Nebrija Universidad

Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas

Source

Building a searchable online french greek parallel corpus for the University of Cyprus

Fryni Kakoyianni-Doa y Eleni Tziafa

University of Cyprus

frynidoa@ucy.ac.cy, tziafa.eleni@ucy.ac.cy

Kakoyianni-Doa, F. y Tziafa, E. (2013). Source: Building a searchable online french greek parallel corpus for the University of Cyprus. *Revista Nebrija de Lingüística Aplicada* (2013) 13.

RESUMEN

Palabras clave:

ABSTRACT

This paper gives an overview of the searchable online French-Greek parallel corpus (SOURCe), which aims to serve the needs of students of French as a foreign language and also to facilitate future linguistic research. This project is led by Fryni Kakoyianni-Doa and is fully funded by the University of Cyprus. We included different registers (Biber 1993), so that students may compare the results and the use of each word or phrase in different contexts (e.g., literature, scientific, official, technical and journalistic language). In general, the corpus comprises of a fiction and a non-fiction part. In this article, we describe the design principles and the properties of this French and Greek linguistically annotated corpus, we also report on tools used for the collection, sentence and word alignment, dictionary extraction and POS tagging for the parallel corpus. Finally, we outline its future perspectives and applications, including ongoing work on adverbs and their properties.

Keywords: parallel corpus, teaching French, data driven learning (DDL)

1. INTRODUCTION

In the present article we will focus on the construction of a parallel corpus, the composition, annotation, encoding and availability of which are meant to serve the needs of students of French as a foreign language and also to facilitate future linguistic research. This project is led by Fryni Kakoyianni-Doa and is fully funded by the University of Cyprus.

Parallel corpora consist of original and translated texts.A parallel corpus, as defined bySinclair & Ball (1996) in EAGLES typology, is “a collection of texts, each of which is translated into one or more other languages than the original”. Most parallel corpora are aligned in the sentence level.

Translation can be an extremely effective way to build up a good working vocabulary. But in the past, uncommunicative texts that were difficult to translate did little for motivation. Nowadays, translation is rediscovered as a tool for language learning and assessment mainly due to the use of electronic parallel corpora. Previously, electronic language corpora have not been widely taken up in traditional teaching contexts, partly because it is thought to require substantial resources and considerable investment on the part of the teachers, as well as being appropriate only for the most advanced learners following specific training (Landoure & Boulton, 2010). Nevertheless, parallel corpora, are proving increasingly influential in language teaching, as they promote autonomy and language awareness.

Based on the firm belief that the days of pencil-and-paper teaching, and not only research (Váradi et al. 2008), are numbered, we decided to proceed to a "data driven learning" approach (Johns 1991) and the building of an electronic resource for a less resourced language, such as the Greek language. New tools are critical to creating a dynamic and engaging learning environment. Furthermore, according to Gabrielatos (2005), "changes in knowledge, skills and attitudes […] are needed for learners and teachers to take advantage of the opportunities offered by the availability of corpus resources". On the other hand, as we observed in class the fact that students tend to rely more and more on machine translating, this led us to take a different approach: every entry in the dictionary and every sentence displayed have been translated by humans, in order to help students finding or making the best possible translation, or understanding how the words are used in certain context.

The goal of the project is to provide high-quality, online content on general education subjects to students and training in a cost-effective manner, as an Open Educational Resource, aiming to be accessible, adaptable, free, high-quality, empowering and relevant. All translated texts are displayed in full sentences. Queries are based not only on words, but also on linguistic annotation.

Another important issue is choosing appropriate topics, texts and material to be included in the corpus. Most parallel corpora are not register-diversified; nevertheless, our objective is to include at least five different registers (Biber 1993), so that students may compare the results and the use of each word or phrase in different contexts (e.g., literature, scientific, official, technical and journalistic language). In general, the corpus comprises of a fiction and a non-fiction part.

Our aim is to describe the design principles and the properties of the French and Greek linguistically annotated corpus that we have created. Then, we will report on tools used for the collection, sentence and word alignment, dictionary extraction and POS tagging for the parallel corpus. Finally, we will outline its future perspectives and applications, discussing how it can be incorporated into effective learning resources.

