Filling in the gaps: what we need from TM subsegment recall
Kevin Flanagan
Dept. of Languages, Translation and Communication, Swansea University
kevin@kftrans.co.uk

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
Translation Memory (TM) has been credited with creating a ‘revolution’ in the translation industry (Robinson, 2003: 31). While Machine Translation (MT) – in particular, Statistical Machine Translation (SMT) – is once again transforming how the industry works, and according to Pym, “expected to replace fully human translation in many spheres of activity” (2013: 1), TM still very much has a place, either when used alongside MT, or for projects where MT is not used. Widely-used CAT tools such as SDL Trados Studio and Wordfast Pro – products built around TM – provide MT system integration allowing translators to benefit from MT and TM translations, reflecting the assertion from Kanavos and Kartsaklis that “MT – when combined properly with Translation Memory (TM) technologies – is actually a very useful and productive tool for professional translation work” (2010: 1). TM results may be valued alongside MT not least because of a distinction noted by Teixeira, that “TM systems show translators the ‘provenance’ and the ‘quality’ of the translation suggestions coming from the memory, whereas MT systems display the ‘best translation suggestion possible’ without any indication of its origin or degree of confidence” (2011: 2). Waldhör describes the implementation of a ‘ recommender’ system intended to exploit such provenance distinctions (Waldhör, 2014). Provenance factors aside, TM can complement MT in providing immediate recall of new translation content in a project, without any SMT retraining requirement or risk of ‘data dilution’ (Safaba, [no date]), and can be used where there is too little (or no) relevant data with which to train an SMT engine.
Nevertheless, the segment-oriented nature of TM has seemed to restrict its usefulness, in ways to which MT 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 and proposed to translators. However, TM technology 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. In principle, TM subsegment recall – automatically finding phrases within segments that have been translated before, when they occur in a new text, and automatically identifying the corresponding translated phrase in the previously-translated segment – should recover all that content, with all the aforementioned TM benefits that complement MT. This functionality (sometimes referred to as ‘advanced leveraging’) is described by Zetzsche as “probably the biggest and the most important development in TM technology” (2014), but in practice, implementations in TM systems of subsegment recall 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” (2012: 51).

This paper focusses on TM subsegment recall in three ways. Results are presented 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 current-generation 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.

The discussion in this paper is expressed in terms of segment-based TM, that is, TM containing Translation Units (TUs), each of which contains 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 such segments, and retrieve the translation of that fragment, rather than of the whole segment in which 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, such that automatic identification of the translation of a fragment within a sentence is not possible. For both segment-level and CSB systems, translators can usually prompt a search for a specific fragment – referred to herein as a concordance search – and so find occurrences of fragment repetitions. Even so, discounting the time required to prompt such a search for all possible fragments (something that some CAT tools will attempt automatically, as discussed below), the results show only the larger segment or sentence within which the translation of the fragment is found, leaving the translator obliged to spend time and effort scanning through it. To aid discussion of these and other considerations, the next section defines a typology for describing subsegment recall implementations and characteristics.
2 Subsegment recall typology

The distinctions between approaches to subsegment recall found in different CAT tools are not all easily discerned by examining vendor documentation alone. To help identify them, the following list defines techniques and characteristics that can be used to describe subsegment recall implementations. This list is discussed at greater length in (Flanagan, forthcoming 2015b).

2.1 Use TM like a TDB (TM-TDB)

One of the most straightforward approaches to providing subsegment recall – described by Reinke as “[t]he simplest approach to subsegment matching” (Reinke, 2013: 33) – is to treat TUs in a TM like entries in a Terminology Database (TDB). TDBs are typically used to store domain-specific terms and their translations (Bowker, 2003: 51). When translating, the CAT tool checks the segment being translated to see if it contains any of the terms in the TDB – as opposed to the way the entire segment is compared to entire TM segments – and if so, the term translation is proposed to the translator. The technique – referred to herein as ‘TM-TDB’ – of treating TUs like TDB entries has the advantage of potentially finding matches and translations for fragments of a segment to be translated, using TM content, where a complete TU segment occurs as a fragment in the segment to be translated. For example, if an English document being translated into French has a section headed ‘Dynamic Purchasing System’, then once that segment is translated, the English-French TM contains a TU such as this:

- **EN**: Dynamic Purchasing System
- **FR**: Système d’acquisition dynamique

Later in the document, when translating the sentence “It is therefore necessary to define a completely electronic dynamic purchasing system for commonly-used purchases”, then even if there is no segment-level TM match for it, the TM-TDB technique would identify ‘dynamic purchasing system’ as a complete segment in the TM, so recall and propose the fragment translation, ‘Système d’acquisition dynamique’.

