

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
11 April 2002 (11.04.2002)

PCT

(10) International Publication Number
WO 02/29621 A1

(51) International Patent Classification⁷: **G06F 17/28**

(21) International Application Number: PCT/US01/30652

(22) International Filing Date: 1 October 2001 (01.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/237,537 4 October 2000 (04.10.2000) US

(71) Applicant: **IDIOM TECHNOLOGIES, INCORPORATED** [US/US]; 200 Fifth Avenue, Waltham, MA 02451 (US).

(72) Inventors: **SLOAN, William, N.**; 200 5th Avenue, Waltham, MA 02451 (US). **MOREHEAD, Kem**; 49

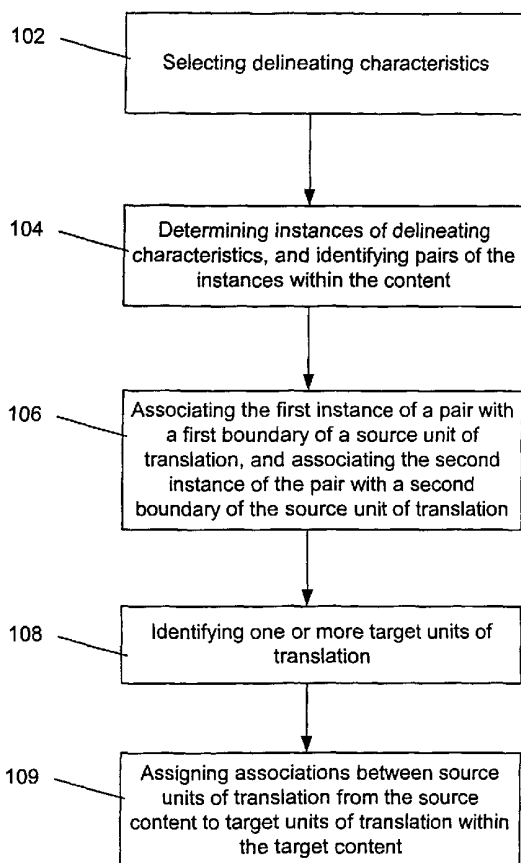
Fresh Pond Place, Cambridge, MA 02138 (US). **WING HIN HO, Herman**; 5 Old Colony Lane, Apartment 7, Arlington, MA 02476 (US). **LIU, Kenneth, Y.**; 403 Washington Street, Apartment 2, Somerville, MA 02143 (US). **MITCHELL, Richard, B.**; 313 Longely Road, Groton, MA 01450 (US). **SHANKAR, Umesh**; 2212 Blake Street, Apartment 303, Berkeley, CA 94704 (US).

(74) Agents: **KUSMER, Toby, H.** et al.; McDermott, Will & Emery, 28 State Street, Boston, MA 02109 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

[Continued on next page]

(54) Title: METHOD OF AND SYSTEM FOR SPLITTING AND/OR MERGING CONTENT TO FACILITATE CONTENT PROCESSING



(57) Abstract: A method of identifying units of translation in a block of source content, so as to segment the block of content into the units of translation, includes selecting (102) one or more delineating characteristics of the source content in addition to lexical characteristics. The method further includes determining (104) instances of the delineating characteristics in the block of source content, and identifying pairs of the instances within the text. The method also includes, for each pair of instances of the delineating characteristics, associating (106) a first instance of the pair with a first boundary of a unit of translation, and associating a second instance of the pair with a second boundary of the unit of translation. One embodiment further includes identifying (108) target units of translation in a block of target content, and assigning (109) associations among the source units of translation and the target units of translation.



WO 02/29621 A1



(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHOD OF AND SYSTEM FOR SPLITTING AND/OR MERGING CONTENT TO FACILITATE CONTENT PROCESSING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 60/327,537 entitled "METHOD OF AND SYSTEM FOR SPLITTING AND/OR MERGING CONTENT TO FACILITATE CONTENT PROCESSING" filed on October 4, 2000, the disclosure of which is entirely incorporated herein by reference.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0002] Not Applicable

REFERENCE TO MICROFICHE APPENDIX

[0003] Not Applicable

BACKGROUND OF THE INVENTION

[0004] The present invention relates to processing information content, and more particularly, to combining and/or separating segments of this content to simplify and otherwise facilitate translation and other processing functions associated with the content. Over the past few decades, opportunities for international relationships have expanded at a staggering rate. Many factors have contributed to this expansion -- improved transportation capabilities, advances in communication and media technologies, opening of once inaccessible cultures, among others. More recently, the Internet (the World Wide Web, in particular) has provided seemingly unlimited access to international audiences. The Internet represents a massive global business opportunity, and has provided the means for a wide range of businesses to deploy a multilingual and multicultural marketing presence, thereby increasing revenue, improving customer loyalty and reinforcing brand recognition.

[0005] As information becomes available globally, the role of translators has shifted away from simple transcription of text into a target language. Translators always had to pay close attention to any attributes and linguistic idiosyncrasies of the target culture, as well as understand and adapt to these differences. Now, however, translators must also ensure the timely deployment of the translated content to the designated site. Translation can be made more

efficient with greater flexibility in software functionality and the ability to save previous translations for future use. Traditionally, translators worked with *hard copy documents*, from which they had the flexibility to translate content at any suitable level. Thus, translators had the ability to look at an entire document and translate it without confines. The increased need for efficient content translation has motivated numerous companies to develop tools that automate at least part of the translation process.

[0006] To increase the overall speed of content translation, tools have been developed to save translations in some type of memory (referred to herein as "translation memory" or "TM"), so that the tool can make automatic substitutions, and the translator will not have to consider further instances of those translations. The TM provides a record of pairs of units of translation that have already been translated. A "unit of translation" is a segment of content that has been delineated by any of several criteria, as is discussed in more detail herein. Each associated pair in the TM includes a unit of translation from the content in the source language (i.e., the language of the content that is to be translated), and the corresponding translation unit from content in the target language (i.e., the language into which the source content is being translated). In order to populate the TM, prior art translation methods segment content into sentences (or other syntactic units, e.g., words, phrases, etc.) based on predetermined criteria so that the translator can focus on translating one sentence (or other syntactic unit) at a time.

[0007] However, differences between the source language and the target language create difficulties in translating directly from one language to another within the constraints of the particular segments chosen. Such differences may include, but are not limited to, differences in grammatical structure, differences in idiomatic expressions, and punctuation differences. Further, segments that are spatially adjacent in the source document may not necessarily be best suited as adjacent in the target content. Content generally cannot be translated word for word, sentence for sentence, paragraph for paragraph, because of these language differences. Another consideration is that competent, efficient translation is typically not deterministic. For example, three translators operating on the same content may well produce three different translations, each of which would be technically correct. Any type of segmentation tool that segments the content based on a rigid set of criteria will force a translator to approach translation of the content on a word for word (etc.) basis.

[0008] Flexibility in content segmentation is important because translators must be able to account for the differences in language structures. For instance, translating content sentence by sentence may populate a translation memory with more specific entries. Storing more specific entries in translation memory is useful because doing so increases the likelihood that future translation instances will make use of those entries. However, as described herein, a sentence-to-sentence translation may not be accurate, depending on the languages being used in the translation. For example, the following sentences in Italian:

Per quanto riguarda la Banca Centrale Europea, un euro debole può essere un problema soltanto se aumenta l'inflazione. Però, a 2.3%, l'inflazione nella zona euro è ancora abbastanza modesta.

would be translated as a single sentence in English:

Yet as far as the ECB is concerned, a weak euro is only really a problem if it pushes up inflation; and at 2.3%, inflation in the euro zone is still rather modest. (*The Economist*, September 23-29, 2000, p. 89)

On the other hand, although translating an entire paragraph as a unit may be more accurate, it can be inefficient for translators because doing so will populate the translation memory with entries that are unlikely to be used again.

[0009] An additional problem with content segmentation is determining the sentence boundaries. Typically, a period denotes a sentence end. Yet, if a word within a sentence is abbreviated and uses a period (e.g., "Mr."), the period following the abbreviation could be interpreted as a sentence end and the sentence would thus be segmented at that point. Likewise, some languages such as Thai do not even use period punctuation.

[0010] It is an object of the present invention to substantially overcome the above-identified disadvantages and drawbacks of the prior art.

SUMMARY OF THE INVENTION

[0011] The present invention provides a method of and system for splitting and merging blocks of information content (e.g., textual blocks) so as to simplify and expedite a translator's task in converting content from one language to another. The method and system of the present invention is referred to herein, in general, as "Split/Merge." The textual information to be translated from one language to another is referred to herein as "content." The Split/Merge

method and system allows a user (i.e., a translator) to decide, in real time, the level at which he or she wishes to translate content. The translator has the ability to “split” a paragraph into separate sentences, allowing for individual translation of each sentence. Thus, the translation memory contains entries at the sentence level, which are more likely to be repeated than entire paragraphs. In addition, the translator can “merge” selected sentences together to form a single segment for translation. Furthermore, the translator can “merge” all sentences of a paragraph into a single textual “chunk,” as well as merge all of the paragraphs into a larger textual “chunk”. This split/merge functionality provides flexibility for source material that is not suitable for sentence-by-sentence translation.

[0012] The utility of the Split/Merge invention may be exploited in a translation system such as Idiom’s WorldServer. In general, WorldServer is a Web-based application that enables enterprises to manage their content while leveraging established Web architecture, content management and workflow systems. A translator uses WorldServer to determine what content he or she needs to translate. The translator can either export the content needing translation to a third party editing tool, or use the Translation Workbench to perform the actual translation. A translator can be an individual contributor, including users that are adapting but not translating content and reviewers who review content.

