*Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts**

Karlheinz Moerth (*ICLTT, Vienna*),

Thierry Declerck (*DFKI, Saarbrücken* & *ICLTT, Vienna*)

Piroska Lendvai & Tamás Váradi (*HASRIL, Budapest*)

* Featuring Alfred, the robot with a passion for Cultural Heritage

Alfred: The reading robot

Alfred discovers an HTML Version of TMI (Thompson's Motif-Index of Folk-Literature)

A. MYTHOLOGICAL MOTIFS

A0–A99. Creator.

A0. **A0. Creator.**--For a general bibliography of creation myths, see Alexander N. Am. 278 n. 15. For bibliographies of North American Indian mythologies arranged by areas, see Thompson Tales 272 n. 1; **Feilberg Skabelses og Syndflodssagn; Jewish: Neuman.--Mexican Indian: (Tarascan) Alexander Lat. Am. 85, (Zapotecan) ibid. 87; GuarayÑŠ: MÐ¹traux RMLP XXXIII 147; Polynesia: Dixon 21 n. 47; Hawaiian: Beckwith Myth 42; Mono-Alu: Wheeler 28, 66f., 70; Easter Is.: MÐ¹traux BMB CLX 313; Marshall Is.: Davenport Folk Tales 221f.; Tahiti: Henry Ancient Tahiti 335ff.; New Hebrides: Codrington II 365.--Armenian: Ananikian 20; African: Werner African 127ff., **Frobenius and Fox, (Loango): PechuÐ»l-Loesche 267; Hindu: Penzer I 10; Buddhist myth: Malalasekera II 338; Icel.: Boberg, MacCulloch Eddic 326; Irish myth: Cross.

A1. A1. *Identity of creator.*

A1.1. A1.1. *Sun-god as creator.*--Egyptian: MÑŒller 69; Persian: Carnoy 260.

A1.2. A1.2. *Grandfather as creator.*--S. Am. Indian (Paressi): MÐ¹traux BBAE CXLIII (3) 359, (GuarayÑŠ): MÐ¹traux RMLP XXXIII 147.

A1.3. A1.3. *Stone-woman as creator.*--Paressi: MÐ¹traux BBAE CXLIII (3) 359.

A1.4. A1.4. *Brahma as creator.*--Buddhist myth: Malalasekera II 338.

A2. A2. *Multiple creators.*

A2.1. A2.1. *Three creators.*--Icel.: Boberg, MacCulloch Eddic 327.--Oceanic: Dixon 24; Hawaii: Beckwith Myth 42.

A2.2. A2.2. *First human pair as creators.* (Cf. A1270.) Chinese: Eberhard FFC CXX 115 No. 70.

A3. A3. *Creative mother source of everything.*--India: Thompson-Balys.

A5. A5. *Reason for creation.*

A5.1. A5.1. *Gods make earth to have place to rest their feet.*--Hawaiian: Beckwith Myth 43.

A7. A7. *Creatorâ€™s descendants.* (Cf. A32.)

A7.1. A7.1. *Creator has two sons.*--GuarayÑŠ: MÐ¹traux RMLP XXXIII 147.

A10. **A10. Nature of the creator.**--India: Thompson-Balys.

A11. A11. *Invisible creator.*--Jewish: Neuman.--Ackawoi: Alexander Lat. Am. 269.

A11.1. A11.1. *Invisibility of creator learned from the impossibility of staring at the sun, his servant.*--Jewish: Neuman.

A12. A12. *Hermaphroditic creator.* The creator is half man and half woman or is thought of as both male and female.--*Lang Myth I 200f., 299; GÑŒEntert 324.--Greek: Eisler 396; Egyptian: MaspÐ¹ro Histoire ancienne des peuples de lâ€˜Orient classique 141; Indian (Hindu): Keith 75.--Aztec: Alexander Lat. Am. 88.

