europeana
connect

Memo

Europeana Language Resources - Evaluation Criteria

Definition of criteria to evaluate the usability of linguistic resources (lemmatizers, translation dictionaries, named entity recognition tools, etc.) for Europeana.

Version	Date	Name
0.1	03/03/2010	Anne Schiller, Xerox
0.2	08/03/2010	Anne Schiller, integrated feedback from Aaron Kaplan

co-funded by the European Union

The project is co-funded by the European Union, through the ***e*Content*plus*** programme

http://ec.europa.eu/econtentplus

Österreichische Nationalbibliothek

EuropeanaConnect is coordinated by the Austrian National Library

Table of Contents

1 Introduction

EuropeanaConnect work package 2 aims to provide multilingual access for international users. One major task is to set up translation modules or services for cross-lingual user queries. Such services rely on language resources. It is not in the scope of EuropeanaConnect to build such resources, but to collect and assess available resources as described in (Bosca, Petras and Schiller 2009) and to adapt and maintain them as necessary.

The quality of a translation module depends undeniably on the quality of the underlying resources, tools or web services (cf. position paper (Luca and Petras 2010)), and we therefore need some criteria to evaluate which resources, tools or services to integrate into Europeana.

One way of evaluating a language resource is to test it in the context of query translation, i.e. setting up a test environment that measures the overall quality of the query results using a given resource. Such end-to-end testing is potentially useful and should be considered in the post-EuropeanaConnect phase (** CELI: any ideas about how this could be set up ? **).

But if we want to evaluate the usability of resources *before* transforming and integrating them into the query translation module, we need a set of criteria for each of the resource types which will provide some measures or indicators for the quality of the final, integrated module.

2 Evaluation Criteria

In the subsequent sections we will distinguish textual language resources and language processing tools as they differ with respect to their applicability and maintainability:

- Textual resources are lists or databases that may serve for different software tools to perform linguistic processing. The maintenance of these resources can be done with basic text or data base editing tools and does not require specific interfaces or compilers.
- NLP tools, on the other hand, include both processing algorithms and language data. The (API or Web Service description) must be well defined. Specific software or programming skills may be necessary to maintain the incorporated linguistic data.

For an evaluation of language resources we can take into account information about the resource itself, such as the number of items, the complexity or granularity of the information provided or the error rate of the data. But it is difficult to determine the “usefulness” of a language resource for the specific application of Europeana without testing on an appropriate corpus.

Multilingual resources in Europeana will mainly apply query translation. Therefore we will need a somehow representative set of user queries for evaluation purposes.[1]

In the ideal case, we would have a gold standard, i.e. a corpus that is (manually) annotated with the target results. Then some automated procedures could evaluate the results provided by a given resource. As we do not have (nor plan to set up) such a gold standard, we suggest to use general measures wherever possible, and perform a detailed analysis on some random samples.

[1] After mail discussion with Sjoerd Siebinga, he offered to extract all the queries for the past two weeks from the ClickStreamLogs with the interface language and frequency.

2.1 Languages

A simple criteria for a language resources to be integrated in Europeana is the language which must be among the main languages to be covered which are divided into two groups:

Core Languages: English, French, German, Italian, Polish, Spanish

Secondary Languages: Dutch, Hungarian, Portuguese, Swedish

2.2 Quantitative Evaluation Criteria

Quantitative criteria are numeric indicators that can be measured without time-consuming testing or the data.

2.2.1 Size

The size of a resource has different aspects: the overall number of entries relates to coverage of a resource and the average size of a single entry may indicate the linguistic coverage and completeness. On the other hand the size may impact the processing efficiency.

2.2.2 Speed

As long as we consider mainly query translation where only few input words are processed, speed should not be an issue for most of the currently available NLP tools that use local resources, but response time will be an important consideration for web services. It may also be important for applications that involve bigger amounts of text, such as mappings of vocabulary, processing of catalogue meta data, or even the analysis of full documents.

2.2.3 Coverage

The coverage of a language resource measures how much of the input data is covered by a given resource. This does not necessarily reveal if the provided results are correct or complete, but it conveys a minimal information id the resource is usable or not. Coverage should be determined with respect to a given “representative” corpus to measure how well the resource performs on real data of Europeana.