2. THE SOURCE PARALLEL CORPUS

2.1 DESIGN PRINCIPLES

The Source Corpus was designed as a parallel corpus, since parallel corpora can be extremely useful in both translation and teaching, as "they can be used for comparisons at different levels of language, from lexis to syntax to discourse […]. Nevertheless,pedagogical applications of parallel and comparable corpora are not confined to translator training or translation teaching.By facilitating the mapping of correspondences between languages, parallel corpora can not only shed light on the commonalities and differences between language pairs, but also improve the accuracy of descriptions of individual languages" (Kenning 1999).

Research has shown that the use of corpora in classroom can have remarkable results as regards foreign language learning (Hadley 2002; Landure & Boulton 2010). It is also an easy resource which does not require a special training, since we all are corpus users, as we are internet users (McCarthy 2008).

Being inspired by great open source projects such as Opus project & Project Gutenberg, we designed this corpus to be open and freely accessible to all interested teachers, learners and researchers. Moreover,for its construction we worked mostly with open source or freely available tools.Following the principles of the Common Language Resources and Technology Infrastructure (CLARIN) (Váradi et al. 2008) we set as design principles of this corpus the following:

- Stability: the resources and services will be offered with a high availability
- Persistency: the resources and services are planned to be accessible for many years so that researchers can rely on them
- Accessibility: the resources and services are accessible via the web
- Extendability: the infrastructure will be open so that new texts, resources and services can be added easily

?2.2 OBJECTIVES OF THE SOURCE CORPUS

The Source Corpus was compiled with teaching primarily in mind, and also in order to extract translation units from authentic data. As it is intended to be used as teaching material in the classroom, we manually selected the content, in order to be appropriate for learners. We made it register-diversified in order to be as representative as possible, and because "linguistic analyses carried out on a multi-domain corpus can uncover differences in the lexicon and morphology, in names and named entity structures, and in lexical semantics, syntactic and discourse structure" (Bentivogli et al. 2003).

In agreement with Krishnamurthy (2001: 83), two chief principles justify corpus integration in our language program: "A corpus can give us accurate statistics" and "a corpus can provide us with a vast number of real examples". Thus, our objective is to provide quantitative information that can reveal what is frequently and typically used in language.

It is also among our objectives that this corpus will form a basis for the construction of awareness-raising activities where the students will be guided towards identifying patterns in language as well as similarities and differences between their native and second language, as the interaction is inevitably occurring in the process of language acquisition.

The corpus is meant to be used as a tool, that students may themselves build their knowledge base, through access to authentic language use, only now accompanied by convenient translation into a known language. For example, while it is possible to infer from a small selection of occurrences of French*toujours pas* and*pas toujours* in context, together with their translations, that the former means*still not/not yet* whereas the latter means*not always*, to work this out without a translation represents a much greater challenge (Kenning 2010).

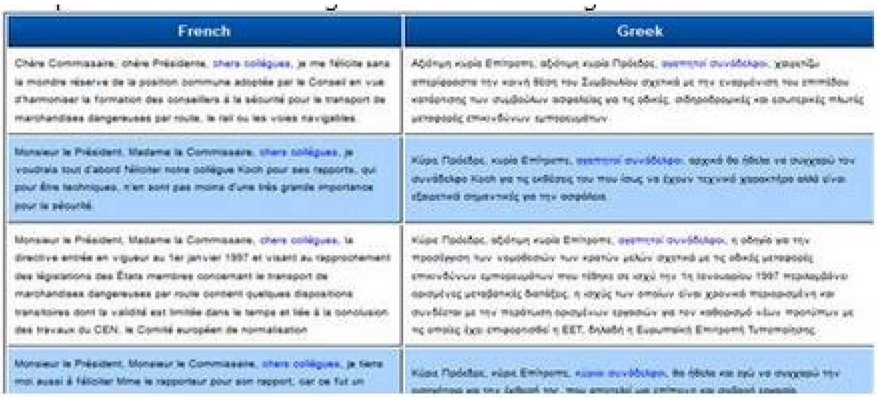
Finally, as this "data-driven learning" (DDL) approach has not been widely taken up in traditional teaching contexts, partly because it is thought to require substantial resources and considerable investment on the part of the teachers, as well as being appropriate only for the most advanced learners following specific training (Landure & Boulton, 2010), our vision is that both the resources for processing language and the data to be processed will be made available in usable formats for teachers and learners.

