Terminology Management Revisited

Nizar Ghoula
University of Geneva
and The Olanto
Foundation

Jaques Guyot
The Olanto Foundation

Gilles Falquet
University of Geneva
and The Olanto
Foundation

ABSTRACT

Large repositories publishing and sharing terminological, ontological and linguistic resources are available to support the development and use of translation. However, despite the availability of language resources within online repositories, some natural languages associations cannot be found (rare languages or non-common combinations, etc.). Consequently, multiple tools for composing linguistic and terminological resources offer the possibility to create missing language associations. These generated resources need to be validated in order to be effectively used. Manually checking these resources is a tedious task and in some cases hardly possible due to the large amount of entities and associations to go through or due to the lack of expertise in both languages. To solve this matter and generate sound and safe content, tools are needed to automatically validate and filter associations that make no sense. Hence, a validation tool is based itself on external resources such as parallel corpora which need to be either collected or created and filtered. To solve these matters we propose a set of tools that generate new terminological resources (**myTerm**) and a filter them using a parallel corpus generated by another tool (**myPREP**). We describe our methodology for terminology management and we describe its implementation within an original framework.

1. Introduction

The translation business has considerably changed over the past decade. Smaller full-time teams must translate larger volumes, the difference being distributed over a network of external translators, which are located worldwide. Besides, deadlines are ever tighter and costs must be reduced. As a result, translation workflows are changing in order to automate every possible step: submitting a document for translation, affecting the translation to a translator, performing the translation, performing the quality control steps, sending back the translation to the customer and feeding the CAT tools with the new document pair and/or related segments. Consequently, a complete suite of CAT tools is needed to support every phase of this new workflow. Within this context, the Olanto foundation[1] proposes and publishes Open Source tools for professionals to face these new challenges.

[1] www.olanto.org

The initial goal of the Olanto Foundation is to build and share a complete suite of professional CAT tools:

- a Concordancer (Bitext-based search engine);
- a Statistical Machine Translation Tool;
- a Terminology Database Management System;
- a Translation Memory Management System.

These tools can be integrated within several Electronic Document Management Systems (EDMS). In particular, a cross-lingual search engine, which may be, associated with other existing search tools (typically Lucene or SharePoint). Despite the existence of a considerable number of open source tools in the CAT field, these tools remain complex and their integration incomplete. Thus, these tools do not meet the complete chain of needs commonly expressed by Translation Services and Language Service Providers. Additionally, they generally don't benefit from a robust distribution and support structure and some of them are not really scalable.

- Based on a previous research work on building a repository of multilingual terminological and ontological resources (Ghoula, Falquet, & Guyot, 2010), we identified the following objectives for such a tool:
- Compatibility of the resources representation models with TBX (basic) (Wright, Melby, Rasmussen, & Warburton, 2010);
- Ability to manage a large number of terminological resources;
- Ability to support a large number of standards and formalisms for resources representations (TBX, UTX, DXDT, GlossML, etc.);
- Availability of XML-based representation models for structured resources that do not correspond to all standards or formalisms (e.g. JIAMCATT[2]).

One of the latest tools in development by Olanto is the **myTerm** terminology manager. In **myTerm**, resources are imported into the terminology manager's repository and attached to a hyper graph where terminological resources from different domains connect languages to each other either directly or by transitivity.

Our main goal is not to generate dictionaries by transitivity but we mainly focus on building a framework and a set of tools for helping translators to validate automatically or semi-automatically their dictionaries using their own corpora or other kinds of corpora.

This approach can also be used to interactively query a large parallel or comparable corpora to compute candidates translation for a given term or multi-word expression. We implemented this idea in the "**How2Say**" interactive tool. This tool allows finding the possible translations of a specific expression from one language to another. We will describe this framework through the different sections of this paper.

