Filling in the gaps: what we need from TM subsegment recall

Kevin Flanagan
Swansea University

ABSTRACT

Alongside increasing use of Machine Translation (MT) in translator workflows, Translation Memory (TM) continues to be a valuable tool providing complementary functionality, and is a technology that has evolved in recent years, in particular with developments around subsegment recall that attempt to leverage more content from TM data than segment-level fuzzy matching. But how fit-for-purpose is subsegment recall functionality, and how do current Computer-Assisted Translation (CAT) tool implementations differ? This paper presents results from the first survey of translators to gauge their expectations of subsegment recall functionality, cross-referenced with a novel typology for describing subsegment recall implementations. Next, performance statistics are given from an extensive series of tests of four leading CAT tools whose implementations approach those expectations. Finally, a novel implementation of subsegment recall, 'Lift', is presented (integrated into SDL Trados Studio 2014), based on subsegment alignment and with no minimum TM size requirement or need for an 'extraction' step, recalling fragments and identifying their translations within the segment even with only a single TM occurrence and without losing the context of the match. A technical description explains why it produces better performance statistics for the same series of tests and in turn meets translator expectations more closely.

1. Introduction

The segment-oriented nature of Translation Memory (TM) can seem to restrict its usefulness, in ways to which Machine Translation (MT) – in particular, Statistical Machine Translation (SMT) – provides an alternative. Bonet explains that, for the TMs at the DGT, "Many phrases were buried in thousands of sentences, but were not being retrieved with memory technology because the remainder of the sentence was completely different" (2013: 5), and that SMT trained on those memories enabled some of that 'buried' content to be recalled. However, TM technology has evolved in recent years, including subsegment recall features that attempt to leverage more content from TM data than segment-level fuzzy matching. In principle, TM subsegment recall – automatically finding phrases within segments that have been translated before and identifying the corresponding translated phrase in the previously-translated segment – should recover all that content. This functionality is described by Zetzsche as "probably the biggest and the most important development in TM technology" (2014), but in practice, implementations in TM systems vary widely, and fall very short of that level of capability, leading to further observations by Zetzsche that "we are still in the infancy of these developments", and that "subsegmenting approaches are almost as varied as the number of tools supporting them" (Zetzsche, 2012: 51).

The discussion in this paper is expressed in terms of segment-based TM, that is, TM containing Translation Units (TUs), each containing an easily-demarcated source text (ST) segment – such as a sentence, heading or list item – and its corresponding target text (TT) translation. However, the principal issue for subsegment recall – how to match fragments of segments, and retrieve the translation of a fragment, rather than of the whole segment where it occurs – applies equally to character-string-in-bitext (CSB) TM systems, where STs and TTs are stored in full, since the ST and TT alignment information available is essentially at the same level of granularity, so automatic identification of the translation of a fragment is problematic. For both segment-based and CSB systems, translators can usually prompt a search for a specific fragment – referred to herein as a *concordance search* – to find occurrences of fragment repetitions. Even so, discounting the time required to do so for all possible fragments (which some CAT tools will attempt automatically), the results show only the larger segment within which the fragment's translation is found, leaving the translator obliged to spend time and effort scanning through it. To aid discussion of these and other considerations, and since the distinctions between approaches to subsegment recall in different CAT tools are not immediately obvious, Table 1 defines a typology of 'behaviours' – different techniques and characteristics – that can be used to describe subsegment recall implementations. These are discussed at greater length in (Flanagan, forthcoming 2015b). A more detailed version of this paper is also available at http://kftrans.co.uk/FillingInTheGaps.pdf. In the next section, the typology will be used to present the views of translators that participated in a subsegment recall survey.

2. Translators' views

To gauge what functionality translators would like from subsegment recall, a controlled multiple-choice survey was conducted of translators from four groups: the Western Regional Group of the Institute of Translation and Interpreting (ITI);[1] translators registered with Wolfestone,[2] a successful language services agency; the ITI's French Network;[3] and students on MA in Translation programmes at Swansea University.[4] In all, 91 responses were received, approximately evenly spread across the four groups. Details of questions and responses can be viewed at http://kftrans.co.uk/benchmarks/Home/Survey and are discussed at greater length in (Flanagan, forthcoming 2015b).

In summary, the responses showed a broad consensus with regard to subsegment recall features. Most expect TM-TDB to be available; there is a fairly equal split between wanting DTA/BFE and wanting ACS; VL is not desirable; requiring a TM to be large for subsegment recall is not desirable, and subsegment recall should be available for fragments occurring only once in the TM. The split between those wanting DTA/BFE and those wanting ACS merits examination. As

[1] http://www.itiwrg.org.uk
[2] http://www.wolfestone.co.uk/
[3] http://www.iti-frenchnetwork.co.uk/
[4] http://www.swansea.ac.uk/translation/

ACS, on the face of it, requires more translator time, since the TU has to be manually examined to locate the corresponding fragment translation, why would this be preferred by some over DTA or BFE? I speculate that this is because experienced translators are more aware of the dangers of decontextualisation, and the DTA/BFE response option did not specify whether context is provided. If another option had been available, like the DTA/BFE option but explaining that the translation suggestion was provided by (say) displaying the target segment from the TU with the translation suggestion highlighted, I suspect this response would have been chosen by the majority of respondents.