2.2 Automatic Concordance Search (ACS)

The description above of ‘concordance search’-type features highlighted how time-consuming it would be for a translator to search exhaustively in this way for all possible fragments matching TM content. Some CAT tools attempt to perform this exhaustive searching automatically, a technique referred to herein as ‘automated concordance search’ (ACS). If the CAT tool displays a matching source text fragment, the translator can examine the target text of the matching TU to find its translation. For example, suppose an English-French TM contains this TU:

- **EN**: A procuring entity may set up a system for commonly-used purchases that are generally available on the market.
- **FR**: L’entité adjudicatrice peut mettre en place un système pour des achats d’usage courant généralement disponibles sur le marché.

If ACS is available while translating the sentence, “It is therefore necessary to define a completely electronic dynamic purchasing system for commonly-used purchases”, then (subject to whatever settings) the fragment match “for commonly-used purchases” will be indicated, but without identifying that its translation was “pour des achats d’usage courant”; the translator must scan the
target segment of the matching TU to locate it. While matches like this could already assist translating, it is arguably more helpful for the CAT tool also to identify the translation of the matching fragment for the translator, saving time and effort. Some CAT tools attempt to do just that, categorised by the following definitions.

2.3 Dynamic TM Analysis (DTA)

Certain CAT tools attempt to identify the translation of a matching fragment using an on-the-fly TM content analysis, herein referred to as dynamic TM analysis (DTA). Like segment-level matching, this has the advantage of making immediate use of whatever is the current TM content, rather than requiring any separate resource to be created. The ‘DeepMiner’ feature of Déjà Vu X2 (and its successor, X3) is an example of this technique, as is the ‘Guess translation’ feature available when using concordance search in memoQ. Both those tools (it appears; commercial secrecy shrouds the details) use what can roughly be called a statistical approach to analysing TM content, whereas the corresponding feature in Similis applies linguistic methods (Planas, 2005: 5). While the immediacy of these techniques is desirable, other approaches require some TM content pre-processing before subsegment recall can be used.

2.4 Bilingual Fragment Extraction (BFE)

Although tools for bilingual terminology extraction attempting to “identify potential terms and their equivalents” (Bowker, 2003: 60) have existed for some time, recent commercial TM functionality is intended to extract more generalised fragments. As an example, the AutoSuggest ™ feature for SDL Trados Studio 2009 was described as adding

a new dimension to the power of translation memory. AutoSuggest maximizes the reuse of
previously translated content, by suggesting possible translations of words or phrases,
known as subsegments, from within the TM (TAUS, 2010: 9).

AutoSuggest also uses a statistical approach for extracting fragments and their corresponding translations, requiring a large TM for extraction to be performed, whereas Similis uses its linguistic approach to implement a corresponding feature with no minimum TM size requirement. However effective BFE implementations may be, they have the disadvantage of being ‘static’ data: if TM content is changed or new content added, subsegment recall matches and translation suggestions will not be adjusted to reflect those TM changes until the extraction step is performed again.

2.5 Decontextualisation

TM has been criticised as imposing a piecemeal, decontextualised approach to translation, as segment matches are recalled from the TM in isolation from the text in which they originally occurred (Bédard, 2000: 2; Christensen & Schjoldager, 2011: 8; Robinson, 2003: 32). Since subsegment recall by definition involves even smaller portions of text, those criticisms may apply yet more strongly, if the translation of a matching fragment is proposed without the translator being able to see the segment in which it originally occurred. That behaviour in CAT tools providing subsegment recall is referred to herein as ‘decontextualisation’.
2.6 Machine Recall (MR) vs. Assisted Recall (AR)

When a CAT tool searches for segment-level matches, no particular intervention by the translator is required – the tool compares the segment to be translated with segments in the TM, and displays any results. Some CAT tools take the same approach for subsegment matches, automatically comparing the segment to be translated with TM content or with the results of a BFE step, then displaying the results. Other CAT tools require the translator to begin translating the segment before displaying any subsegment matches, intercepting the translator’s keystrokes and proposing subsegment translations that begin with the same letter or letters typed. Automatic display of subsegment recall results is referred to herein as ‘Machine Recall’ (MR), while display that requires initial translator intervention is referred to as ‘Assisted Recall’ (AR).