2. Context and research issues

As a consequence of the availability of large repositories publishing and sharing terminological, ontological and linguistic resources on the Web, we notice a significant

[2] www.jiamcatt.org

improvement in the quality of automatic and semi-automatic translation systems. Nevertheless, these resources are not yet available for all possible combinations of pairs of languages. For example, to run or enhance a translation process from a language A to a language B, there is a need for a dictionary or a terminology associating both languages. Despite the existence of these types of resources for language A and language B within online repositories, none of them may directly associate the pair of languages (particularly in the case of rare languages or non-common combinations, etc.). One way to address this issue consists in using available language resources to generate the missing ones. Hence, automatically deriving terminologies by transitivity has become a common procedure to produce resources for language services.

For example, if we have a glossary EN→FR and another glossary FR→DE, using composition, we can generate a new glossary EN→DE. It is well known that polysemy within both resources can produce associations between pairs of terms that do not make sense. For example, starting from the associations *time→temps* in EN→FR, *temps→Zeit* and *temps→Wetter*, in FR→DE, the composition produces two term associations: *time->Zeit* and *time→Wetter*[*] for EN→DE. Consequently, this kind of operations on terminological resources is not completely safe in terms of sense. Therefore, the resulting terminological resource has to be filtered to detect and remove meaningless term associations. Manually checking these resources is a tedious task and in some cases hardly possible due to the large amount of entities and connections to go through or due to the lack of expertise in both languages.

In the context of composing ontology alignments, we encountered the issue of inconsistent mappings, which can be solved using reasoning and combination of confidence measures to filter mappings (Ghoula, Nindanga, & Falquet, 2013). Unfortunately, for terminological resources associations, there are no standards or use cases allowing the application of confidence measures. However, it is possible to use a parallel corpus of (aligned) sentences between both languages to assign a confidence measure to associations between pairs of words. This measure is based on the co-occurrence of both terms in the sentences of the corpus.

In this paper, we describe our approach and present the architecture of **myTerm** repository and define operations for producing, managing and filtering terminological resources for validating languages associations. We explain in detail the computation of correlation measures that filter term associations based on their co-occurrence.

3. State of the art

New technologies for the Web and knowledge sharing made language resources more accessible. Thus, the volume of data to process for training or evaluating machine translation tools is more important. The scalability issues to process more data and more resources are often revisited and algorithms are becoming more efficient thanks to the rising hardware power and the efficiency of new software and services architectures. Consequently, we can aim for interactive systems that process large corpora (public and private) and offer a real time and flexible interaction with translators. This also opens the door for creating more language services and better interaction between different CAT tools. Multiple tools for creating language resources such as dictionaries by transitivity have been proposed in the literature (Paik, Shirai, & Nakaiwa, 2004), (Zhang, Ma, & Isahara, 2007).

Our concern is about the approaches that use parallel or comparable corpora to validate the result of transitivity. An approach for automatically generating dictionaries between languages was proposed by (Nerima & Wehrli, 2008) in order to reduce the number of linguistic resources used as an input for a multilingual translation system. Since such a system requires a lexical database for each pair of language combination, then for a number of **n** languages, there is a need for **n*(n-1)/2** dictionaries. Even if these dictionaries were available, which is not always the case, the large number of bilingual dictionaries might affect the performances of such a system. Thus, the authors propose to derive a bilingual dictionary by transitivity using existing ones and to check the generated translations in a parallel corpus. The quality of the result relies on multiple parameters such as the quality of the input, which have to be manually validated, the attribution of a preference and the usage of tagging. Consequently, the approach is language dependent and there are needs to have multiple language models for an effective tagging.

Another approach proposed by (Tao & Zhai, 2005) close to our methodology is based on using correlation measures between words within comparable corpora to build a cross-lingual text mining framework that can exploit these bilingual text corpora to discover mappings between words and documents in different languages. This approach is based on the hypothesis that the words that tend to co-occur more frequently in comparable corpora are either translations of each other or related to the same topic. Thus, the authors use comparable corpora to extract associations of words in multiple languages. The authors use the Pearson's correlation coefficient to compute associations between words, which are used to create a similarity score between documents. The correlation measure is combined with information retrieval techniques in order to match documents between languages but does not go further into matching sentences inside the associated documents.