2.3 DESCRIPTION OF THE SOURCE CORPUS

The Source corpus is an electronic, online tool, to be used as a pedagogical tool. The translation outputs included are of high quality, based on human understanding of textual relations, which is not the case for machine translation (yet), despite the fact that students tend to rely more and more on it.

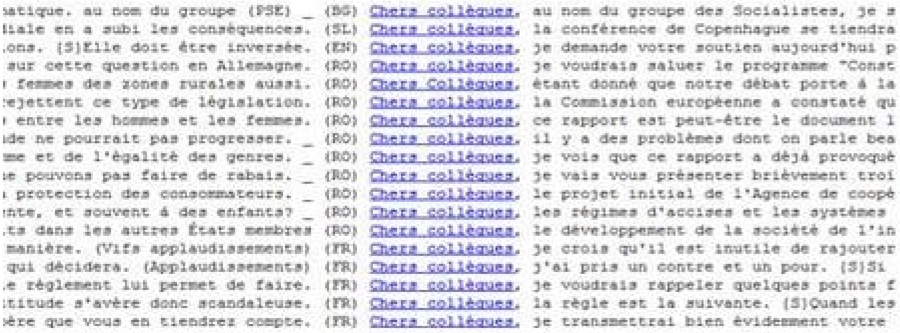
The user has the ability to choose between seeing the translated texts displayed in full sentences or in their context, as key words in context (KWIC). This was considered to be necessary, since, according to Danielsson & Mahlberg (2010) "while a text is normally read horizontally, it is one of the features of a concordance that it needs to be read vertically (Tognini-Bonelli 2001: 3). Thus, the fact that concordance lines are generally not displayed as full sentences is not purely a technical matter. Full sentences might encourage students to focus too much on each line and overlook the patterns that only become evident in relation to the other concordance lines. Sentence segments focus the view on the more general patterns and the possibilities of sorting on the surrounding words".

?



French	Greek
Chère Commissaire, chère Présidente, chers collègues, je me félicite sans la moindre réserve de la position commune adoptée par le Conseil en vue d'harmoniser la formation des conseillers à la sécurité pour le transport de marchandises dangereuses par route, le rail ou les voies navigables.	Αξιότιμη κυρία Επίτροπε, αξιότιμη κυρία Πρόεδρε, αγαπητοί συνάδελφοι, χαιρετίζω ανεπιφύλακτα την κοινή θέση του Συμβουλίου σχετικά με την εναρμόνιση του επιπέδου κατάρτισης των συμβούλων ασφαλείας για τις οδικές, σιδηροδρομικές και εσωτερικές πλωτές μεταφορές επικινδύνων εμπορευμάτων
Monsieur le Président, Madame la Commissaire, chers collègues, je voudrais tout d'abord féliciter notre collègue Koch pour ses rapports, qui pour être techniques, n'en sont pas moins d'une très grande importance pour la sécurité.	Κύριε Πρόεδρε, κυρία Επίτροπε, αγαπητοί συνάδελφοι, αρχικά θα ήθελα να συγχαρώ τον συνάδελφο Koch για τις εκθέσεις του που ίσως να έχουν τεχνικό χαρακτήρα αλλά είναι εξαιρετικά σημαντικές για την ασφάλεια.
Monsieur le Président, Madame la Commissaire, chers collègues, la directive entrée en vigueur au 1er janvier 1997 et visant au rapprochement des législations des États membres concernant le transport de marchandises dangereuses par route contient quelques dispositions transitoires dont la validité est limitée dans le temps et liée à la conclusion des travaux du CEN, le Comité européen de normalisation.	Κύριε Πρόεδρε, αξιότιμη κυρία Επίτροπε, αγαπητοί συνάδελφοι, η οδηγία για την προσέγγιση των νομοθεσιών των κρατών μελών σχετικά με τις οδικές μεταφορές επικινδύνων εμπορευμάτων που τέθηκε σε ισχύ την 1η Ιανουαρίου 1997 περιλαμβάνει ορισμένες μεταβατικές διατάξεις, η ισχύς των οποίων είναι χρονικά περιορισμένη και συνδέεται με την περάτωση ορισμένων εργασιών για τον καθορισμό κάποιων προτύπων με τις οποίες έχει επιφορτισθεί η ΕΕΤ, δηλαδή η Ευρωπαϊκή Επιτροπή Τυποποίησης.
Monsieur le Président, Monsieur le Commissaire, chers collègues, je tiens moi aussi à féliciter Mme le rapporteur pour son rapport, car ce fut un	Κύριε Πρόεδρε, κύριε Επίτροπε, κύριοι συνάδελφοι, θα ήθελα και εγώ να συγχαρώ την εισηγήτρια για την έκθεσή της, που αποτελεί μια επιτυχή και σοβαρή εργασία

