SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

Key Node in Context (KNIC) Concordances: Improving Usability of an Old French Treebank

Rainsford, T. M.[*], & Heiden, Serge[+]

University of Oxford[*]
CNRS (ICAR & University of Lyon)[+]

tmr740-ac@yahoo.co.uk
slh@ens-lyon.fr

1 Introduction[1]

While much research concentrates on the challenges inherent in the creation and annotation of treebanks, there are also a number of less well-documented challenges presented when the annotated data is used in subsequent linguistic research. Search engines for unannotated corpora, or corpora annotated only at the word level (e.g. PhiloLogic 4[2]) usually present query results in the form of a Key Word In Context (KWIC) concordance: a convenient synoptic overview which can be easily exported to a spreadsheet or sorted in a variety of ways. By way of contrast, most treebank corpus search engines return tree fragments or tree representations as their default representation of query results (e.g. CorpusSearch[3], TigerSearch[4], TrED 2.0[5]). While these preserve the syntactic annotation, enabling the user to verify that their query has returned the desired results, it can be very difficult to obtain a synoptic view of the data, and it is virtually impossible to export the results to common software environments (word-processors, spreadsheets) in a format appropriate for subsequent analysis.

In order to tackle this problem, we propose the creation of Key Node In Context (KNIC) concordances, and present a first implementation of such concordances for an Old French treebank. Our implementation combines the search functionalities of the TigerSearch engine with the versatile TXM text analysis platform[6]. The work was carried out as part of the ANR/DFG funded Syntactic Reference Corpus of Medieval French (SRCMF) project, which has created a 300 000-word treebank of Old French texts (Stein and Prévost, 2013)[7]. A demo version of our implementation of the KNIC concordances is available for the GRAAL corpus at http://txm.textometrie.org/demo?locale=en.

2 Using concordances in linguistic research

2.1 The KWIC concordance

For corpus searches at the lexical level, results are frequently represented in table form in a KWIC concordance, with the keyword matching the search term in the central column. The example in table 1 is based on a KWIC concordance produced by PhiloLogic 4[8] using the MCVF historical French corpus[9].

auvains, et si i fu mes sire	Yvains	, et avoec ax Qualogrenanz, uns chevaliers
Par mon chief, fet mes sire	Yvains	, vos estes mes cosins germains; si nos dev
Or tost, por Deu, mes sire	Yvain	, movroiz vos enuit ou demain? Feites le nos

Table 1: Schematic representation of KWIC concordance produced by PhiloLogic on the MCVF corpus, regex query “Yvains?”.

The concordance form has a number of key advantages for researchers working with the corpus data. Firstly, it presents each search term in its textual context. Secondly, since results are arranged vertically, parallels between occurrences and contexts of occurrence are clearly visible, particularly when the

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801250

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

concordance is sortable (e.g. alphabetically by keyword.) — for example, in the concordance above, it is clear that the name 'Yvain' is preceded by the title 'mes sire' in each occurrence. Thirdly, the tabular form of the concordance makes it an ideal source of data to export to spreadsheet software for more advanced analysis. Spreadsheet software permits additional, more specific annotations to be added to the results of the corpus search[10], which may be suitable for individual studies.

2.2 Case study: Old French flexional *-s*

However, if the nature of the linguistic query requires treebank annotation, it is far less simple to obtain such a user-friendly output. For example, suppose we wish to study the use of flexional *-s*, the Old French masculine singular nominative case marker, on proper nouns. The French case system disappears by the end of the Old French period (mid-14th century), and flexional *-s* is not necessarily marked consistently in texts composed in earlier periods. Our hypothetical corpus user wishes to get a quick overview as to whether the flexional *-s* is consistently used on proper nouns in a particular text.

The most straightforward way of studying this using a treebank is to search for all proper nouns contained in noun phrase subjects[11] and then to check the lexical forms extracted manually. The two main Old French treebanks (the MCVF corpus and the SRCMF corpus) do not contain morphological tagging beyond part-of-speech, so it is impossible to search tags for 'nominative case' or even to restrict the search to masculine nouns only. Consequently, the results returned by the search engine will include some noise, and must be checked manually.