Having established a baseline for translators' expectations for subsegment recall functionality, the next section will compare those expectations with actual CAT tool capabilities.

3. CAT tool comparison

Table 2 compares the subsegment recall functionality for all CAT tools that provide such a feature and were available at time of writing for trial (or free) use by translators, representing the range of software available to a translator evaluating tools before purchase. A tick means the CAT tool supports the feature, and any term used for it appears below the tick. This gives a high-level view of how varied is the functionality in different CAT tools providing subsegment recall. DTA and BFE implementations merit further examination, since approaches and results vary much more than for (say) the comparatively straightforward TM-TDB feature. Furthermore, the expectations from translators include the ability to recall translations of fragments even if they only occur once in a TM, and without needing the TM to be large. The following section examines this more closely.

Behaviour	TM content	Example query	Response
Use TM like a TBD (TM-TDB)	**EN:** Dynamic Purchasing System **FR:** Système d'acquisition dynamique	We will define a completely electronic dynamic purchasing system for commonly-used purchases	Highlights 'dynamic purchasing system' in query, displays "Système d'acquisition dynamique".
Automatic Concordance Search (ACS)	**EN:** A procuring entity may set up a system for commonly-used purchases. **FR:** L'entité adjudicatrice peut mettre en place un système pour des achats d'usage courant.	(as above)	Automatically highlights 'commonly-used purchases', displays complete FR segment from TM where match found.
Dynamic TM Analysis (DTA)	(as above)	(as above)	Retrieves translation for 'commonly-used purchases', i.e. 'achats d'usage courant'.
Bilingual Fragment Extraction (BFE)	(as above)	(as above)	Same as DTA, but requires TM content to be extracted beforehand.
Decontextualisation	(as above)	(as above)	Retrieves translation for 'commonly-used purchases', i.e. 'achats d'usage courant', but does not show context,

			i.e. "L'entité adjudicatrice […]".
Machine Recall (MR)	(as above)	(as above)	Displays translation for 'commonly-used purchases' automatically.
Assisted Recall (AR)	(as above)	(as above)	Only displays translation for 'commonly-used purchases' when user starts to type it, i.e. types 'a' or 'ac'.
Variation Loss (VL)	**EN:** The company can therefore be qualified as a firm in difficulty. **FR:** C'est pourquoi elle est considérée comme une entreprise en détresse. […] **EN:** The firm in difficulty may benefit from aide. **FR:** L'entreprise en difficulté peut bénéficier d'une aide.	It is doubtful whether a firm generating profits so quickly can be deemed to be a firm in difficulty.	Only one of the translations of 'firm in difficulty' is retrieved.

Table 1: Typology examples

	TM-TDB	ACS	DTA	BFE	Min TM size	Min. occurrences	Decontextualisation	Recall type
SDL Trados Studio 2014	-	-[6]	-	✓ 'AutoSuggest Creator'	**10,000**	-[3]	**Yes**	**AR**
MetaTexis v3.17	✓ 'use TM as TDB	-	-	-	-	-	-	-
memoQ 2013[8] R2	✓[7] 'LSC	✓[7] 'LSC'	✓[1]	✓ 'Muse'	-	**(ACS)2[2]** **(BFE)5**	**(ACS)No** **(BFE)Yes**	**(ACS)MR[4]** **(BFE)AR**
MemSource v3.148	✓ 'Subsegment match'	-	-	-	-	-	-	**MR**
Déjà Vu X2[5] v8	✓ 'Assemble'	-	✓ 'DeepMiner'	-	-	-[3]	**Yes**	**MR[4]**
Similis Freelance v2.16	-	-	✓	✓ 'Glossary'	-	-	**Yes**	**MR**

Table 2: Subsegment recall types by CAT tool

1. if ‘Guess translation’ activated.
2. Can be configured for just one occurrence, though DTA results less reliable (see later analysis in this paper).
3. No minimum specified, but with few occurrences or only one, results may be poor (see later analysis in this paper).
4. AR suggestions are also available.
5. Déjà Vu X3 was released in February 2014; initial testing indicates this functionality is essentially unchanged.
6. The Concordance Search option “Perform search if the TM lookup returns no results” is not an implementation of ACS.
7. The same ‘LSC’ feature names covers both TM-TDB and ACS when – say - enabling/disabling this functionality, even though they give rise to different behaviours; TM-TDB matches show the translation in the results pane, ACS matches don’t.
8. memoQ 2014 was released in June 2014; initial testing indicates this functionality is essentially unchanged.

(Note: Fluency 2013 includes BFE, but this was not functional at time of writing, something the vendor confirmed would be addressed (Tregaskis, 2014). Across Language Server provides BFE functionality, but unlike Personal Edition there is no trial or free version available.)