Figure 1. Full sentence display



```
atique. au nom du groupe (PSE) _ (BG) Chers collègues, au nom du groupe des Socialistes, je s
itale en a subi les conséquences. (SL) Chers collègues, la conférence de Copenhague se tiendra
ons. [S]Elle doit être inversée. (EN) Chers collègues, je demande votre soutien aujourd'hui p
sur cette question en Allemagne. (RO) Chers collègues, je voudrais saluer le programme "Const
 femmes des zones rurales aussi. (RO) Chers Collègues, étant donné que notre débat porte à la
ejettent ce type de législation. (RO) Chers collègues, la Commission européenne a constaté qu
 entre les hommes et les femmes. (RO) Chers collègues, ce rapport est peut-être le document l
ide ne pourrait pas progresser. _ (RO) Chers collègues, il y a des problèmes dont on parle bea
me et de l'égalité des genres. _ (RO) Chers collègues, je vois que ce rapport a déjà provoqué
e pouvons pas faire de rabais. _ (RO) Chers collègues, je vais vous présenter brièvement troi
 protection des consommateurs. _ (RO) Chers collègues, le projet initial de l'Agence de coopé
nte, et souvent à des enfants? _ (RO) Chers collègues, les régimes d'accises et les systèmes
ts dans les autres États membres (RO) Chers collègues, le développement de la société de l'in
manière. (Vifs applaudissements) (FR) Chers collègues, je crois qu'il est inutile de rajouter
qui décidera. (Applaudissements) (FR) Chers collègues, j'ai pris un contre et un pour. [S]Si
e règlement lui permet de faire. (FR) Chers collègues, je voudrais rappeler quelques points f
titude s'avère donc scandaleuse. (FR) Chers collègues, la règle est la suivante. [S]Quand les
ère que vous en tiendrez compte. (FR) Chers collègues, je transmettrai bien évidemment votre
```

Figure 2. Keywords in Context display

Queries are based not only on words, but also on linguistic annotation. The texts are part-of-speech tagged, so that the learner can perform queries based on part of speech and not only on words or phrases. For a start, a user can make a query for all adverbs by inserting the acronym ADV in the search box.

The first version of Source Corpus will comprise of 1,000,000 words, but it will continue growing. It is a diachronic corpus, as the time period covers at least six centuries, from the 15th to 21st century. The texts are copyright free texts, parallel, for the moment in French and Greek language.

The texts under study are instances of different domain-specific registers. A variety of language according to use is acknowledged in Systemic Functional Linguistics as being a register. Usually, parallel corpora are not register-diversified; nevertheless, our objective is to include at least five different registers (Biber, 1993), so that students may compare the results and the use of each word or phrase in different contexts (e.g., literature, scientific, official, technical and journalistic language). Therefore, we will include parts of commonly used parallel corpora like EUROPARL (Koehn, 2005), the JRC Acquis corpus (Steinberger et al., 2006) and other corpora from the Opus open parallel corpus (Tiedemann, 2012). The first release (scheduled before the end of 2012) will include also literary works available by Project Gutenberg. In general, the corpus comprises of a fiction and a non-fiction part.

?

2.4 COLLECTION OF TEXTS

In accordance to design principles and committed to the notion of open source tools and resources, we started searching for copyright free texts, in order to avoid or reduce risks incurred in possible violations of intellectual property rights (IPR) or basic ethical rules.