See: http://www.ruthenia.ru/folklore/thompson/. TMI seems to be a good resource for annotating folktale text. But it needs to be converted onto RDF, for allowing Alfred to use it as a semantic resource

Once TMI is in RDF, Alfred will use the *lemon* model (MONNET project*), for encoding the links between labels and concepts in TMI
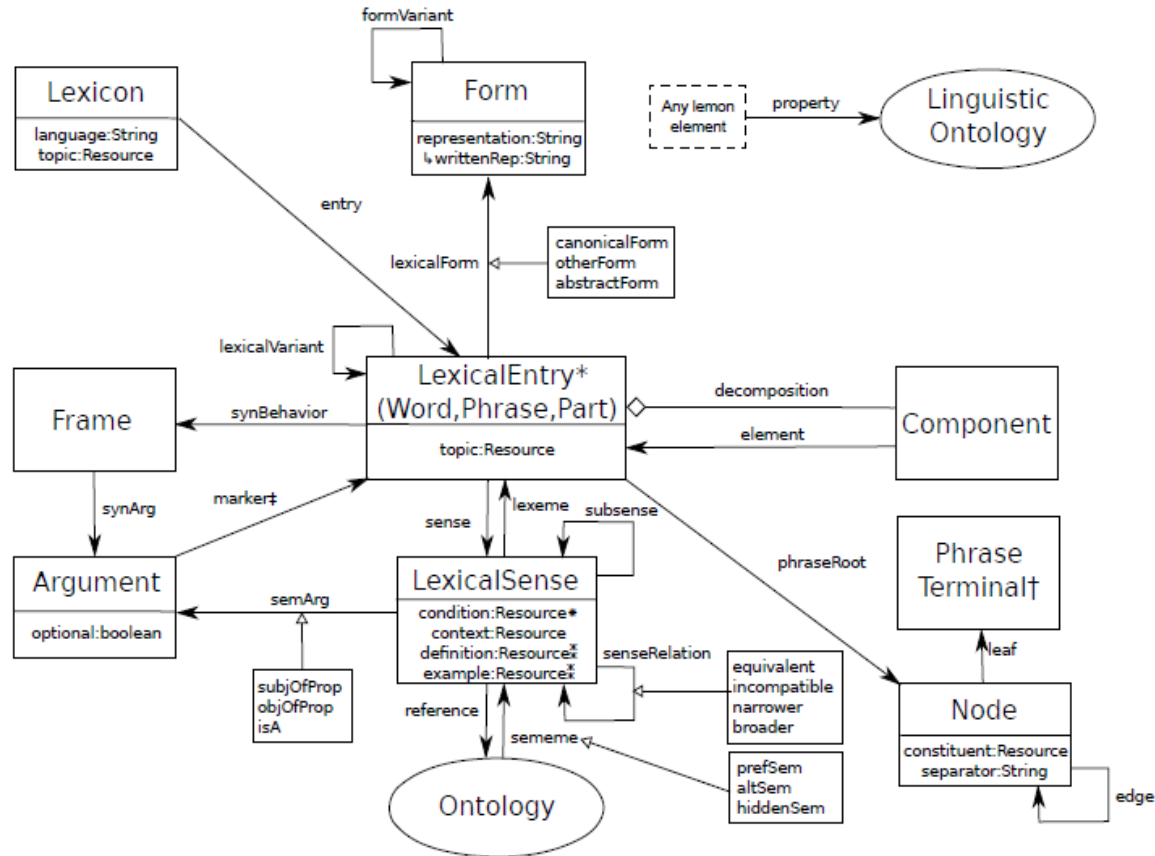
Alfred parse the labels (here using NooJ*) for creating Linguistic Objects that can be encoded in *lemon*

A1. Identity of creator.

A1.1. Sun-god as creator.

A1.2. Grandfather as creator..

A1.3. Stone-woman as creator.

A2. Multiple creators.

A2.1. Three creators

A2.2. First human pair as creators.

.....

A21. Creator from above.

A21.1. Woman who fell from the sky.
....

Extract from Thompson Motif Index for Folktales

Lexical analysis

```
woman,N+Nb=s+Distribution=Hum
who,PRO+Distribution=RelQ
fell,fall,V+Tense=PT+Pers=3+Nb=s
from,PREP
the,DET
sky,N+Nb=s
```

Syntactic analysis

```
<NP>
<NP><HEAD><REFOF XREF="396.2">Woman</HEAD>
</NP>
<SENT><RELCLAUSE>
<SUBJ><XREF>who</XREF></SUBJ>
<PRED>fell</PRED>
<PP><PPOBJ>from
<NP><SPEC>the</SPEC><HEAD>sky</HEAD></NP>
</PPOBJ></PP>
</RELCLAUSE></SENT></NP>
```

Lexical semantics + syntax =>
   Apply to linguistically annotated tales
      for automatic index annotation

* http://www.nooj4nlp.net/pages/nooj.html

Alfred transforms the NooJ XML results into lemon and can start to annotate folktales in English