We propose an original approach that is language and corpus independent and a framework for indexing parallel corpora and calculating correlation measures between n-grams at the level of sentences. We also propose a real-time application that retrieves n-grams and their translations from voluminous corpora.

4. Approach

Our approach relies on the validation of new glossaries, generated by transitivity, using parallel and comparable corpora between the two languages. For instance, going back to the example for the introduction we can find in the EN→DE corpus a number of co-occurrences of "*time*" and *"Zeit"* in a sentence and its translation that confirm the "*time→Zeit*" association (*« Members shall furnish statistics and information within a reasonable* ***time...*** *»→ « Die Mitglieder legen Statistiken und Angaben innerhalb einer angemessenen* ***Zeit*** *...* »), whereas almost no co-occurrence confirms the "*time→Wetter*" association.

For implementing and preparing a corpus independent framework we assembeled different components of the Olanto's suite as described in figure 1. The reference parallel corpora are produced by **myPREP**, Olanto's text aligner tool. This tool automatically aligns pairs of documents from a multilingual corpus at the sentence level and generates a translation memory in TMX format. Each set of TMX is then indexed using the **myCAT** indexer (Guyot, Falquet, &

Benzineb, 2006). The indexer generates two types of vectors of values for each term within the corpora:

- ***idx$_j$*** =(o_1, o_2, ..., o_n) is the index of the n-gram g_j for a given corpora *C* containing n documents where o_k defines the number of occurrences of g_j within the document d_k.
- ***pos$_{jk}$***=(p_1, p_2, ...p_{mk}) is the vector of positions of the n-gram g_j within a given document $d_k$where $p_m$defines the position of the m^{th} occurrence of g_j in the document d_k.

We developed a module that calculates the correlation between two n-grams based on the generated index and used it to build the **How2Say** web application on top of the indexed corpora. This application let the user enter an n-gram, then it computes, in real-time, the best translation in the target language (the maximal n-grams with the hights correlations), and displays these translations together with example sentences.

In order to automatically validate the generated translation (dictionary entries) we created a module that generates dictionaries by transitivity. Finally we added a component that takes as input the generated dictionary and generates as output for each entry a triple (n-gram, n-gram, correlation).