4. Performance comparison

DTA and BFE subsegment recall implementations in CAT tools are very varied and require close examination to determine how well they meet translators’ functionality expectations. This section presents a suite of tests used to measure their performance in this regard, starting with a TM containing known subsegment fragments and their translations, querying the TM with sentences to translate containing one of the fragments, then checking whether the fragment translation is recalled.

4.1. Data preparation

To select test fragments and their translations for use in such a performance evaluation, a 40,000 TU French-English section of the DGT-TM (Steinberger, Eisele, Klocek, Pilos, & Schlüter, 2013) was processed to select some frequently-occurring fragment pairs, shown in Table 3, along with codes used to refer to them herein. For each fragment pair, 100 ‘fragment-bearing’ TUs (TUs containing the fragment pair) were extracted. A further 10,000 ‘padding’ TUs containing none of the fragments was extracted for creating test TMs. To simulate translating a source text that includes a test fragment also found in a test TM, example sentences – hereafter, ‘queries’ – were created by adapting fragment-bearing TUs. Each query TU was compared to the 10,000 ‘padding’ TUs and the relevant fragment-bearing TUs to ensure that neither French nor English segment constituted a ‘fuzzy match’ with any TU segment.

Code	French	English
1	Règlement	Regulation
1a	Établi	Established
2	conclut que	concludes that
2a	État membre	Member State
3	modifiée comme suit	amended as follows
3a	les autorités polonaises	the Polish authorities
4	intégrée dans l'accord	incorporated into the Agreement
6	Journal officiel de l'Union européenne	Official Journal of the European Union

Table 3: Test fragment pairs

For each query TU, a TM was created for different combinations of padding-TU quantity and fragment-bearing TU quantity; 100, 1,000 or 10,000 padding TUs combined with 1, 100 or 1,000 fragment-bearing TUs, making nine TMs per query TU (and nine further reversed-language-way TMs). Two documents per query TU were created – one containing the French query sentence; the other, the English – and presented for translation by each CAT tool using each of the nine TMs in turn. Subsegment translation suggestions were recorded and scored as described below. Further details of test data and queries, plus discussion of the motivation behind them and their preparation can be found in (Flanagan, forthcoming 2015b).

4.2. Scoring

For these tests, suggestions were scored in terms of *precision* and *recall*. Precision is the percentage of words in the suggestion that occur in the expected fragment translation (expressing how much is relevant), while recall is the percentage of words in the expected fragment translation found in the suggestion (expressing how complete the recall is). Where there are several suggestions, these values are averaged. This is discussed further in (Flanagan, forthcoming 2015b)

For certain CAT tools, subsegment recall precision cannot be evaluated, since the interface does not show which part of the source text the suggestion words are meant for. These cases are shown below as 'Precision unavailable'. The exact procedures for recording results for each CAT tool vary according to their very different approaches; the specifics for each can be found in (Flanagan, forthcoming 2015b). Results show how each tool performed under varying conditions (TM size, number of fragment occurrences, fragment length). Due to their approaches to providing subsegment recall suggestions, results for different tools are not directly comparable, but do show how the variables concerned affect performance in different ways and give some indication of how performance may differ between tools. Results are based on AR suggestions only if the tool does not offer any MR implementation.

4.3. Results

The graphs in Table 4 show recall and precision for each CAT tool tested, averaged over all test queries for that tool (eight English queries and eight French queries), where the X-axis shows fragment frequency (the number of fragment-bearing TUs in the TM), and the different lines show the number of padding TUs in the TM. Note: for Similis, varying volumes of TM padding make no difference to results, so they were all obtained using the same amount of TM padding. For memoQ, DTA subsegment recall was evaluated, since its AR-based BFE implementation cannot be configured to recall fragments with fewer than 5 occurrences. Detailed results for the individual queries and CAT tools can be found at http://kftrans.co.uk/benchmarks.

The averaged results show that very different results are produced by the tools tested. Results from a given CAT tool for individual queries show an interesting lack of consistency. With memoQ, for fragment 3, recall is consistently high, and precision tends to increase with frequency, regardless of padding volume, while for fragment 3a, recall and precision drop sharply as frequency increases, depending on padding volume. Performance is generally comparable when the language direction is reversed, but in some cases differs noticeably. With Déjà Vu X2, performance in individual cases is very varied, with noticeable differences dependent on language direction. Graphs for Similis tend to be flat – if Similis can recall a fragment suggestion, frequency usually makes no difference to whether it is recalled, generally with consistent precision. However, recall seems to be affected by the grammatical category of the fragment sought (per the results for the two different three-word fragments, for example), so that for certain fragments, no translation suggestion is produced regardless frequency. With SDL Trados Studio, performance overall is quite consistent, with 100% recall usually achieved at a frequency of 10, though it has a large TM requirement, and the implementation is AR rather than MR.