While AR may help the translator who has already begun typing the required translation (enabling the fragment to be completed with fewer keystrokes), it clearly does not help inform a translator that (for instance) “ready-to-eat foods” has previously been translated as “denrées alimentaires prêtes à être consommées”. The translator may – perhaps should – already know this, but unless ‘d’ is typed to begin a word, that translation will not be displayed, and even when it is, may be hidden amongst other suggestions for words or fragments beginning with ‘d’ that the subsegment recall engine considers of possible relevance for the segment to be translated.

Another type of translator intervention may occur with CAT tools that provide an initial subsegment translation suggestion automatically, but also allow a list to be invoked with further translation suggestions for the same source text fragment. This perhaps helps a translator more than requiring the first letter(s) of the translation to be typed, but selecting the most relevant translation from the list is obviously not an automatic process. Suggestion provision of that kind is also referred to herein as AR.

2.7 Variation Loss (VL)

If a segment to be translated closely resembles several different TM segments during segment-level TM matching, CAT tools will typically present all the different matching segments, so that the translator can choose between them and edit accordingly. Similarly, if the segment exactly matches several TM segments with identical source segment text but different target segment translations, CAT tools will present all the matching segments, so that the translator can choose the most appropriate translation, or add another. A TM can legitimately contain different translations of the same source segment for several reasons, such as lexical ambiguity in the source language that is disambiguated in the target language, different wording to maximise cohesion with the occurrence context, etc. Seeing these different translations can be advantageous for the translator. This can be contrasted with typical use of a TDB where a certain term or expression is entered, along with a single translation for it (notwithstanding less typical use, where multiple TDB entries are added for the same term, or synonym fields are used, and so on). TDB proposals then show just this single translation, tending to ‘fix’ the translation of that term or expression, with segment-level matching unable to recall alternative translations that may occur in a TM. A subsegment recall implementation may be able to recall and display all these subsegment TM variations, or may be able to recall and display (say) only the most common translation. This latter behaviour is referred to herein as ‘Variation Loss’ (VL).
Other properties can be defined to characterise subsegment matching, such as whether matching is ‘fuzzy’ enough to deal with inflections, or whether discontinuous spans of words can be matched and recalled, etc., which are beyond the scope of this paper.

In the next section, this typology will be used to present the views of translators that participated in a subsegment recall survey.

3 Translators’ views

In order to gauge what functionality translators would ideally like from subsegment recall features, a controlled multiple-choice survey was conducted of translators with various levels of CAT tool experience. Responses were invited from four groups of translators: 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. The details of questions and responses can be viewed on a results web page at [http://kftrans.co.uk/benchmarks/Home/Survey](http://kftrans.co.uk/benchmarks/Home/Survey) and are discussed at greater length in (Flanagan, forthcoming 2015b).

The varying experience levels of the translators concerned were of interest from point of view of analysing whether experienced translators tended to give responses different from those of less experienced translators. If notable differences were found, this could mean a question needed examining more closely – perhaps responses from less experienced translators fail to take into account important factors, or perhaps more experienced translators were too conditioned by the ‘status quo’ of segment-level recall to appreciate where subsegment recall might be useful. Conversely, responses giving rise to broad consensus could be read as a strong result.

With these considerations in mind, the first survey question asked respondents to grade their familiarity with TM as one of five levels from ‘No familiarity’ to ‘Very familiar’. About two-thirds placed themselves in the upper two levels (nearly half responding ‘Very familiar’, while most of the others placed themselves in the middle category). This would seem to imply most respondents had at least a reasonable understanding of TM. To validate respondents’ self-descriptions and check basic response quality, a calibration question was used, showing the content of a very small TM and a sentence to translate that closely matched one in the TM, and asking what TM match result would be expected. The overwhelming majority correctly identified the expected match result, confirming their understanding of TM, and that they were taking care to provide suitable responses.

In summary, the remaining responses showed a broad consensus with regard to subsegment recall features. Most translators 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 so subsegment recall to be available 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 ACS, on the face of it, requires more translator time, since the TU has to be

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

4 CAT tool comparison

Table 4-1 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 wishing to evaluate tools before making a purchase decision. A tick indicates that the CAT tool supports the feature, and any term it uses to refer to the feature appears below the tick.