Using the web-based Corpus Search[12] interface provided for the MCVF corpus[13], the query is simple to construct:

search domain: IP-MAT
query: (NP-SBJ* Doms NPR*)

However, for each proper noun NP subject, the results page returns the full bracketed tree, with the subject tags highlighted, for example:

```
Mes sire Yvains cele nuit ot molt boen ostel, (YVAIN,25.793)
1 IP-MAT:  2 NP-SBJ, 8 NPRS
( (IP-MAT    (NP-SBJ (DZ Mes)
                     (NCS sire)
                     (NP-PRN (NPRS Yvains)))
             (NP-TMP (D cele) (NCPL nuit))
             (VJ ot)
             (CODE <milestone%%%unit="folio"%%%n="[82a]"$$>)
             (CODE <lb%%%n="792"$$>)
             (NP-ACC (ADJP   (Q molt)  (ADJ boen))
                     (NCS ostel))
             (PONFP ,))
(ID YVAIN,25.793))
```

In terms of providing a rapid initial answer to the research question "do most inflecting proper nouns show case inflection?", this output is of limited use, as there is no simple way of browsing the lexical form of the proper nouns returned. Moreover, the output of a local installation of CorpusSearch is similar, and while it would be in principle be possible to post-process the output file to produce results in a more readable form, the software required to do this has to the best of our knowledge not been implemented in either the MCVF or other major Penn-format corpora.

Using the TigerSearch search engine on the SRCMF corpus, the query is equally simple to formulate[14]:

```
#sj:[cat = "SjPer" & type = "nV"] >* #npr:[pos = "NOMpro"]
```

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

TigerSearch provides a "Graph Viewer" which, rather like the Corpus Search output file, allows the user to visualize all matching trees in the treebank with the proper noun highlighted in red. Additionally, the built-in statistics viewer can be used to provide a list of the proper nouns matched[15], but without any context. However, TigerSearch additionally allows the treebank to be exported in Tiger-XML format, with nodes matched by the query identified by a <matches> element (*cf.* König *et al.* 2003: 115-117). As a large XML file export is similarly of little use in itself to many researchers, TigerSearch includes a number of XSL stylesheets which can be applied to the Tiger-XML as it is exported in order to produce a more 'readable' output in a plain text file.

Perhaps the most useful of the default stylesheets for the current case study is the 'variables and their tokens' stylesheet, which produces the following style of result for each sentence:

```
------------------------------------------------------------
YvainKu_pb:82_lb:789Tom%20%-%20%1/YvainKu_03_1319815824.9:
Messire Yvains cele nuit ot Mout boen ostel
#npr: Yvains
#sj: Messire Yvains
```

The result shows the context sentence and lists the full NP subject constituent (#sj) as well as the proper noun keyword (#npr). Nevertheless, with around 200 hits in this text alone, this is still not a particularly user-friendly way to browse the data. Above all, the file is in text order and cannot be sorted in any other way.