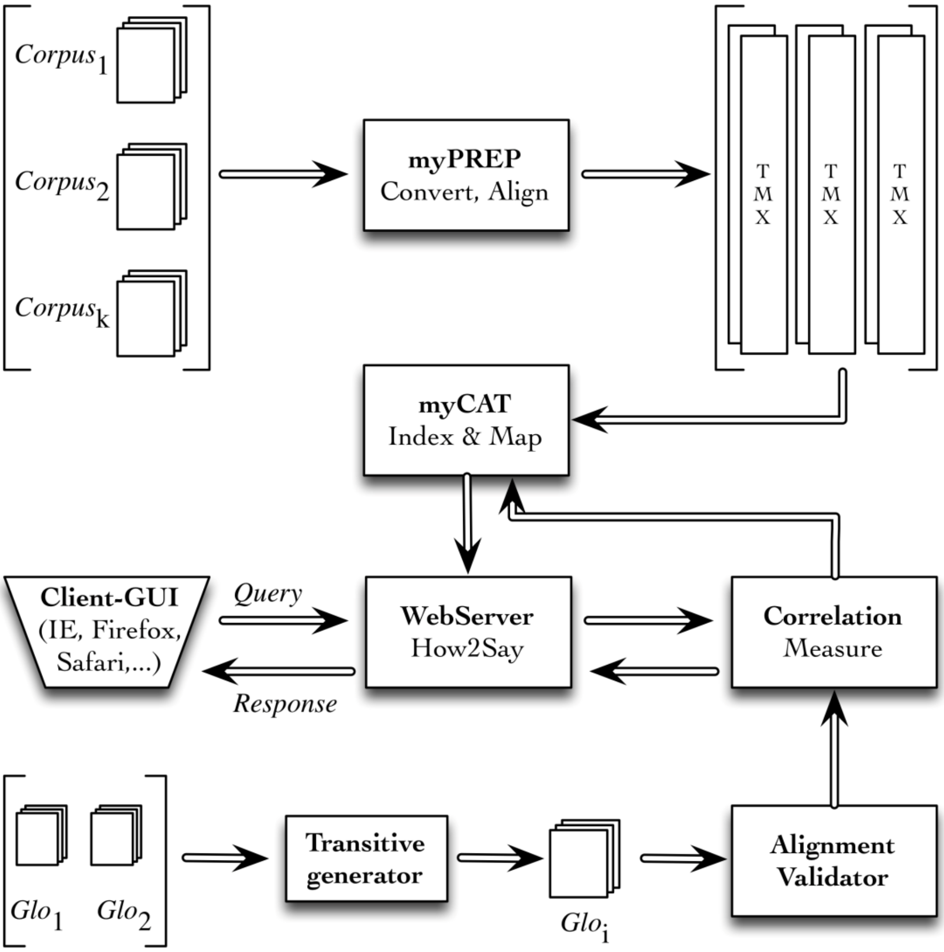


Figure 1: Architecture of the bilingual dictionary automatic validation framework

Each correlation is measured based on a specific corpus. The whole approach is supported by the framework of processing parallel and comparable corpora to compute correlation masures for pairs of n-grams in a given source and target language.

5. Correlation

A correlation-based technique computes a correlation measure between two terms or expressions based on their co-occurrences in aligned sentences. Based on our indexer, the calculation of correlation measures is quite fast. If g1 and g2 are two n-grams (d-grams) in the source and target languages respectively, the similarity between g1 and g2 is obtained as the correlation between the occurrence vectors x and y, where x_i (resp. y_i) = 1 if g1 (resp. g2) occurs in sentence no. i of the source (resp. target) language, and 0 otherwise.

The correlation between x and y is defined as:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Since x and y are binary vectors, r_{xy} can be reduced to

$$r_{xy} = \frac{nn_{12} - n_1 n_2}{\sqrt{nn_1 - {n_1}^2}\sqrt{nn_2 - {n_2}^2}}$$

where:

- *n* is the number of aligned sentences;
- n_1 is number of sentences in the source language containing g1;
- n_2 is number of sentences in the target language containing g2;
- n_{12} is the number of aligned sentences containing g1 in the source language and g2 in the target language (co-occurrences).

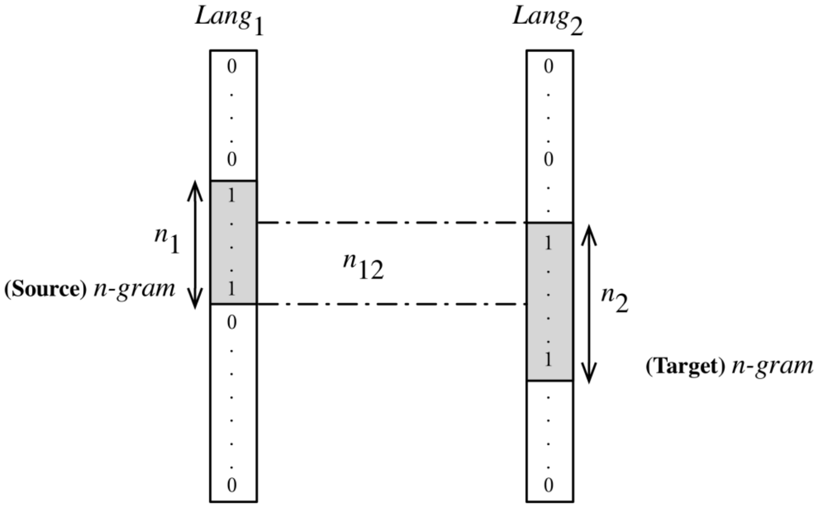


Figure 2: Correlation measures between n-grams within the indexed corpora

6. Testing our approach for validating bi-lingual dictionaries

In order to test the usefulness of the corpora in the process of automatic validation of terms associations within a bilingual dictionary, we conducted two types of experiments.