- But Alfred wants also to „translate" TMI labels in German and Hungarian, for annotating the tales in those languages
  - How?
    - Alfred searches for multilingual lexical resources in RDF format in the Linked Open Data framework
    - Alfred finds the Lexvo* Services
      - In lexvo he sends the query „creator" (for English) =>

* See http://www.lexvo.org/   („Lexvo.org brings information about languages, words, characters, and other human language-related entities to the Linked Data Web and Semantic Web")

Lexvo answer to the query term „creator"

Resource: term/eng/creator

This Lexvo.org page describes the entity referred to by the URI http://lexvo.org/id/term/eng/creator. A machine-readable RDF version of this description is provided here.

rdf:type	lvont:Term
lvont:language	lexvo:iso639-3/eng
skosxl:literalForm	creator ('en' language string)
rdfs:seeAlso	http://en.wiktionary.org/wiki/creator
lvont:means	lexvo:wordnet/30/noun/creator_1_18_00
lvont:means	http://sw.opencyc.org/2009/04/07/concept/en/Creator_Person
lvont:means	http://sw.opencyc.org/2009/04/07/concept/en/_http___dbpedia_org_ontology_creator_
lvont:means	http://sw.opencyc.org/2009/04/07/concept/en/createdBy
lvont:translation	lexvo:term/ang/metod
lvont:translation	lexvo:term/ang/ordfruma
lvont:translation	lexvo:term/ang/wyrhta
lvont:translation	lexvo:term/bul/%D1%81%D1%8A%D0%B7%D0%B4%D0%B0%D1%82%D0%B5%D0%BB
lvont:translation	lexvo:term/ces/stvo%C5%99itel
lvont:translation	lexvo:term/ces/tv%C5%AFrce
lvont:translation	lexvo:term/chu/%D0%B7%D1%8C%D0%B4%D0%B0%D1%82%D1%94%D0%BB%D1%8C
lvont:translation	lexvo:term/chu/%D1%80%D0%BE%D0%B4%D0%BE%D1%82%D0%B2%D0%BE%D1%80%D1%8C%D1%86%D1%8C
lvont:translation	lexvo:term/chu/%D1%82%D0%B2%D0%BE%D1%80%D1%8C%D1%86%D1%8C
lvont:translation	lexvo:term/chu/%D1%85%D1%AB%D0%B4%D0%BE%D0%B6%D1%8C%D0%BD%D0%B8%D0%BA%D1%8A
lvont:translation	lexvo:term/cym/cr%C3%ABwr
lvont:translation	lexvo:term/deu/Sch%C3%B6pfer
lvont:translation	lexvo:term/ell/%CE%B4%CE%B7%CE%BC%CE%B9%CE%BF%CF%85%CF%81%CE%B3%CF%8C%CF%82
lvont:translation	lexvo:term/fin/luoja

HTML output for the human reader. We see that Lexvo uses Wiktionary als its basic lexicon (equivalent to list of wordforms that can be queried by the user)

Lexvo answer to the query „creator"

```
        This information is available under an open source license.
        For detailed license information, please refer to:
        http://www.lexvo.org/legal.html
    -->
- <rdf:Description rdf:about="http://lexvo.org/id/term/eng/creator">
    <rdf:type rdf:resource="lvont:Term"/>
    <lvont:language rdf:resource="http://lexvo.org/id/iso639-3/eng"/>
    <skosxl:literalForm rdf:datatype="xsd:string" xml:lang="en">creator</skosxl:literalForm>
    <rdfs:seeAlso rdf:resource="http://en.wiktionary.org/wiki/creator"/>
    <lvont:means rdf:resource="http://lexvo.org/id/wordnet/30/noun/creator_1_18_00"/>
    <lvont:means rdf:resource="http://sw.opencyc.org/2009/04/07/concept/en/Creator_Person"/>
    <lvont:means rdf:resource="http://sw.opencyc.org/2009/04/07/concept
    /en/_http___dbpedia_org_ontology_creator_"/>
    <lvont:means rdf:resource="http://sw.opencyc.org/2009/04/07/concept/en/createdBy"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/ang/metod"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/ang/ordfruma"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/ang/wyrhta"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/bul/%D1%81%D1%8A%D0%B7%D0%B4%D0
    %B0%D1%82%D0%B5%D0%BB"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/ces/stvo%C5%99itel"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/ces/tv%C5%AFrce"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/chu/%D0%B7%D1%8C%D0%B4%D0%B0%D1
    %82%D1%94%D0%BB%D1%8C"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/chu/%D1%80%D0%BE%D0%B4%D0%BE%D1
    %82%D0%B2%D0%BE%D1%80%D1%8C%D1%86%D1%8C"/>
    <lvont:translation rdf:resource="http://lexvo.org/id/term/chu/%D1%82%D0%B2%D0%BE%D1%80%D1
    %8C%D1%86%D1%8C"/>
```