3 The TXM platform

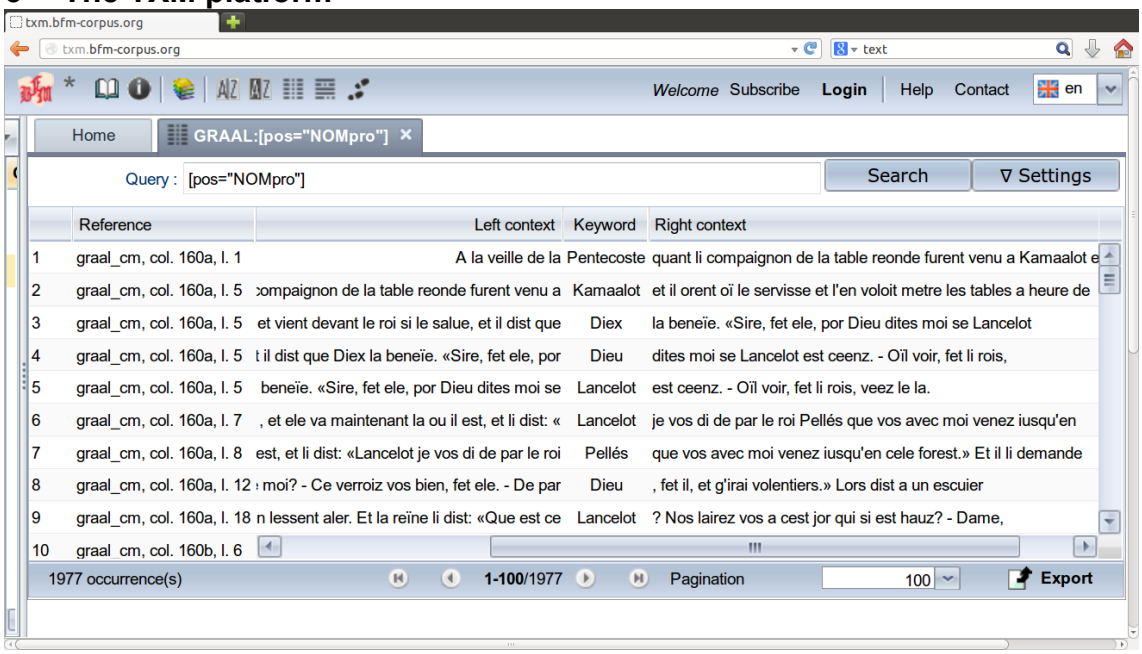

Figure 1: KWIC concordance from CQP query in TXM web portal within Mozilla Firefox web browser

3.1 Overview

The TXM platform is a powerful tool for working with large, richly annotated corpora, and is designed to form a modular and open-source framework for corpus analysis.

The platform is built around four core modules:

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

- the CQP search engine[16], designed for high-performance lexical pattern searches on large corpora (up to a billion words) tagged at the word level;
- R statistics software[17], enabling statistical analysis of corpus queries to be carried out within the TXM platform;
- a web-based or desktop GUI providing access to the CQP and R modules;
- a rich Java/Groovy and XSLT based import subsystem which enables users to manage and to import their own corpora from a variety of file formats (for example, plain text Unicode, XML, XML-TEI, XML-TMX, XML-Transcriber, etc.) while allowing application of NLP tools on the fly (like TreeTagger for example).

Based on the query results returned by the CQP module, TXM provides KWIC concordances combined with a number of sort options (e.g. sort by keyword combined with sort by left or right context, using lexical forms or part-of-speech tags as sort key), all of which are integrated into the GUI, as shown in figure 1. Concordances are also exportable in CSV format. The TXM platform also provides a complete HTML edition of the full text; thus by double-clicking a line of the KWIC concordance, the user has immediate access to an on-screen edition of the text with the keyword highlighted in its full original context.

The platform is available both as a desktop application (for Windows, Mac and Linux) and as a web-based portal. The desktop version is targeted principally at users wishing to import and work with their own corpora. The web-based portal is intended for giving direct access to corpora online or for corpora for which the source files are not freely downloadable. It offers full user subscription and authentification, and allows corpus administrators to restrict corpora to particular users or groups of users, and to block any of the functions of TXM for particular texts. This last feature is extensively used by the *Base de Français Médiéval* corpus to block access to the full online HTML edition of certain texts[18], while allowing concordances with limited contexts to be generated. TXM, and its sources, is downloadable for free at http://sf.net/projects/txm.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