[0013] The Split/Merge feature of the present invention provides value for translators by giving them greater flexibility of how to translate content before performing the translation. In addition, increased flexibility in segmentation will populate the TM with more utilizable entries.

[0014] The foregoing and other objects are achieved by the invention which in one aspect comprises a method of identifying one or more source units of translation in a block of source content, so as to segment the block of content into the one or more source units of translation. The method includes selecting one or more delineating characteristics of the source content in addition to lexical characteristics. The method further includes determining instances of the delineating characteristics in the block of source content, and identifying one or more pairs of the instances within the text. The method also includes, for each pair of instances of the delineating characteristics, associating a first instance of the pair with a first boundary of a source unit of translation, and associating a second instance of the pair with a second boundary of the source unit of translation.

[0015] Another embodiment of the invention further includes identifying one or more target units of translation in a block of target content, and assigning associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

[0016] Another embodiment of the invention further includes translating content in the source units of translation to the associated target units of translation.

[0017] In another embodiment of the invention, the delineating characteristics include syntactic characteristics. The method further includes determining pairs of instances of syntactic characteristics of the source content.

[0018] In another embodiment of the invention, the delineating characteristics include formatting characteristics. The method further includes determining pairs of instances of formatting characteristics of the source content.

[0019] In another embodiment of the invention, the document formatting characteristics include HTML code markers.

[0020] In another embodiment of the invention, the delineating characteristics include conceptual characteristics. The method further includes determining pairs of instances of conceptual characteristics of the source content.

[0021] In another embodiment of the invention, the conceptual characteristics include spatial adjacency.

[0022] In another embodiment of the invention, the delineating characteristics include sound-based characteristics. The method further includes determining pairs of instances of sound-based characteristics of the source content.

[0023] In another embodiment of the invention, the sound based characteristics include voice inflections.

[0024] In another embodiment of the invention, the delineating characteristics include one or more markers manually inserted by a user. The method further includes determining pairs of instances of markers within the source content.

[0025] Another embodiment of the invention further includes translating the one or more source units of translation into a target language so as to form target units of translation, and merging the target units of translation into one or more blocks of target content.

[0026] In another embodiment of the invention, the source units of translation are characterized by a first adjacency pattern. The method further includes merging the target units of translation so as to follow the first adjacency pattern.

[0027] In another embodiment of the invention, the source units of translation are characterized by a first adjacency pattern. The method further includes merging the target units of translation so as to follow a second adjacency pattern different from the first adjacency pattern.

[0028] In another embodiment of the invention, at least one of the source units of translation corresponds with two or more target units of translation.

[0029] In another embodiment of the invention, two or more of the source units of translation corresponds with a single target unit of translation.

[0030] In another embodiment of the invention, each one of the source units of translation corresponds with a single target unit of translation.

[0031] Another embodiment of the invention further includes merging the target units of translation into a hierarchical structure.

[0032] Another embodiment of the invention further includes providing one or more predetermined hierarchy criteria. The characteristics of the hierarchical structure are defined by the predetermined hierarchy criteria.

[0033] In another aspect, the invention comprises a system for computer assisted identification one or more source units of translation in a block of source content, so as to segment the block of content into the one or more source units of translation. The system includes a user interface for allowing a user to select one or more delineating characteristics of the source content in addition to lexical characteristics. The system further includes a content processor for determining instances of the delineating characteristics in the block of source content, and identifying one or more pairs of the instances. The system also includes, for each pair of instances of the delineating characteristics, a segment processor for associating a first instance of the pair with a first boundary of a source unit of translation. The segment processor also associates a second instance of the pair with a second boundary of the source unit of translation.

[0034] In another embodiment of the invention, the content processor further identifies one or more target units of translation in a block of target content. The content processor also

assigns associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

[0035] In another embodiment of the invention, the content processor further translates content in the source units of translation to the associated target units of translation.

[0036] In another embodiment of the invention, the delineating characteristics include syntactic characteristics, and the content processor further determines pairs of instances of syntactic characteristics of the source content.

[0037] In another embodiment of the invention, the delineating characteristics include document formatting characteristics, and the content processor further determines pairs of instances of document formatting characteristics of the source content.

[0038] In another embodiment of the invention, the document formatting characteristics include HTML code.

[0039] In another embodiment of the invention, the delineating characteristics include conceptual characteristics, and the content processor further determines pairs of instances of conceptual characteristics of the source content.

[0040] In another embodiment of the invention, the conceptual characteristics include spatial adjacency.

[0041] In another embodiment of the invention, the delineating characteristics include sound-based characteristics, and the content processor further determines pairs of instances of sound-based characteristics of the source content.

[0042] In another embodiment of the invention, the sound based characteristics include voice inflections.

[0043] In another embodiment of the invention, the delineating characteristics one or more markers manually inserted by a user, and the content processor further determines pairs of instances of markers within the source content.

[0044] In another embodiment of the invention, the segment processor further translates the source units of translation into a target language so as to form target units of translation, and merges the target units of translation into one or more blocks of target content.

[0045] In another embodiment of the invention, the source units of translation are characterized by a first adjacency pattern, and the segment processor further merges the target units of translation so as to follow the first adjacency pattern.

[0046] In another embodiment of the invention, the source units of translation are characterized by a first adjacency pattern, and the segment processor further merges the target units of translation so as to follow a second adjacency pattern different from the first adjacency pattern.

[0047] In another embodiment of the invention, at least one of the source units of translation corresponds with two or more target units of translation.

[0048] In another embodiment of the invention, two or more of the source units of translation correspond with a single target unit of translation.

[0049] In another embodiment of the invention, each one of the source units of translation corresponds with a single target unit of translation.

[0050] In another embodiment of the invention, the segment processor further merges the target units of translation into a hierarchical structure.

[0051] In another embodiment of the invention, the segment processor further receives one or more predetermined hierarchy criteria, and the characteristics of the hierarchical structure are defined by the predetermined hierarchy criteria.

[0052] In another aspect, the invention comprises a system for computer assisted identification one or more source units of translation in a block of source content, so as to segment the block of text into the one or more source units of translation. The system includes means for allowing a user to select one or more delineating characteristics of the source content in addition to lexical characteristics. The system also includes means for determining one or more pairs of instances of the delineating characteristics in the block of source content. The system further includes, for each pair of instances of the delineating characteristics, means for associating a first instance of the pair with a first boundary of a source unit of translation, and means for associating a second instance of the pair with a second boundary of the source unit of translation.

[0053] In another aspect, the invention comprises a method of dynamically selecting one or more segmentation criteria used to identify source units of translation in a block of source content, wherein the segmentation criteria identifies delineation characteristics of the source content for defining boundaries of the source units of translation. The method includes providing two or more source segmentation criteria associated with the block of source content. The method also includes selecting one of the source segmentation criteria from the two or more

segmentation criteria as an initial source criterion, and using the initial source criterion for defining boundaries of the source units of translation. The method further includes dynamically selecting, as a function of one or more external factors, subsequent source segmentation criteria from the two or more source segmentation criteria, as the boundaries of the source units of translation are defined.

[0054] Another embodiment of the invention further includes providing two or more target segmentation criteria associated with a block of target content. The method further includes selecting one of the target segmentation criteria from the two or more target segmentation criteria as an initial target criterion, and using the initial target criterion for defining boundaries of the target units of translation. The method also includes dynamically selecting, as a function of one or more external factors, subsequent target segmentation criteria from the two or more target segmentation criteria, as the boundaries of the target units of translation are defined. The method also includes assigning associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

[0055] In another embodiment of the invention, the one or more external factors includes the associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

[0056] In another embodiment of the invention, the one or more external factors includes input from a user translating from the source units of translation to the target units of translation.

[0057] In another embodiment of the invention, the one or more external factors includes data relating to characteristics of the source content.

[0058] In another embodiment of the invention, the data relating to characteristics of the source content includes HTML code.

[0059] In another aspect, the invention comprises a system for computer assisted dynamic selection of one or more segmentation criteria used to identify source units of translation in a block of source content. The segmentation criteria identifies delineation characteristics of the source content for defining boundaries of the source units of translation. The system includes a user interface for providing two or more source segmentation criteria associated with the block of source content, and for selecting one of the source segmentation

criteria from the two or more segmentation criteria as an initial source criterion. The system also includes a content processor for using the initial source criterion for defining boundaries of the source units of translation. The system further includes a segment processor for dynamically selecting, as a function of one or more external factors, subsequent source segmentation criteria from the two or more source segmentation criteria, as the boundaries of the source units of translation are defined.

[0060] In another embodiment of the invention, the user interface further provides two or more target segmentation criteria associated with a block of target content, and selects one of the target segmentation criteria from the two or more target segmentation criteria as an initial target criterion. The content processor further uses the initial target criterion for defining boundaries of the target units of translation. The segment processor further dynamically selects, as a function of one or more external factors, subsequent target segmentation criteria from the two or more target segmentation criteria, as the boundaries of the target units of translation are defined. The segment processor further assigns associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

[0061] In another embodiment of the invention, the one or more external factors includes the associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

In another embodiment of the invention, the one or more external factors includes input from a user translating from the source units of translation to the target units of translation.

[0062] In another embodiment of the invention, the one or more external factors includes data relating to characteristics of the source content.

[0063] In another embodiment of the invention, the data relating to characteristics of the source content includes HTML code.

[0064] In another aspect, the invention comprises a system for computer assisted dynamic selection of one or more segmentation criteria used to identify source units of translation in a block of source content. The segmentation criteria identify delineation characteristics of the source content for defining boundaries of the source units of translation. The system includes means for providing two or more source segmentation criteria associated with the block of source content, and for selecting one of the source segmentation criteria from

the two or more segmentation criteria as an initial source criterion. The system further includes means for using the initial source criterion for defining boundaries of the source units of translation. The system also includes means for dynamically selecting, as a function of one or more external factors, subsequent source segmentation criteria from the two or more source segmentation criteria, as the boundaries of the source units of translation are defined.

[0065] Another embodiment of the invention further includes means for providing two or more target segmentation criteria associated with a block of target content. The system further includes means for selecting one of the target segmentation criteria from the two or more target segmentation criteria as an initial target criterion, and using the initial target criterion for defining boundaries of the target units of translation. The system also includes means for dynamically selecting, as a function of one or more external factors, subsequent target segmentation criteria from the two or more target segmentation criteria, as the boundaries of the target units of translation are defined. The system also includes means for assigning associations among the source units of translation in the block of source code and the target units of translation in the block of target code.

BRIEF DESCRIPTION OF DRAWINGS

[0066] The foregoing and other objects of this invention, the various features thereof, as well as the invention itself, may be more fully understood from the following description, when read together with the accompanying drawings in which:

[0067] FIG. 1 shows a flow diagram of a method for splitting and merging blocks of information content according to the present invention;

[0068] FIG. 2A illustrates the sample paragraph and a list of selected delineating characteristics;

[0069] FIG. 2B shows the delineating instances determined within the content of FIG. 2A;

[0070] FIG. 2C shows the first boundary and the second boundary of each unit of translation corresponding to the delineating instances of FIG. 2B;

[0071] FIG. 3A shows source content as paragraphs represented as separate, contained segments or bigger paragraphs that contain sub-paragraphs;

[0072] FIG. 3B shows the content of FIG. 3A in hierarchical form; and,

[0073] FIG. 4 illustrates a computer-based system for splitting and merging blocks of information content according to the method of FIG. 1.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0074] One embodiment of a method 100 for splitting and merging blocks of information content according to the present invention is shown in flow-diagram form in FIG.1. In one aspect, the method identifies units of translation in a block of source content (hereinafter referred to as "source units of translation") based on one or more delineating characteristics associated with the source content. Within the block of source content, individual source units of translation exist bounded by delineating instances. Examples of delineating instances include (but are not limited by) syntactic characteristics (i.e., the relationships among characters or groups of characters within the content), lexical characteristics (e.g., punctuation, character case, white space between characters), conceptual characteristics (e.g., semantics, such as characters or character groups that should be spatially adjacent to be proper in a particular language), multimedia content characteristics (e.g., proper relationships among various multi-media components), sound based characteristics (e.g., voice inflection), markup/formatting characteristics (e.g., an HTML document, or the row/column boundaries of a table), and markers manually inserted into the content by the translator.

[0075] In general, a translator (i.e., a person with knowledge of the target language and the source language; also referred to herein as a "user") translates the source content by translating the individual source units of translation into target units of translation. A competent translator may desire to have the source units of translation formed by different delineating instances, as described above, depending upon the nature of the document. For example, in one area of content, a translator may wish to have the units of translation formed by lexical instances such as punctuation. Thus, in that area of content the units of translation could simply be phrases delineated by periods, commas, semicolons, etc. In another area of content, the translator may desire to have the units of translation formed by syntactical instances, such as grammatical structural boundaries. In yet another situation, a translator may be translating an Internet web page, so that the translator may want to use HTML markup characters as the boundary delineators for the units of translation. In other cases, the translator may desire to have a mixture

of boundary delineators for the units of translation (e.g., HTML character on one boundary, and a lexical instance on the other boundary).

[0076] As FIG. 1 shows, the method 100 begins by selecting 102 at least one, but possibly several, delineating characteristics in addition to lexical characteristics. The delineating characteristics are chosen in addition to lexical characteristics because the use of lexical characteristics alone is nearly trivial. In other words, segmenting the content by sentences delineated by periods, or phrases delineated by commas, is a relatively common practice. However, segmenting the content using other criteria, either alone, in combination with lexical characteristics, or combinations thereof is novel and an important aspect of the present invention. The method 100 then determines 104 of instances of the selected delineating characteristics within the source content, and identifying at least one pair of the instances in the content. The pairs of instances do not need to be consecutive within the content, although in many cases they will be consecutive. The method 100 then evaluates individual pairs of instances of delineating characteristics. The method 100 associates 106 the first instance of a pair with a first boundary of a source unit of translation, and associates the second instance of the pair with a second boundary of the source unit of translation. Thus, the method 100 identifies the source units of translation within the content by using the various delineating characteristics. In general, the pairs of instances will both be the same type of characteristic, i.e., both lexical, or both syntactical, etc. In other embodiments, the method may identify units of translation via a hybrid pair of instances, i.e., the first instance may be syntactic, and the second instance could be conceptual, as described herein. In some embodiments of the invention, the translator (i.e., the user) selects the delineating characteristics (step 102) and the remaining steps (steps 104 and 106) are completely automatic. In other embodiments, the translator provides input to steps 104 and 106 as well. For example, in some embodiments the actual instances within the content may be automatically determined, and the translator may manually select (via keystrokes or mouse clicks on the computer, for example) which instances should be paired.

[0077] FIGs. 2A, 2B and 2C illustrate a sample source paragraph segmented into source units of translation by the method 100. FIG. 2A illustrates the sample paragraph 110 and the list 112 of selected delineating characteristics. For simplicity in this example, only "lexical" characteristics are included, although other characteristics, alone or in combination, may also be included in the list. FIG. 2B shows the delineating instances 114 determined within the content,

and the pairs 116 the method 100 identifies. FIG. 2C shows the first boundary 118 and the second boundary 120 of each unit of translation 122 as determined by the method 100.

[0078] In one embodiment of the invention, the method 100 further identifies 108 one or more target units of translation, and assigns 109 associations between source units of translation from the source content to target units of translation within the target content. In general, the associations among the target units and the source units of translation are dictated by the language translation the translator performs from the source content to the target content. The association may be, but is not necessarily, a one to one relationship between the source unit of translation and the target unit of translation. In some cases, two or more source units of translation may be associated with a single target unit of translation. In other cases, a single source unit of translation may be associated with two or more target units of translation. In general, the translator defines the initial association during the act of translation. The source-to-target associations are typically stored in translation memory (TM), so that future occurrences of the source unit of translation in the source content can be automatically associated to a target unit of translation.

[0079] The source units of translation are typically characterized by an adjacency pattern, i.e., the first source unit of translation is adjacent to the second source unit of translation, the second source unit of translation is adjacent to the third source unit of translation, etc. In some embodiments, the method enforces this adjacency pattern during the association with the target units of translation, so that the target content follows the adjacency pattern of the source content. In other embodiments, the nature of the language dictates that the target units of translation do not follow the adjacency pattern of the source units of translation. One example of this is when the source content is interspersed with comments or other non-essential content segments. In this case, the translator may wish to delete the comments from the target content, and would simply not associated the source units of translation that contain the comments with any target units of translation.

[0080] In one embodiment of the invention, the method merges the target units of translation into a hierarchical structure. FIGs. 3A, 3B and 3C illustrate an example of hierarchical structuring. FIG. 3A shows source content as paragraphs (and the sentences that form the paragraphs) represented as separate, contained segments or bigger paragraphs that contain sub-paragraphs. FIG. 3A also shows the content broken down into the first level of

hierarchy -- the three main paragraphs in the content. FIG. 3B shows the hierarchical breakdown of the first paragraph, and FIG. 3C shows the hierarchical breakdown of the third paragraph. Note that the second paragraph is a single sentence, and so no further hierarchical breakdown is necessary. In the most extreme case, the entire content could be represented as one segment. With the present invention, the user (i.e., the translator) has the ability to configure how paragraphs can be merged. For example, if the server upon which the Split/Merge is resident is configured such that paragraphs with blank lines between them can be merged, the user may segment the Example shown in FIG. 1 as depicted in FIG. 2. A box represents one segment that can be translated separately (assuming that it is not a Tag only segment). The user can merge segments that are siblings (sentences within a paragraph) and translate the result as one segment. Likewise, the user can split a segment that contains children and translate each child separately. While the user can dynamically split and merge the segments, they can only view one level for any given node at a time. For example, the user can not translate at the paragraph and sentence level for the same paragraph at the same time. They can, however, decide to translate one segment as a paragraph, and a different paragraph at the sentence level. In one embodiment, the translator may provide one or more hierarchy criteria, such that the method merges the target units of translation into a hierarchical structure defined by the criteria. In other embodiments, a fixed, predetermined set of hierarchy criteria may be accessible by the method for hierarchical structuring of the target units of translation.

[0081] Another aspect of the invention includes a method 150 of selecting the one or more segmentation criteria that are used to identify source units of translation in a block of source content. The segmentation criteria identify the delineation characteristics described herein that define the boundaries of the source units of translation. The method 150 includes providing 152 a set of two or more source segmentation criteria that are associated with the block of source content. The method 150 also includes selecting 154 one of the source segmentation criteria as an "initial" source criterion that is used for defining boundaries of the first source units of translation within the source content. The method 150 further includes dynamically selecting 156 subsequent source segmentation criteria from the set of source segmentation criteria as the boundaries of the source units of translation are defined. The dynamic selection is done as a function of one or more external factors, typically including inputs from the translator as he or she performs the translation. Thus, in one embodiment of the

method 150, the translator designates several segmentation criteria that identify delineation characteristics that may be useful in segmenting a particular block of source content. In some cases, the segmentation criteria may simply be the delineation criteria themselves, as defined herein for the method 100. In other cases, the segmentation criteria may include higher-level translation goals that imply the use of a particular set of delineation characteristics. In other cases, the segmentation criteria may be real time input from the translator (such as keystrokes, "point and click" via a mouse interface, or voice inflections). Once a set of segmentation criteria have been designated, the translator selects one (or possibly more) of the criteria to be used initially. As the translation proceeds, the translator may change the criteria "on the fly," i.e., dynamically selecting criteria from the designated list as the translator deems appropriate for the content.

[0082] In one embodiment of the invention, the method 150 further includes selecting 158 one or more segmentation criteria that are used to identify target units of translation for a block of target content. Similar to the source segmentation criteria, the target segmentation criteria identify the delineation characteristics that define the boundaries of the target units of translation. In some embodiments, the set of target criteria and source criteria may be identical, i.e., a common set of criteria may be used for both target and source criteria. The method 150 further includes providing 160 a set of two or more target segmentation criteria that are associated with the block of target content. The method 150 also includes selecting 162 one of the target segmentation criteria as an "initial" target criterion that is used for defining boundaries of the first target units of translation within the target content. The method 150 further includes dynamically selecting 164 subsequent target segmentation criteria from the set of target segmentation criteria as the boundaries of the target units of translation are defined. As with the source content, the dynamic selection is done as a function of one or more external factors, typically including inputs from the translator as he or she performs the translation. The external factors may also include, among other things, information related to the content itself, such as HTML code that describes the layout of a web page, or data file that provides the structural layout of the source document. The method 150 also assigns associations among the source units of translation and the target units of translation. In general, the associations among the target units and the source units of translation are dictated by the language translation the translator performs from the source content to the target content. The association may be, but is not

necessarily, a one to one relationship between the source unit of translation and the target unit of translation. In some cases, two or more source units of translation may be associated with a single target unit of translation. In other cases, a single source unit of translation may be associated with two or more target units of translation. In general, the translator defines the initial association during the act of translation. The source-to-target associations are typically stored in translation memory (TM), so that future occurrences of the source unit of translation in the source content can be automatically associated to a target unit of translation.

[0083] A computer-based system 200 for splitting and merging blocks of information content according to the present invention is conceptually illustrated in FIG. 4. In one embodiment, the system 200 includes a user interface 202, a content processor 204, a segment processor 206, and a translation memory (TM) 208, all resident on a computer system 210 such as a personal computer, workstation or similar system known in the art. The user interface 202 provides a mechanism for a translator to select the delineating characteristics in addition to lexical characteristics, as described herein. In one embodiment, the translator selects from a group of predetermined characteristics. In other embodiments, the translator is provided the option of entering his or her own characteristics that may be unique to the source and/or target content. The user interface 202 also provides the translator a mechanism for manually inserting delineating markers so that the translator can manually segment units of translation. The user interface may include a keyboard input for keystrokes, a mouse input for point-and-click input, or a voice recognition processor for recognizing voice commands and inflections. The user interface also provides an output interface to the user from the system 200, so that the translation procedure is interactive.

[0084] The content processor 204 then analyzes the source content and determines instances of the selected delineating characteristics within the source content. The content processor also identifies one or more pairs of instances from all of the instances it finds. The pairs of instances do not need to be consecutive within the content, although in many cases they will be consecutive. The segment processor 206 then evaluates individual pairs of instances of delineating characteristics. The segment processor 206 associates the first instance of a pair with a first boundary of a source unit of translation, and associates the second instance of the pair with a second boundary of the source unit of translation. Thus, the segment processor 206 identifies the source units of translation within the content by using the various delineating characteristics,

as described herein. In one embodiment of the invention, the segment processor 206 further identifies one or more target units of translation, and assigns associations between source units of translation from the source content to target units of translation within the target content, as described herein. In general, the translator defines the initial association during the act of translation. The source-to-target associations are stored in TM 208, so that future occurrences of the source unit of translation in the source content can be automatically associated to a target unit of translation.

[0085] The system 200 also allows the translator to dynamically select and vary the segmentation criteria the system uses to identify source and target units of translation, and to dynamically vary the criteria the system uses to merge the target units of translation after they have been formed. The translator provides a set of source and target segmentation criteria to the system via the user interface 202. The translator selects, also via the user interface 202, initial source and target segmentation criteria, and the user interface 202 provides these selections to the content processor 204 and the segment processor 206, which in turn utilize the selections to segment and merge the units of translation. As the translator performs the translation of the source content into the target content, he or she may decide at some point in the translation that the current segmentation criteria is not suitable, and that different criteria would be more appropriate. The translator may, "on the fly," provide new segmentation criteria to the system 200 via the user interface 202, selected from the set of criteria entered earlier. The segment processor 206 subsequently uses the new segmentation criteria to segment and merge content thereafter.

[0086] The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The present embodiments are therefore to be considered in respects as illustrative and not restrictive, the scope of the invention being indicated by the appended claims rather than by the foregoing description, and all changes which come within the meaning and range of the equivalency of the claims are thus intended to be embraced therein.

What is claimed is:

- 1 1. A method of identifying one or more source units of translation in a block of source
2 content, so as to segment the block of content into the one or more source units of translation,
3 comprising:
4 selecting one or more delineating characteristics of the source content in addition to
5 lexical characteristics;
6 determining instances of the delineating characteristics in the block of source content,
7 and identifying one or more pairs of the instances within the text;
8 for each pair of instances of the delineating characteristics, associating a first instance of
9 the pair with a first boundary of a source unit of translation, and associating a second instance of
10 the pair with a second boundary of the source unit of translation.
- 1 2. A method according to claim 1, further including identifying one or more target units of
2 translation in a block of target content, and assigning associations among the source units of
3 translation in the block of source code and the target units of translation in the block of target
4 code.
- 1 3. A method according to claim 2, further including translating content in the source units
2 of translation to the associated target units of translation.
- 1 4. A method according to claim 1, wherein the delineating characteristics include syntactic
2 characteristics, the method further including determining pairs of instances of syntactic
3 characteristics of the source content.
- 1 5. A method according to claim 1, wherein the delineating characteristics include
2 formatting characteristics, the method further including determining pairs of instances of
3 formatting characteristics of the source content.
- 1 6. A method according to claim 5, wherein the document formatting characteristics include
2 HTML code markers.

- 1 7. A method according to claim 1, wherein the delineating characteristics include
2 conceptual characteristics, the method further including determining pairs of instances of
3 conceptual characteristics of the source content.
- 1 8. A method according to claim 7, wherein the conceptual characteristics include spatial
2 adjacency.
- 1 9. A method according to claim 1, wherein the delineating characteristics include sound-
2 based characteristics, the method further including determining pairs of instances of sound-based
3 characteristics of the source content.
- 1 10. A method according to claim 9, wherein the sound based characteristics include voice
2 inflections.
- 1 11. A method according to claim 1, wherein the delineating characteristics include one or
2 more markers manually inserted by a user, the method further including determining pairs of
3 instances of markers within the source content.
- 1 12. A method according to claim 1, further including translating the one or more source
2 units of translation into a target language so as to form target units of translation, and merging
3 the target units of translation into one or more blocks of target content.
- 1 13. A method according to claim 12, wherein the source units of translation are
2 characterized by a first adjacency pattern, the method further including merging the target units
3 of translation so as to follow the first adjacency pattern.
- 1 14. A method according to claim 12, wherein the source units of translation are
2 characterized by a first adjacency pattern, the method further including merging the target units
3 of translation so as to follow a second adjacency pattern different from the first adjacency
4 pattern.

- 1 15. A method according to claim 12, wherein at least one of the source units of translation
2 corresponds with two or more target units of translation.
- 1 16. A method according to claim 12, wherein two or more of the source units of translation
2 corresponds with a single target unit of translation.
- 1 17. A method according to claim 12, wherein each one of the source units of translation
2 corresponds with a single target unit of translation.
- 1 18. A method according to claim 12, further including merging the target units of translation
2 into a hierarchical structure.
- 1 19. A method according to claim 18, further including providing one or more predetermined
2 hierarchy criteria, wherein the characteristics of the hierarchical structure are defined by the
3 predetermined hierarchy criteria.
- 1 20. A system for computer assisted identification one or more source units of translation in a
2 block of source content, so as to segment the block of content into the one or more source units
3 of translation, comprising:
4 a user interface for allowing a user to select one or more delineating characteristics of
5 the source content in addition to lexical characteristics;
6 a content processor for determining instances of the delineating characteristics in the
7 block of source content, and identifying one or more pairs of the instances; and,
8 for each pair of instances of the delineating characteristics, a segment processor for
9 associating a first instance of the pair with a first boundary of a source unit of translation, and
10 associating a second instance of the pair with a second boundary of the source unit of translation.
- 1 21. A system according to claim 20, wherein the content processor further identifies one or
2 more target units of translation in a block of target content, and assigns associations among the

3 source units of translation in the block of source code and the target units of translation in the
4 block of target code.

1 22. A system according to claim 21, wherein the content processor further translates content
2 in the source units of translation to the associated target units of translation.

1 23. A system according to claim 20, wherein the delineating characteristics include syntactic
2 characteristics, and the content processor further determines pairs of instances of syntactic
3 characteristics of the source content.

1 24. A system according to claim 20, wherein the delineating characteristics include
2 document formatting characteristics, and the content processor further determines pairs of
3 instances of document formatting characteristics of the source content.

1 25. A system according to claim 24, wherein the document formatting characteristics include
2 HTML code.

1 26. A system according to claim 20, wherein the delineating characteristics include
2 conceptual characteristics, and the content processor further determines pairs of instances of
3 conceptual characteristics of the source content.

1 27. A system according to claim 26, wherein the conceptual characteristics include spatial
2 adjacency.

1 28. A system according to claim 20, wherein the delineating characteristics include sound-
2 based characteristics, and the content processor further determines pairs of instances of sound-
3 based characteristics of the source content.

1 29. A system according to claim 28, wherein the sound based characteristics include voice
2 inflections.

1 30. A system according to claim 20, wherein the delineating characteristics one or more
2 markers manually inserted by a user, and the content processor further determines pairs of
3 instances of markers within the source content.

1 31. A system according to claim 20, wherein the segment processor further translates the one
2 or more source units of translation into a target language so as to form target units of translation,
3 and merges the target units of translation into one or more blocks of target content.

1 32. A system according to claim 31, wherein the source units of translation are characterized
2 by a first adjacency pattern, and the segment processor further merges the target units of
3 translation so as to follow the first adjacency pattern.

1 33. A system according to claim 31, wherein the source units of translation are characterized
2 by a first adjacency pattern, and the segment processor further merges the target units of
3 translation so as to follow a second adjacency pattern different from the first adjacency pattern.

1 34. A system according to claim 31, wherein at least one of the source units of translation
2 corresponds with two or more target units of translation.

1 35. A system according to claim 31, wherein two or more of the source units of translation
2 correspond with a single target unit of translation.

1 36. A system according to claim 31, wherein each one of the source units of translation
2 corresponds with a single target unit of translation.

1 37. A system according to claim 31, wherein the segment processor further merges the target
2 units of translation into a hierarchical structure.

1 38. A system according to claim 37, wherein the segment processor further receives one or
2 more predetermined hierarchy criteria, and the characteristics of the hierarchical structure are
3 defined by the predetermined hierarchy criteria.

1 39. A system for computer assisted identification one or more source units of translation in a
2 block of source content, so as to segment the block of text into the one or more source units of
3 translation, comprising:
4 means for allowing a user to select one or more delineating characteristics of the source
5 content in addition to lexical characteristics;
6 means for determining one or more pairs of instances of the delineating characteristics in
7 the block of source content; and,
8 for each pair of instances of the delineating characteristics, means for associating a first
9 instance of the pair with a first boundary of a source unit of translation, and associating a second
10 instance of the pair with a second boundary of the source unit of translation.

1 40. A method of dynamically selecting one or more segmentation criteria used to identify
2 source units of translation in a block of source content, wherein the segmentation criteria
3 identifies delineation characteristics of the source content for defining boundaries of the source
4 units of translation, comprising:
5 providing two or more source segmentation criteria associated with the block of source
6 content;
7 selecting one of the source segmentation criteria from the two or more segmentation
8 criteria as an initial source criterion, and using the initial source criterion for defining boundaries
9 of the source units of translation; and,
10 dynamically selecting, as a function of one or more external factors, subsequent source
11 segmentation criteria from the two or more source segmentation criteria, as the boundaries of the
12 source units of translation are defined.

1 41. A method according to claim 40, further including:
2 providing two or more target segmentation criteria associated with a block of target
3 content;
4 selecting one of the target segmentation criteria from the two or more target
5 segmentation criteria as an initial target criterion, and using the initial target criterion for
6 defining boundaries of the target units of translation;

7 dynamically selecting, as a function of one or more external factors, subsequent target
8 segmentation criteria from the two or more target segmentation criteria, as the boundaries of the
9 target units of translation are defined, and,
10 assigning associations among the source units of translation in the block of source code
11 and the target units of translation in the block of target code.

1 42. A method according to claim 41, wherein the one or more external factors includes the
2 associations among the source units of translation in the block of source code and the target units
3 of translation in the block of target code.

1 43. A method according to claim 41, wherein the one or more external factors includes input
2 from a user translating from the source units of translation to the target units of translation.

1 44. A method according to claim 41, wherein the one or more external factors includes data
2 relating to characteristics of the source content.

1 45. A method according to claim 44, wherein the data relating to characteristics of the
2 source content includes HTML code.

1 46. A system for computer assisted dynamic selection of one or more segmentation criteria
2 used to identify source units of translation in a block of source content, wherein the segmentation
3 criteria identifies delineation characteristics of the source content for defining boundaries of the
4 source units of translation, comprising:

5 a user interface for providing two or more source segmentation criteria associated with
6 the block of source content, and for selecting one of the source segmentation criteria from the
7 two or more segmentation criteria as an initial source criterion;

8 a content processor for using the initial source criterion for defining boundaries of the
9 source units of translation; and,

10 a segment processor for dynamically selecting, as a function of one or more external
11 factors, subsequent source segmentation criteria from the two or more source segmentation
12 criteria, as the boundaries of the source units of translation are defined.

1 47. A system according to claim 46, wherein (i) the user interface further provides two or
2 more target segmentation criteria associated with a block of target content, and selects one of the
3 target segmentation criteria from the two or more target segmentation criteria as an initial target
4 criterion, (ii) the content processor further uses the initial target criterion for defining boundaries
5 of the target units of translation, and (iii) the segment processor further dynamically selects, as a
6 function of one or more external factors, subsequent target segmentation criteria from the two or
7 more target segmentation criteria, as the boundaries of the target units of translation are defined,
8 and (iv) the segment processor further assigns associations among the source units of translation
9 in the block of source code and the target units of translation in the block of target code.

1 48. A system according to claim 47, wherein the one or more external factors includes the
2 associations among the source units of translation in the block of source code and the target units
3 of translation in the block of target code.

1 49. A system according to claim 47, wherein the one or more external factors includes input
2 from a user translating from the source units of translation to the target units of translation.

1 50. A system according to claim 47, wherein the one or more external factors includes data
2 relating to characteristics of the source content.

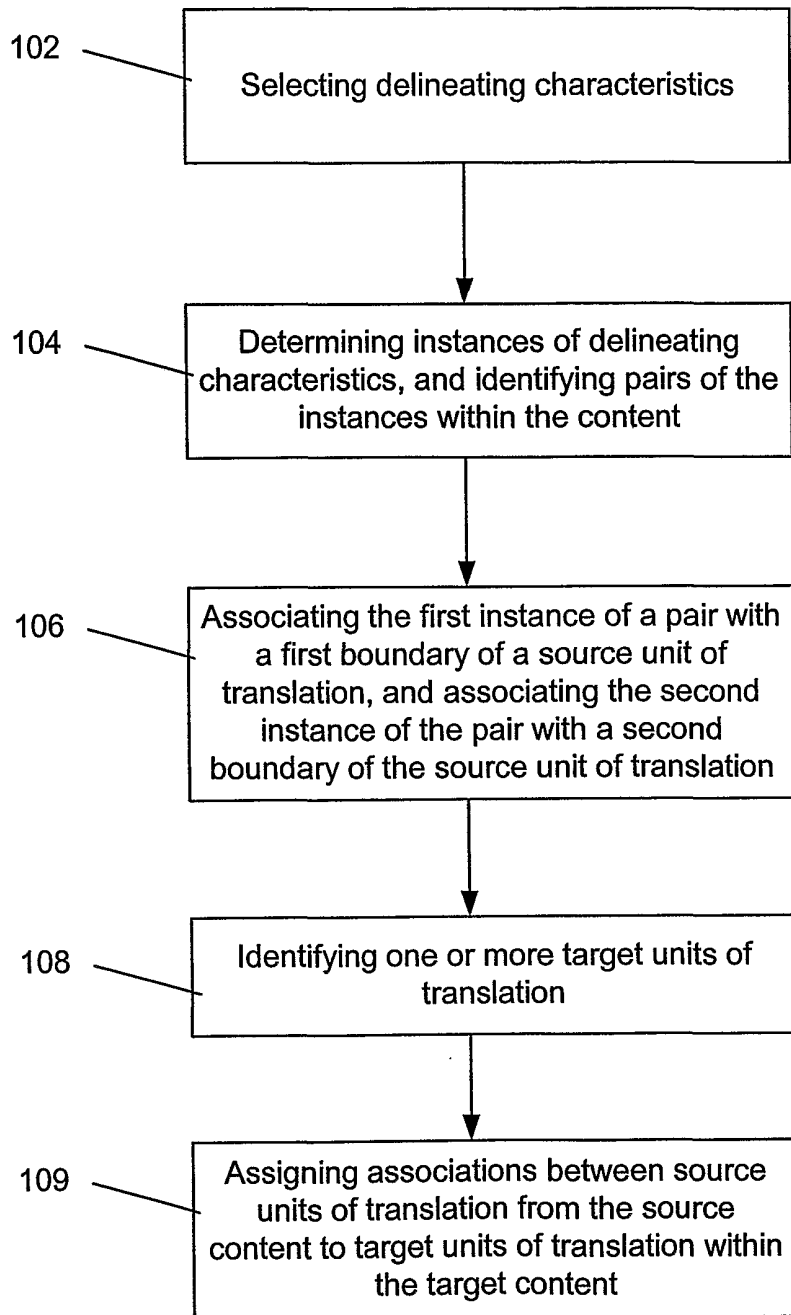
1 51. A method according to claim 50, wherein the data relating to characteristics of the
2 source content includes HTML code.

1 52. A system for computer assisted dynamic selection of one or more segmentation criteria
2 used to identify source units of translation in a block of source content, wherein the segmentation
3 criteria identifies delineation characteristics of the source content for defining boundaries of the
4 source units of translation, comprising:

5 means for providing two or more source segmentation criteria associated with the block
6 of source content, and for selecting one of the source segmentation criteria from the two or more
7 segmentation criteria as an initial source criterion;

8 means for using the initial source criterion for defining boundaries of the source units of
9 translation; and,
10 means for dynamically selecting, as a function of one or more external factors,
11 subsequent source segmentation criteria from the two or more source segmentation criteria, as
12 the boundaries of the source units of translation are defined.

