JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Flanagan, K. (2015). Subsegment recall in translation memory – perceptions, expectation and reality. *The Journal of Specialised Translation, 23*, 64-88.
https://doi.org/10.26034/cm.jostrans.2015.340

Subsegment recall in Translation Memory — perceptions, expectations and reality

Kevin Flanagan, Swansea University

ABSTRACT

Some CAT tool vendors describe their software as providing 'subsegment' matching, sometimes called 'advanced leveraging.' The descriptions can seem quite similar, but different tools in fact provide very different subsegment matching techniques and performance, with no typology available to distinguish implementations. This article first describes subsegment matching, then proposes just such a typology. The results of the first survey of translators to gauge their interpretations of vendor descriptions and their expectations of the software are presented and analysed in terms of the typology. A matrix of all CAT tools providing subsegment matching and available for trial (or free) use by translators is used to cross-reference their features with the typology. Finally, for the four CAT tools that provide the more advanced functionality, the first detailed analysis of their subsegment matching performance is presented, using an extensive series of tests. The overall findings show that interesting subsegment matching functionality is available, but that performance could be improved to meet translator expectations better. This in turn helps highlight the functionality to translators who may benefit from it, and the differences so as to help them make informed CAT tool purchase decisions.

KEYWORDS

TM, translation memory, CAT tools, subsegment recall, advanced leveraging.

1 Introduction

Although Machine Translation (MT) has improved in recent years, with post-editing of MT results becoming more common, Translation Memory (TM) remains a key feature of Computer-Assisted Translation (CAT) tools. Some CAT tool vendors describe their TM system as including 'subsegment' matching, sometimes called 'advanced leveraging.' Since TM segment-level matching tends to give more-or-less the same results, regardless of CAT tool, translators may believe that subsegment matching exhibits the same uniformity, if they have time to wonder at all. But in fact, CAT tools providing subsegment matching do so in very different ways with very different results. Since subsegment matching is meant to address a long-standing TM weakness — though not as well as might be assumed — these differences are potentially important for translators hoping to realise speed or consistency benefits from it, or needing information for a purchasing decision. So, just what is subsegment matching for?

2 Segment-level matching versus subsegment matching

Using TM involves segmenting texts into sentences or other readily-identifiable items like headings — either during translation or using alignment tools with existing source and target text pairs — which are

then stored in a TM with their translations, as pairs known as Translation Units (TUs). New texts are segmented for comparison against those stored, so that matching segments and their translations can be retrieved. Segments of new text may match TM segments exactly, or be 'fuzzy' matches, where they are not identical but similar enough for the translation to be worth recalling for the translator, and a similarity threshold can be applied as a filter. While TM technology has been credited with bringing "a revolution in the translation profession" (Robinson 2003:31), this segment-level matching process can miss useful TM content. If only a fragment of a new text segment — say, a six-word clause in a longer sentence — matches something in the TM, no translation for that part is suggested. This is because the segment-comparison algorithms consider those segments overall not to be very similar, and "users are generally advised not to set the similarity coefficient too low, to avoid being swamped by dissimilar and irrelevant examples" (Macklovitch and Russell 2000: 141).

Does it matter if TM matching misses results for fragments like those? Using TM is meant to help avoid translating the same segment twice, thereby saving time and making translations more consistent. Segments generally correspond to sentences, so those benefits are only realised for identical or similar sentences, which may be relatively rare, while fragments may recur much more often, for which Grönroos and Becks assert that "there is in all text types much more repetition than on the sentence level" (2005: 2). When translating the English sentence "Ensure participation rate data is recorded correctly and of the highest quality", it may be very helpful for the CAT tool to detect that "participation rate data" already exists in the TM, and recall its translation. (Conversely, detecting that "and of the" already exists in the TM may be less interesting to the translator.) Most CAT tools provide a facility known as 'concordance search' or similar that allows the translator to search for a particular text fragment in the TM. So, a translator suspecting that "participation rate data" already exists in the TM can prompt such a search to find any TU containing that fragment. This arguably does not satisfactorily provide the aforementioned TM benefits, however, since it takes up translator time (especially if performed exhaustively so as to maximise re-use and consistency), and results recall the whole segment in which the translation of "participation rate data" occurs, requiring the translator to scan the target sentence to locate the corresponding fragment.

Whether or not this is acceptable may depend on individual translators' preferences, but by 2007, enough CAT tool vendors had introduced subsegment matching features for the Translation Automation User Society (TAUS) to produce an *Advanced Leveraging* report, described by its author as presenting a "new generation of translation tools that builds on older principles of Example Based and Statistical MT and resolves deficiencies in classic TM [using] statistical analysis and linguistic

intelligence tools" (Kuhns 2007). How are these features described by CAT tool vendors?

3 CAT tool vendor descriptions

Vendor descriptions of subsegment recall features can seem to describe largely the same functionality. These are some examples:

> it doesn't just use the individual terms and sentences in your databases, but also carries out sophisticated cross-analyses of those databases on the fly to "mine" translations of the building block words and phrase segments embedded in them. [This brings] you enhanced productivity, even for texts with few or no database segment matches. [It] works with you as you translate, automatically proposing a series of terms, phrases and sentences that are mined from your databases and interactively assembled (Atril 2013).

> [User query:] Is subsegment matching available in [the system]? For instance a part of a segment has already been translated in another segment. Will it get suggested to the translator? [Vendor response:] Yes, it will (MemSource 2014).

> [The feature] monitors what you are typing and, after you have typed the first few characters of a word it presents you with a list of suggested words or phrases in context and in your target language. [Data for the feature] can be created by extracting words and phrases for your translation memory (SDL 2014).

> [The system] includes a linguistic analysis engine that uses 'chunking' technology to split sentences into intelligent terminological groups, so as to generate domain-specific glossaries automatically (Lingua et Machina 2014).