<table>
<thead>
<tr>
<th>CAT tool</th>
<th>TM-TDB</th>
<th>ACS</th>
<th>DTA</th>
<th>BFE</th>
<th>Min. TM size</th>
<th>Min. occurrences</th>
<th>Decontextualisation</th>
<th>Recall type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDL Trados Studio 2014</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓ 'AutoSuggest Creator'</td>
<td>10,000</td>
<td>-</td>
<td>Yes</td>
<td>AR</td>
</tr>
<tr>
<td>MetaTexis v3.17</td>
<td>✓ 'use TM as TDB'</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>memoQ 2013² R2</td>
<td>✓ 'LSC'</td>
<td>✓ 'LSC'</td>
<td>✓ 'Muse'</td>
<td>-</td>
<td>-</td>
<td>(ACS)2² (BFE)5</td>
<td>(ACS)No (BFE)Yes</td>
<td>(ACS)MR² (BFE)AR</td>
</tr>
<tr>
<td>MemSource v3.148</td>
<td>✓ 'Subsegment match'</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Déjà Vu X2² v8</td>
<td>✓ 'Assemble'</td>
<td>-</td>
<td>✓ 'DeepMiner'</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>Yes</td>
<td>MR³</td>
</tr>
<tr>
<td>Similis Freelance v2.16</td>
<td>-</td>
<td>-</td>
<td>✓ 'Glossary'</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>MR</td>
</tr>
</tbody>
</table>

Table 4-1: 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.
This gives a high-level view of how varied is the functionality in different CAT tools providing subsegment recall, which may not be obvious to translators reading similar-sounding vendor descriptions. DTA and BFE implementations merit further examination, since approaches and results vary much more than for (say) the comparatively straightforward TM-TDB feature, and because these are the implementations that attempt to identify the translation of a fragment within its containing segment. Furthermore, the expectations from translators as described above include the ability to recall translations of fragments even if they only occur once in a TM, and without a requirement for the TM to be large. The following section will examine CAT tool capability in those regards across a range of cases.

5 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. To that end, this section presents a suite of tests used to measure their performance in this regard. Measuring TM performance in general terms is a complex undertaking, which may encompass evaluating the core matching algorithms, or quantifying productivity gains, or assessing how easy the software is to use (Somers, 2003: 42). For the purposes of evaluating subsegment recall implementations, however, a much narrower and more controlled approach can be taken: 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.

5.1 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 extract the most frequent n-grams of order 1 to 6, from which a selection of n-grams to use as test case fragments was taken that satisfied the following criteria:

- To be of interest to the translator, no more than 50% of the words in a test fragment are ‘stop’ words (i.e. prepositions, articles, etc.)
- Test cases use fragments where corresponding pairs of English and French fragments have the same length in words, to allow the reverse case to be tested with the same data
- More than one test case is used for shorter fragments, so as to test different parts of speech.

The fragments pairs chosen are shown in Table 5-1, along with codes used to refer to them herein. The 40,000 TUs were then processed to select 10,000 TUs that contained none of the fragments shown in either language, for use as ‘padding’ to create test TMs of different sizes. For each fragment pair, the 40,000 TUs were further processed to extract up to 100 fragment-bearing TUs – that is, containing the fragment pair – meeting the following criteria (where fewer than 100 were...
found, the difference was made up with randomly-chosen copies of the available pairs, prefixed with a unique alphanumeric key):

- For fragment occurrences in TMs to test, fragments constitute less than 50% of the words in the segment
- The minimum segment length for TUs in test TMs is four words.

For each fragment pair, two subsets of the 100 fragment-bearing TUs were created by random selection, of size 10 and 1. Two subsets of the 10,000 ‘padding’ TUs were created by random selection, of size 1,000 and 100.