3.2 Adaptation of TXM for treebank corpora

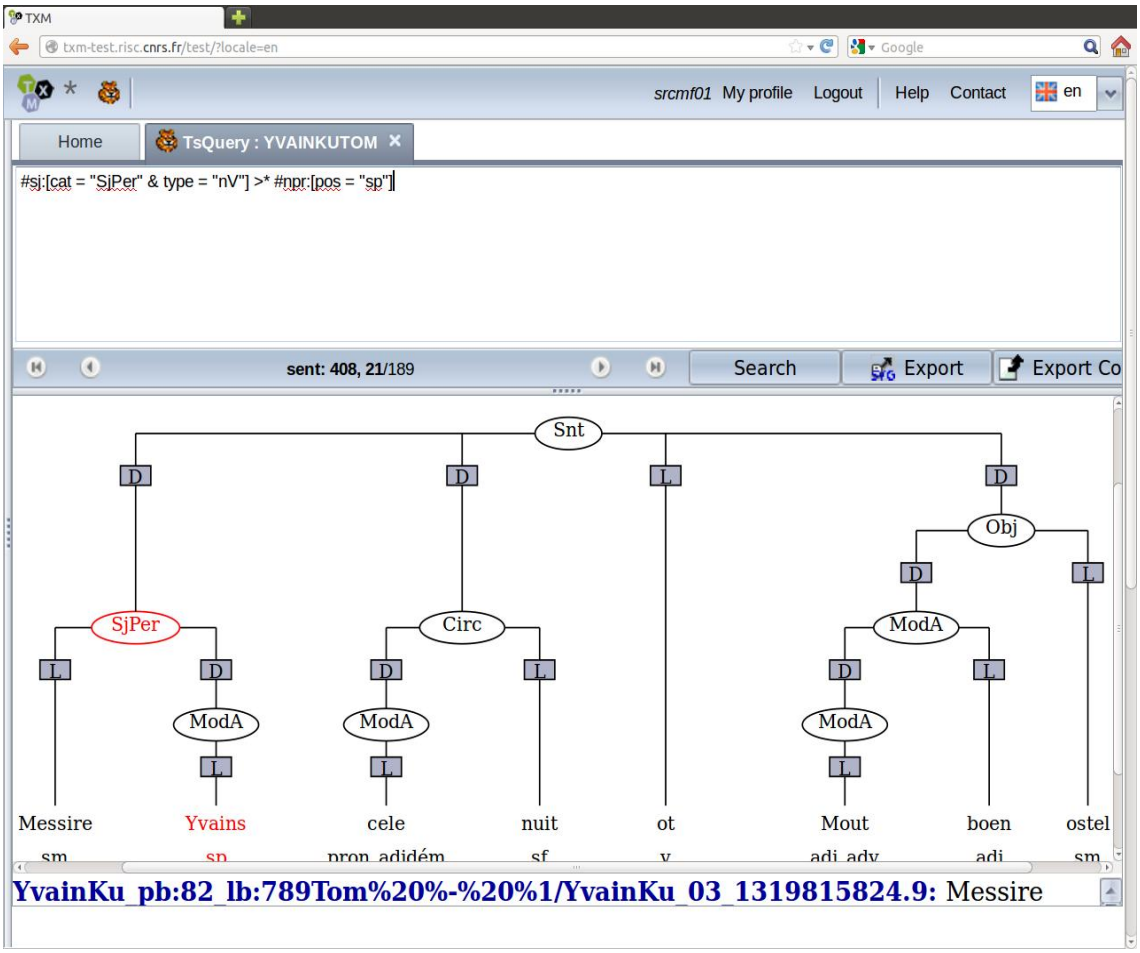

Figure 2: TigerSearch interface within TXM web portal

In order to enable treebank queries in TXM, the TigerSearch search engine and tree drawing components was plugged in to the platform, their GUIs integrated with that of TXM. At present, only the web portal version of TXM includes a UI for TigerSearch. Figure 2 shows TigerSearch within TXM: TigerSearch queries are entered as plain text in the top panel and the resulting trees are shown in the bottom panel. While integration remains relatively low-level, this combination of TigerSearch with TXM provides immediate advantages:

- An online interface for TigerSearch and Tiger corpora;
- Both TigerSearch and CQP are available for the same corpora. Users can thus develop both treebank queries relating to syntactic structure and more efficient lexically oriented queries[19];
- The opportunity for corpus administrators to work with the software developers to improve the visualisation and the export of query results.

A similar online interface for TigerSearch queries is offered by the INESS platform (Rosén *et al.* 2012)[20], although the underlying engine is a re-implementation rather than an integration of the original software. The primary objective of the INESS platform is to provide an online infrastructure for the hosting of

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

treebanks in a variety of formats (dependency, constituency, LFG, HPSG, etc.). However, while it offers advanced functionality for treebank queries, the platform is similar to other treebank search engines in that results are always expressed in the form of trees (with highlighted nodes) or full sentences. To the best of our knowledge, no concordance-style output has been implemented. Moreover, since the software itself is not released under an open-source licence, such functionality cannot be added by third-party developers.

In the final section of this article, we will focus on how the TXM platform is used to provide a user-friendly interface to a number of scripts and stylesheets designed to produce concordance-style results from TigerSearch queries.