1 53. A system according to claim 52, further including:
2 means for providing two or more target segmentation criteria associated with a block of
3 target content;
4 means for selecting one of the target segmentation criteria from the two or more target
5 segmentation criteria as an initial target criterion, and using the initial target criterion for
6 defining boundaries of the target units of translation;
7 means for dynamically selecting, as a function of one or more external factors,
8 subsequent target segmentation criteria from the two or more target segmentation criteria, as the
9 boundaries of the target units of translation are defined, and,
10 means for assigning associations among the source units of translation in the block of
11 source code and the target units of translation in the block of target code.



100

FIG. 1

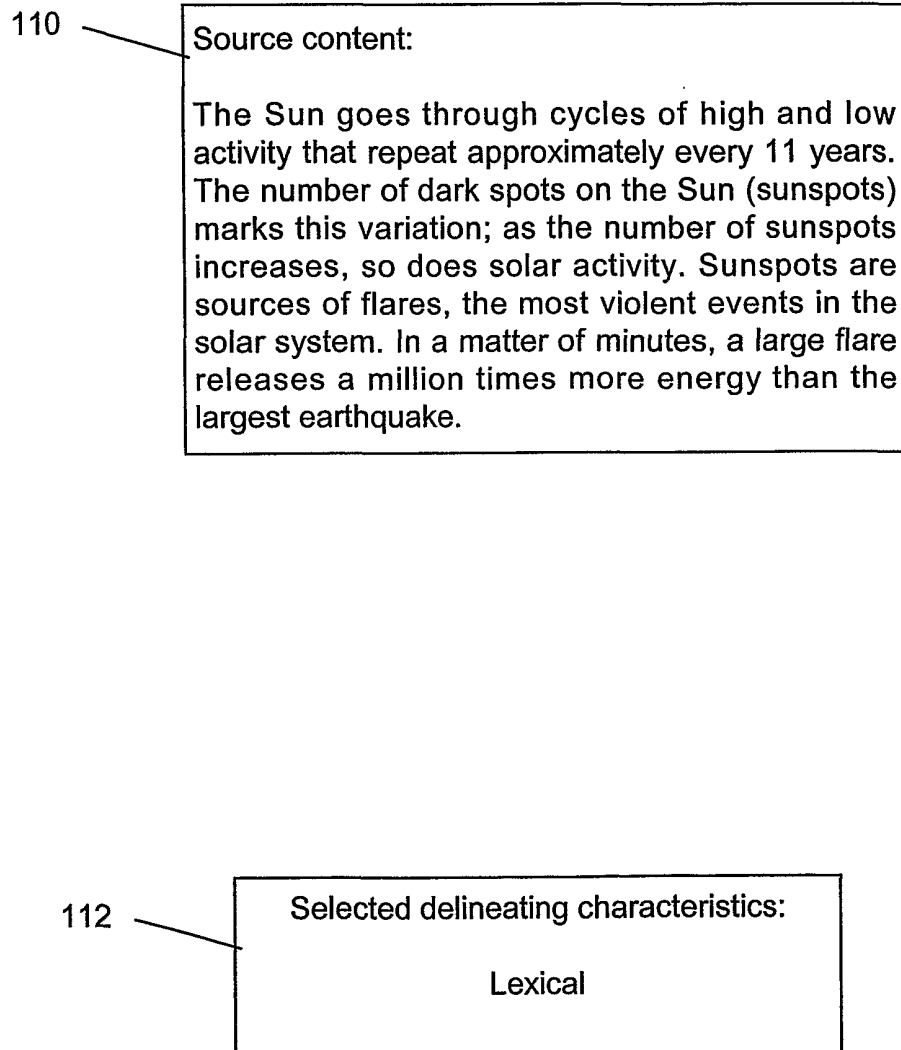


FIG. 2A

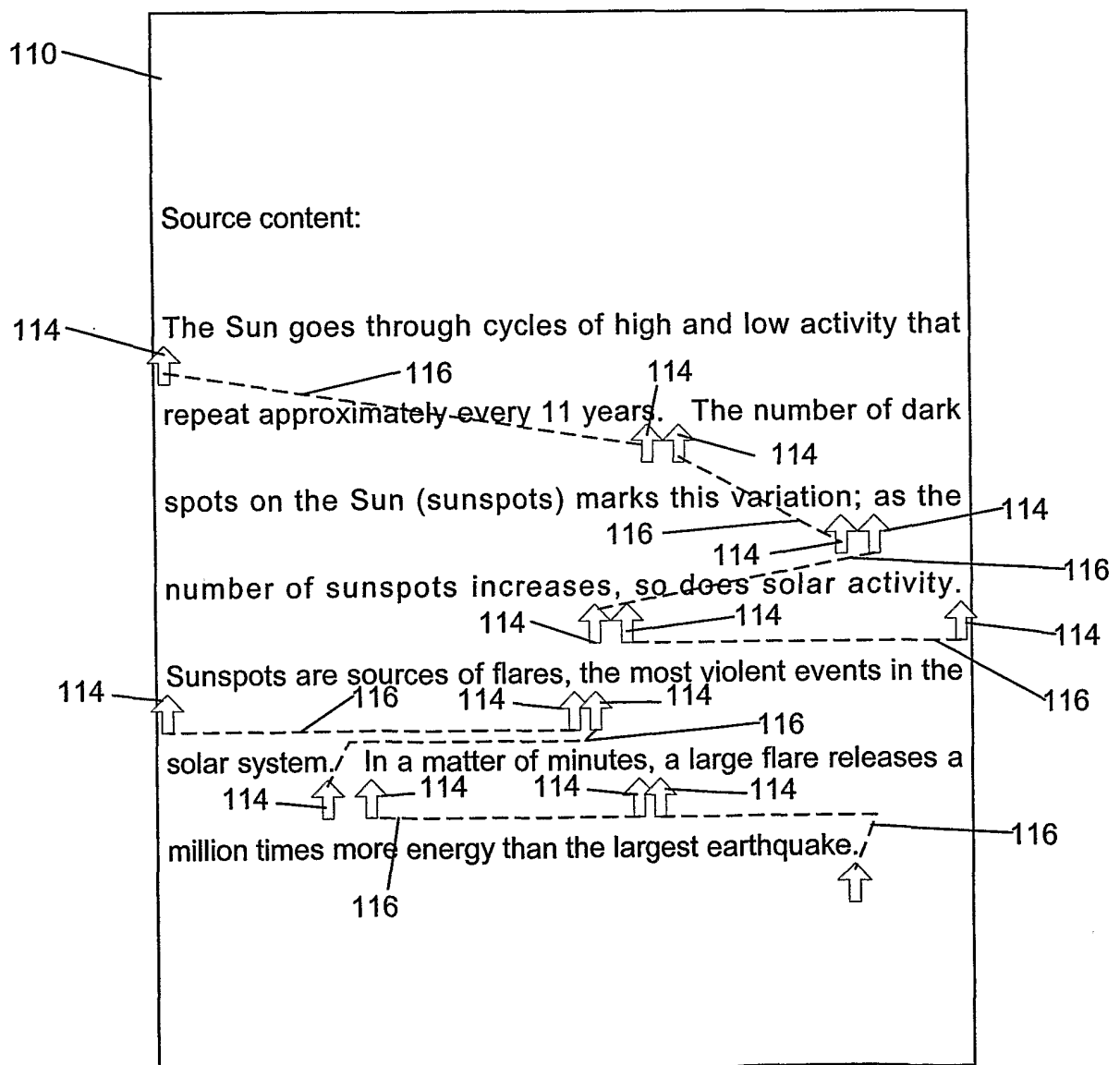


FIG. 2B

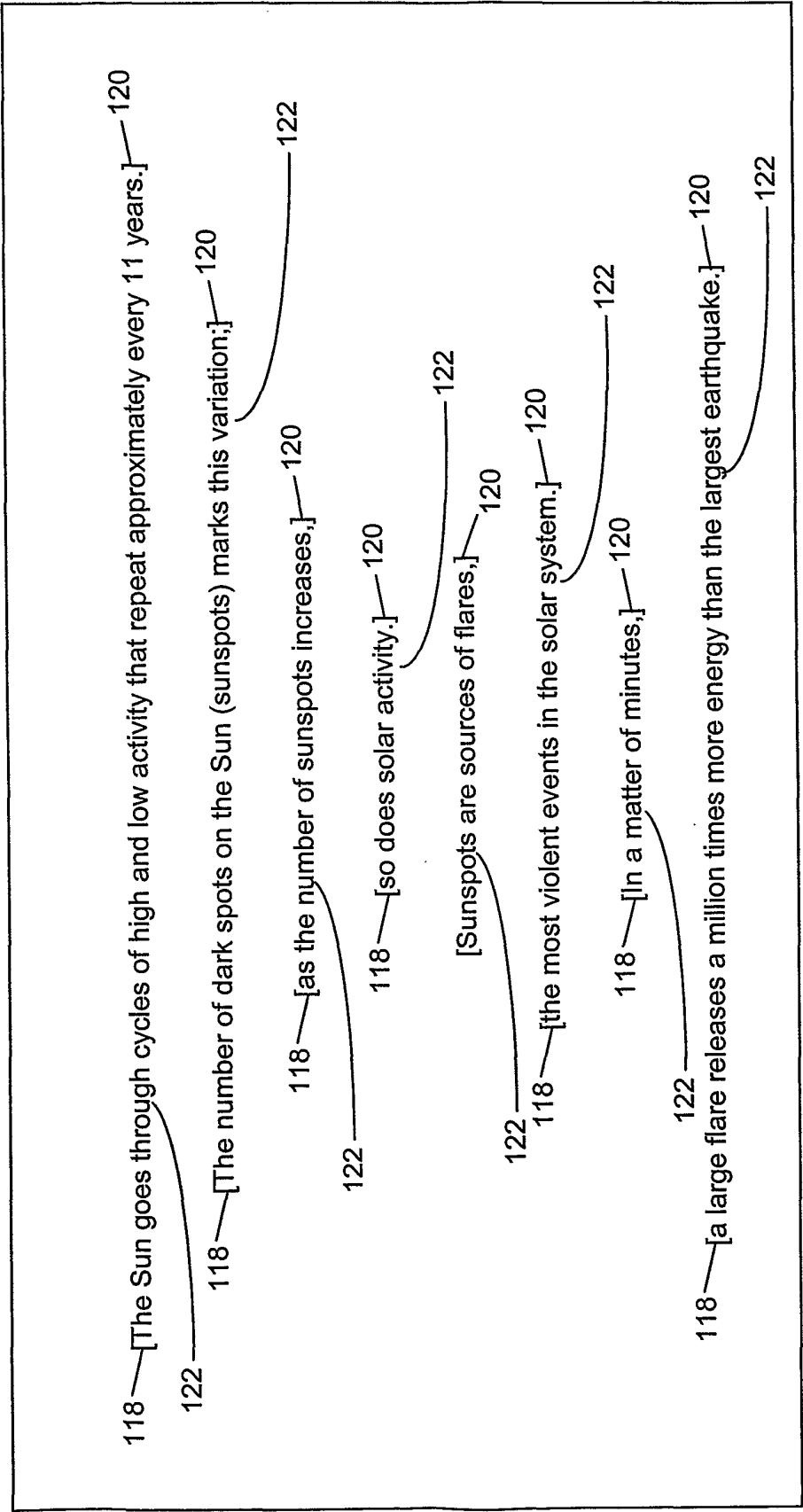


FIG. 2C

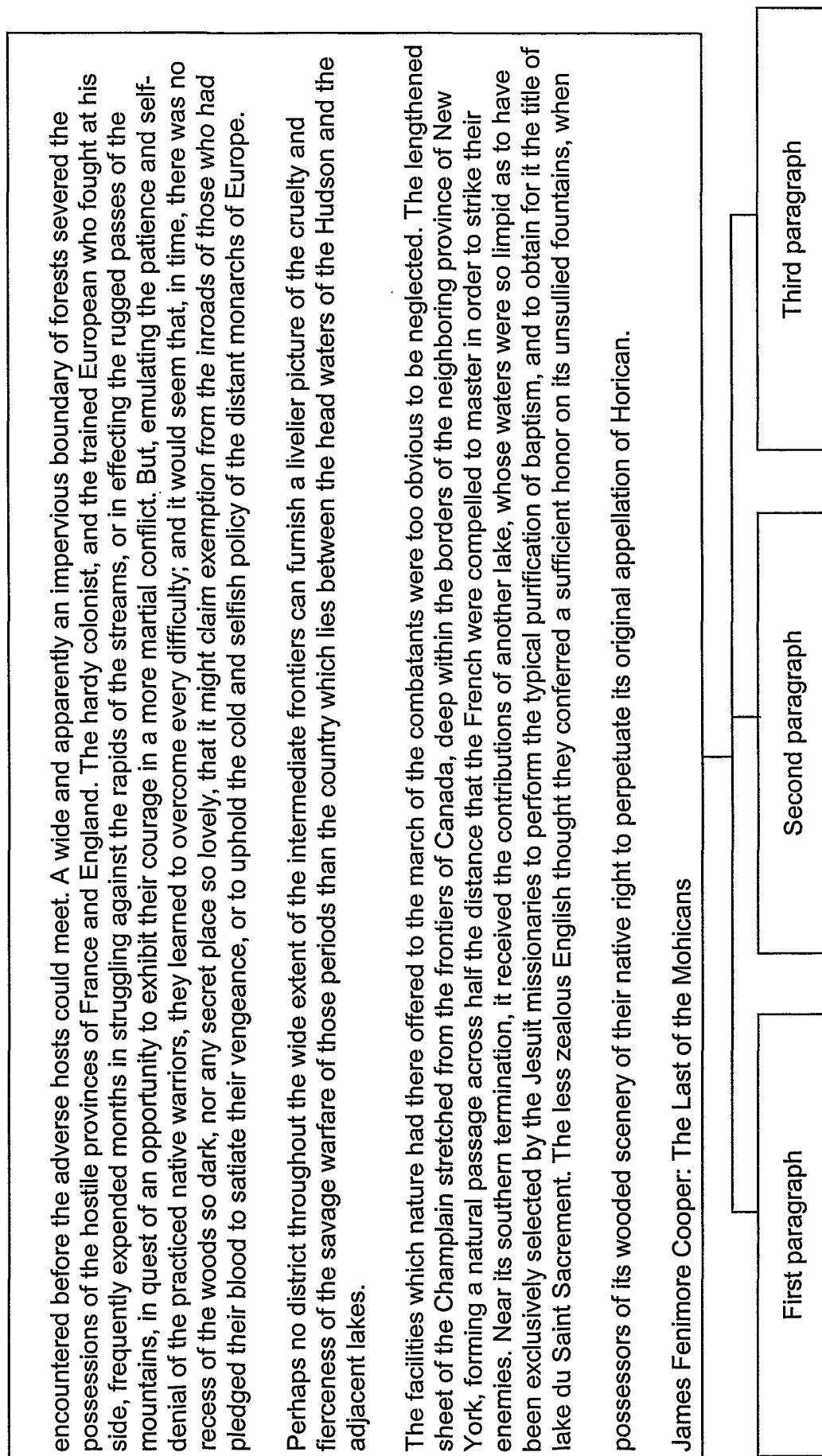