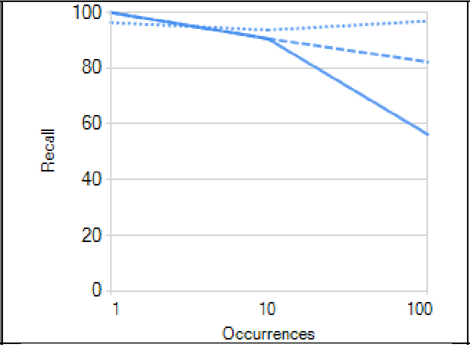
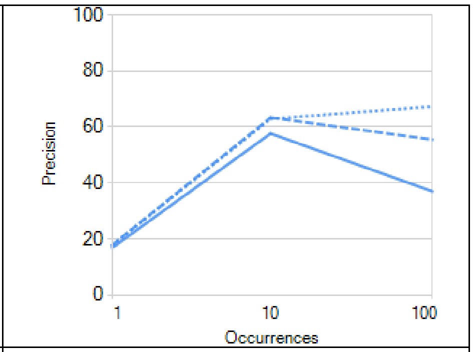
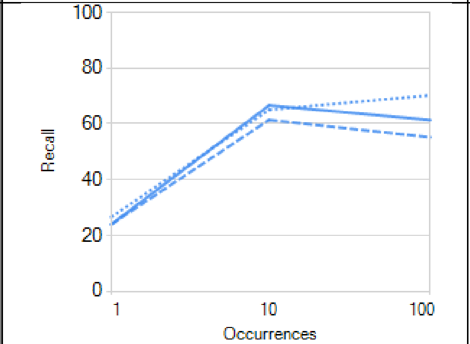
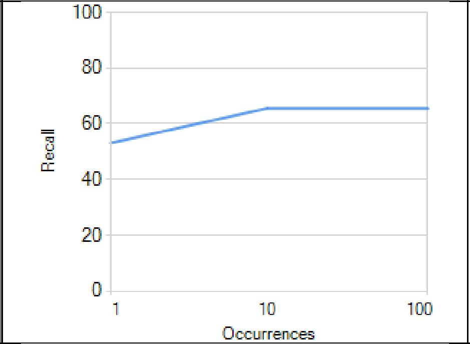
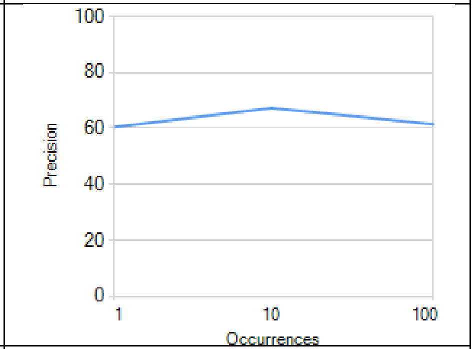
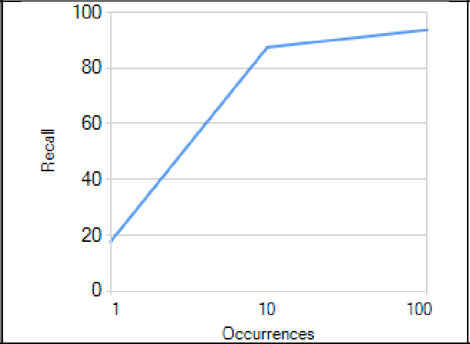
memoQ		
DéjàVu X2		Precision unavailable
Similis		
SDL Trados Studio		Precision unavailable

Table 4: Averaged performance statistics by CAT tool

TM padding:	
───	100 TUs
- - -	1000 TUs
······	10000 TUs

4.4. Discussion

At least some CAT tools provide implementations which – under the right circumstances – provide subsegment translation suggestions with good recall and precision levels, though performance may be inconsistent, with identical texts and data giving different results if the language direction is reversed, for instance. Translators surveyed have some clear preferences about subsegment recall functionality, including wanting it available even for small TMs, and even for fragments occurring only once. Of the DTA/BFE systems tested, Similis had the best average performance under those circumstances, recalling translations of single-occurrence fragments about half the time, with average precision around 60%. However, its BFE methodology decontextualises the translations, arguably aggravating still further a weakness in segment-level TM, and in different circumstances (more fragment occurrences, sufficiently large TM) it can be out-performed by other systems.

Although weaker in other areas, Similis meets the aforementioned preferences better because it is the only system not reliant on statistical analysis or repetitions, instead aligning ‘chunks’ of source and target language segments, “as long as the languages processed are parallel enough for it to do so” (Planas, 2005: 5). The next section presents Lift, a TM system intended to meet translator expectations better by also taking an ‘aligning’ approach, but developed for more consistent results, and with a DTA rather than BFE methodology so as not to decontextualise translations recalled, as well as to reflect TM content changes immediately.

5. Lift

5.1. Overview

Lift is a TM system implementing DTA subsegment recall based on fine-grained alignment of segment pairs, or *subsegment alignment*. To enable subsegment alignment, Lift uses a range of bilingual dictionary resources to establish tentative lexical alignments, then an iterative hypothesis evaluation algorithm plus some deductive heuristics to produce a hierarchically-arranged set of word span alignments. The alignment algorithm is described in detail and compared with other approaches in (Flanagan, 2014). Figure 1 shows a high-level view of the alignment process. (For the aligned sentence pair, connecting lines show alignments between words, while parallelograms show alignment between spans of words.) The effects of using the optional components are described in (Flanagan, forthcoming 2015a).