4 From TigerSearch query to KNIC concordance

4.1 Transforming TigerSearch exports

Once a query has been run, the TigerSearch engine allows the user to re-export the corpus as a Tiger-XML in which structures matching the user's query are marked by <matches> nodes. The default stylesheets, such as the 'variables and their tokens' example above, take this Tiger-XML as their input. Matches nodes con-tain one or more <match> nodes which identify each subgraph within the sentence which matches the whole Tiger-XML query. Moreover, each <match> node contains one or more <variable> nodes, which correspond to each node identified within the query. Returning to our query from section 2.2, the Tiger-XML representation of the match identifies the values assigned to query variables #npr and #sj:

```
<matches>
    <match subgraph="[...]#Tom%20%-%20%1/YvainKu_03_1319815832.29">
                <variable name="#npr"idref="[...]#w_YvainKu_5490" />
                <variable name="#sj" idref="[...]#Tom%20%-
%20%1/YvainKu_03_1319815832.29" />
     </match>
</matches>
```

As with the TigerSearch default stylesheets, concordances are generated by the TXM platform by applying an XSL to this exported Tiger-XML file. However, three key innovations in the design of the export module permit a much more sophisticated treatment of the data than that offered by the default stylesheets.

Firstly, the concordance XSL is sensitive to the query variable name #pivot. By identifying a single node within the query, terminal or non-terminal, as #pivot, the user is able to specify the node within on which the concordance will be centred. The XSL uses the name attribute of the <variable/> element in the exported Tiger-XML to identify this node.

Secondly, the transfomation of the Tiger-XML file is managed by the TXM platform rather than by TigerSearch. Although TigerSearch also permits new XSL stylesheets to be integrated into the export procedure (*cf.* König *et al.* 2003: 108-109), there are a number of technical limitations:

- Sentences are piped one by one through the XSL, so the context shown for each result in the concordance is limited to a single sentence;
- No support is provided for XSLT version 2;
- It is impossible to pass parameters to the XSL (e.g. 'show syntactic function in addition to lexical content').

Thirdly, with particular reference to the SRCMF corpus, the corpus text present in the annotated Tiger-XML file does not always respect the source text. Discrepancies are caused mainly by the removal of

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

punctuation from the Tiger-XML file, as this had been found to impede TigerSearch queries based on the linear ordering of constituents. Using the unmodified source text from the CQP corpus, TXM is able to post-process TigerSearch query results to ensure that the punctuation is visible in KNIC concordances, making them much easier and quicker to read.

The TXM platform permits a seamless integration of the KNIC concordance with TigerSearch, in addition to corpus-specific post-processing such as the re-injection of punctuation. However, we have also made the XSL stylesheets used to generate the concordances available independently of TXM[21], so that advanced users can manually apply them to XML files exported by local installations of TigerSearch using a XSLT 2.0 processor such as SAXON 9[22]. In this way, KNIC concordances can be produced for *any* TigerSearch-compatible corpus.

4.2 Creating a simple concordance

To apply the concordance stylesheet, the flexional *-s* query above must be modified to apply the concordance stylesheet with the introduction of the #pivot variable:

```
#sj:[cat = "SjPer" & type = "nV"] >* #pivot:[pos = "NOMpro"]
```

From the exported Tiger-XML, the XSL produces a table in CSV format, with all proper noun subjects aligned under 'pivot'. Sorting the concordance by pivot in a spreadsheet editor allows the linguist to summarize the use of the case system in the Yvain text in a matter of few minutes: all masculine singular nouns carry nominative *-s* inflection (sometimes spelt <z>) when in subject position. False hits are easily identified thanks to the context: for example, the only nouns without inflection in the results are (a) non-inflecting feminine nouns or (b) proper nouns more deeply embedded within the subject, in a prepositional phrase or as a genitive (e.g. *la suer Monseignor Gauvain*, 'my lord Gawain's sister').

LeftCx	Pivot	RightCx
Messire	Yvains	ne sejorna [...]
[...] Que n'ot conté	Calogrenanz	
Messire	Yvains	cele nuit ot Mout boen ostel
[...] Au chevalier messire	Yvains	
Et messire	Yvains	de randon [...]

Table 2: Schematic representation of basic treebank concordance, proper nouns within the subject.

