40
 - Use of uninitialized value in open at /home/kepler/bin/moses-scripts/scripts-20080310-0437/training/clean-corpus-n.perl line 38.
 - Use of uninitialized value in concatenation (.) or string at /home/kepler/bin/moses-scripts/scripts-20080310-0437/training/clean-corpus-n.perl line 38.
 - Can't open " at /home/kepler/bin/moses-scripts/scripts-20080310-0437/training/clean-corpus-n.perl line 38.

Original Pashto Source Text:

د نړۍ طاقتونو د عربي ملکونو نه غوښتي دي چې د اسرائیلو سره د سولي په خبرو کې دی د فلسطینیانو سره د ملاتړ د خپلو وعدو احترام وکړی. د منځني ختیځ په باب څلورو طاقتونو نن د جمعي په ورځ په لندن کې یوه بیانیه خپره کړه او په هغې کې یې د عرب ملکونو نه غوښتي دي چې د سولي د جریان په ملاتړ کولو کې دي خپلي سیاسي او مالي وعدې پوره کړی .

GOTS MT Output Text:

World strenght Arabian civil not want he/is that wastefulness <-- with/red greed in inform KI he/is scales <-- with/red support get promise respect would do. The Middle East in door four strenght today collection in day in London KI one explanation spot character and then KI he/it Arab civil not want he/is that greed movement in support victim KI he/is own political and unskilled laborer promise whole did .

Kepler Pashto Stat MT Output Text:

the arab countries of the world طاقتونو people with the اسرائیلو peace talks in the the middle east , four طاقتونو today on friday in london a statement issued and in the arab countries of the people of the peace process in in are سیاسي and financial have completely .

Original Pashto Source Text:

د نړۍ طاقتونو د عربي ملکونو نه غوښتي دي چې د اسرائیلو سره د سولي په خبرو کې دی د فلسطینیانو سره د ملاتړ د خپلو وعدو احترام وکړی. د منځني ختیځ په باب څلورو طاقتونو نن د جمعي په ورځ په لندن کې یوه بیانیه خپره کړه او په هغې کې یې د عرب ملکونو نه غوښتي دي چې د سولي د جریان په ملاتړ کولو کې دي خپلي سیاسي او مالي وعدې پوره کړی .

GOTS MT Output Text:

World strenght Arabian civil not want he/is that wastefulness <-- with/red greed in inform KI he/is scales <-- with/red support get promise respect would do. The Middle East in door four strenght today collection in day in London KI one explanation spot character and then KI he/it Arab civil not want he/is that greed movement in support victim KI he/is own political and unskilled laborer promise whole did .

Kepler Pashto Stat MT Output Text:

the arab countries of the world طاقتونو people with the اسرائیلو peace talks in the the middle east , four طاقتونو today on friday in london a statement issued and in the arab countries of the people of the peace process in in are سیاسي and financial have completely .

wastefulness

political

promise

strength

scales

strength

- Applying Kepler to other MT workflows
 - Pre-processing
 - Post-editing
 - Parallelization
- Support for dynamically modifying workflow components
- Identification of MT domain-specific Kepler objects/actors

- Moses Toolkit
 - <http://www.statmt.org/moses/>
 - Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. "Moses: Open source toolkit for statistical machine translation." In Proceedings of the Annual Meeting of the ACL, Demonstration and poster session, Prague, Czech Republic (2007).
- Kepler Project
 - <http://kepler-project.org/>
 - Ludaescher, B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E.A. Lee, J. Tao, Y. Zhao (2005) "Scientific Workflow Management and the KEPLER System." Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows.
- MLC Publications
 - Voss, C., Aguirre, M., Micher, J., Chang, R., Laoudi, R., Hobbs, R.(2008), "Boosting Performance of Weak MT Engines Automatically: Using MT Output to Align Segments & Build Statistical Post-Editors", In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation (EAMT '09)*, Hamburg, Germany, September 22 - 23, 2008. (To Appear)
 - Voss, C., Laoudi, J. and Micher, J. (2008), "Exploitation of an Arabic Language Resource for Machine Translation Evaluation: using Buckwalter-based Lookup Tool to Augment CMU Alignment Algorithm", In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 28-30, 2008.