During translation, Lift uses a longest-common-substring algorithm coupled with indexing techniques and configurable parameters (such as minimum fragment length and proportion of ‘stop’ words) to match fragments of a query (that is, a sentence to translate) with fragments of TM content, and uses the alignment information to recall and propose the translations of those fragments to the translator. The recall process is described at greater length in (Flanagan, forthcoming 2015a).

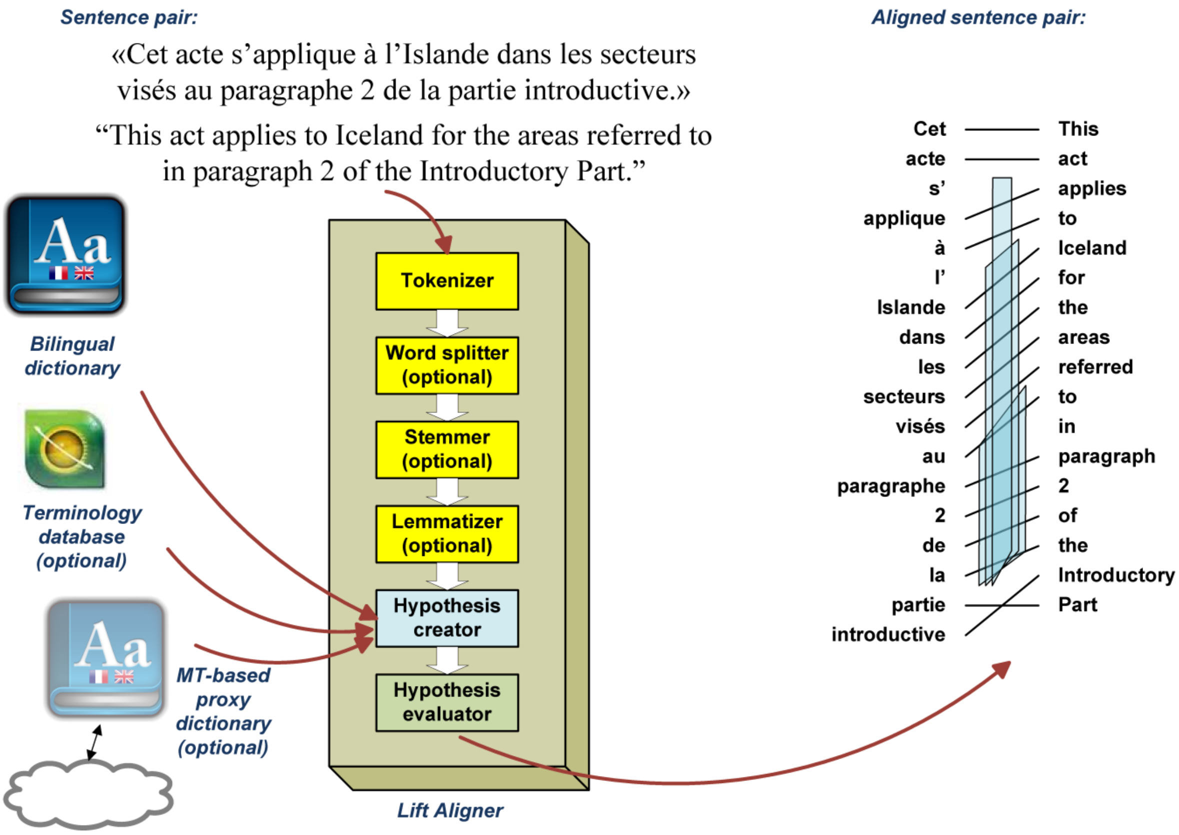
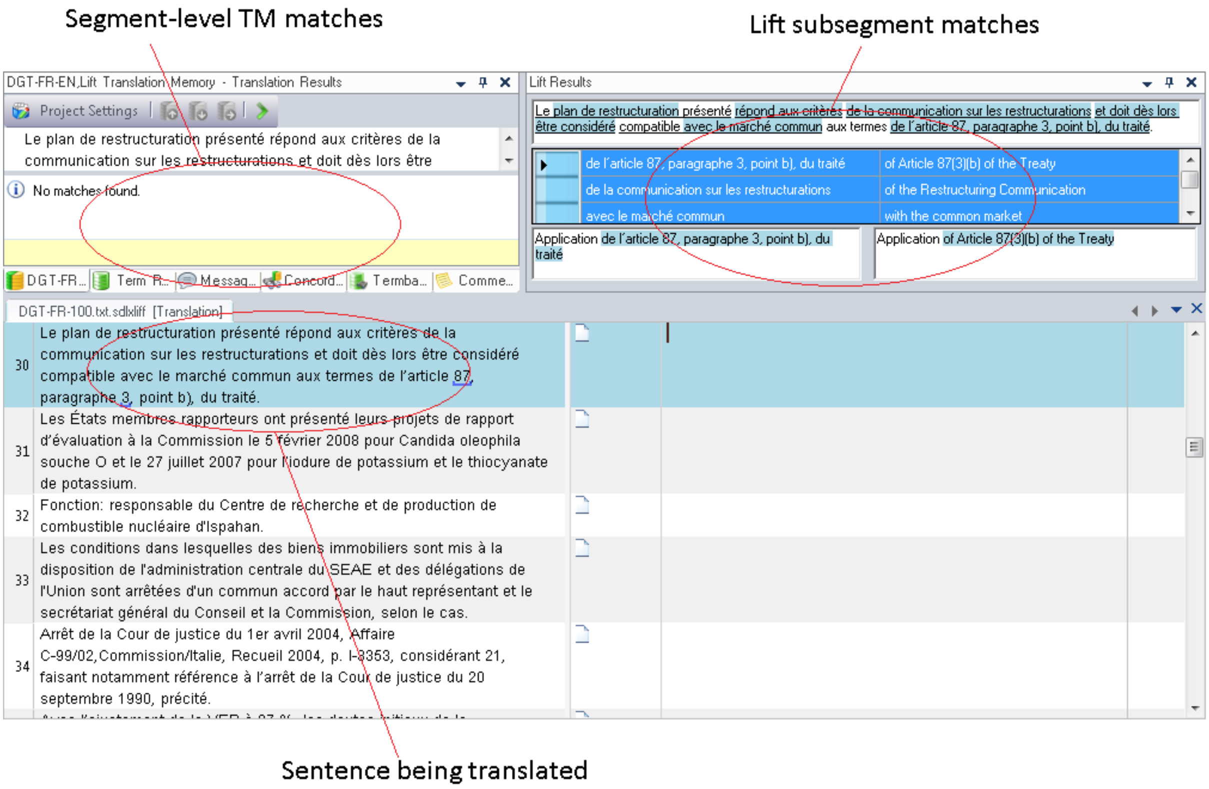