2.3 Qualitative Evaluation Criteria

In the context of EuropeanaConnect we cannot perform an exhaustive qualitative evaluation of all language resources, but define some basic criteria to verify the quality for each of them.

2.3.1 Accuracy (Precision)

Check if the provided results are correct. In addition the coverage test (cf. section 2.2.3) the given result must not only exist, but also be correct and adequate for subsequent processing.

2.3.2 Completeness (Recall)

Check if there are gaps in the results. Whereas a simple coverage test just verifies if a resource provides some result, the completeness test checks if all relevant results are present.

Example: *Evaluation of a German-English translation dictionary or module*

- **Test corpus**: DE: (1) “Schloss”, (2) “Garten”, (3) “Anlage”
- **Results of translation module**: EN: (1) “lock”, (2) “garden”, (3) -
- **Evaluation**:
  - The *coverage* is 66%, as the 3rd item has no translation
  - *Accuracy* is 100% as all (provided) translations are correct
  - *Completeness* is only 40%, considering that the translation “castle” is missing for (1) and two translations are missing for (3).

2.4 Further Aspects for Evaluation

2.4.1 Compatibility or interoperability

The evaluation of an individual resource must take into account aspects of compatibility or interoperability (cf. (Witt, et al. 2009)). For example, if the output of a lemmatizer is used for dictionary lookup, the base forms must correspond.

2.4.2 Availability and support

Both legal aspects (copyrights, licensing issues) and availability in time should be taken into account.

2.4.3 Maintainability

Indicate what is required to maintain (remove, add, correct, extend) entries of the resource.

3 Language Resources

3.1 Stop Word Lists

Purpose:

- Identify non-content words in a text that are irrelevant for further processing such as indexing, semantic analysis or translation.

Content:

- List of (inflected) word forms as they appear in the text

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	Number of items in a list of stop words	Informative (see below)
Coverage	Percentage of words in the corpus that are recognized as stop words	Informative (see below))
Accuracy	Error rate = Percentage of content words in the stop word list	
Completeness	Gaps = Percentage of stop words that were not identified in the corpus	
Compatibility		
Availability	(should not be an issue)	
Maintainability	no issue, because of the relatively small size and simple format of stop word lists.	

Potential issues:

- The size and coverage depend on the language. If a list is bigger it is not necessarily better than another one.
- A stop word list for a specific application may be inadequate for another, as the relevance of a given word may vary.

3.2 Vocabularies

Purpose:

- Identify specific terms or named entities (person names, geographic names, etc.) in a text that may require a special treatment for semantic analysis or translation.

Content:

- List of single or multi-word terms
- *Plus:* associated with morphosyntactic information
- *Plus:* associated with a (semantic) type.

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	Number of entries	Informative (depending on application and domain)
	Ambiguity rate = number of readings for a term or name	This provides some information how difficult it will be to identify a name in

		a query or text.
Coverage	Percentage of identified names or terms in user queries	important
Accuracy	Error rate = Percentage of errors in a list of identified names or terms	Important
Completeness	Gaps = missed items	Important
Compatibility	Entries in vocabularies must be compatible with lemmatizer and dictionaries	
Availability	important because of the changeability and productivity of proper names and terms.	
Maintainability	important because of the changeability and productivity of proper names and terms.	

Potential issues:

- *Ambiguity*: many proper names are ambiguous, either with common words ("Bush") or with proper names of other types ("Paris"). Terms may have different meanings depending on the domain where they are used ("key" in music or mechanics).
- *Variation*: some names (especially when transliterated from another language) are spelled in different ways.
- *Inflection:* the word list generally contains base forms while a term or proper name may occur in an inflected form when it occurs in a query or text (cf. lemmatization).

3.3 Translation Lexicons

Purpose:

- Provide the translation of a word in one language into another language.

Content:

- List of word pairs (or n-tuples)
- Plus: associated with a (syntactic) category
- Plus: associated with a (semantic) type.