4.3 Representing keynodes heading discontinuous structures

Unlike traditional concordances, there is no guarantee that the structure headed by the #pivot variable is a single word nor, in a dependential corpus, that it even denotes a contiguous sequence of words. For example, suppose we wished to extract all proper nouns modified by relative clauses, placing the whole structure (noun and dependents) within a tabular concordance[23]:

```
#pivot:[type = "nV"] >1,2 #npr:[pos = "NOMpro"]
& #pivot >D #moda:[type = "VFin"]
```

The #pivot may head a non-contiguous structure:

Je meïsmes **cil Yvains** sui **Por cui vos estes an esfroi**
'I myself am that Yvain for whom you are crying out.'

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

In the first example here, the subject #pivot is divided into two parts (NP and relative clause) by the finite verb *sui*.

LeftCx	Pivot
Je meïsmes	cil Yvains **[sui]** Por cui vos estes an esfroi
S'an est or entrez an grant painne	Messire Gauvains qui la quiert

Table 3: Basic treebank concordance, proper nouns with relative clause modifiers.

Words which split the lexical content headed by a key node into two or more parts are marked in the concordance by square brackets, as shown in table 3. This ensures that reading from left to right, the concordance still represents word order of the sentence as written, while also providing a clear visual indication that the syntactic structure is not contiguous.

4.4 Multiple keynodes: Pivot and blocks

While the basic concordances are extremely useful, studies of word order often need to identify a number of nodes which are of interest to the user, not only one. For example, suppose we wished to study all main clause sentences with a NP subject and an NP object in order to study which word orders are attested (SVO, OVS, etc.). It is straightforward to create a corpus query which returns sentences with an NP subject and an NP object, for example:

```
#snt:[cat = "Snt"] >D #suj:[cat = "SjPer" & type="nV"]
& #snt >D #obj:[cat = "Obj" & type="nV"]
& #suj >L [pos = /NOM.*/]
& #obj >L [pos = /NOM.*/]
& #snt >L #pivot
```

The #pivot here is the finite verb, but the #suj and #obj nodes are also of interest. A more advanced form of KNIC concordance allows the user to name secondary nodes of interest as 'blocks', using variable names #blocka, #blockb etc. in the query. The resulting concordance presents a table in which each 'block' is assigned a separate column either preceding or following the pivot column, depending on the block's position within the sentence.

Block-2	Block-1	Pivot	Block+1	Block +2
	Sj li preudons	resgarde	Obj les letres	
	Obj escu {vos}	envoiera	Sj Diex	
		oste	Sj li chevaliers	Obj son hiaume

Table 4: Schematic representation of pivot and block KNIC concordances, main clause verb with NP subject and NP object.

The examples in table 4 are representative of the three combinations of subject, object and verb (SVO, OVS and VSO) found in the prose *Queste du saint Graal* text as shown by the pivot and block concordance (*cf.* Table 5 in the appendix for a more complete view of these complex concordances). The syntactic function of each block is shown as well as the text, allowing quick sorting of the concordance in a spreadsheet editor in order to group sentences with similar word order together. Where blocks are not immediately adjacent to the pivot or to each other, intervening words (such as the indirect object pronoun *vos* in the second example) are marked with curly brackets.

Feedback from users of the SRCMF corpus indicates that this form of output is the most helpful form of result visualization. It permits the highlighting of a particular constituent (or constituents) as 'blocks' while at the same time aligning results by another common element (in this case the finite verb). Users of the SRCMF corpus have made use of this type of concordances in a number of studies on Old French:

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

- elements preceding the finite verb in main clauses (Rainsford *et al.* 2012)
    - pivot: finite verb
    - blocks: preverbal elements
- development of the neuter demonstrative pronoun CE used other than as a subject (Glikman and Rainsford 2012)
    - pivot: governing verb
    - block: CE and dependents
- relative order of objects and complements (current research project within the Labex "Empirical Foundations of Language")
    - pivot: governing verb
    - blocks: object(s) and complements

Due to the computational complexity of this concordance, it would be difficult to implement and maintain in XSL. A further strength of the TXM platform is that it can also apply scripts written in Groovy, a Java-based scripting language, to the XML outputted by TigerSearch. We plan to develop several concordancing scripts, similar to existing TigerSearch XSL stylesheets.

5 Evaluation and Future Developments

In the course of the development of an Old French treebank, we found that researchers wishing to use the corpus were not able to export and analyse results in such a way as to be able to answer core research questions in Old French syntax. The KNIC concordance presented in this paper is a response to this problem, providing the user with a convenient synoptic view of query results which is easy to export and analyse further in spreadsheet software. A number of recent research projects and papers based on the SRCMF corpus have relied on data presented in the form of a KNIC concordance.