Figure 1: Lift alignment overview

5.2. SDL Trados Studio 2014 integration

Lift exposes an Application Programming Interface (API) allowing it to be integrated into CAT tools. An example integration has been developed for SDL Trados Studio 2014. The screen capture in Figure 2 gives an overview of how Lift's functionality is provided while translating.Figure 3 shows a larger view of the 'Lift Results' pane. The top section of the pane shows the sentence being translated, with underlining on all words that have been found in a subsegment match. This allows the translator to see at a glimpse for which parts of the sentence translations have been located, without having to scroll through the list of matches. The matches and their corresponding translation suggestions are shown in a list immediately below. The user can quickly insert a selected suggestion into the target text for the segment by double-clicking on the highlighted text it matches in the sentence being translated (right-clicking highlighted text produces a fly-out display of the translation suggestion, to save scrolling the list of matches if it is not visible). Double clicking an item in the list will also insert the translation. Alternatively, the user can begin typing one of the translation suggestions, then use auto-complete functionality for the rest of it, as shown in Figure 4. To review additional matches, the user can scroll down the list and click an item to see details and context, as shown in Figure 5.

To examine whether this functionality would better meet translators' expectations, the next section describes the results achieved using Lift for the same suite of tests used above to compare CAT tool performance.


Segment-level TM matches
Lift subsegment matches
Sentence being translated
DGT-FR-EN.Lift Translation Memory - Translation Results
Project Settings
Lift Results
DGT-FR… Term R… Messag… Concord… Termba… Comme…
DGT-FR-100.txt.sdlxliff [Translation]

Le plan de restructuration présenté répond aux critères de la communication sur les restructurations et doit dès lors être

No matches found.

Le plan de restructuration présenté répond aux critères de la communication sur les restructurations et doit dès lors être considéré compatible avec le marché commun aux termes de l'article 87, paragraphe 3, point b), du traité.

de l'article 87, paragraphe 3, point b), du traité	of Article 87(3)(b) of the Treaty
de la communication sur les restructurations	of the Restructuring Communication
avec le marché commun	with the common market

Application de l'article 87, paragraphe 3, point b), du traité

Application of Article 87(3)(b) of the Treaty

30 Le plan de restructuration présenté répond aux critères de la communication sur les restructurations et doit dès lors être considéré compatible avec le marché commun aux termes de l'article 87, paragraphe 3, point b), du traité.