The first experiment is intended to use a valid bilingual dictionary to test the quality of the used corpora and their relevance to the dictionary based on calculating correlation measures for valid terms associations.

The second experiment is intended to test the usage of correlation measures for the validation of the generated dictionary.

Corpus	Size (# of sentences)	Number of languages
Wikipedia: comparable built using **myPREP**	1,000,000	3
MultiUN parallel corpora[3]	69,300,000	7
DGT2014 parallel corpora[4]	84,561,191	23
EuroBook parallel corpora[5]	173,200,000	48

Table 1: Parallel and comparable corpora, parsed, indexed and mapped for How2Say

6.1. Evaluating the correlation method

The first step of our approach is to use an existing valid dictionary to test the coverage of the corpora used in the experimentation. We selected the Dictionary "Wiktionary 2008" containing 15'000 entries between English and French. The main characteristic of this dictionary is its general aspect covering multiple domains.

This experiment is used in order to define the terminological signature of corpora and the utility of the approach in general. There are two interpretations of the weak correlation values:

- The corpora's terminology does not support the domain of the dictionary;
- The dictionary contains false associations between terminological entities due to polysemy;

The horizontal axis represents the term number and the vertical axis shows the correlation values. The terms are sorted according to their correlation values. For instance the ONU curve (red) shows that approx. 6500 terms in this corpus have a translation with a correlation higher than 0.1.

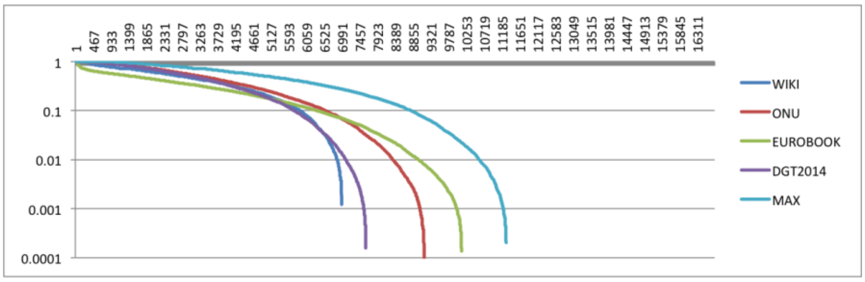


Figure 3: Coverage of the used corpora for "Wikitionary" for English and French

[3] http://www.lrec-conf.org/proceedings/lrec2012/pdf/641_Paper.pdf
[4] http://optima.jrc.it/Resources/DGT-TM-2014/DGT-TM_Statistics.pdf
[5] http://opus.lingfil.uu.se/EUbookshop.php

The coverage of the used corpora varies from 40% to 65% taken separately. In order to maximize the coverage of the corpora, we used a maximum aggregation of the correlation measures. Thus, the resulting coverage of all the corpora combined for the used dictionary is 75%.

In general, based on this experiment we realized that:

- the correlation is different depending on the corpora;
- a given corpus does not always cover the dictionary;
- a corpus has a specific terminological signature;
- the maximum aggregation of correlation measures allows to enlarge the coverage of a corpus;
- using the correlation-based method we can also determine the correlation between corpora for a given dictionary.

6.2. Validating a transitive bi-lingual dictionary using out method

For the second experiment we used two corpora, EuroBook and MultiUN. These corpora contain the biggest number of entries for German (among the four processed corpora). In order to validate the generated dictionary using transitivity, there are two possibilities for interpreting a weak correlation; the first interpretation is that the entry is not covered by the corpora or that the entry is invalid due to polysemy.