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	Number of entries	Important
	Ambiguity rate = number of translations per word	Important
Coverage	Percentage of words that can be translated	Very important
Accuracy	Error rate = Percentage of translation errors	Very important
Completeness	Gaps = number of missing translations	Very important
	Gaps = number of missing readings	Very important
Compatibility	Base forms must be compatible with lemmatizer and vocabularies	
Availability	Crucial	
Maintainability	Important, especially if we want to react to user feedback for translations	

Potential issues:

- *Ambiguity*:
- *Inflection*: (cf. lemmatization)

4 NLP Tools

4.1 Language Guesser

Purpose:

- Identify the language of a query string or piece of text.

Input:

- A sequence of words.

Output:

- an identifier of the most likely language
- **plus**: a list of languages and associated probabilities

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	??	
Speed	??	
Coverage	Number of covered languages	must include the 10 Europeana Connect languages
Accuracy	Percentage of correctly identified user queries	Important
Completeness	??	
Compatibility	Character encodings; language codes[2]	
Availability	important especially if language information is not provided with user queries.	
Maintainability	important if other languages should be added at a later stage of Europeana	

Potential issues:

- accuracy may depend on input length

4.2 Lemmatization

Purpose:

- Identify the base form of an inflected word. The base form could be used for indexing, but indexing is outside the scope of EuropeanaConnect. In the context of EuropeanaConnect, it is mainly necessary to access dictionaries or vocabularies.

Input:

- A single word as it appears in a text (or query)

[2] Considering the rather small sets of Europeana languages it should be easy to provide simple mappings if language codes do not correspond.

Output:

- The corresponding *lemma* or "citation form" as it occurs in a standard dictionary.
- Plus: associated part-of-speech
- Plus: associated morphosyntactic features
- Plus: associated probability in case of ambiguous output

Examples:

- EN: flies -> fly (noun) (pl) ; fly (verb) (3rd sg present)
- FR: chevaux -> cheval (Noun) (masc pl)
- DE: Bremsen -> Bremse (noun) (fem pl) ; bremsen (verb) (inf)

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	Number of base forms Number of inflected forms	Important indicator of the overall language coverage.
Speed	??	
Coverage	Percentage of lemmatized words	Very important Every input word should be mapped to a base form for further processing.
Accuracy	Percentage of correctly lemmatized input words	Very Important
Completeness	Number of missing lemmas	Very important All relevant readings of a word must be covered
Compatibility	Lemmas must be compatible with translation dictionaries or modules	
Availability	important	
Maintainability	depending on the complexity of the language, modifications may consist in simply adding/removing a word and its inflected forms or it requires deeper understanding of the morphological model and implementation.	

Potential issues:

- Standard definition of "citation form"
- Inflection or derivation of proper names (e.g. DE "Marias", "goethesche")
- Spelling conventions (cf. section 4.2.3)

4.2.1 **De-compounding**

Decompounding may be seen as an extension of lemmatization. Languages like German or Hungarian make use of productive compounding to form new words which are usually not present as an entry in a dictionary. For translation each component must be looked up separately if the dictionary contains no entry for the compound.

Purpose:

- Identify the segments of a compound word.

Input:

- An inflected compound word (as it appears in a text or query) or
- A compound base form (as a result of the lemmatizer)

Output:

- The *lemmas* or “citation form” for all segments
- **Plus**: associated probability in case of ambiguous output

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	n/a	n/a
Speed	??	
Coverage	Percentage of lemmatized words	Very important Every input word should be mapped to a base form for further processing.
Accuracy	error rate = number of incorrect segmentations	Very Important (as wrong segmentations will yield wrong translations)
Completeness	Number of missing segmentations	Very important All relevant readings of a word must be covered
Compatibility	Lemmas must be compatible with translation dictionaries or modules	
Availability	important	
Maintainability	different levels of modifications are possible, from as adding/removing segmentations tor changing compounding paradigms.	

Potential issues:

- Some segments appear only in compounds (e.g. “anglo-“, “-teilig”)
- Ambiguous segmentation (“Staub#ecken” vs. “Stau#becken”)
- Oversegmentation (“Mini#ster”, “Komm#uni#kat#ion”, “Verb#raucher”)

4.2.2 Multiword Recognition

Whereas decompounding splits a word into segments, compounding or multi-word recognition groups words (separated by white space) into a single token for further processing such as dictionary lookup.

Purpose:

- Identify multi-words that should be interpreted or translated as single unit.

Input:

- A sequence of inflected words *or*
- A sequence of base form (as a result of the lemmatizer)

Output:

- The *lemma* or “citation form” for the word group
- ***Plus:*** associated part-of-speech
- **Plus:** associated morphosyntactic features

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	Number of multi-words	informative
Speed	??	

Coverage	Percentage of corpus words that are part of multi words	informative
Accuracy	error rate = number of incorrect groupings	??
Completeness	Number of missed multi-words	Important for translation
Compatibility	Tightly linked to translation dictionaries or modules	
Availability	important	
Maintainability	??	

Potential issues:

- Discontinuous multi-words, especially with verbs (e.g. “cut … off”)
- Overlapping multi-words (e.g. “[paper work] flow” vs. “paper [work flow]”)
- Ambiguous with single words (e.g. “(not) at all” vs. “at all (stages)”)

4.2.3 Normalization and spelling correction

Users often type thier queries in a non-standard way, for example by using only lower case characters or omitting. A query specific normalization may be necessary to provide a lemmatizer with “normal” input strings. As user queries are likely to also contain typos, a resource for spelling correction may be necessary.

Purpose:

- Normalize and correct input queries

Input:

- user input strings

Output:

- Normalized input string
- **Plus**: corrected input string

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	??	
Speed	??	
Coverage	Percentage of corpus words that require normalization (or correction)	informative
Accuracy	error rate = number of incorrect groupings	??
Completeness	??	??
Compatibility	normalized (or corrected) forms must correspond to lemmatizer input	
Availability	important	
Maintainability	??	

Potential issues:

- Spelling correction may require user interaction.

4.3 Named Entity Recognition

The recognition of named entity relies on lexical information (e.g. from vocabularies, cf. section 3.2), but it also includes methods and other strategies to classify input words within a given context.

Purpose:

- Identify named entities (person names, geographic names, etc.) in a text or query.

Input:

- A sequence of inflected words or
- A sequence of base form (as a result of the lemmatizer)

Output:

- A "citation form" for named entity
- Plus: associated semantic type or features

Evaluation criteria:

Criteria	*Measure*	*Relevance*
Size	??	??
Speed	??	
Coverage	Percentage of recognized named entities in queries	informative
Accuracy	Percentage of correctly recognized named entities	Very Important
Completeness	Number of missed named entities	Very important
Compatibility	Lemmas must be compatible with translation dictionaries or modules	
Availability	important	
Maintainability	Because of the variability of named entities maintenance is important.	

Potential issues:

- Ambiguity with common words or between different types of names.
- Variations due to transliteration or spelling conventions
- *Accuracy:* Error detection for named entities (e.g. "Miller" identified as country name) is difficult without specific domain knowledge. In case of a "closed" domain, such as country names, it may be possible, in most open domains, however, it is hard to state if a name is correct or not. For example, "Africa" may as well be a person or a product name. But it may not be evident from the user query what he/she was looking for.

5 Evaluation Procedure

*** *TBD* ***

6 Bibliography

Bosca, Alessio, Vivien Petras, and Anne Schiller. "Europeana - Language Resources. necessary Language Resources for a Multilingual Query Translation System in Europeana." *Europeana Connect WP2.2 - Liferay.* 2009. https://version1.europeana.eu/c/document_library/get_file?p_l_id=16989&folderId=26117&name=DLFE-5613.doc.

Luca, Dini, and Vivien Petras. "The Challenge of Multilinguality in Europeana: Web Services as Language Resources." *FlAReNet Forum, Barcelona.* February 2010. http://www.flarenet.eu/sites/default/files/S2_Dini-Petras_Position_Paper.pdf.

Witt, Andreas, Ulrich Heid, Felix Sasaki, and Gilles Sérasset. "Multilingual language resources and interoperability." *Language Resources and Evaluation* (Springer) 43, no. 1 (March 2009): http://www.springerlink.com/content/pp7x18343662g5k2/fulltext.pdf.