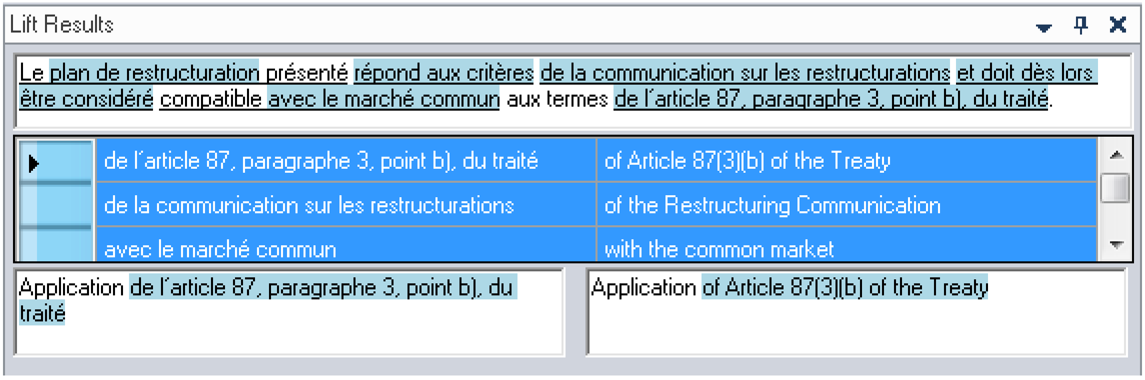
31 Les États membres rapporteurs ont présenté leurs projets de rapport d'évaluation à la Commission le 5 février 2008 pour Candida oleophila souche O et le 27 juillet 2007 pour l'iodure de potassium et le thiocyanate de potassium.

32 Fonction: responsable du Centre de recherche et de production de combustible nucléaire d'Ispahan.

33 Les conditions dans lesquelles des biens immobiliers sont mis à la disposition de l'administration centrale du SEAE et des délégations de l'Union sont arrêtées d'un commun accord par le haut représentant et le secrétariat général du Conseil et la Commission, selon le cas.

34 Arrêt de la Cour de justice du 1er avril 2004, Affaire C-99/02,Commission/Italie, Recueil 2004, p. I-3353, considérant 21, faisant notamment référence à l'arrêt de la Cour de justice du 20 septembre 1990, précité.

Figure 2: Trados integration overview



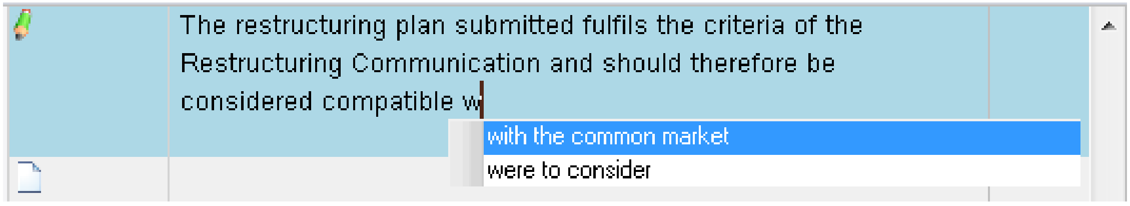
Le plan de restructuration présenté répond aux critères de la communication sur les restructurations et doit dès lors être considéré compatible avec le marché commun aux termes de l'article 87, paragraphe 3, point b), du traité.

de l'article 87, paragraphe 3, point b), du traité	of Article 87(3)(b) of the Treaty
de la communication sur les restructurations	of the Restructuring Communication
avec le marché commun	with the common market

Application de l'article 87, paragraphe 3, point b), du traité

Application of Article 87(3)(b) of the Treaty

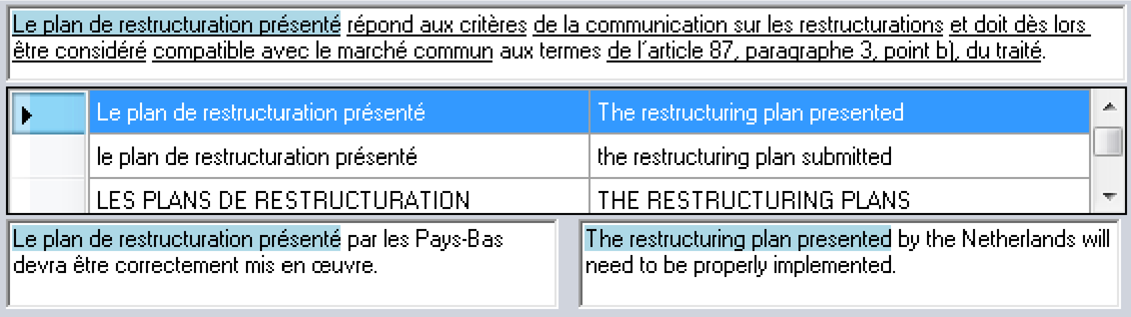
Figure 3: Lift results pane



The restructuring plan submitted fulfils the criteria of the Restructuring Communication and should therefore be considered compatible w

with the common market
were to consider

Figure 4: Auto-complete functionality



Le plan de restructuration présenté répond aux critères de la communication sur les restructurations et doit dès lors être considéré compatible avec le marché commun aux termes de l'article 87, paragraphe 3, point b), du traité.

Le plan de restructuration présenté	The restructuring plan presented
le plan de restructuration présenté	the restructuring plan submitted
LES PLANS DE RESTRUCTURATION	THE RESTRUCTURING PLANS

Le plan de restructuration présenté par les Pays-Bas devra être correctement mis en œuvre.