In addition to providing an architecture in which post-processing of TigerSearch's exported results files can easily be implemented to create these concordances, the TXM platform's corpus administration features and web interface allow Tiger treebanks to be uploaded and used online without distributing the corpus source files (if necessary). However, currently the integration of TigerSearch and TXM remains relatively low-level. A number of key improvements are envisaged in the future:

- The TigerSearch interface within TXM will be developed to include more features of the TigerSearch GUI, particularly coloured query syntax and lists of available values for each node feature.
- KNIC concordances are currently available only for export, but should be integrated into the TXM's KWIC concordance GUI. This requires a higher-level integration of the Tiger module within TXM.
- Higher-level integration of TigerSearch will also allow TXM's R-based statistical tools to be used on treebank query results, as it is already available for the CQP query results.

Bibliography

Glikman, J. and Rainsford, T. M. (2012) CE objet ou attribut : Étude diachronique. Presentation at Diachro VI, KU Leuven, 17–19 October 2012.

Heiden, S. (2010) The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto and Y. Harada (eds) *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLING 24)* Institute for Digital Enhance-ment of Cognitive Development, Waseda University, p. 389–398.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

Heiden, S., Magué, J.-P., Pincemin, B. (2010) TXM: Une plateforme logicielle open-source pour la textométrie — conception et développement. In S. Bolasco, I. Chiari, and L. Giuliano (eds) *Statistical Analysis of Textual Data: Proceedings of 10th International Con-ference JADT 2010.* Rome : Edizioni Universitarie di Lettere Economia Diritto.

König, E., Lezius, W. and Voormann, H. (2003) *TIGERSearch 2.1: User's Manual*, IMS, University of Stuttgart. Published online at http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html.

Lezius, W. (2002) *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*, Ph.D. thesis, IMS, University of Stuttgart, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4.

Pincemin, B., Heiden, S., Lay, M.-H., Leblanc J.-M. and Viprey, J.-M. (2010) Fonctionnalités textométriques: Proposition de typologie selon un point de vue utilisateur. In S. Bolasco, I. Chiari, and L. Giuliano (eds) *Statistical Analysis of Textual Data: Proceedings of 10th International Conference JADT 2010.* Rome : Edizioni Universitarie di Lettere Economia Diritto.

Rainsford, T. M., Guillot, C., Lavrentiev, A., and Prévost, S. (2012) La zone préverbale en ancien français : apport de corpus annotés. *SHS Web of Conferences, 1,* 159-176 <DOI: 10.1051/shsconf/20120100246>.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012) An open infrastructure for advanced treebanking. In J. Hajič, K. De Smedt, M. Tadić, and A. Branco (eds.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012,* Istanbul, Turkey, May 2012, p. 22-29.

Stein, A. and Prévost, S. (2013) Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF). In P. Bennett, M. Durrell, S. Scheible, and R. J. Whitt (eds) *New Methods in Historical Corpora, Corpus Linguistics and International Perspectives (CLIP) vol. 3*. Tübingen : Narr, p. 275-282.

---

[1] This work was initially carried out as part of the 'Syntactic Reference Corpus of Medieval French (SRCMF)' project (2009-2012), jointly funded by the ANR (France) and the DFG (Germany) with principal investigators Sophie Prévost (Lattice, CNRS & ENS) and Achim Stein (Stuttgart). Subsequent development has formed part of T. M. Rainsford's British Academy Post-Doctoral Fellowship (2012-2015) at the University of Oxford.

[2] Project homepage: https://sites.google.com/site/philologic3/.

[3] Created by Beth Randall. Homepage: http://corpussearch.sourceforge.net/.

[4] Lezius (2002). Project homepage http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml.

[5] http://ufal.mff.cuni.cz/tred/.

[6] Heiden (2010); Heiden *et al.* (2010); Pincemin *et al.* (2010). Project homepage: http://textometrie.ens-lyon.fr/?lang=en.

[7] Project homepage: http://www.srcmf.org.

[8] Project homepage: https://sites.google.com/site/philologic3/.

[9] Project directed by France Martineau, project homepage http://www.voies.uottawa.ca/index.html.