RDF output for Alfred (and Humans working in the Semantic Web field)

Alfred clicks on the link to Wiktionary,
a multilingual lexical resource on the Web

But Alfred can not read HTML! See later for possible help we can offer him.

The Human clicking on the „plural creator" gets:

Much less information as for the singular form, and no semantic linking between both forms

In Lexvo, Alfred checks the DE and HU translations
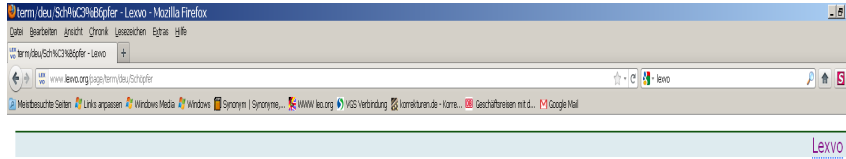
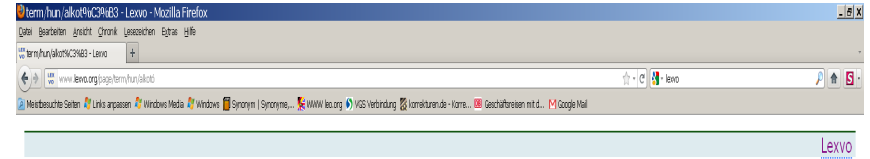

Resource: term/deu/Sch%C3%B6pfer

This Lexvo.org page describes the entity referred to by the URI http://lexvo.org/id/term/deu/Sch%C3%B6pfer. A machine-readable RDF version of this description is provided here.

rdf:type	lvont:Term
lvont:language	lexvo:iso639-3/deu
skosxl:literalForm	Schöpfer ('de' language string)
rdfs:seeAlso	http://en.wiktionary.org/wiki/Schöpfer
lvont:translation	lexvo:term/eng/Creator
lvont:translation	lexvo:term/eng/creator

Lexvo.org 2008-2011 Gerard de Melo. Contact Data Sources Legal Information / Imprint

Resource: term/hun/alkot%C3%B3

This Lexvo.org page describes the entity referred to by the URI http://lexvo.org/id/term/hun/alkot%C3%B3. A machine-readable RDF version of this description is provided here.

rdf:type	lvont:Term
lvont:language	lexvo:iso639-3/hun
skosxl:literalForm	alkotó ('hu' language string)
rdfs:seeAlso	http://en.wiktionary.org/wiki/alkotó
lvont:translation	lexvo:term/eng/author
lvont:translation	lexvo:term/eng/composer
lvont:translation	lexvo:term/eng/constructive
lvont:translation	lexvo:term/eng/creative
lvont:translation	lexvo:term/eng/creator
lvont:translation	lexvo:term/eng/maker
lvont:translation	lexvo:term/eng/productive

Lexvo.org 2008-2011 Gerard de Melo. Contact Data Sources Legal Information / Imprint

Alfred notices that the translation links point to the DE and HU entries within the English Wiktionary page. Is this OK?

To understand the issue, have a look at both EN and DE Wiktionary entries for „Schöpfer"

EN Wiktionary   DE WIktionary ———→

Wouldn't it be better for Lexvo translation link to point to the DE Wiktionary?

Or to compare the EN and HU entries for the Hungarian word „alkotó"

alkotó

Contents [hide]

Hungarian

Etymology

alkot + *-ó*

Pronunciation

- IPA: /ˈɒlkotoː/
- Hyphenation: al·ko·tó

Adjective

alkotó (*comparative* **alkotóbb**, *superlative* **legalkotóbb**)

1. creative, constructive, productive

Noun

alkotó (*plural* **alkotók**)

1. creator, maker, author

Declension

declension of *alkotó*	[show ▼]
possessives of *alkotó*	[show ▼]

Verb

alkotó

1. *present participle of* **alkot**

Szótár | Vitalap

alkotó

Tartalomjegyzék [elrejtés]

Magyar

Melléknév

alkotó

Fordítások

- ANGOL: **creative** (en)
- NÉMET: **schöpferisch** (de)

Képzett szavak

- (*összetételek*): **alkotóelem**

Főnév

alkotó

Fordítások

- ANGOL: **creator** (en)
- NÉMET: **Schöpfer** (de) *(m.)*

Szinonimák

- **alakító**, **építő**, **formáló**

Etimológia

alkot *folyamatos* melléknévi igeneve.