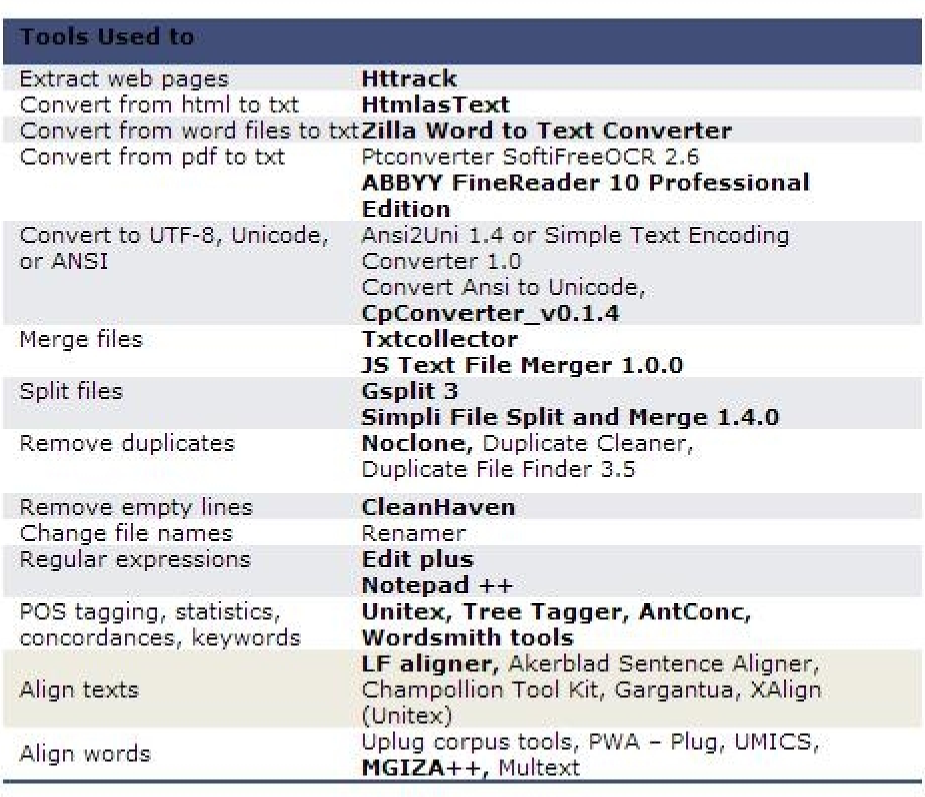
For this reason, we turned to already existing parallel corpora and available parallel texts in French and Greek language (e.g. United Nations Corpora, DGT Multilingual Translation Memory of the Acquis Communautaire), French and Greek aligned texts from Opus project (e.g. ECB, EUconst, EUROPARL, OpenSubs, which contain French and Greek texts), French and Greek literature from Project Gutenberg. We also added new texts, such as technical texts, manuals (e.g. Linux, Php).

In general, we searched and collected texts from online (Gallica-http://gallica.bnf.fr, Wikisource-http://fr.wikisource.org) and physical libraries, after consulting the Index Translationum and bibliographical lists of translated works in both languages. We also scanned available and copyright free books. Finally, we asked for donations by publishing houses.

2.5 TOOLS USED FOR THE CONSTRUCTION OF THE SOURCE CORPUS

A comprehensive review of all the tools used in order to build corpora is beyond the scope of this paper. Nevertheless, in order to obtain control over the texts and also to gain a deeper insight in the process involved, we used individual tools, freeware or open source, along with a few commercial products; for keywords and statistical measures we used Wordsmith tools.

Each tool was selected among many others, depending on its function and the ability to support languages with other than latin alphabets, such as Greek. Moreover, tools for very large and batch files were preferred. Therefore, texts were extracted from the Web, converted to plain text, cleaned from duplicates, merged in larger documents, aligned in text level and in word level (to obtain bilingual dictionaries, but with limited entries). The procedure included also manual text treatment.