These descriptions evidently concern features intended to address the issue identified above — recalling fragments of segments from the TM — but the differences between them are not obvious, and neither are the circumstances required for them to recall fragments. A typology to be used for categorising these features will enable the differences and requirements to be discussed.

4 Subsegment recall typology

A later version of the above-mentioned TAUS report described several subsegment recall implementations in terms of certain capabilities (TAUS 2010: 18). However, for the purposes of this paper, a more fine-grained typology will help provide a fuller picture. To that end, the following list defines techniques and characteristics that can be used to describe subsegment recall implementations.

4.1 Use TM like a TDB (TM-TDB)

One of the simplest approaches to providing subsegment recall 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. While TM content can be created during translation with a CAT tool or using alignment tools, TDB content is usually more labour-intensive, with terms being chosen and entered manually, or generated by an extraction tool requiring considerable manual intervention.

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, but without the work required to create TDB content. (Of course, matches will only be found if the TM contains suitable TUs, so this by no means makes TDB content irrelevant.) For example, suppose you have an English document you are translating into French, with a section headed 'Dynamic Purchasing System'. Once that segment is translated, the English-French TM contains a TU as shown in Figure 1.

Figure 1. TM-TDB example TU

Later in the document, the sentence "It is therefore necessary to define a completely electronic dynamic purchasing system for commonly-used purchases" is found. 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, then recall and propose the fragment translation, 'Système d'acquisition dynamique,' as shown in Figure 2.

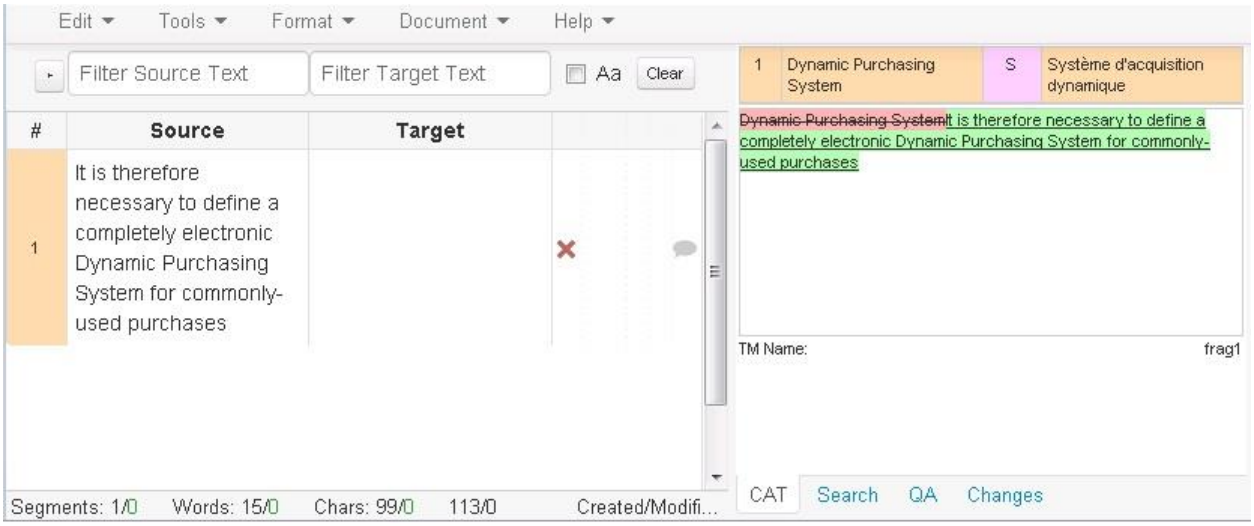

Figure 2. MemSource showing a TM-TDB match

4.2 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). As noted above, a sentence to be translated may have fragments that match TM content, but which are of no real interest to the translator. ACS implementations therefore try to be selective about fragments sought, such as by applying a minimum fragment length. 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 the English-French TM considered above contains the TU shown in Figure 3.

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

Figure 3. ACS example TU

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 to locate it, as shown in Figure 4.

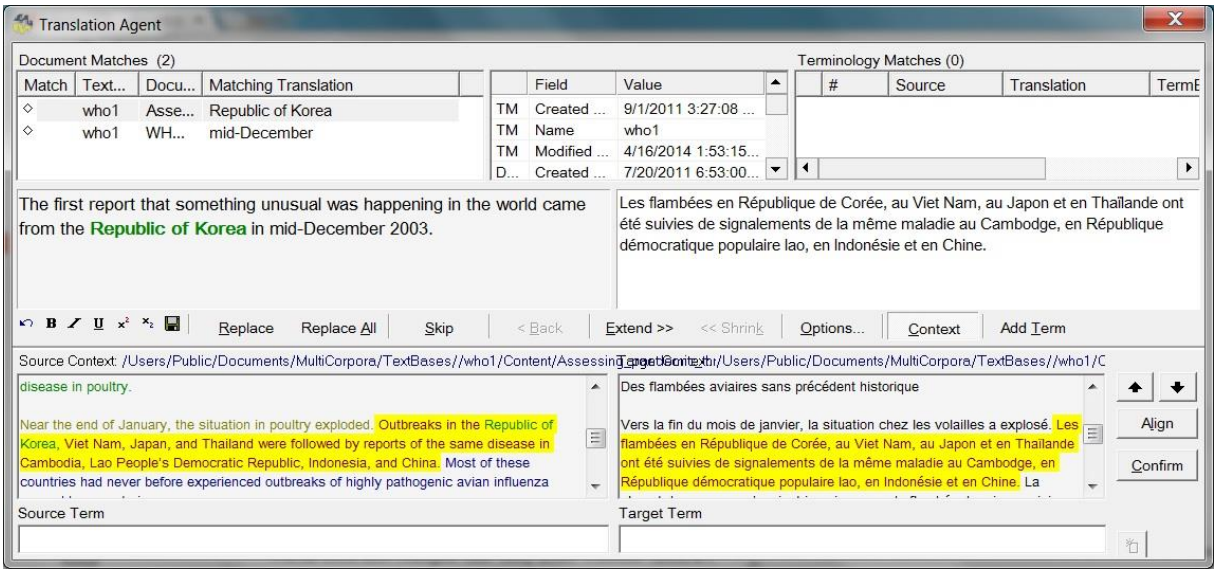

Figure 4. MultiTerm Prism ACS matching (middle pane showing sentence to translate, lower pane showing match and context)

While matches like this could already assist translating, it is 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, per the following definitions.

4.3 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, while 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.

4.4 Bilingual fragment extraction (BFE)

While tools for bilingual terminology extraction already existed that attempted to "identify potential terms and their equivalents" (Bowker 2003: 60), the 'advanced leveraging' wave included features intended to extract more generalised fragments. 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).

AutoSuggest also uses a statistical approach for extracting fragments and their corresponding translations, requiring a large TM for extraction to be performed, while 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.

4.5 Machine Recall (MR) versus 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.

4.6 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 (Robinson 2003: 32, Christensen and Schjoldager: 127). 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.'

4.7 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 using a TDB

where a certain term or expression is entered, along with single translation for it. 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. (It is possible to add multiple TDB entries for the same term, so as to see multiple TDB proposals, though typically TDBs are compiled specifically to mandate a single, consistent translation of a given term.) 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, whether discontinuous spans of words can be matched and recalled, etc. — but these are beyond the scope of this paper.

The definitions above can help describe subsegment matching implementations in CAT tools. Descriptions of that kind could be used to clarify the different ways in which subsegment matching is provided, and so identify which best meet translator needs. Those goals raise two questions, however: how do translators currently understand subsegment matching, and how would they like it to work?

5 Subsegment matching survey

In order to gauge how translators interpret vendor descriptions of subsegment matching features, and to determine how translators would ideally like them to work, a controlled multiple-choice survey of translators with various levels of CAT tool experience was used. Responses were invited from four groups of translators: the Western Regional Group of the Institute of Translation and Interpreting (ITI) (http://www.itiwrg.org.uk); translators registered with Wolfestone (http://www.wolfestone.co.uk), a successful language services agency; the ITI's French Network (http://www.iti-frenchnetwork.co.uk); and students on MA in Translation programmes at Swansea University (http://www.swansea.ac.uk/translation). In all, 91 responses were received, 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, while a summary follows here.

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. For the purposes of further analysis, having established how respondents were distributed across TM experience levels, two categories of respondent were used, 'experienced' (the two-thirds of respondents in the upper two levels of familiarity) and 'less experienced' (all other respondents).

The remaining questions were split into two sections. The first (labelled 'Vendor description interpretation – 1' on the results web page) examined general interpretation of subsegment matching functionality. Given vendor descriptions of tools providing subsegment matching, an example case to consider, and a selection of possible results that might be expected from those tools (question 1A on the results web page), 49% of respondents expected the result corresponding to DTA (or BFE once extraction has been performed), while 29% chose the result corresponding to ACS. When asked which result they would actually want from such a tool (question 1B), fewer respondents chose DTA (particularly those with more TM familiarity), preferring ACS. As 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 so? I speculate that this is because experienced translators are more aware of the dangers of decontextualisation, and the DTA/BFE 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 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. The final question (1C) in the section adjusted the example case by addition of a TU in the TM that would allow TM-TDB to provide a subsegment match. 73% of respondents then selected the expected result option corresponding to TM-TDB.

The second section (labelled 'Vendor description interpretation – 2' on the results web page) examined interpretations of BFE techniques. Given vendor descriptions of tools providing subsegment matching via a BFE step, an example case to consider, and a selection of possible results that might be expected from those tools (question 2A), respondents overwhelmingly expected the result corresponding to BFE without VL, that is, subsegment recall via an extraction step that retains and proposes variant translations of a fragment. When asked if it was acceptable for a minimum TM size to be required before extraction could occur (question 2B), 51% of respondents chose 'No acceptable minimum' as the response, with another 21% viewing a minimum TM size of 100 TUs as acceptable, and 13% viewing minimum TM size of 1000 TUs as acceptable. A further question (2C) asked if it was acceptable to require that a fragment occur a minimum number of times in the TM before it could be recalled. 67% chose just one occurrence (the lowest figure) as the point at which subsegment recall should be available for a fragment, with a further 24% choosing five occurrences (the next-lowest figure).

Taken as a whole, these responses begin to provide a picture of how translators interpret vendor descriptions of subsegment recall features, and what functionality they would prefer from those features. In brief, based on responses to questions shown on the results web page and the consensus found across translator experience levels, most translators expect TM-TDB to be available (question 1C); there is a fairly equal split between wanting DTA/BFE and wanting ACS (question 1B – though I speculate more might want DTA/BFE if 'including context' were specified); VL is not desirable (question 2A); requiring a TM to be large is not desirable (question 2B), and subsegment recall should be available for fragments occurring only once in the TM (question 2C). How does this compare with functionality in available CAT tools?

6 CAT tool comparison

The following table compares the subsegment recall functionality for all CAT tools that provide such a feature and are available at time of writing for trial (or free) use by translators. A tick indicates that the CAT tool supports the feature, and any term it uses to refer to the feature appears below the tick.

	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	No	(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 1. CAT tool comparison

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 do not.
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, pers. comm.). Across Language Server provides BFE functionality, but unlike Personal Edition there is no trial or free version available.

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. Four of the tools shown above provide functionality corresponding to the DTA/BFE survey options, while just one offers ACS functionality. I speculated above that respondents expressing a preference for ACS in the survey may have chosen DTA/BFE if 'including context' had been specified. Just one of the tools above providing DTA/BFE does so without decontextualising the recalled translations. Whether or not that would make it more desirable for those respondents expressing ACS preference, translators seeking DTA/BFE functionality may wish to know about the different strengths and weaknesses of the four very different implementations shown above, as

results vary much more than for the comparatively straightforward TM-TDB feature. These four implementations will be examined in the next section.

7 Detailed performance analysis of DTA/BFE CAT tools

The results from the survey described above seem to indicate that at least some translators would like subsegment recall functionality corresponding to the DTA or BFE definitions, depending on the requirements in terms of TM size and fragment frequency. But which tools best meet translator expectations? A comprehensive test suite for evaluating CAT tool performance in this respect would need to consider many independent variables, including:

- The language pair(s) concerned (some may correspond more closely at subsegment level than others)
- The morphology of each language (e.g. a single word form in English may correspond to many translated forms in an inflected language)
- The length in words (and perhaps in letters) of the source fragments to match
- The length in words (and perhaps in letters) of the target fragment translations to recall
- The length of the source fragment relative to the segment in which it occurs.
- The length of the target fragment relative to the segment translation.
- (If statistical implementation) The number of occurrences of the fragment.
- The TM size.
- (If statistical implementation) The consistency with which the fragment is translated.
- The tokenisation rules used by the CAT tool (e.g. whether "l'État" is one word or two)
- (If linguistic implementation) The part(s) of speech represented by the fragment and its translation.

Even testing with only these factors, and even with only a few values for each, the combinations multiply up to a very large number of tests. Nevertheless, a more restricted set of tests can at least serve to give some indication of the strengths and weaknesses of CAT tool DTA/BFE implementations. Such a set of tests is described below.

7.1 Testing parameters

In order to generate some indicative test results for DTA/BFE implementations, a test suite was devised based on the following principles:

- Only French-English and English-French examples are tested
- Tests use a range of fragment word lengths: 1, 2, 3, 4 and 6 words

- Test fragments can be relatively everyday language; the goal is to test the CAT tools, not compile a domain-specific glossary
- 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
- Fragment recall is tested against TMs using a range of fragment frequencies: 1, 10 and 100 occurrences
- Fragment recall is tested against TMs with different numbers of other TUs that contain neither the fragment sought nor its translation: 100, 1,000 and 10,000 other TUs
- For both fragment occurrences in TMs to test, and test sentences to query against them, fragments constitute less than 50% of the words in the segment
- The minimum segment length for TUs in test TMs is four words.

For translators of text types with long sentences, meaningful subsegment recall might involve matching fragments of fifteen or twenty words, while some translators might most benefit from single-word recall. The fragment lengths above were chosen to try to cover a range of cases, while also providing useful matches with the test data described below.

7.2 Test data and queries

Test data was created using a section of the DGT-TM (Steinberger et al. 2013), with punctuation normalised. 40,000 English-French TUs were extracted, and their N-grams of order 1 to 6 counted using SRILM (Stolcke 2002). The most frequent N-grams were examined manually to select suitable candidates for subsegment recall testing, eliminating N-grams with too many stop words, using the stop-word lists for French and English included in the Snowball stemmer suite (Porter 2001). The following fragments shown in Table 2 were chosen for use as test cases.

French	English
règlement	Regulation
établi	established
conclut que	concludes that
État membre	Member State
les autorités polonaises	the Polish authorities
modifiée comme suit	amended as follows
intégrée dans l'accord	incorporated into the Agreement
Journal officiel de l'Union européenne	Official Journal of the European Union

Table2. Recall test fragments

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 processed to extract up to 100 TUs containing the fragment pair and meeting the criteria above. Where fewer than 100 were found, the difference was made up with randomly-chosen copies of the available pairs. The 100 pairs were prefixed with a unique alphanumeric key, to ensure CAT tools did not consider any to be repetitions and so filter them out during import into a TM.

For each fragment pair, two subsets of the 100 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.

To simulate translation of a source text that includes a test fragment also found in a test TM, example sentences were created by adapting TUs in the test data containing the fragment pairs. These example sentences are referred to herein as 'queries'.

Each query was compared to the 10,000 'padding' TUs to ensure that neither French nor English segment constituted a 'fuzzy match' with any TU segment. An edit distance percentage value was computed, comparable to the fuzzy match values assigned by CAT tools, using 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%.

Test data and queries can be downloaded from http://kftrans.co.uk/benchmarks/benchmarkdata.zip. The queries used are also shown at http://kftrans.co.uk/benchmarks/Home/Queries and are referred to herein using the numbering shown there ('query 1a', etc.)

7.3 Testing and evaluation

TM system performance can usefully be expressed in terms of *precision* and *recall* (Whyman and Somers 1999: 1270). Roughly speaking, recall measures what proportion of possible relevant matches in the TM was actually displayed in the TM results, while precision measures what proportion of results displayed was relevant for the query. In these tests, the 'relevant matches' could be defined as the TUs in the TM that contain the fragment matching part of the query, and recall could be defined in terms of how many of these TUs are displaying in response to the query. However, this measure could not be applied to decontextualising systems — all but one of the four tools examined — since they do not display the TU(s) from which matches have been drawn. In any event, for DTA/BFE systems, a recall measure that is more indicative in practice is one based on the words contained in any subsegment translation suggestion, where the possible relevant words are those found in the translation of the

fragment in the query, and recall measures how many of those words are suggested. For the purposes of the testing described here, formulae for precision and recall were defined in those terms. 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 in Figure 5.

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

Figure 5. Precision equation

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 shown in Figure 6.

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

Figure 6. Recall equation

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 are defined as the average recall and precision of the suggestions.

Ideally, DTA/BFE performance for the four CAT tools concerned would be tested using identical procedures and measured in exactly the same way, thereby producing directly comparable results. However, the tools take such different approaches to subsegment recall that this is not feasible. In the tests described below, the specifics of how the formulae above were applied are given for each tool. 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. Results for different tools are not directly comparable, but nevertheless 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. This decision is motivated by the response to survey question 1B, where no AR mechanism is described, as well as my own assertion that a DTA/BFE system is of more use if it is an MR implementation.

7.4 Results

Detailed results for the individual queries and CAT tools can be found at http://kftrans.co.uk/benchmarks. The results for each CAT tool are discussed below.

7.4.1 memoQ 2013 R2

With the 'Guess translation' feature enabled, the user can invoke a dialog displaying the DTA translation suggestions for a subsegment match, as shown in Figure 7.

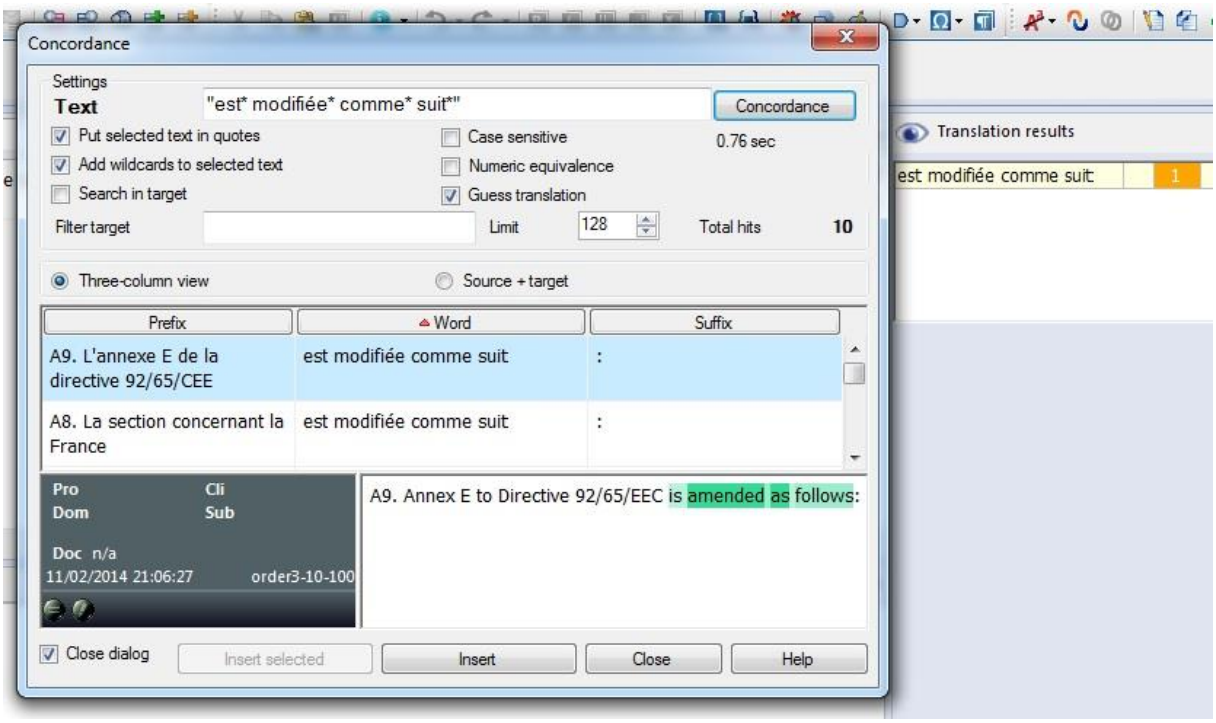

Figure 7. memoQ 'Guess translation'

(The memoQ BFE implementation of subsegment recall was not evaluated, since it cannot be configured to recall fragments with fewer than 5 occurrences.)

The dialog presents a list of TUs where a subsegment match has been found. The user can click between these TUs to see the target segment and proposed translation (highlighted), which may vary between TUs. With the test data used, the TU target segment always contains the same corresponding fragment translation, but the highlighted words vary. A user may choose to click between TUs to find the most relevant translation, but per the definitions above, this constitutes AR rather than MR. Precision and recall values for these tests were calculated using the selected-by-default first suggestion[1]. Suggestion display is further finessed to the user by giving stronger highlighting to more probable translation words. Based on the same AR/MR reasoning, precision and recall were calculated using all highlighted words, regardless of strength.

The graphs in Figure 8 show recall and precision averaged over all test queries, where the X-axis shows the frequency in the TM of the fragment pair to be matched, for a total of nine TMs (1 occurrence, 10 occurrences and 100 occurrences, each in TMs with 100, 1,000 and 10,000 padding TUs).

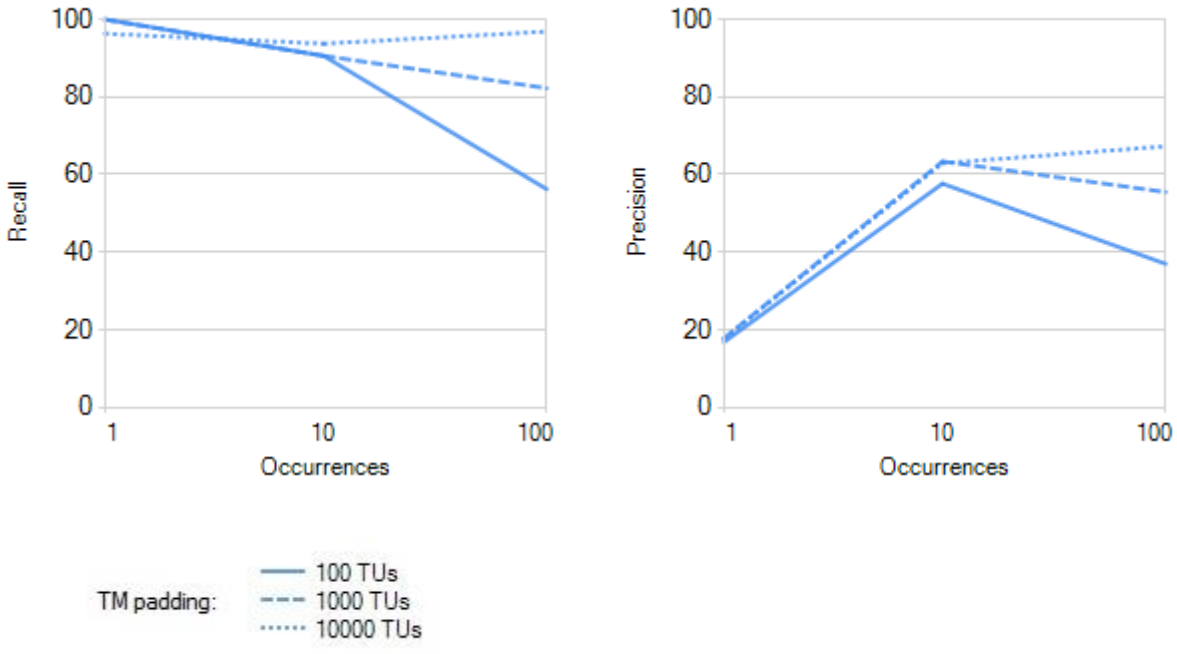

Figure 8. memoQ performance

The average results help summarise that with the fragments and TMs described above, when the fragment pair occurs 10 or fewer times in the TM, memoQ almost always highlights the correct words in the translation — but with just one occurrence, also highlights many more wrong words. The increase in precision between 1 and 10 occurrences is shown in the graph as linear; further testing could show that precision increases much more steeply with increased frequency, or less so. With more occurrences, both precision and recall tend to drop. The detailed results show that

performance in individual cases varies. For query 3, recall is consistently high, and precision tends to increase with fragment frequency, regardless of padding volume. For query 3a, however, 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.

7.4.2 Déjà Vu X2 v8

Results for Déjà Vu X2 (DVX2) were obtained using the 'Pretranslate' function, with the 'Assemble from portions' and 'Use DeepMiner statistical extraction' options enabled. An example of how Déjà Vu X2 presents subsegment matches is shown in Figure 9.

Figure 9. Déjà Vu X2 DeepMiner

Translation suggestions for whatever source text fragments are inserted together into the target text segment. This presents a difficulty in evaluating precision, since there is no way to determine which suggestion words have been presented in response to which source text fragment[2]. As a result, only recall has been calculated here for DVX2. (The complete translation suggestions for each query are shown at http://kftrans.co.uk/benchmarks/Home/dvx2. Reading these gives an impression of translation suggestion precision, especially when compared with the query texts and their translations.) The underlining DVX2 shows on certain words in translation suggestions indicates that the user can choose to display further suggestions. As explained above, this is herein considered a type of AR. Recall has been calculated using only the suggestion automatically inserted.

Figure 10 shows recall averaged over all test queries:

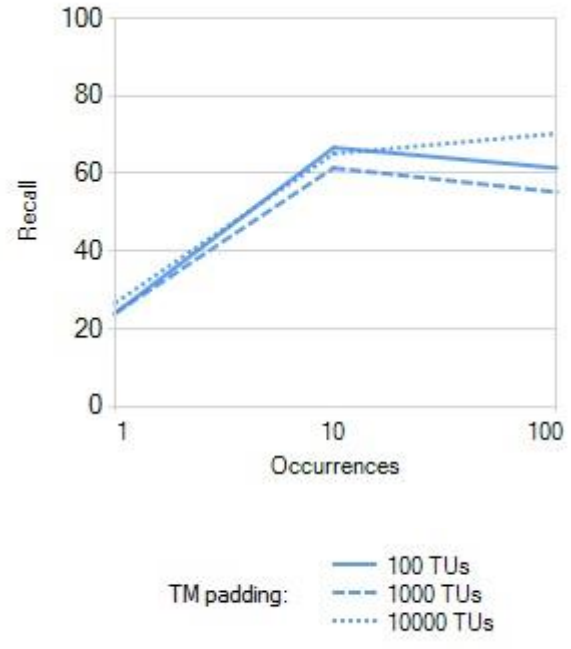

Figure 10. Déjà Vu X2 performance

The average results help summarise that with the fragments and TMs described above, when the fragment pair occurs just once in the TM, DVX2 is unlikely to recall it (or will recall only part of it), but is much more likely to do so with ten occurrences. Again, the increase in recall between the two is shown above as linear; in practice, it may not be. The detailed results show that performance in individual cases is again very varied, with noticeable differences dependent on language direction.

7.4.3 Similis Freelance v2.16

Similis provides subsegment translation suggestions using both DTA and BFE, as shown in Figure 11.

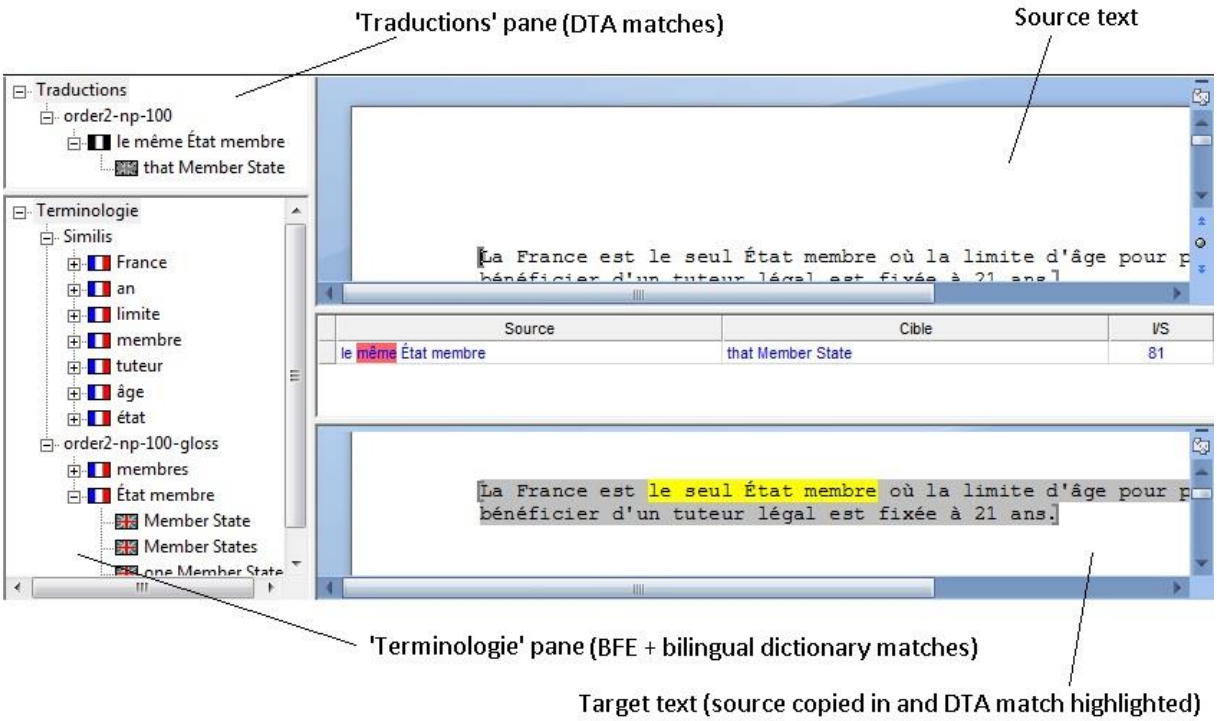

Figure 11. Similis glossary

The 'Traductions' pane shows translations suggested using DTA (with source text copied to the target pane, match highlighted in yellow, and a 'match difference' display in the pane above), while the 'Terminologie' pane shows translations suggested using BFE (by 'glossary', as well as a set of translations headed 'Similis' from a standard bilingual dictionary). In all tests described here, DTA suggestions were either absent or less complete than BFE suggestions, so performance has been measured using the latter. 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.

The graphs in Figure 12 show recall and precision averaged over all test queries.

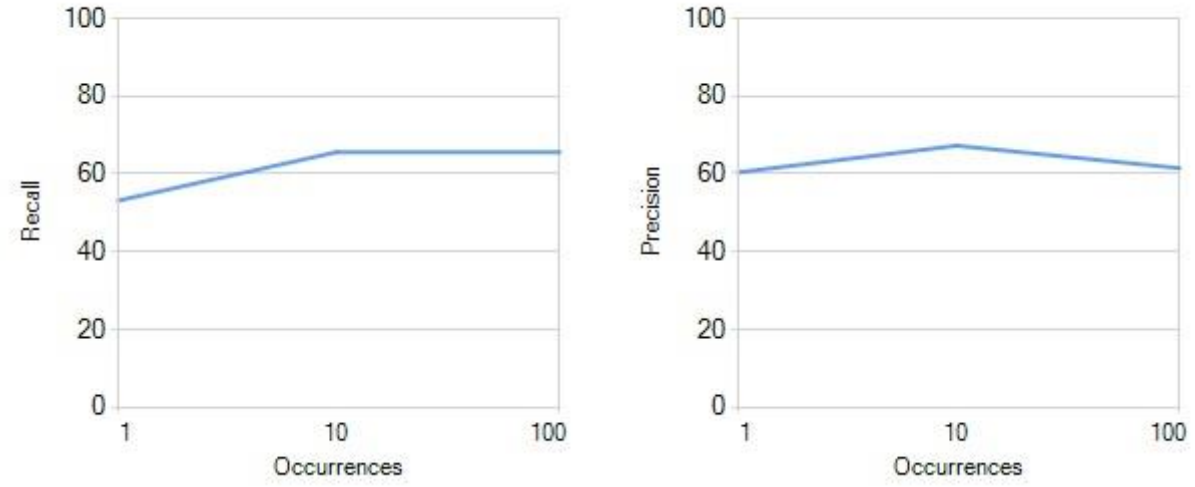

Figure 12. Similis performance

The average results help summarise that with the fragments and TMs described above, when the fragment pair occurs just once in the TM, Similis is quite likely to recall a translation for it, with quite good precision. The detailed results show that recall graphs tend to be flat – if Similis can recall a fragment suggestion, the number of occurrences usually makes no difference to whether it is recalled. Precision graphs also tend to be flat, but precision can reduce with more occurrences, as less accurate linguistic alignments are more likely to arise (and this reduction varies depending on the translation direction). Recall seems to be affected by the grammatical category of the fragment sought (per the results for the two different three-word fragments), so that for certain fragments, no translation suggestion is produced regardless of how many occurrences are in the TM. Conversely, translation direction seems to have relatively little effect on performance.

7.4.4 SDL Trados Studio 2014

Subsegment translation suggestions are only available in Trados Studio 2014 by means of AR. Figure 13 shows an example.

Figure 13. Trados Studio AutoSuggest

As with DVX2, this approach presents a difficulty in evaluating precision, since there is no way of determining for which fragment of the source

sentence the suggestions are meant to be of relevance. Consequently, only recall has been calculated for Trados Studio 2014.

Figure 14 shows recall averaged over all test queries.

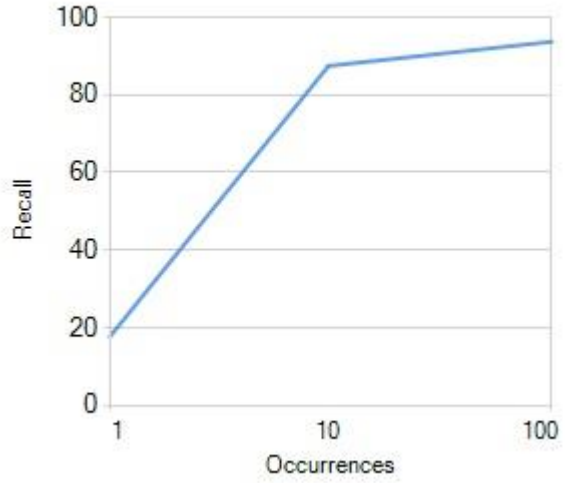

Figure 14. Trados performance

The average results help summarise that with the fragments and TMs described above, when the fragment pair occurs just once in the TM, Trados is unlikely to recall it (or will recall only part of it), but is very likely to do so with ten occurrences or more. The detailed results show that performance varies, but 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 query 1, expecting a suggestion for 'règlement', the recalled translation 'Regulation' is only displayed if upper-case 'R' is typed. Similarly, when testing using query 2a, expecting a suggestion for 'Member State', the recalled translation 'État membre' is only displayed if upper-case 'E' with an acute accent is typed.

8 Conclusion

Implementations of subsegment recall in CAT tools are much more varied than might be assumed. 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. This functionality may be important, since it should allow translators to benefit more from TM content re-use. As an example, a study conducted by RR Donnelley found

that subsegment recall provided word coverage that was an order of magnitude greater than segment-level recall (TAUS 2010: 11).

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.

Given the potential for increased TM leverage, translators should certainly be aware and make use of these subsegment recall implementations. How might a CAT tool vendor provide an implementation that meets translator expectations even better? While weaker in other areas, Similis meets the aforementioned preferences better because it is the only system not reliant on statistical analysis, 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). It would be very interesting for a vendor to develop an implementation that also takes an 'aligning' approach, but with more consistent results, and with a DTA rather than BFE methodology so as not to decontextualise translations recalled. In a recent webinar, Jost Zetzsche described subsegmenting methods as "probably the biggest and most important development in [TM] technologies" (Zetzsche 2014a), and has since reported on research into such an implementation (Zetsche 2014b) - so subsegment recall may become even more useful for translators before long.

Bibliography

- **Atril.** (2013). *Déjà Vu X2 Professional.* http://www.atril.com/software/d%C3%A9j%C3%A0-vu-x3-professional (consulted 10.12.2013).

- **Bowker, Lynne** (2003). "Terminology tools for translators." Harold Somers (ed.) (2003) *Computers and Translation: A Translator's Guide*. Amsterdam/Philadephia: John Benjamins, 49–65.

- **Christensen, Tina and Schjoldager, Anne** (2011). "The Impact of Translation-Memory (TM) Technology." Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds) (2011) *Human-Machine Interaction in Translation.* Copenhagen Studies in Language 41. Copenhagen: Samfundslitteratur, 119–130.

- **Grönroos, Mickel and Becks, Ari** (2005). "Bringing Intelligence to Translation Memory Technology." *Proceedings of the International Conference Translating and the Computer 27.* London: Aslib.

- **Kuhns, Bob** (2007). "It's not MT, and it's not TM." https://www.taus.net/articles/its-not-mt-and-its-not-tm (consulted 10.12.2013).

- **Levenshtein, Vladimir** (1966). "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady* 10(8), 707–710.

- **Lingua et Machina** (2014). *Lingua et Machina : L'écrit multilingue dans l'entreprise.* http://similis.org/linguaetmachina.www/index.php?afficher=10&info=Similis (consulted 2.3.2014).

- **Macklovitch, Elliott and Russell, Graham** (2000). "What's been forgotten in translation memory." John White (ed.) (2000) *Envisioning Machine Translation in the Information Future: Proceedings of Fourth Conference of the Association for Machine Translation in the Americas (AMTA-2000)*, Cuernavaca, Mexico: Springer, 137–146.

- **MemSource.** (2014). "Is subsegment matching available in MemSource?" http://support.memsource.com/topic/is-subsegment-matching-available-in-memsource. (consulted 2.3.2014).

- **Planas, Emmanuel** (2005). "SIMILIS Second-generation translation memory software." *Proceedings of the International Conference Translating and the Computer 27.* London: Aslib.

- **Porter, Martin** (2001). "Snowball: A language for stemming algorithms." http://snowball.tartarus.org/texts/introduction.html (consulted 2.3.2014).

- **Robinson, Douglas** (2003). *Becoming a Translator: An introduction to the theory and practice of translation*. London: Routledge.

- **SDL** (2014). "SDL AutoSuggest - Subsegment Matching Suggestions." http://www.translationzone.com/products/sdl-trados-studio/sdl-autosuggest.html (consulted 2.3.2014).

- **Steinberger, Ralf, Eisele, Andreas, Klocek, Szymon, Pilos, Spyridon and Schlüter, Patrick** (2013). "DGT-TM: A freely available translation memory in 22 languages." *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*. Istanbul, 454-459.

- **Stolcke, Andreas** (2002). "SRILM-an extensible language modeling toolkit." *Proceedings of International Conference on Spoken Language Processing*, Denver, CO, vol. 2. 901–904.

- **TAUS** (2010). "How to Increase Your Leveraging." https://www.taus.net/reports/how-to-increase-your-leveraging (consulted 2.3.2014).

- **Whyman, Edward and Somers, Harold** (1999). "Evaluation metrics for a translation memory system." *Software-Practice and Experience* 29(14), 1265–1284.

- **Zetzsche, Jost** (2014a). *Translation Technology - What's Missing and What Has Gone Wrong*. eCPD [webinar].

- — (2014b). *232nd Tool Box Journal*. [email distribution]

Biography

Kevin Flanagan is a freelance software developer and translator currently completing a PhD at Swansea University, researching improvements to Translation Memory. He teaches the CAT unit on the University of Bristol's MA Translation programme, and his other research involvements include an online tool for studying variation in translation, shown at http://www.delightedbeauty.org/vvv. E-mail: kevin@kftrans.co.uk.

Endnotes

[1] An alternative approach would have been to calculate those values for all suggestions in the list, and give an average, but since many of the tests use TMs with 100 occurrences, giving rise to 100 suggestions, and since there is no way to extract them programmatically, this was not feasible. Manual inspection of a selection of cases suggests that while the selected-by-default first suggestion is not necessarily the 'best' in the list, it generally gives a good representation of the average suggestion quality.

[2] In the example shown, the underlining groups some words together, which might allow an association to be made with a source text fragment, but that underlining is not always present, is not necessarily found on all words (as above) and does not constitute a reliable means of making such an association.