<table>
<thead>
<tr>
<th>Code</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>règlement</td>
<td>Regulation</td>
</tr>
<tr>
<td>1a</td>
<td>établi</td>
<td>established</td>
</tr>
<tr>
<td>2</td>
<td>conclut que</td>
<td>concludes that</td>
</tr>
<tr>
<td>2a</td>
<td>État membre</td>
<td>Member State</td>
</tr>
<tr>
<td>3</td>
<td>modifiée comme suit</td>
<td>amended as follows</td>
</tr>
<tr>
<td>3a</td>
<td>les autorités polonaises</td>
<td>the Polish authorities</td>
</tr>
<tr>
<td>4</td>
<td>intégrée dans l'accord</td>
<td>incorporated into the Agreement</td>
</tr>
<tr>
<td>6</td>
<td>Journal officiel de l'Union européenne</td>
<td>Official Journal of the European Union</td>
</tr>
</tbody>
</table>

Table 5-1: Test fragment pairs

To simulate translation of a source text that includes a test fragment also found in a test TM, example sentences – hereafter, ‘queries’ – were created by adapting TUs in the test data containing the fragment pairs. 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. To do so, an edit distance percentage value was computed, which gives a measure of text string similarity by counting the minimum number of edit operations required to transform one string into another, and is comparable to the fuzzy match values assigned by CAT tools. The computation used an implementation of the Levenshtein distance algorithm (Levenshtein, 1966), where 100% corresponded to two identical strings. Query sentences were adjusted to ensure none matched any padding TU with a value higher than 60%.

For each query TU, a TM was created for all combinations of padding-TU set size and fragment-bearing TU set size; 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.

Test data and queries can be downloaded from [http://kftrans.co.uk/benchmarks/benchmarkdata.zip](http://kftrans.co.uk/benchmarks/benchmarkdata.zip). The queries used are also shown at [http://kftrans.co.uk/benchmarks/Home/Queries](http://kftrans.co.uk/benchmarks/Home/Queries). Further discussion of the motivation behind and preparation of the test data and queries can be found in (Flanagan, forthcoming 2015b).

5.2 Scoring

For the purposes of the testing described here, formulae for precision and recall were defined as follows. Given a test fragment whose corresponding translated fragment is expressed as a set of words $F_t$ (all words being unique in the fragments shown above) and a subsegment match
translation suggestion expressed as a set of unique words \( S \), the precision \( P_s \) of that proposal expressed as a percentage is defined as follows:

\[
P_s = \frac{|S \cap F_t|}{|S|} \times 100
\]

In other words, the precision value for the suggestion is the percentage of unique words in the suggestion that can be found in the expected fragment translation, so expressing how accurately the suggestion recalls the expected fragment translation. This precision definition ignores the word ordering of suggestion \( S \) compared with expected translation words \( F_t \), as discussed with the results below, and suggestions with multiple repetitions of a word found in \( F_t \) do not have lower precision.

Recall \( R_s \) expressed as a percentage is defined as follows:

\[
R_s = \frac{|S \cap F_t|}{|F_t|} \times 100
\]

In other words, the recall value for the suggestion is the percentage of unique words in the expected fragment translation that can be found in the suggestion, so expressing how much of the expected fragment translation has been recalled. In a given test case, where several translation suggestions may be proposed for the test fragment, overall recall \( R \) and precision \( P \) is defined as the average recall and precision of the suggestions.

Certain CAT tools’ subsegment recall implementations present a difficulty in evaluating precision, since when providing subsegment translation suggestions for a query sentence, there is no way to determine which part of the source text the suggestion words have been presented in response to, and so to determine which are intended to be translation words for the test fragment. These cases are shown in the results as ‘Precision unavailable’. More generally, the exact procedures for recording results for a given CAT tool are conditioned by the very different interfaces and approaches taken by each tool. The specifics of how the formulae above were applied to each tool can be found in (Flanagan, forthcoming 2015b). Test results show how each tool performed under varying conditions (TM size, number of fragment occurrences, fragment length). The different results for a given tool are directly comparable, and give an indication of how its performance is affected by the variables used. 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, especially when the differences are stark. With regard to AR versus MR, results are based on AR suggestions only if the tool does not offer any MR implementation.

5.3 Results

The graphs in Table 5-2 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 the number of fragment-bearing TUs in the TM, and the different lines show the number of padding TUs in the TM. Note: the technical descriptions of Similis (Planas, 2005: 5) suggest – and experimentation confirms – that varying volumes of TM padding make no difference to subsegment recall results, since it uses linguistic rather than statistical measures. Results for Similis were therefore 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.