Tools Used to	
Extract web pages	**Httrack**
Convert from html to txt	**HtmlasText**
Convert from word files to txt	**Zilla Word to Text Converter**
Convert from pdf to txt	Ptconverter SoftiFreeOCR 2.6 **ABBYY FineReader 10 Professional Edition**
Convert to UTF-8, Unicode, or ANSI	Ansi2Uni 1.4 or Simple Text Encoding Converter 1.0 Convert Ansi to Unicode, **CpConverter_v0.1.4**
Merge files	**Txtcollector** **JS Text File Merger 1.0.0**
Split files	**Gsplit 3** **Simpli File Split and Merge 1.4.0**
Remove duplicates	**Noclone,** Duplicate Cleaner, Duplicate File Finder 3.5
Remove empty lines	**CleanHaven**
Change file names	Renamer
Regular expressions	**Edit plus** **Notepad ++**
POS tagging, statistics, concordances, keywords	**Unitex, Tree Tagger, AntConc, Wordsmith tools**
Align texts	**LF aligner,** Akerblad Sentence Aligner, Champollion Tool Kit, Gargantua, XAlign (Unitex)
Align words	Uplug corpus tools, PWA – Plug, UMICS, **MGIZA++,** Multext

Table 1. Tools used for the construction of the corpus (in bold those mostly used)

2.6 PROBLEMS ENCOUNTERED

A vital aspect of handling translated texts is the inherent intellectual property rights (IPR) issues, which can be complex and delicate. Whatever type of corpus is being constructed, it is essential to investigate who owns the copyright, bearing in mind that there may be two copyrights in the case of translated work, one protecting the original author and the other the translator.

Other issues encountered were the rarity of parallel texts, given that we have also a rare pair of languages (out of 19 opus project corpora only four are available in both French and Greek).

Moreover, we had to deal with alignment errors and omissions. A distinctive aspect of the preparation of parallel corpora is the need to line up texts so that equivalent segments in the two (or more) languages can be compared. Because translators do not always translate one sentence by one sentence and do not always keep to the order of the source text, identifying corresponding segments is not a straightforward matter. For this reason, we had to align texts automatically at paragraph or sentence level, using alignment software, and then to inspect the output to correct any misalignment manually.

In many cases, the available tools for corpora were not suitable for Greek language due to different alphabet. Additional problems were caused by the duality in Greek language (katharevousa and Modern Greek): an almost ancient version of language with different writing symbols (?,?,?,?) and a modern one, but still with a non-latin alphabet (not compatible with many of NLP tools).

3. FUTURE PLANS AND PERSPECTIVES

The corpus is in its way to realization. All its overall structure has been designed and its annotation scheme has been developed. We are collecting materials from different sources and we are devising semi-automatic procedures to speed up its construction. Our work will go on until the corpus will be entirely created and all the levels of linguistic annotation will be performed. In the future, we hope to include as many languages as possible.

With our effort, we hope that we may encourage teachers to start looking into the possibilities corpora can offer them.

REFERENCIAS BIBLIOGRÁFICAS

Biber, D. (1993). Representativeness in corpus design.*Literary and Linguistic Computing* 8/4: 243-257.

Danielsson, P., M. Mahlberg (2003). There is more to knowing a language than knowing its words: using parallel texts in the bilingual classroom, in*English for Specific Purposes World. Online Journal for Teachers*, 3 (6), Vol. 2.

http://www.esp-world.info/articles_6/DanielssonMahlberg2003.htm

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells?*Teaching English as a Second Language – Electronic Journal*, 8/4: 1-35. http://tesl-ej.org/ej32/a1.html.

Hadley, G. (2002). Sensing the Winds of Change: An Introduction to Data-Driven Learning. *RELC Journal* 33 (2): 99-124.

Johns, T. (1991). Should you be persuaded: two examples of data-driven learning, in T. Johns et P. King (dir.)*Classroom Concordancing. English Language Research Journal* 4: 1-16.

Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them?, in: Michael McCarthy and Anne O'Keeffe (eds)*The Routledge Handbook of Corpus Linguistics*, Abingdon: Routledge, 487-501.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation.*Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, 79-86.

Krishnamurthy, R. (2001). The Science and Technology of Corpus. Corpus for Science and Technology, in G. Aguado and P. Durán (Eds.)*La investigación en lenguas aplicadas: enfoque multidisciplinar*. Madrid: Universidad Politécnica, 79-114.

Landure & A. Boulton (2010). Corpus et autocorrection pour l'apprentissage des langues.*ASp*, 57: 11-30.

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context,*Language Teaching* 41/4: 563-574.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., & Tufis, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.*Proceedings of LREC*, Genoa, Italy, 2142-2147

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS.*Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, Instanbul, Turkey, 2214-2218.

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure,*Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 1244-1248.

http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf