Term-spotting with TransCheck: A Capital Idea

Elliott Macklovitch, Guy Lapalme
Laboratoire RALI
Département d'informatique et de recherche opérationnelle
Université de Montréal
macklovi@iro.umontreal.ca
lapalme@iro.umontreal.ca

Nelida Chan
Government Translation Service
Ontario Ministry of Government and Consumer Services
77 Grenville Street
Toronto, ON M5S 1B3
Nelida.Chan@ontario.ca

Abstract

TransCheck, the RALI's automatic translation checker, has recently undergone a field trial at the Government Translation Service of Ontario, where the system was used not only to detect inconsistent terminology, but also to find new source language terms in texts sent to outside translation suppliers. We describe a specialized term-spotting module developed in the course of that trial to assist the terminologists identify new official names to be added to ONTERM, the Ontario government's online terminology database.

1 The TransCheck system

The RALI Laboratory has been developing the TransCheck system (henceforth abbreviated as TC) for some years now.[2] As its name suggests, TransCheck was originally conceived as a translation checker, i.e. as a system that would be used by a translator or a reviser to detect possible errors in a draft translation, somewhat like a spell checker. Unlike a spell checker, however, which detects errors of form in a single monolingual text, TransCheck is designed to detect errors of correspondence that occur between two texts – a source text in one language and its translation in another. So, for example, the French word 'librairie' would not be flagged as an error by a monolingual spell checker, since it is a correct form of the French language; however, that same form could be flagged as an error of correspondence by TC if it appeared in a target text as the translation of the English word 'library'.

Roughly speaking, TC works as follows. The user begins by specifying a source and a target text file, which the system first tokenizes (i.e. segments into words and sentences) and then automatically aligns. The latter is a crucial step in which TC determines which target sentence(s) correspond to each source sentence. To these aligned regions, TC then applies its various error detection modules. The system currently detects the following types of errors: inconsistent terminology, or terms that diverge from those required by the user (which we call the 'positive' terminology); source language interference, including false cognates like 'library/librairie' (called 'negative' terminology); and paralinguistic expressions, like numbers, dates and monetary expressions. Of course, a draft translation may well contain many other types of errors, but to automatically detect these will often require a deep understanding of the two texts that are in a translation relation. For the time being, TC limits itself to the aforementioned set of errors, all of which can in principle be detected by relatively simple, purely formal means. TC flags a potential error when, in an aligned region, it either detects a certain source item (e.g. a numerical expression) without finding its obligatory target correspondent, or when a source item is detected along with its prohibited target correspondent (e.g. a false cognate). After the complete bi-text has been processed in this way, the

[2] For a detailed description of an earlier version of TransCheck, see (Macklovitch, 1994).

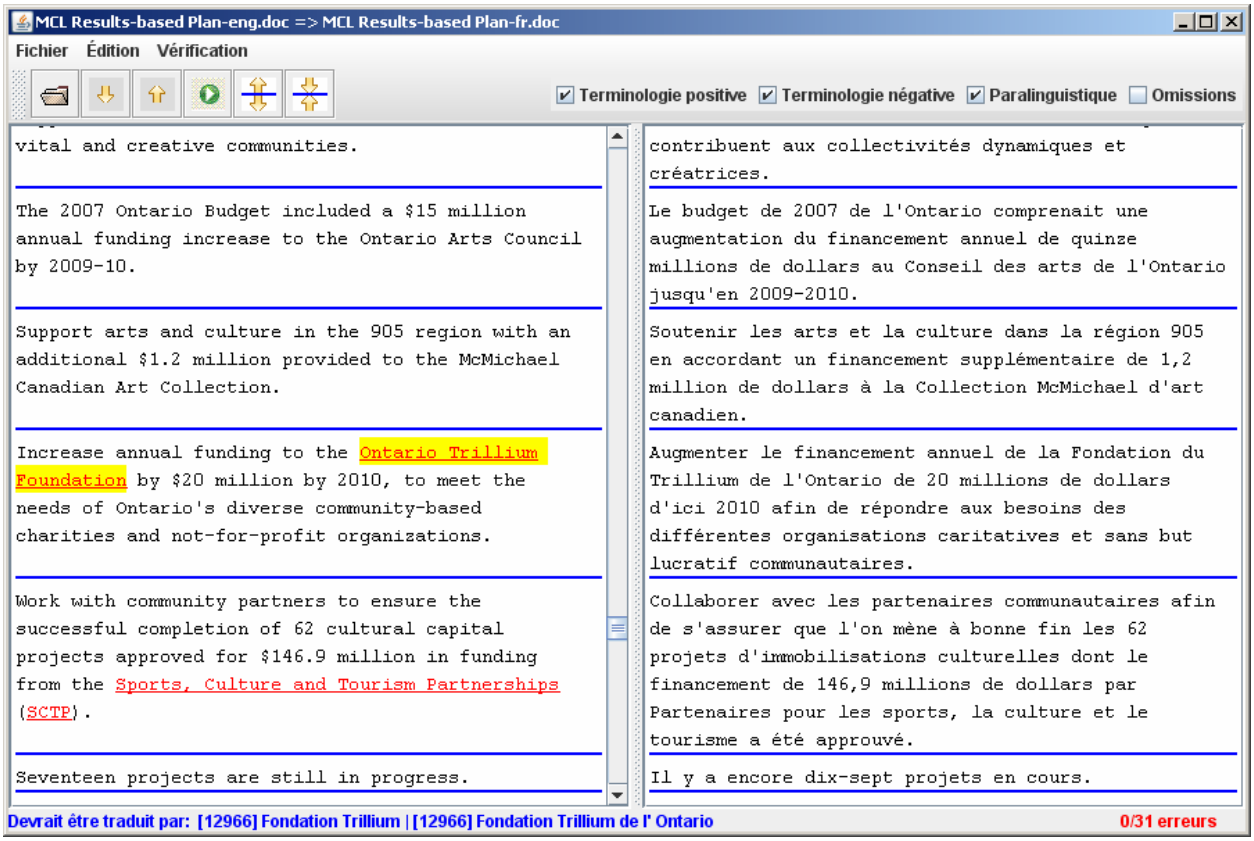

Figure 1: An example of TransCheck output

The source text appears on the left and the target text on the right; the dark horizontal lines mark the alignments that TC has automatically calculated. The term 'Ontario Trillium Foundation' is highlighted here because TC has not found in the aligned target segment one of the desired French terms, which are indicated on the bottom of the window.

results are output in a graphic user interface for the user to review. Figure 1 provides an example of TC output illustrating a potential error of positive terminology.

Now suppose that the user does not specify a target file when asking the system to check the positive terminology. TC will verify every source term that appears in its glossary and, not finding any target equivalents, will systematically flag each one. Odd as this may at first appear, it could in fact be very useful, especially when the output is saved in HTML format, since this HTML file can then be sent to outside translation suppliers, informing them of the terminology that the client requires in their translation. See Figure 2 below for an example of such output.

TC recently underwent an extensive field trial at the Government Translation Service (GTS) of Ontario, where the great majority of texts are outsourced to freelancers and other service providers who are contractually obliged to respect the terminology in ONTERM, the Ontario government's online terminology database. For various administrative reasons, GTS did not send its suppliers HTML files like that in Figure 2. However, the terminologists at GTS did supply them with a list of new terms not yet in ONTERM, which they extracted from the source texts using a specialized term-spotting module that the RALI had added to TC at their request. We will describe this new module in some detail below. First, however, we turn to a brief description of ONTERM, which provided the terminology that was used in the TC field trial at GTS.

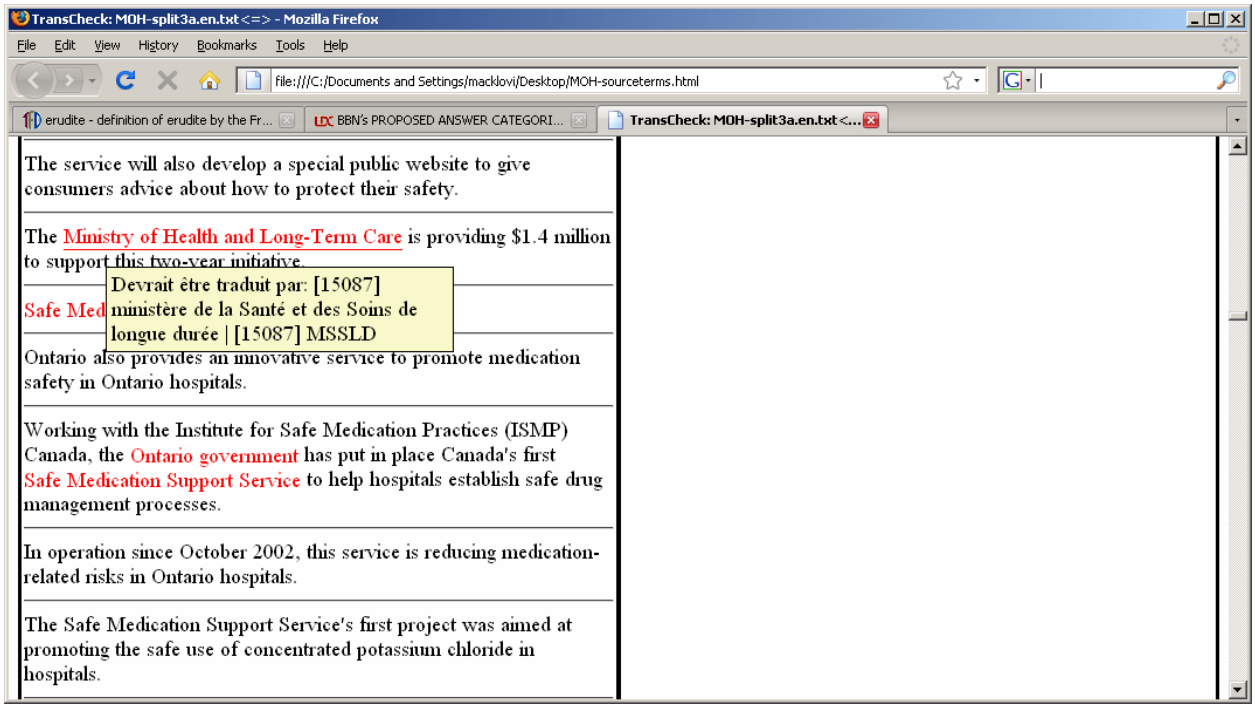

Figure 2: An example of monolingual output in HTML format

The user has previously specified a source text file and a glossary, and asked the system to verify the positive terminology. TC responds by highlighting all the terms that are present in the source text. The user saves the output in HTML format, which can then be sent to outside translation suppliers and viewed in any browser.

2 ONTERM

The demand for French translation within the Ontario government has been increasing steadily since 1986, when the government passed the French Services Act. In 1997, the Government Translation Service moved from an in-house translation model to the current outsource model which farms out almost all translations to private-sector suppliers. An in-house team of corporate translation advisors (CTAs) are responsible for managing a roster of suppliers and overseeing quality assessment and control. The Terminology Unit within GTS was given the mandate to create French equivalents for official Ontario Government English names, to ensure the consistent use of official names across Ontario Public Service (OPS) documents and to make names and terminology available to the OPS, translation suppliers and the public at large. To assist in accomplishing this mandate, the ONTERM database and website were created.[3] The ONTERM database focuses on official Ontario government names which include: the names of Ministries, agencies and organizational units; the names of councils, committees, working groups; the names of plans, programs, projects, policies and strategies; position titles; the names of IT systems and applications, conferences, exhibits, commemorative events, awards, distinctions, scholarships, catch phrases and political geographic entities.[4] ONTERM currently contains over twenty-six thousand terminology records.

[3] http:// www.onterm.gov.on.ca

[4] It will be noted that all of these, with the exception of the catch phrases, correspond to a particular set of *named entities*, whose defining characteristic is that the entity must fall within the jurisdiction of the Ontario government. For further discussion, see note 7 below, as well as Section 5.

3 Term checking in TransCheck

Term checking was a central component in the field trial of TC at GTS. It therefore behooves us to describe in somewhat more detail precisely how this operation proceeds in TC, beginning with the format of the entries in the system's terminological glossary.

3.1 Format of the entries in TransCheck's glossary

The following is a typical example of an entry in TC's terminological glossary:

```
(i)  EN: [1619 ; 1619] Government Translation Service ; GTS
     FR: [1619 ; 1619] Service de traduction du gouvernement ; STG
```

The terminological equivalents that are to be verified by TC must be listed in a plain text file which the user loads at run time, and all the entries in the file must conform to this simple format: two fields, one for the English term, the other for the French term, with alternate terms (i.e. synonyms, abbreviations or other shortened forms) being separated by a semicolon. The system interprets this entry as follows: every time an occurrence of 'Government Translation Service' or 'GTS' is encountered in a segment of an English text, either 'Service de traduction du gouvernement' or 'STG' must be found in the aligned French segment; otherwise, a potential terminological inconsistency will be flagged. Note that such entries are bi-directional, i.e. they can be applied equally well to a French source text and an English translation, although the errors will always be flagged in the source. As for the numbers in square brackets, they refer to the original record in ONTERM and were inserted to help GTS' terminologists interpret TC's output; the system itself does not use them.

When TransCheck was originally conceived, the developers assumed (somewhat naively) that the users of the system would manually create the term glossaries, composed of entries in the above format, which the system would apply to draft translations. While this is certainly possible, it has the undesirable consequence of limiting the scope of the terminology that TC can verify; for few users are likely to manually key in thousands of entries like the one above. This is one reason why we reacted so favorably when we were approached by GTS regarding a possible field trial of TC; the terminologists there wanted to import into TC all the records in ONTERM and use the resulting system to help them vet outsourced translations.

Now obviously, many of the records in a bona fide term database will be far more complex than the one shown in (i) above, containing fields for the subject domain, definitions, observations and other notes, author and date of the record, etc.; and the same is true (though perhaps to a somewhat lesser extent) for the smaller in-house glossaries that are maintained by translation agencies and individual translators. Because we only import into TC the term equivalents from such records, some of the information on the records will inevitably be lost. In some cases, this will not be overly serious; in others, however, the information that is discarded may be crucial in helping a human translator decide if, or in what context, a particular term equivalence applies. Consider in this regard the following (non-hypothetical) example:

```
(ii) EN: aim
     FR: objet
      NOTE: Title of paragraph found in Course Management Details. It
          gives the main objectives of a course.
```

For a human translator, the note in (ii) is highly informative: it tells him/her in precisely what context the English term 'aim' can and should be translated as 'objet' in French. This information will be lost, however, when this record is imported into TC; and even if it were retained, the system would have no way of determining whether the text it was currently processing corresponded to a course lesson. Consequently, every time it encounters the word 'aim' in an English text, TC will seek the word 'objet' in the aligned French segment; and often it will not find it, because the word 'aim' allows for many other fully acceptable French equivalents, e.g. 'but', 'objectif', 'cible', and 'mire', to cite just a few.

3.2 Term invariability

As we saw in section 2, the majority of the records in ONTERM correspond to official government names. An important property of such appellations is that they tend to be invariant, in two senses; first, their form is relatively frozen, in not admitting the insertion of modifiers, for example; and second, these terms and their translation are unlikely to vary according to the linguistic context in which they appear. Consider the following two entries, taken from ONTERM and provided here in their simplified TC format:

```
(iii)  EN:[588] Chief Election Officer
       FR:[588 ; 588] directeur général des élections ; directrice géné-
           rale des élections

 (iv)  EN:[2882 ; 21329] Regional Director ; Area Director
       FR:[2882,21329 ; 2882,21329] directeur régional ; directrice ré-
           gionale
```

As we can see from the entry in (iii), it is not quite true to say that official names never vary, for the French designation for this position does change, according to whether the incumbent is a man or a woman. This being said, it is quite improbable that this multi-word term could accept other inflections, or admit an interposed modifier in either language. Notice too that the first letter of each English word in the term is capitalized, indicating that we are dealing here with the name of a particular official position. Hence, we are not likely to find the head noun of this term in the plural, because, like all proper nouns, it has a unique reference.[5] Synonymy with such terms is usually limited to abbreviations, acronyms or other shortened forms of the full term, as in the example in (i) above.

The entry in (iv) is slightly more complex; one way of interpreting it is as follows: the French term 'directeur regional' (or its inflectional variant 'directrice régionale') may be translated either as 'Regional Director' or as 'Area Director' – these last being two distinct terms. Notice as well that two different record numbers are included in the square brackets of entry (iv), indicating that the first English term comes from record #2882, while the second comes from record #21329. When we consult these records in ONTERM, what we find is that 'Regional Director' is the generally recommended term in the Ontario Public Service, while 'Area Director' is the term that is preferred in the ministry of Municipal Affairs and Housing. In other words, these two English terms name two distinct entities (which is why they have two distinct records), although both happen to use the same name in French. So here too we need to qualify the claim made above that official names tend to be invariant. The English designation for 'directeur régional' does indeed vary, although the conditioning factor is not *linguistic* but rather the organization to which the position belongs. The only solution open to TC in cases like this is to relax the condition used to flag terminological inconsistencies. Applying the entry in (iv) to a pair of aligned texts, TC will not flag a potential error upon encountering the French term if either 'Regional Director' or 'Area Director' is found on the English side; and this, even though one of those English terms may perhaps be incorrect in a text coming from a particular ministry. Because government organizational structures and position titles are often named the same way in one or other language in the various ministries across the Ontario government, though not always translated the same way, the TC ONTERM glossary contains a relatively high number of such entries: 1225 to be exact, or approximately 5% of the total entries, although many of the alternate terms turn out to be minor inflectional or orthographic variants, e.g. 'Board of Negotiation ; board of negotiation' or 'Pesticides Residue Section ; Pesticide Residue Section'.

[5] See (Quirk et al., 1972) for a succinct description of the key properties of proper nouns. Regarding the invariability of the English term in (iii), we do occasionally find instances in which the internal noun 'election' is pluralized: 'Chief Elections Officer'. While this may be correct in other jurisdictions or in other countries, ONTERM tells us that the correct form of the term, in Ontario at least, always has this noun in the singular.

3.3 Glossaries for human and machine use

There are a number of important lessons for automatic term checking that can be drawn from this discussion. The first, and perhaps the most general, is that the requirements of a term glossary intended for humans and one designed for automatic term checking are quite different. For the former, the comprehensiveness and detail of the information provided on each record are important attributes, because terminologists can rightly assume that humans are capable of intelligently interpreting such information. For a program like TC, on the other hand, the non-ambiguity of the entries is primordial, and this, for similar though inverse reasons. The program simply does not have the requisite intelligence to correctly interpret such information – neither the linguistic intelligence that would allow it to infer the grammatical dependencies that are often specified in the observations, and certainly not the real-world knowledge upon which humans instinctively rely to make the necessary distinctions that allow them to select the correct term. Entries like that in (ii) above are therefore anathema for TC. In general, we should not ask the system to verify the translation of vague or polysemous words that allow for multiple, equally correct target language equivalents. The system will fare far better when the terms it is asked to verify are linguistically complex (i.e. multi-word expressions), technical and highly specific, because such terms and their translations tend to remain invariable across different contexts.

Another lesson that the field trial at GTS served to underscore was obliquely alluded to above when we mentioned that TC was used there to help vet outsourced translations. Again, when TC was first conceived, the idea was that it would be primarily used to assist revisers. However, the manner in which the system was employed at GTS was more akin to quality control than to revision. The distinction between the two activities may not at first be apparent, but is in fact fundamental. In revision, a (normally senior) translator seeks to improve a draft translation by making changes to it.[6] The objective in quality control, on the other hand, is simply to determine whether or not a translated text meets certain quality standards; if it does not, it is refused or returned to the translator, without the quality controller necessarily having to make any changes to it. In its current state, our TC prototype lends itself better to quality control than to revision. The principal reason for this is that, while the user of TC can make changes to the target text and save those changes, the resulting modified text may not accurately replicate the format of the original source text, especially if the latter contains figures, tables or other elaborate layout.

4 Term-spotting with TransCheck

In addition to managing the Ontario government's translation needs, GTS is also responsible for maintaining ONTERM, the provincial government's online terminology database. A substantial part of this work involves keeping ONTERM up-to-date by adding to it new terms and their official equivalents. In principle, there are two ways in which GTS terminologists might go about this: they could either scour government documents on their own, with a view to identifying new terms that need to be added to ONTERM; or new terms could be brought to their attention by outside translation suppliers, who request that they provide an official French equivalent. Neither of these procedures is ideal: the first is labour-intensive and time consuming; the second is essentially reactive, whereas GTS would much prefer to be proactive.

In the course of the TC field trial, the terminologists at GTS asked the RALI if it could help them with this task by developing a specialized term-spotting module that would be integrated within TransCheck. Recall that a majority of the entries in ONTERM are official names that designate government positions, administrative units, programs, legislation, etc.; as such, most of these are proper nouns that begin in English with a capital letter. The suggestion of the lead terminologist at GTS was that the new term-spotting module in TC use this simple formal property as a diagnostic for identifying official names in English texts. Furthermore, since TC already incorporates a glossary containing all the terms in ONTERM, the new module should be able to determine which of the potential terms identified in a given text do not yet have entries in the database and hence might need to be added.

---

[6] Of course, a preliminary translation may also be reviewed and improved by the same translator who drafted it.

The RALI agreed to implement this suggestion in a new term-spotting module that was added to TC and subsequently tested at GTS. As before, TC begins by identifying all the terms in the English source text that are present in ONTERM. In a second pass, the new module then scans the text and locates all sequences of two or more words that begin with a capital letter. (Isolated, single words that begin with a capital are simply too ambiguous; this would result in the first word of every sentence being proposed as a potential term and would significantly increase the program's noise level.) The capitalized words in these sequences are generally contiguous, although the program does allow for a small number of lower-case 'skip words', in the form of certain common prepositions (e.g. 'of' and 'for'), the articles 'the' and 'a', and the conjunction 'and'. When the entire text has been processed in this way, TC outputs an HTML version that is colour-coded in the following way: terms already present in ONTERM are highlighted in blue; potential new terms are highlighted in yellow; those multi-word expressions that begin with a capital and already have an entry in ONTERM are highlighted by both routines and so appear in green. TC also produces a table at the end of this HTML file listing all the potential new terms and their frequency. Entries in the table are hyperlinked to their occurrences in the text, so that a user can easily inspect the candidate terms in their context. An example of this HTML output is given in Figure 3 below.

A quick glance through this list shows that many of the proposed multi-word sequences are not in fact true terms and hence should not be added to ONTERM. Among them are proper names, either of persons (e.g. 'Dan Strasbourg' or 'Dr. Joe Clarke') or of companies (e.g. 'CNW Group Ltd') which do not belong in a term bank. Other dubious entries in the table arise from problems in properly segmenting the source text. For example, at the beginning of each of the press releases that make up this text, there is a call to a particular set of addressees: 'Attention News/Health Editors'. Our program does not properly parse the scope of the slash as a conjunction and so this sequence is mistakenly segmented into two potential terms, the first of which is incoherent. A similar problem occurs with the period at the end of 'Myth No', which should normally include the following number, e.g. 'Myth No. 5'. (Whether this should be considered a true term is another question.) And then there are instances of noise which are simply unavoidable, given the simplicity of our term-spotting criteria. For example, 'The McGuinty' comes from the beginning of a sentence that continues '…government is encouraging Ontarians to protect themselves'. Here we have a sequence of two words that begin with a capital letter, although this is certainly no term.

On the other hand, this short portion of the table shown in Figure 3 also contains a fair number of entries which do seem to constitute bona fide terms, and hence might have to be added to ONTERM: 'Complex Continuing Care', 'Emergency Department Care', 'Ontario Hospital Association', 'Canadian Institute for Health Information', and 'Hospital for Sick Children'.[7] The status of other entries, e.g. 'Hospital Reports' is perhaps less clear and would require the considered judgment of a qualified terminologist who would begin by verifying the occurrences of this expression within the text – something s/he could easily do by means of the inserted hyperlinks. The important point is this: despite the noise that such lists may contain, qualified terminologists have no trouble running through them and rapidly picking out new terms for inclusion in the database.

[7] In point of fact, the last three terms would not be added to ONTERM, because they either designate bodies that are outside the Ontario government's jurisdiction (the *Canadian* Institute for Health Information), or because they are not actually part of the government itself (both terms involving hospitals). This is a good illustration of the subtly of the knowledge that is required to distinguish between terms that should or should not be added to ONTERM.

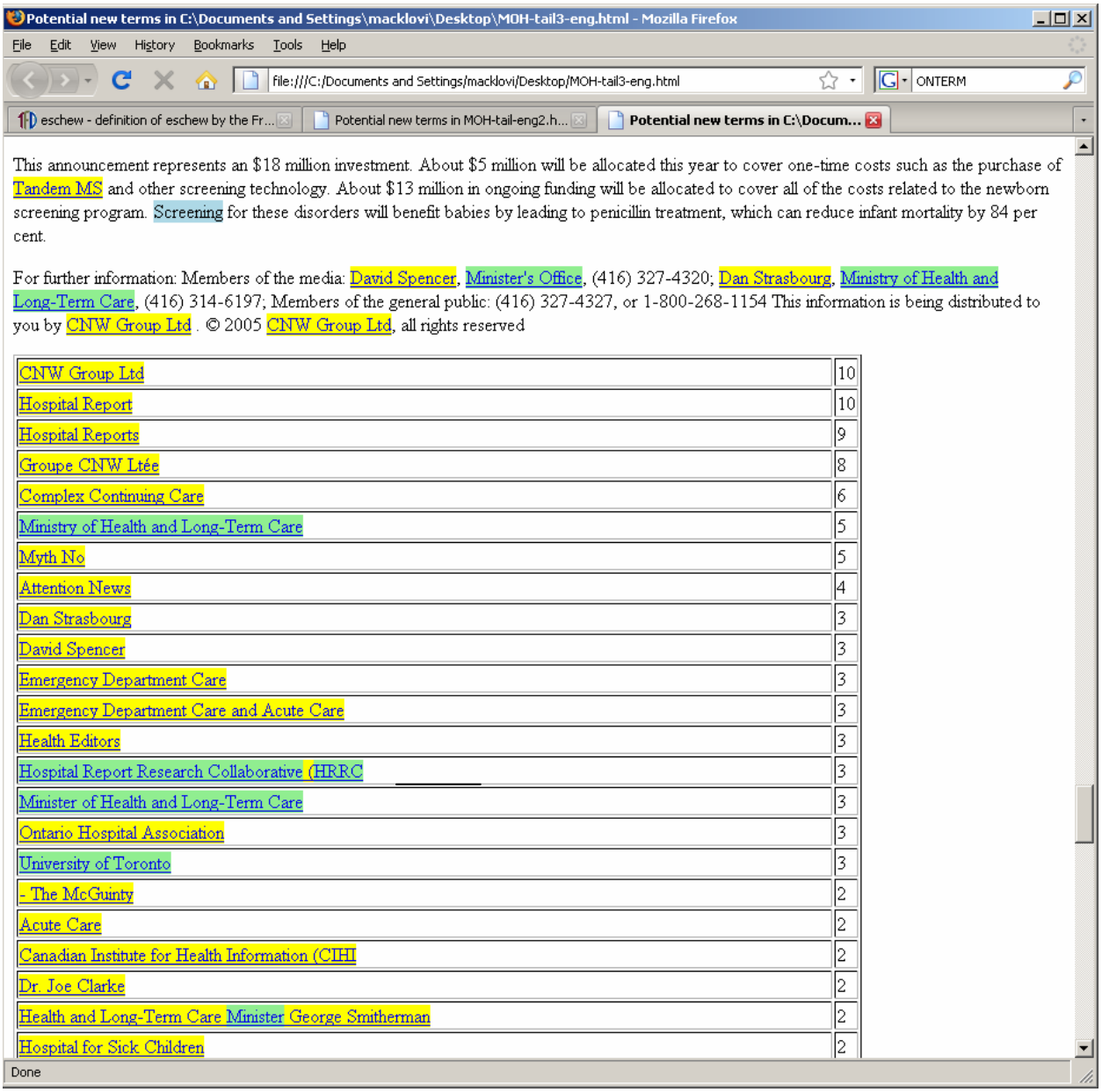

Figure 3: Output of TransCheck's term-spotting module

The table lists all multi-word sequences that begin with a capital letter. Those highlighted in yellow are not found in ONTERM and so correspond to potential new terms; their frequency is given in the right-hand column of the table. Those terms that are found in ONTERM are highlighted in blue; we see one occurrence of 'screening' in the text above the table. Capitalized multi-word sequences that are identified by our term-spotting module and also appear in ONTERM receive both blue and yellow highlighting, and so they appear as green. Each entry in the table is hyperlinked, allowing the terminologist to quickly peruse all its occurrences in the text.

5 Discussion

Automatic term identification and extraction – what we have been calling term-spotting – has a relatively long history, going back at least to the seminal paper of (Justeson & Katz, 1993).[8] What these researchers demonstrated is that a large proportion of the technical terminology in English texts corresponds to nominal groups (i.e. nouns and their pre- and post-modifiers) that can be accurately identified using regular expressions defined over part-of-speech categories. Moreover, a simple but surprisingly effective way of distinguishing technical terms from ordinary noun phrases is to use the criterion of full repetition. Multi-word technical terms tend to reappear verbatim in a text, whereas non-technical, descriptive noun phrases do not; upon repetition, they are often pronominalized or truncated in various ways. Subsequent research has elaborated on this basic approach. (Daille, 1994), for example, extended it to French and showed that lengthy complex terms could profitably be analysed as combinations of simpler term sequences that are generated by means of various devices such as co-ordination. In an effort to increase the precision of the extracted candidate terms, (Drouin, 2003) compares the frequencies of nouns and adjectives in a technical corpus under analysis with their frequency in a non-technical, general corpus of the language – the idea being to extract only those terms made up lexical items that are highly specific to the technical domain. The results are quite encouraging, particularly for single-word terms, which are often ignored by term extraction programs because their recognition is problematic. (Patry & Langlais, 2005) propose an alternative to a handcrafted static definition of what constitutes a legitimate term, specified as possible sequences of POS tags; from a corpus of term examples supplied by the user, they train a language model that automatically generates these POS patterns.

Compared to these approaches, the term-spotting algorithm that we have implemented in TransCheck and described above appears extremely simple, not to say simplistic. We do no part-of-speech tagging, do not calculate any statistical correlate of term likelihood (e.g. mutual information), and do not compare the frequency of the components of our candidate terms with their frequency in a non-technical reference corpus. Instead, we define one simple pattern for our candidate terms, based on whether their component words begin with a capital letter. The reason we can do this, of course, is that the terms we are looking for are all official names, and official names are proper nouns which all begin with a capital letter – at least in English.[9] The problem, however, is that there exist other types of proper nouns which do not designate official government appellations, e.g. personal names, temporal names and geographical or place names. These too begin with a capital letter and are the source of much of the noise in the table of candidates terms that TC produces.[10]

Now in principle, it would be possible to automatically filter out some of this noise by incorporating within TC additional linguistic machinery. Consider personal names, for example, several of which appear in the list reproduced in Figure 3 above. There has been much work on named entity recognition in recent years, and many programs now exist which can reliably identify personal names in running text. The problem is that ONTERM contains many records for terms that include personal names, e.g. 'Sir John A. Macdonald Highway', 'Philip Shrive Memorial Bridge', 'Lincoln M. Alexander Building', Dr. Albert Rose Bursary'. This probably explains the less-than-enthusiastic response of GTS terminologists to our suggestion of adding such a filtering component to TC. Given their ability to rapidly scan the candidate terms and accurately pick out those that should be added to ONTERM, they prefer to retain the noise rather than run the risk of missing a potential new term. And so we have left TC as is, imperfect in theory perhaps, but quite adequate in practice.

---

[8] Although this paper was published in 1993, it was actually written and disseminated several years earlier.

[9] Hence, our term-spotting algorithm wouldn't work with source texts in French, where proper nouns do not follow the same capitalization rules as in English, or in German, where all nouns are capitalized. But at GTS, the overwhelming majority of source texts are in English.

[10] Actually, some geographical names are included in ONTERM, e.g. those (like 'Thousand Islands Parkway') that serve to designate government buildings, highways, historical sites, etc. Moreover, the ONTERM Web site provides access to the GeoNames database, which lists over fifty-seven thousand geographic names in English and French.

6 Conclusion

A recognized feature of good writing and good translation, especially in electronic environments, is terminological consistency. Urgent texts frequently have to be divided up among several translators or outside translation suppliers; the reviser – or in the case of GTS, the corporate translation advisor – is then left with the task of merging these parts into a coherent whole. Ensuring terminological consistency is an important part of this job, and it can be quite onerous. The trial at GTS has shown that TransCheck can help by automating this task to a considerable extent. Indeed, CTAs found that simply by aligning the source and target texts in side-by-side format, TC allowed them to perform quality control on the entire text (not just the terminology) more efficiently and thoroughly than in the past.

For the Ontario government, this question of terminological consistency is particularly important. In today's electronic environment, the consistent use of correct terms, and especially names, is essential for accessing information. We all know how helpful it is, in conducting a successful Web search, to be able to query the correct designation of what one is looking for. But terminological consistency is also important for developing a brand and a brand personality. While in the private sector, the focus is mainly on company and product names, within a government context, the official names used across ministry websites play a special role in projecting the government's brand personality and fostering a relationship with its citizenry. Since ministry home pages increasingly act as the electronic doorways to virtual government offices, it is crucial that clear and consistent names in both English and French, and between English and French, be used to project the government's personality, to find information and to navigate effectively among its numerous websites and pages.

The latest version of TransCheck with its term spotting module has allowed the Terminology Unit in GTS to make more efficient use of its research time. During the period of the field trial, the Terminology Unit handled a 53% increase in the number of terms rendered into French, in large part on account of the use of the new TC system to pre-process the government's results-based plans. Not only does TC give the terminology service the ability to be more systematic and thorough in the detection of new Ontario government terminology, it allows it to be more proactive, thereby reducing the amount of after-the-fact checking. Making corrections after translations have been submitted is both costly and time-consuming. Sometimes, they are so costly that the corrections are never made. Being proactive promotes the philosophy of 'doing it right the first time'.

References

Daille, Béatrice. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, Paris : Université de Paris 7.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99-115.

Justeson, John & Slava Katz. 1993. Technical terminology: some linguistic properties and an algorithm for identification in text. Technical Report RC 18906, IBM Research Division.

Macklovitch, Elliott. 1994. Using Bi-textual Alignment for Translation Validation: the TransCheck System. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, pp.157-168.

Patry, Alexandre & Philippe Langlais. 2005. Corpus-Based Terminology Extraction. *7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Danemark, pp. 313-321.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1972. *A Grammar of Contemporary English.* Longman Group Ltd., London.