Two types of experimentations have been driven for the validation of the generated dictionary from French to German through English:

- Generate the transitive dictionary FR -> DE and then validate it using the maximum of correlation from both corpora (EuroBook[FR-DE], DGT2014 [FR-DE]). The result of this operation gave a dictionary of 27'183 entries where 11'662 have a not null correlation based on the maximum from both corpora. This is a result of a maximum aggregation of the correlation values;
- A more radical approach is to compose only the dictionary entries from FR to EN, that are covered by the corpora DGT2014[FR-EN], with the dictionary entries from EN to DE that are covered by the corpora DGT2014[EN-DE] and then validate the resulting dictionary FR to DE using the corpora DGT2014[FR-DE]. The resulting dictionary contains only 3'800 terms associations that are considered as valid based on their correlation measures.

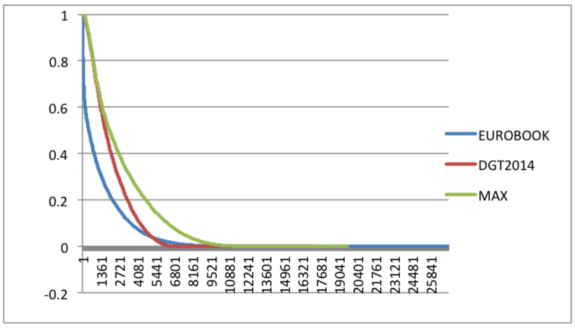
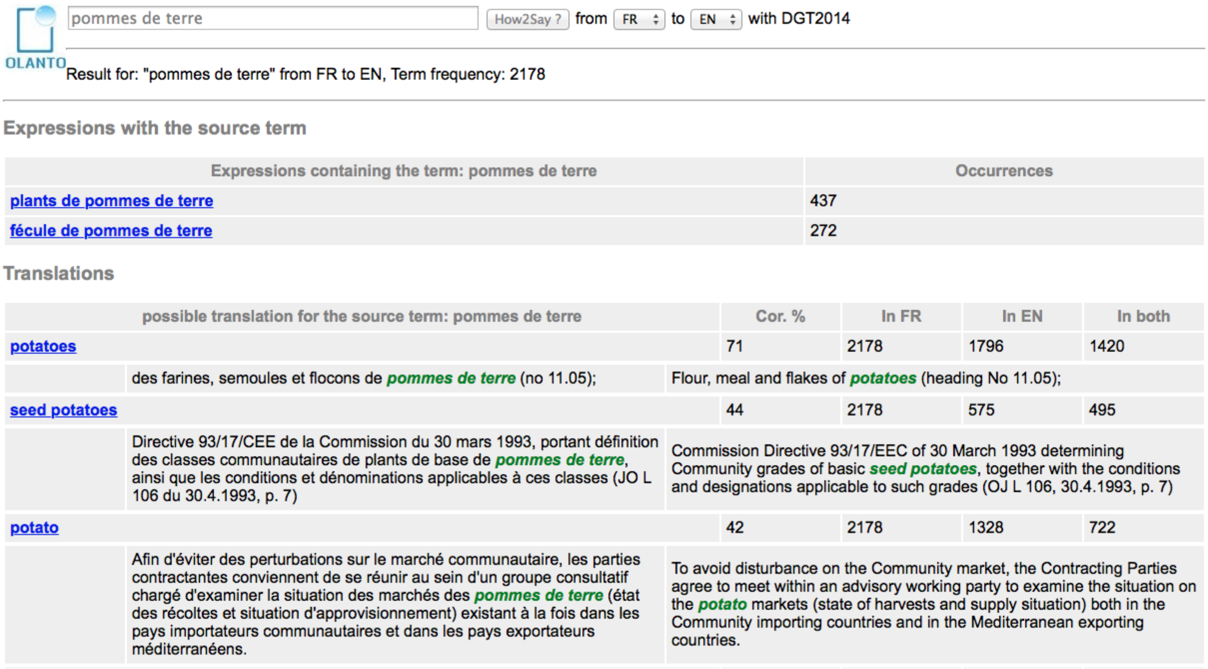


Figure 4: Validation of the FR-DE dictionary using two corpora

The terminological signature of the used corpora is not as general as the used dictionary for transitivity. For each couple of terminological entries there are different possible translations depending on the context. An automatic translation system imposes the translations using a translation memory. In our context, we propose a diversified approach for proposing translations based on correlation measures.