Here the amount of linguistic information seems to be better shared across entries

What is still lacking in Lexvo, since Wiktionary is not (yet) delivering its data in RDF (but see http://code.google.com/p/wikokit/wiki/d2rqMappingSPARQL)

Lexvo is not displaying the hyphenation, pronunciation, senses, etymology, morphological or lexical semantic information associated with a Wiktionary entry

Making Wiktionary readable for Alfred:
Mapping DE-Wiktionary to TEI (www.tei-c.org/) -- 1

ICLTT showcase

- Home
- Corpora
- Toolbox
- Lexicography
- Cooperations
- Resources

- Papers and ...
- Events

German Wiktionary

Query Schöpfer   Search   ☑ TEI (P5)

```
<div>
<entry id="entry_029899">
    <form type="lemma">
        <orth>Schöpfer</orth>    <hyph>Schöp|fer</hyph>    <pron>'ʃœpfɐ</
    <etym>[3] Substantivierung des Verbs [[schöpfen]] mit dem [[Deriva

    <gramGrp>
        <gram type="pos">noun</gram>
        <gram type="gender">masculine</gram></gramGrp>

    <form type="inflected" ana="#n_nom_sg">
        <orth>Schöpfer</orth></form>

    <form type="inflected" ana="#n_nom_pl">
        <orth>Schöpfer</orth>    <hyph>Schöp|fer</hyph>    <pron>'ʃœpfɐ</

    <form type="inflected" ana="#n_gen_sg">
        <orth>Schöpfers</orth></form>

    <form type="inflected" ana="#n_gen_pl">
        <orth>Schöpfer</orth>    <hyph>Schöp|fer</hyph>    <pron>'ʃœpfɐ</

    <form type="inflected" ana="#n_dat_sg">
        <orth>Schöpfer</orth></form>
```
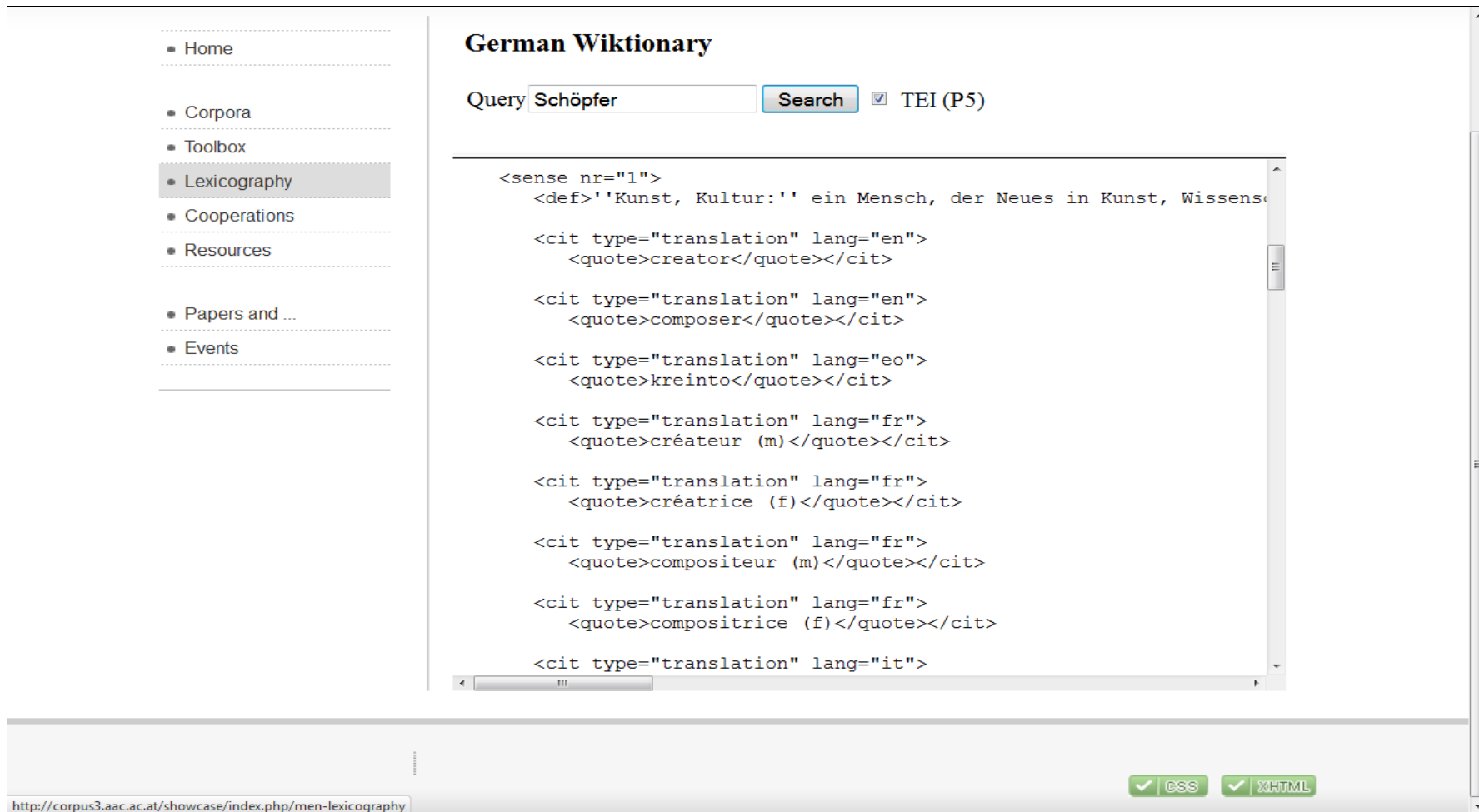
In our work at ICLTT we include the morphological information (using MAF features) in the mapping of Wiktionary to the TEI format, as well as the senses etc. (see next slide). See http://corpus3.aac.ac.at/showcase/index.php/wiktionary001