As the averaged results help summarise, with the fragments and TMs described above, 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 fragment frequency, regardless of padding volume, while for fragment 3a, recall and precision drop sharply as fragment 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, the number of occurrences 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 of how many occurrences are in the TM. With SDL Trados Studio, performance overall is quite consistent, with 100% recall usually achieved when the TM contains ten fragment occurrences. Alongside the large TM requirement, and the fact that the implementation is AR rather than MR, it should be noted that suggestion display is sensitive to case and to diacritics. For instance, when testing using French fragment 1, expecting a suggestion for ‘règlement’, the recalled translation ‘Regulation’ is only displayed if upper-case ‘R’ is typed.

5.4 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.
Table 5-2: Averaged performance statistics by CAT tool
6 Lift

6.1 Overview

Lift is a TM system implementing DTA subsegment recall based on fine-grained alignment of segment pairs, that is, alignment between spans of words or individual words in a segment pair, hereafter referred to as subsegment alignment. In order to enable subsegment alignment without the need for a large TM over which to apply statistical methods, nor by using syntactic information in ways that may inhibit recall for certain grammatical categories, 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 for the segment pair. The alignment algorithm is described in detail and compared with other approaches in (Flanagan, 2014). Figure 6-1 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 6-2 shows a high-level view for an example query with a fragment match against the aligned sentence pair shown in Figure 6-1.

Figure 6-2: Lift query overview

6.2 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 6-3 gives an overview of how Lift’s functionality is provided while translating. The ‘Lift Results’ pane can be resized, docked in different positions, made to float over the main window, and so on, according to user preference. Figure 6-4 shows a larger view of the 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 6-5. To review additional matches, the user can scroll down the list and click an item to see details and context, as shown in Figure 6-6.

Figure 6-3: Trados integration overview

Figure 6-4: Lift results pane

Figure 6-5: Auto-complete functionality
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.

### 6.3 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 configured with the resources shown in Table 6-1. (Note: the Stanford and Berkeley parsers are only used for minimal processing, that is, tokenization with their per-language pattern-matching rules.) French-English and English-French dictionaries used were all from the XML Dictionary Exchange Format project.\(^5\) (The files exhibited some corruption and omission, repaired manually.)

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tokenizer</strong></td>
<td>Stanford parser (Klein &amp; Manning, 2003)</td>
<td>Berkeley parser (Petrov, Barrett, Thibaux, &amp; Klein, 2006)</td>
</tr>
<tr>
<td><strong>Stop list</strong></td>
<td>Snowball(^6)</td>
<td>Snowball</td>
</tr>
<tr>
<td><strong>Stemmer</strong></td>
<td>Snowball</td>
<td>Snowball</td>
</tr>
<tr>
<td><strong>Lemmatizer</strong></td>
<td>Morfologik 1.6(^7)</td>
<td>purpose-built; uses data from the Morphalou project (Romary, Salmon-Alt, &amp; Francopoulo, 2004)</td>
</tr>
</tbody>
</table>

Table 6-1: Test resources

The graphs in Table 6-2 show recall and precision for Lift, averaged over all test queries (eight English queries and eight French queries), where the X-axis shows the number of fragment-bearing TUs in the TM. What was established for Similis is a ‘known’ for Lift – that varying volumes of TM padding make no difference to subsegment recall results. Results for Lift were therefore all obtained using the same amount of TM padding. Detailed results for the individual queries can be found at [http://kftrans.co.uk/benchmarks](http://kftrans.co.uk/benchmarks).

---

\(^5\) [http://sourceforge.net/projects/xdxf/](http://sourceforge.net/projects/xdxf/)

\(^6\) The Snowball-based family of stemmers at [http://snowball.tartarus.org/](http://snowball.tartarus.org/) includes stop lists for the languages concerned

Table 6-2: 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. For example, one TU in the test data contains the following sentence pair:

**EN**: A32. if it is not a printing works, its manufacturing site is located in a Member State or in an EFTA Member State.

**FR**: A32. s'il ne s'agit pas d'une imprimerie, son site de fabrication est situé dans un État membre ou dans un pays membre de l'AELE.