The restructuring plan presented by the Netherlands will need to be properly implemented.

Figure 5: Additional matches

5.3. Performance comparison

In order to compare Lift's subsegment recall performance with that of the CAT tools evaluated above, the same data and suite of tests were used with a Lift installation. The graphs in Table 5 show recall and precision for Lift, averaged over all test queries (eight English queries and eight French queries). For Lift, varying volumes of TM padding make no difference to subsegment recall results, so they were therefore all obtained using the same amount. Detailed results for the individual queries can be found at http://kftrans.co.uk/benchmarks.

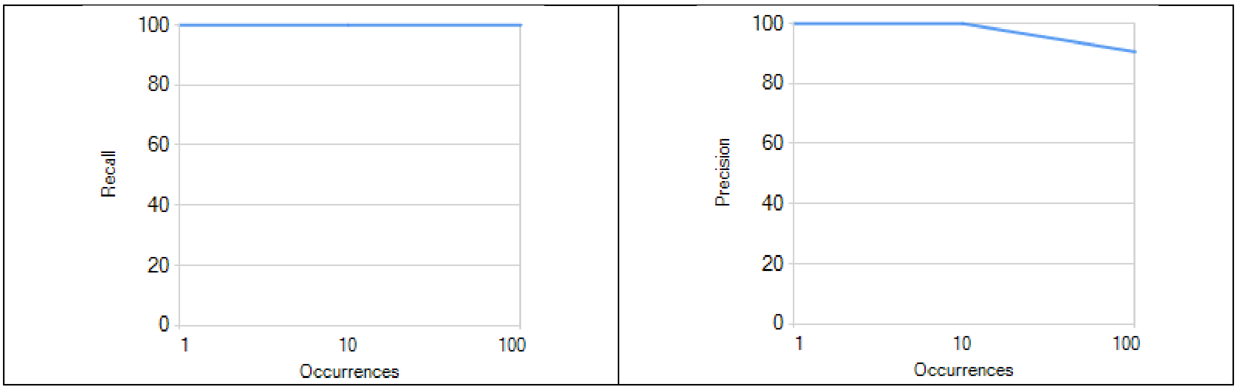


Table 5: Averaged performance statistics for Lift

The average results help summarise that with the fragments and TMs described above, Lift recalls their translations regardless of the number of occurrences, with generally very good precision, as well as neither decontextualising the translations nor exhibiting variation loss. The detailed results show that incorrect translation suggestions can be produced when TUs have not been correctly aligned by Lift. An example is discussed in the longer version of this paper mentioned at the end of the introduction.

6. Conclusion

The survey of translators' expectations of subsegment recall functionality found that around half expected functionality corresponding to DTA or BFE per the typology presented above (and I speculate more would do so if it were clear that the implementation would not be decontextualising). In particular, they expected recall to be available even for fragments occurring only once in the TM, and without any requirement for the TM to be large. A test suite was used to analyse subsegment recall capability for a range of CAT tools, showing that performance did not meet translators' expectations well. When used to analyse Lift's capability, results showed it met those expectations much better. Notwithstanding the small number of problematic alignment cases, these results seem very encouraging. Nevertheless, the suite of tests used involves a limited number of variables and carefully-controlled test data. A wider-ranging evaluation covering English, French, German, Spanish and Welsh, using much more extensive testing, is described in (Flanagan, forthcoming 2015a), where results indicate that performance is also good for those languages and with more comprehensive test cases. Nevertheless, even if controlled experiments suggest that new TM technology performs well when measured using whatever metrics, the success or failure of any attempt to develop and improve TM can ultimately only be judged by providing the developments to translators for real-world use, so that translators themselves can return a verdict.

References

BONET, J. (2013). No rage against the machine. Languages and Translation (6), EU Directorate-General for Translation, Brussels.

FLANAGAN, K. (2014). Bilingual phrase-to-phrase alignment for arbitrarily-small datasets. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas, Vancouver, BC, Canada.

FLANAGAN, K. (forthcoming 2015a). Methods for improving subsegment recall in Translation Memory. (PhD), Swansea University.

FLANAGAN, K. (forthcoming 2015b). Subsegment recall in Translation Memory – perceptions, expectations and reality. Journal of Specialised Translation (23).

PLANAS, E. (2005). SIMILIS - Second-generation translation memory software. Proceedings of the 27th International Conference on Translating and the Computer, London, United Kingdom.

STEINBERGER, R., A. EISELE, S. KLOCEK, S. PILOS AND P. SCHLÜTER (2013). DGT-TM: A freely available translation memory in 22 languages. Proceedings of the 8th international conference on Language Resources and Evaluation, Istanbul.

TREGASKIS, R. (2014). RE: Fluency: Mine terminology from TMs - never commits? [email]

ZETZSCHE, J. (2012). Translation technology comes full circle. Multilingual, 23(3), 50.

ZETZSCHE, J. (2014). Translation Technology - What's Missing and What Has Gone Wrong: eCPD [webinar]