[10] For example, new annotations can be placed in additional spreadsheet columns. These annotations can then be projected back into the source corpus, a workflow already available in TXM for lexically oriented corpora.

[11] It is true that nominative forms also surface in a few other environments in Old French, most notably as subject attributes in copular constructions. However, in order to keep the demonstration as simple as possible, we will restrict the query to subject environments only. Moreover, such a query would be perfectly adequate for our hypothetical corpus user who wishes to get a rough idea

[12] Created by Beth Randall. Homepage: http://corpussearch.sourceforge.net/.

[13] http://www.voies.uottawa.ca/corpus_pg_en.html. Registered users only.

[14] A guide to the tags used in the SRCMF corpus is available at http://www.srcmf.org/fiches/index.hml [in French].

[15] The word property of the terminal node #npr is used.

[16] Part of the IMS Open Corpus Workbench (CWB), see http://cwb.sourceforge.net.

[17] http://www.r-project.org/.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

[18] This is in order to comply with rights holders' requirements for including the text in the corpus.

[19] Note that the CQP search engine can also query syntactic annotations directly, provided that the treebank uses a non-projecting dependential annotation model which can be encoded in the CoNLL format. Simpler syntactic queries can thus make use of the greater processing speed of the CQP engine. For example, one could search for all subjects headed by a proper noun using a CQP query such as: [pos = "NOMpro" & deprel = "SjPer"]. However, it would not be straightforward to replicate the treebank queries from section 2, which search for all subjects containing proper nouns.

[20] http://clarino.uib.no/iness/main-page.

[21] http://sourceforge.net/projects/knicconcordances/

[22] http://saxon.sourceforge.net/

[23] Gloss of query: a non-verbal structure (type = "nV") immediately contains (i.e. has as its head) or contains at distance 2 a proper noun (pos = "NOMpro", and also contains a clause (type = "VFin")). Note that in sequence of title plus proper noun (e.g. messire Yvains), proper nouns are dependent on their titles in the SRCMF corpus, which makes the query more complex.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801250

Appendix

sId	LeftCx	Block-1	Type	Pivot	Block+1	Type	Block+2	Type	RightCx
[1]	et	li baron	SjPer	resgardoient	les letres qui * disoient	Obj	--	--	
[2]	Et	li preudons	SjPer	resgarde	les letres	Obj	--	--	
[3]	car	cele aventure {ne}	Obj	volt	{onques} {mes} nus hons [achever] qui * n' i fust [#] morz ou mehaigniez ainz qu' il l' eust menee a fin	SjPer	--	--	
[4]	Biax sire fet li rois	escu {vos}	Obj	envoiera	Diex	SjPer	--	--	d' aucune part ausi com il a fet espee
[5]	Quant tuit furent assemblez es * prez de Kamaalot # li grant et li petit	Galaad {par} {la} {proece} {#} {dou} {*} {roi} {et} {de} {la} {reïne}	SjPer	mist	{#} son hauberc	Obj	{en} {son} {dos} {et} son hiaume	Obj	en sa teste
[6]	Et # quant il fu auques anuitié et il fu hore de dormir	li rois	SjPer	prist	Galaad	Obj	--	--	
[7]	et	li solaux	SjPer	ot	{ja} {auques} {abatue} la rousee	Obj	--	--	
[8]	Si	--	--	jurroient	li compaignon	SjPer	tel serement come cil font qui * en queste doivent entrer	Obj	
[9]	Et quant il furent aporté devant les mestres dois	li rois	SjPer	apela	mon seignor Gauvain	Obj	--	--	
[10]	aprés	--	--	jura	Lancelot	SjPer	tout autretel serement com il avoit fet	Obj	
[11]	Lors	--	--	oste	mes sires Gauvains	SjPer	son hiaume	Obj	de sa teste
[12]	et mout	--	--	honorerent	li frere	SjPer	Galaad	Obj	quant il oïrent le tesmoign que * li dui chevalier li portoient
[13]	et	li frere de laienz {li}	SjPer	baillierent	.i. escuier [por] [fere] [li] [compaingnie] qui * raportera arriere l' escu s' i le covient a fere	Obj	--	--	

Table 5: First 13 main clauses with both a nominal subject and a nominal object in the *Queste du saint Graal* (SRCMF corpus) presented in a pivot and block concordance.