FIG. 3A

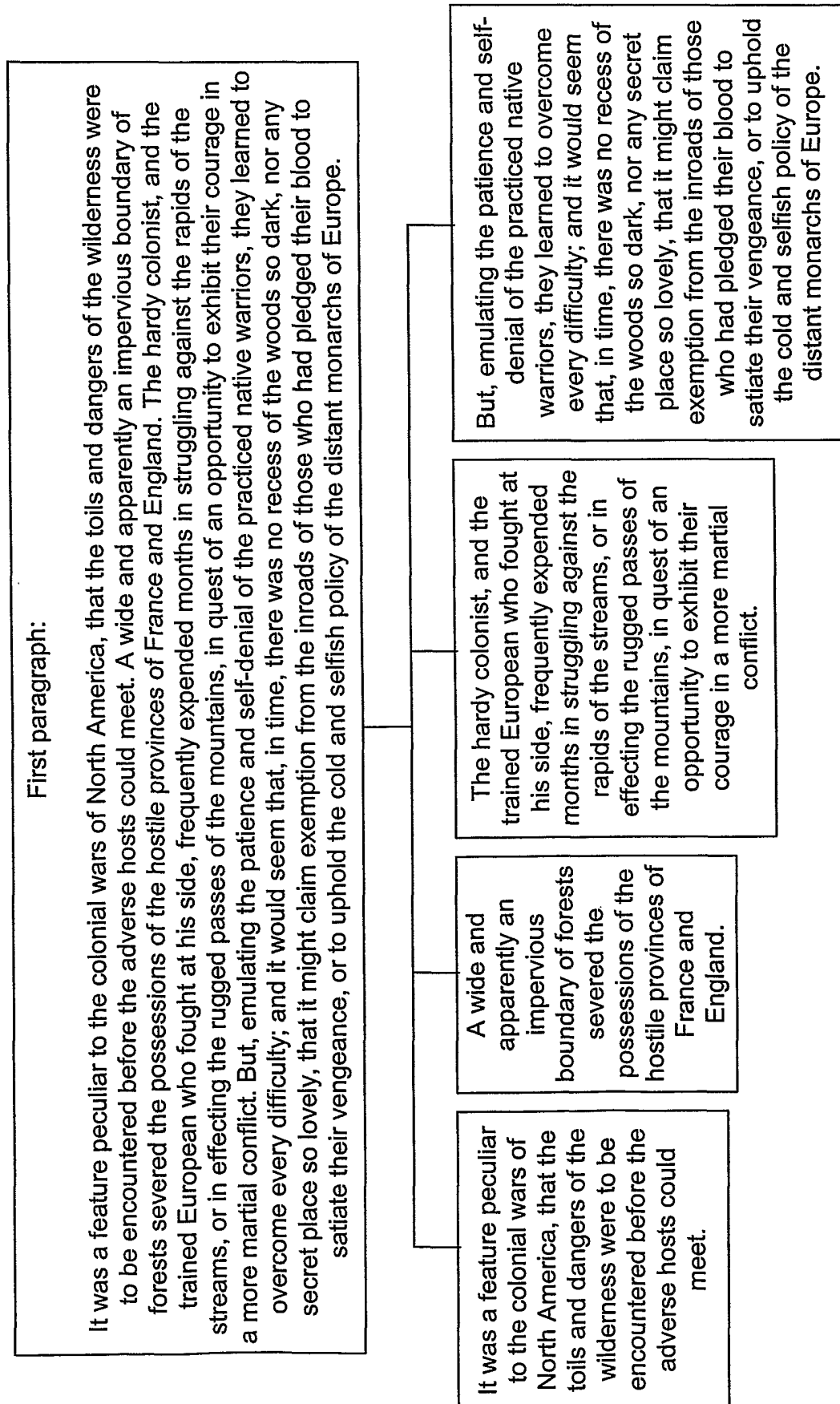


FIG. 3B

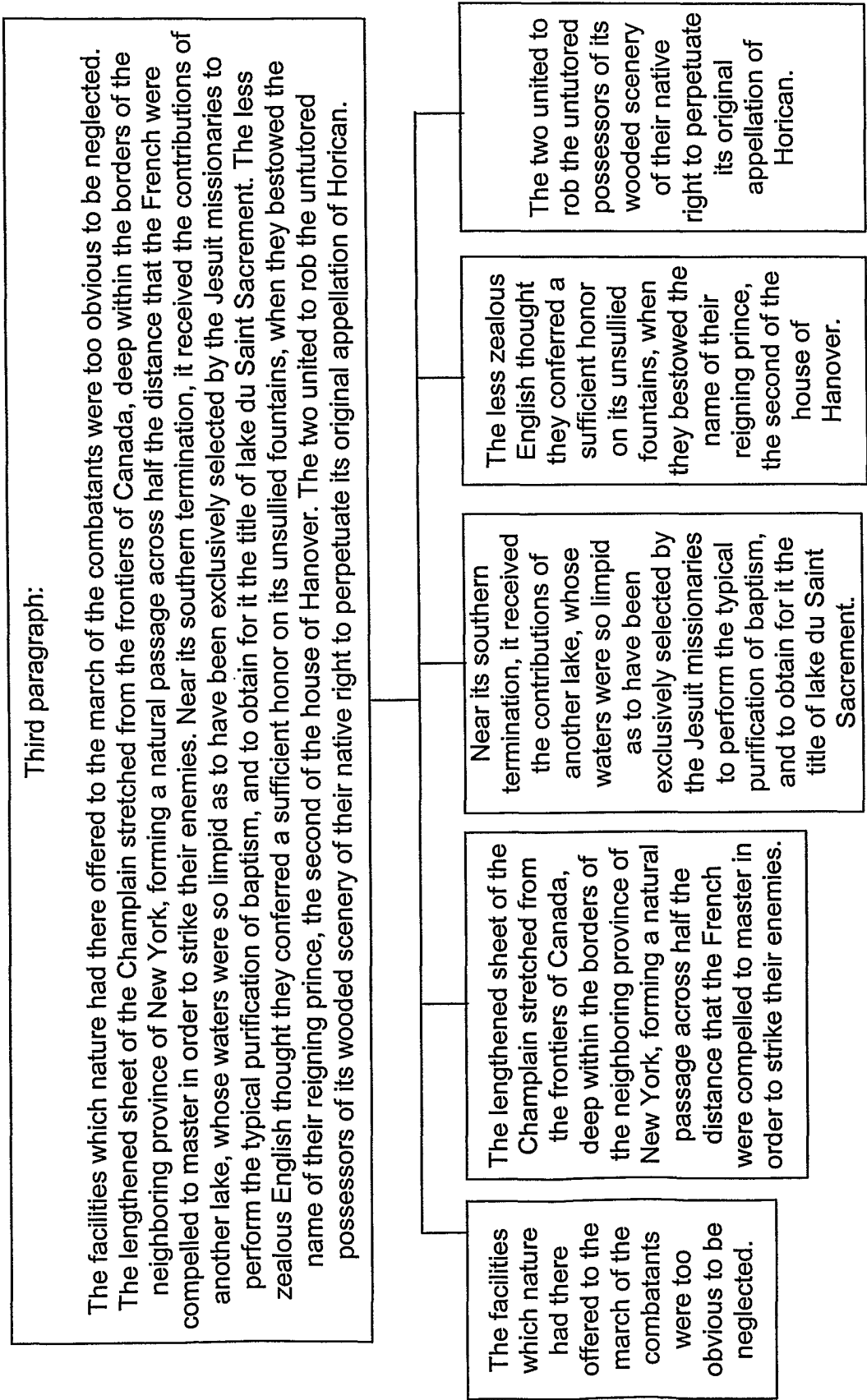


FIG. 3C

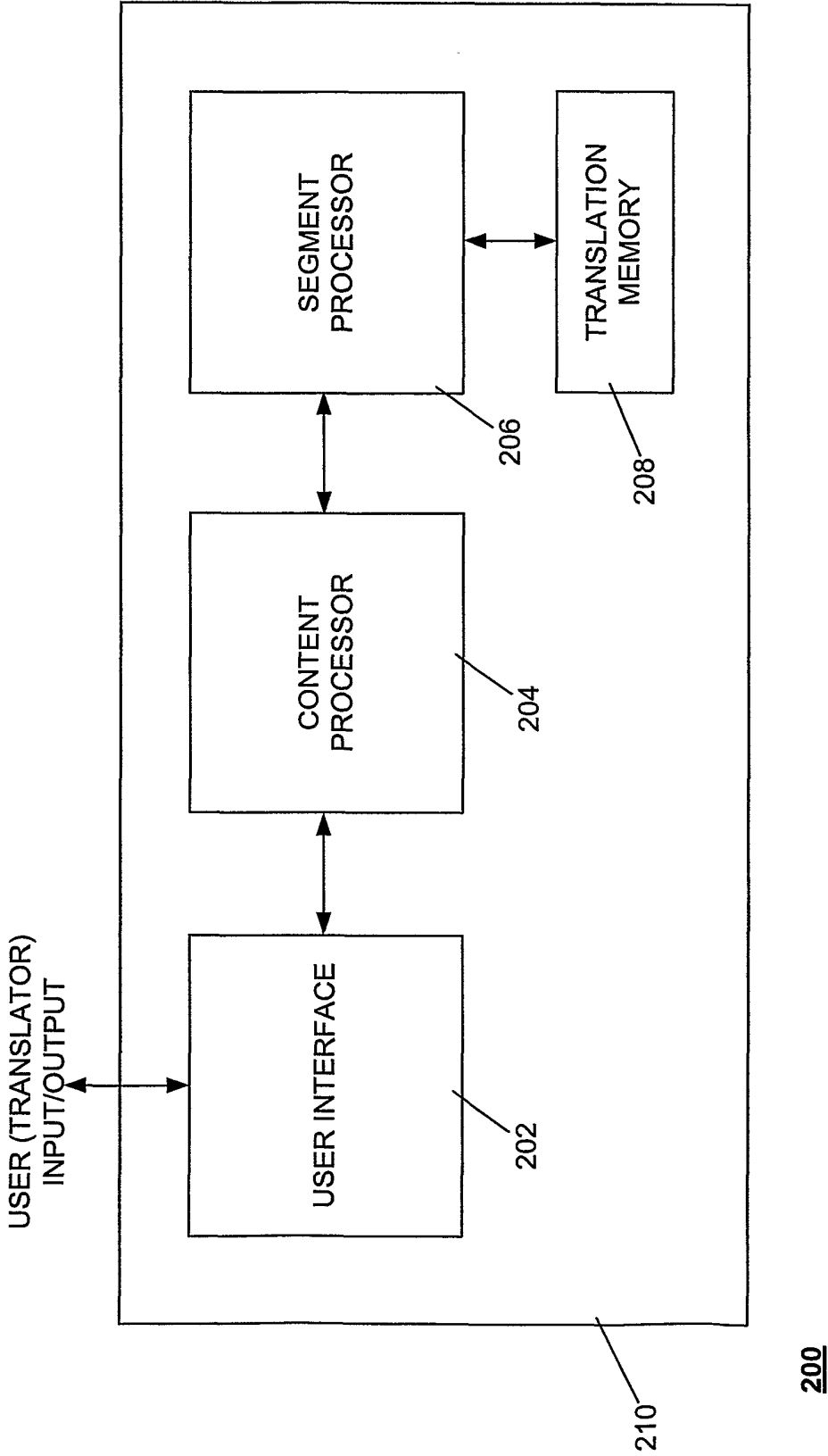


FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US01/30652

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 17/28

US CL : 704/2

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/2, 3, 4, 5, 6, 7, 8, 277

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,349,368 A (TAKEDA et al) 20 September 1994 (20.09.1994), abstract; figs. 3, 7, 10, & 12A-12C; col. 1, line 11 to col. 2, line 26; col. 3, line 37 to col. 9, line 43.	1-53
A	US 5,708,825 A (SOTOMAYOR) 13 January 1998 (13.01.1998), abstract; figs. 8-10; col. 11, line 60 to col. 13, line 42; and col. 15, line 31 to col. 16, line 61.	1-53
Y	US 5,848,386 A (MOTOYAMA) 08 December 1998 (08.12.1998), abstract; figs. 3, 9A-9B, & 15-16; col. 1, line 20 to col. 2, line 63; col. 5, line 28 to col. 15, line 35.	1-53
A	EP 0 887 748 A2 (LANGE at al) 30 December 1998 (30.12.1998) abstract; col. 5, line 20 to col. 12, line 23; and figs. 4-6.	1-53
Y	US 5,987,403 A (SUGIMURA) 16 November 1999 (16.11.1999) abstract; col. 1, line 16 to col. 4, line 36; and col. 5, line 20 to col. 13, line 5.	1-53

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:	
"A" document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"B" earlier application or patent published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

12 December 2001 (12.12.01)

Date of mailing of the international search report

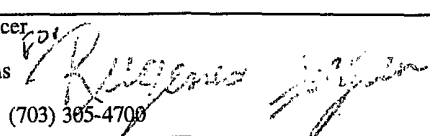
28 December 2001 (28.12.01)

Name and mailing address of the ISA/US

Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703)305-3230

Authorized officer

Joseph Thomas 

Telephone No. (703) 365-4700

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US01/30652

Continuation of B. FIELDS SEARCHED Item 3:

EAST search - files: USPAT, US-PGPUB, EPO, JPO, DERWENT, IBM-TDB

search terms: translation/interpretation, HTML, markup, boundary/border, syntax/syntactic, adjacent