The approach that we propose is flexible and simple; it offers an original and efficient framework for validating transitivity translation (**How2Say**). While testing this approach we realized that the corpora's coverage is very important and the results depends highly on it. We created a system to explore expressions and n-grams. This system explores the parallel corpora and classifies the list of n-grams. For each query retrieving an expression in a source language, we classify the corresponding n-grams in a target language using the correlation measure.



Result for: "pommes de terre" from FR to EN, Term frequency: 2178

Expressions with the source term

Expressions containing the term: pommes de terre	Occurrences
plants de pommes de terre	437
fécule de pommes de terre	272

Translations

possible translation for the source term: pommes de terre			Cor. %	In FR	In EN	In both
potatoes			71	2178	1796	1420
	des farines, semoules et flocons de *pommes de terre* (no 11.05);	Flour, meal and flakes of *potatoes* (heading No 11.05);				
seed potatoes			44	2178	575	495
	Directive 93/17/CEE de la Commission du 30 mars 1993, portant définition des classes communautaires de plants de base de *pommes de terre*, ainsi que les conditions et dénominations applicables à ces classes (JO L 106 du 30.4.1993, p. 7)	Commission Directive 93/17/EEC of 30 March 1993 determining Community grades of basic *seed potatoes*, together with the conditions and designations applicable to such grades (OJ L 106, 30.4.1993, p. 7)				
potato			42	2178	1328	722
	Afin d'éviter des perturbations sur le marché communautaire, les parties contractantes conviennent de se réunir au sein d'un groupe consultatif chargé d'examiner la situation des marchés des *pommes de terre* (état des récoltes et situation d'approvisionnement) existant à la fois dans les pays importateurs communautaires et dans les pays exportateurs méditerranéens.	To avoid disturbance on the Community market, the Contracting Parties agree to meet within an advisory working party to examine the situation on the *potato* markets (state of harvests and supply situation) both in the Community importing countries and in the Mediterranean exporting countries.				

Figure 5: How2Say Interface

7. Conclusion

We propose a framework and a browser (How2Say) offering the possibility of validating transitive language associations based on correlations measures calculated using parallel or comparable corpora. The proposed framework is generic and language independent offering multiple possibilities such as:

- The usage of a sophisticated a query language for finding expressions ("AND", "OR");
- The openness for multiple corpora and dynamic support and processing of new corpora;
- multiple options for translating expressions based on their co-occurrence within the corpora.

We evaluated our approach on multiple voluminous corpora. The feasibility of this approach has been proven by experiments and evaluation. We developed an online demo offering a real-time interrogation of large parallel and comparable corpora supported by correlation measures. The values of correlation measures allow exploring parallel corpora for mining n-gram associations. The correlation measure within this framework is used for automatic validation of associations between expressions from pairs of languages. This approach is a first step for an automatic system of validating glossaries and dictionaries that are created using transitive tools. Parallel corpora use specific terminologies and do not cover all domains.

References

GHOULA, N., FALQUET, G., & GUYOT, J. (2010). Tok: A meta-model and ontology for heterogeneous terminological, linguistic and ontological knowledge resources. *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM* (pp. 297-301). IEEE.

GHOULA, N., NINDANGA, H., & FALQUET, G. (2013). A meta-model and ontology for managing heterogenous alignment resources. *International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM. 3*, pp. 167-170. IEEE.

GUYOT, J., FALQUET, G., & BENZINEB, K. (2006). Construire un moteur d'indexation. *Ingénierie des Systèmes d'Information , 11*, 99-131.

NERIMA, L., & WEHRLI, E. (2008). Generating Bilingual Dictionaries by Transitivity. In P. o. Evaluation (Ed.). Marrakech: European Language Resources Association.

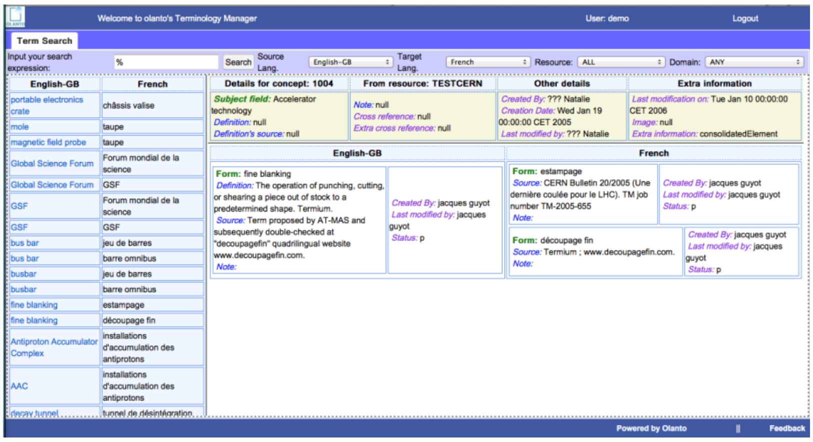
PAIK, K., SHIRAI, S., & NAKAIWA, H. (2004). Automatic construction of a transfer dictionary considering directionality. *Proceedings of the Workshop on Multilingual Linguistic Ressources* (pp. 31-38). Association for Computational Linguistics.

TAO, T., & ZHAI, C. (2005). Mining comparable bilingual text corpora for cross-language information integration. In ACM (Ed.), *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, (pp. 691-696).

WRIGHT, S., MELBY, A., RASMUSSEN, N., & WARBURTON, K. (2010). TBX Glossary: A Crosswalk between Termbase and Lexbase Formats. *Association for Machine Translation in the Americas (AMTA) Conference.*

ZHANG, Y., MA, Q., & ISAHARA, H. (2007). Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. *MT Summit XI* , 551-557.

Annex



Input your search expression: %

Source Lang. English-GB

Target Lang. French

Resource: ALL

Domain: ANY

English-GB	French
portable electronics crate	châssis valise
mole	taupe
magnetic field probe	taupe
Global Science Forum	Forum mondial de la science
Global Science Forum	GSF
GSF	Forum mondial de la science
GSF	GSF
bus bar	jeu de barres
bus bar	barre omnibus
busbar	jeu de barres
busbar	barre omnibus
fine blanking	estampage
fine blanking	découpage fin
Antiproton Accumulator Complex	installations d'accumulation des antiprotons
AAC	installations d'accumulation des antiprotons
decay tunnel	tunnel de désintégration

Details for concept: 1004	From resource: TESTCERN	Other details	Extra information
Subject field: Accelerator technology Definition: null Definition's source: null	Note: null Cross reference: null Extra cross reference: null	Created By: ??? Natalie Creation Date: Wed Jan 19 00:00:00 CET 2005 Last modified by: ??? Natalie	Last modification on: Tue Jan 10 00:00:00 CET 2006 Image: null Extra information: consolidatedElement

English-GB		French	
Form: fine blanking Definition: The operation of punching, cutting, or shearing a piece out of stock to a predetermined shape. Termium. Source: Term proposed by AT-MAS and subsequently double-checked at "decoupagefin" quadrilingual website www.decoupagefin.com. Note:	Created By: jacques guyot Last modified by: jacques guyot Status: p	Form: estampage Source: CERN Bulletin 20/2005 (Une dernière coulée pour le LHC). TM job number TM-2005-655 Note:	Created By: jacques guyot Last modified by: jacques guyot Status: p
		Form: découpage fin Source: Termium ; www.decoupagefin.com. Note:	Created By: jacques guyot Last modified by: jacques guyot Status: p

Figure 6: myTerm browsing interface