This pair contains the English fragment ‘Member State’ twice, and when used with the corresponding test query for English to French, produces two suggestions. Using the language resources described above, part of the sentence is aligned as shown in Figure 6-7. The tokens ‘EFTA’ and ‘AELE’ were not found in the standard bilingual dictionaries used, causing the aligner to make incorrect deductions and for the second occurrence of ‘Member State’ to be incorrectly aligned. As an experiment separate from the results above, when a terminology database is added that includes the term ‘EFTA’/’AELE’, the alignment changes accordingly, as shown in Figure 6-8. This causes “EFTA Member” to be correctly aligned with “membre de l’AELE”. As the dictionaries used do not include ‘pays’ as a translation of ‘State’, nor vice versa, no alignment seed is created between those words, and the aligner does not deduce the alignment between ‘pays’ and the second occurrence of ‘State’, since it is not unambiguous due to unaligned words earlier in the sentence. With the improved alignment, the second occurrence of ‘Member State’ does not recall a translation suggestion for the English test query, as the alignment coverage of the second ‘Member State’ occurrence is considered insufficient. Bearing in mind that Lift is intended to show variant translations per the translator expectations described above, it would be desirable for the translation ‘pays membre’ to be recalled, even if it may be a less useful translation suggestion. For cases such as these, that could be achieved with more optimistic alignment heuristics or additional language resources. These issues and the effects of domain-specific resources such as TDBs are discussed further in (Flanagan, forthcoming 2015a).
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. An analysis of subsegment recall functionality in a range of CAT tools found that implementations available vary widely, not necessarily providing DTA or BFE recall. A suite of tests was used to evaluate the performance of DTA and BFE recall in four CAT tools, across a range of cases, showing that performance did not meet translators’ expectations well.

The approach taken by the CAT tool best meeting those expectations – although not consistently – indicated that a TM system implementing subsegment recall based on more robust fine-grained alignment could provide better results. A TM system designed on those principles, Lift, was tested using the same suite, and shown to perform significantly better. Notwithstanding the small number

---

**7 Conclusions**

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. An analysis of subsegment recall functionality in a range of CAT tools found that implementations available vary widely, not necessarily providing DTA or BFE recall. A suite of tests was used to evaluate the performance of DTA and BFE recall in four CAT tools, across a range of cases, showing that performance did not meet translators’ expectations well.

The approach taken by the CAT tool best meeting those expectations – although not consistently – indicated that a TM system implementing subsegment recall based on more robust fine-grained alignment could provide better results. A TM system designed on those principles, Lift, was tested using the same suite, and shown to perform significantly better. Notwithstanding the small number
of problematic alignment cases – one of which was described above – these results seem very encouraging from the point of view of providing a subsegment recall implementation that improves on existing TM systems and better meets translator expectations. 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.

How a translator would use this functionality is subject to a number of variables, not least the translator’s own preferences, and the amount of subsegment leverage that is available beyond any segment-level leverage for the text and TM’s in question. The wider-ranging tests described in (Flanagan, forthcoming 2015a) measure performance for a range of minimum-length-of-match values, from 2 to 6, with precision tending to increase in line with that value, while recall decreases. It is interesting to consider the results at the mid-point value of 4. Here, the potential coverage shown for the test sentences – each of which has no fuzzy match with TM content equal to or over 70% – is in the 40-55% range for all languages concerned, with translations recalled and suggested for 70-85% of that coverage, with 70-85% precision. By no means all of those suggestions would necessarily be used by a translator, but by providing recall in those cases, given those figures, the potential for increasing translators’ productivity and translation consistency appears significant, particular with an implementation that meets translators’ expectations in terms of not requiring a large TM, nor a minimum number of fragment occurrences, while preserving translation variation, and also making the context of translation suggestions available. Although effective subsegment recall for TM by no means serves the same purpose as using SMT and TM together (which can be very complementary technologies, as noted above), it constitutes an alternative means to retrieve “phrases […] buried in thousands of sentences” (Bonet, 2013: 5) where SMT does not perform well. In doing so, it may help avoid less desirable implementations of post-editing-machine-translation workflow being put in place as a matter of course, if subsegment recall provides better leverage of limited domain-specific content and thereby better translation suggestions, or if translators simply prefer the cognitive load of exploiting subsegment leverage over that of post-editing.

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. Unlike the more abstract and well-defined methods for comparing translation suggestions to reference translations, such as the suite of tests used above, this is an activity that has received less attention; Watkins notes that “no-one appears to have attempted to measure efficiency in the context of time saved in the context of TM recall evaluation, which, considering that the time which could potentially be saved is one of the major selling points of TM, is rather puzzling” (2013: 199), before providing an extensive analysis of the very many factors that may come into play for the purposes of real-world trials. In that respect, trials to evaluate Lift would have to be designed very carefully, but should produce valuable information.
References


TAUS. (2010). *How to Increase Your Leveraging.*