Making Wiktionary readable for Alfred:
Mapping DE-Wiktionary to TEI (www.tei-c.org/) -- 2

In this screenshot from the ICLTT tools, the reader can see that we also include the translation information of the Wiktionary within the associated sense. See http://corpus3.aac.ac.at/showcase/index.php/wiktionary001

The TEI harmonization work is a first but important step for making Wiktionary readable for Alfred

- Contrary to other approaches for accessing and transforming Wiktionary (for example lexvo, or http://code.google.com/p/wikokit/wiki/d2rqMappingSPARQL), we access the different Wiktionary lexicons and do not consult only the EN Wiktionary.
  - Towards the building of direct bi-lingual wiktionaries, not using necessarily English as the pivot language
- We export the full morpho-syntactic information included in Wiktionary, so that this one can be used for NLP applications like translation or NL generation
- The TEI mapping proposes an harmonization of the Wiktionaries, so that a subsequent mapping of all Wikionaries in RDF could be more easily done (?)

Or maybe starting with RDF/lemon, aggregate Information from various lexical sources and then mapping to TEI?

Links

- Projects:
  - AMICUS: Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (http://ilk.uvt.nl/amicus/)
  - „Totentanz" project at ICLTT (http://www.oeaw.ac.at/icltt/, see also http://www.springerlink.com/content/p704452403885454/
  - Monnet: Multilingual Ontologies for Networked Knowledge (www.monnet-project.eu/)
- Resources
  - TMI: http://www.ruthenia.ru/folklore/thompson/
  - Lexvo: http://www.lexvo.org/
  - Lemon: http://www.monnet-project.eu/lemon
  - ICLTT: http://corpus3.aac.oeaw.ac.at/showcase/index.php/wiktionary
- Papers
  - Declerck, T. and Lendvai, P. Towards a standardized linguistic annotation of the textual content of labels in Knowledge Representation Systems. LREC 2010
  - McCrae J., G. Aguado-de-Cea, P. Buitelaar, P. Cimiano,T. Declerck, A. Gomez-Perez, J. Gracia, L.Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner (forthcoming) Interchanging lexical resources on the SemanticWeb Accepted for publication in LanguageResources and Evaluation, Springer.
  - Thierry Declerck and Piroska Lendvai. Linguistic and Semantic Representation of the Thompson's Motif-Index of Folk-Literature. TPDL 2011

Enjoy!